Hypertextsorten
Definition – Struktur – Klassifikation

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie
des Fachbereiches 05 Sprache, Literatur, Kultur
der Justus-Liebig-Universität Gießen

vorgelegt von
Georg Rehm

aus Gießen

2005

Dekan:	Prof. Dr. Hartmut Stenzel
1. Berichterstatter:	Prof. Dr. Henning Lobin
2. Berichterstatter:	Prof. Dr. Gerd Fritz
Tag der Disputation:	23. Januar 2006

Kapitelübersicht

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Listings

Vorbemerkungen zur Typografie

Die nachfolgenden Bemerkungen zur Typografie beziehen sich nicht auf wörtliche Zitate; in diesen wurde die vom jeweiligen Verfasser gewählte Form der typografischen Auszeichnung übernommen, wobei jedoch im Original z. B. durch Kapitälchen oder Unterstreichungen realisierte Hervorhebungen in der vorliegenden Arbeit aus Gründen der typografischen Konsistenz durch Kursivschrift wiedergegeben werden. Vom Verfasser hinzugefügte Hervorhebungen wurden entsprechend gekennzeichnet. Aus typografischen Gründen wurden einfache (‚. . .‘) und doppelte Anführungszeichen („. . .“), die innerhalb von wörtlichen Zitaten aus deutschsprachigen Quellen verwendet werden, durch ›. . .‹ und ». . .« ersetzt. Dies gilt nicht für englischsprachige Zitate; in diesen wurden die ursprünglichen Anführungszeichen beibehalten. Im laufenden Text werden diese ebenfalls in doppelte Anführungszeichen gesetzt. Längere, typografisch abgesetzte Zitate werden in einer um einen Punkt verkleinerten Schrift dargestellt und nicht durch Anführungszeichen markiert.

Etablierte englischsprachige Termini werden in kursiver Schrift dargestellt: *Semantic Web*, *World Wide Web*, *Crawler*, *Document Type Definition*, *Uniform Resource Locator*.

Datenbankfelder, Programmcode, URLs sowie die Bezeichnungen der Elemente und Attribute von Auszeichnungssprachen werden in einer um einen Punkt verkleinerten, dicktengleichen Schrift dargestellt: `head`, `body`, `http://www.uni-giessen.de`.

Die Namen von kommerziell ausgerichteten Software-Paketen, Dienstleistern oder Firmen sowie von Open-Source-Anwendungen werden in kursiver Schrift dargestellt: *Google*, *Altavista*, *Powerpoint*, *Pavuk*, *Linux*, *Apache*, *Perl*.

Abkürzungen werden im Fließtext in einer um einen Punkt verkleinerten Schrift dargestellt: HTML, WWW, CMC, DTD, XML. Diese Verkleinerung wird in Abbildungen und in Tabellen nicht angewendet. Ein Verzeichnis häufig verwendeter Abkürzungen befindet sich in Anhang C (S. 789 ff.).

Zusammenfassung

Suchmaschinen im *World Wide Web* indexieren und durchsuchen Dokumente in großer Geschwindigkeit. Trotz der quantitativ beeindruckenden Ergebnisse lässt die Qualität der Treffer jedoch oft zu wünschen übrig. Die vorliegende Arbeit zielt darauf ab, die theoretischen und praktischen Grundlagen für strukturelle Verbesserungen der Funktionsweise von Suchmaschinen zu liefern. Der Schlüssel hierfür liegt in der maschinellen Identifikation von *Hypertextsorten*. Dieser Begriff bezeichnet den generellen Typ eines WWW-basierten Hypertextes im Sinne eines funktional-thematisch markierten Kommunikats. Hypertextsorten sind – ebenso wie traditionelle Textsorten – auf verschiedenen Ebenen von Konventionen gekennzeichnet, die rekurrent in zugehörigen Text- bzw. Hypertextexemplaren beobachtet werden können. Diese Konventionen resultieren aus einem zyklischen, produzentenseitigen Prozess, der aus der Rezeption verwandter Webangebote und der Übernahme ausgewählter Gestaltungs- oder Strukturierungsmerkmale besteht, die vom Autor für die kommunikative Funktion des eigenen Hypertextes als gewinnbringend eingeschätzt werden. Eine Komponente zur maschinellen Erkennung von Hypertextsorten mit computerlinguistischen Methoden könnte unter anderem im Rahmen einer Suchmaschine Verwendung finden. Ihren Anwendern stünde hierdurch eine weitere Ebene des Zugriffs auf Dokumente zur Verfügung, so dass nach Dokumenten recherchiert werden kann, die die angegebenen Stichwörter enthalten und darüber hinaus den vom Benutzer spezifizierten Hypertextsorten zugehörig sind, z. B. *persönliche Homepage*, *Produktkatalog* oder *Kochrezept* (Kapitel 1).

Die Arbeit nähert sich dieser Thematik aus einer theoretischen Perspektive und behandelt zunächst die für die Charakterisierung des Begriffs Hypertextsorte benötigten Grundlagen (Teil I), die sich auf die Textlinguistik (Kapitel 2) und die linguistischen Spezifika des Konzepts Hypertext stützen (Kapitel 3). Anschließend werden die Kerneigenschaften von Hypertextsorten bzw. Web-Genres dargestellt (Kapitel 4). Teil II erläutert den Rahmenansatz und die Methodologie. Auf Basis des ersten Teils wird in Kapitel 5 ein Hypertextsortenmodell entwickelt, das sowohl für textlinguistische Analysen als auch für texttechnologische Anwendungen ausgelegt ist und zwischen den drei konzeptionellen Ebenen *Hypertextsorte* (betrifft den gesamten Hypertext), *Hypertextknotensorte* (betrifft einzelne HTML-Dokumente) und *Hypertextsortenmodul* (betrifft Bausteine von HTML-Dokumenten) differenziert. Kapitel 6 stellt die Untersuchungsdomäne der universitären Webangebote vor, auf die sich die nachfolgenden Analysen beziehen. Für diese Untersuchungsdomäne wurde ein Korpus von etwa vier Millionen HTML-Dokumenten angefertigt, die sich in einer Korpusdatenbank befinden und sowohl manuell – über eine Web-Oberfläche – als auch automatisch analysiert werden können (Kapitel 7). Teil III umfasst fünf empirische Analysen als exemplarische Anwendungen des Hypertextsortenmodells auf die Untersuchungsdomäne. Die ersten drei Analysen betreffen die quantitative Auswertung persönlicher Homepages (Kapitel 8) und Untersuchungen

von Exemplaren der Hypertextsorten *private Homepage eines Studierenden* (Kapitel 9) sowie *persönliche Homepage eines Wissenschaftlers* (Kapitel 10). In Kapitel 11 wird die Hypertextknotensorte *Einstiegsseite eines universitären Webauftritts* anhand von 35 Einstiegsseiten analysiert; es schließt sich eine Untersuchung der 692 Dokumente, die in den Einstiegsseiten mittels Hyperlinks referenziert werden, hinsichtlich ihrer Hypertextknotensorten und übergeordneten Hypertextsorten an. In der fünften Analyse werden 750 zufällig ausgewählte Dokumente ausgewertet (Kapitel 12). Während die ersten drei Analysen unterschiedliche Hypertextsortenvarianten des Hypertexttyps *Homepage einer Person* fokussieren, wird mit den beiden abschließenden Untersuchungen das Ziel der Sammlung und Identifizierung der zugehörigen Hypertextknotensorten sowie ihrer übergeordneten Hypertextsorten verfolgt. Ein weiteres Ziel betrifft die Typologisierung der jeweils mehr als 100 ermittelten Hypertextsorten und Hypertextknotensorten. Die fünf Analysen weisen eine Vielzahl von Konventionen nach und belegen die Existenz zahlreicher Hypertextsorten, die nicht nur dem universitären Bereich zuzurechnen sind. Teil IV geht auf die technologische Umsetzung ein. Kapitel 13 präsentiert die Hypertextsortenontologie, die eine Anwendung des Hypertextsortenmodells darstellt, auf der *Web Ontology Language* basiert und von einem Domänenmodell sowie einer Ontologie wissenschaftlicher Themen und Fachgebiete flankiert wird. In die Hypertextsortenontologie werden die Ergebnisse der empirischen Analysen (Teil III) integriert, wobei auch die Frage der Typologisierung wieder aufgegriffen wird. Die maschinelle Identifizierung von Hypertextsorten als Grundlage sprachtechnologischer Anwendungen ist Gegenstand von Kapitel 14. Ausgehend von einer kritischen Betrachtung der vorliegenden Arbeiten zur automatischen Identifizierung von Genres bzw. Web-Genres wird eine Architektur mit einzelnen Komponenten entwickelt, die – wiederum in Bezug auf das Hypertextsortenmodell – für die maschinelle Erkennung von Hypertextsorten benötigt werden, um die im WWW existenten realen Gegebenheiten der Kommunikation erfassen zu können. Die Hypertextsortenontologie fungiert in dieser Architektur als Wissensbasis, die die empirisch beobachteten Zusammenhänge zwischen Hypertextsorten, Hypertextknotensorten und Hypertextsortenmodulen beinhaltet. Es wird die prototypische Implementierung eines Textparsers für arbiträre HTML-Dokumente vorgestellt, der innerhalb dieser Architektur einen zentralen Stellenwert einnimmt, um die Komponenten der Textoberfläche auf Hypertextsortenmodule abzubilden.

1 Einleitung

1.1 Einführung in das Untersuchungsgebiet

Suchmaschinen stellen zwar sehr effektive Werkzeuge zur Informationsrecherche im *World Wide Web* dar, ihre Benutzung ist jedoch mit verschiedenen Restriktionen verbunden. Diese beziehen sich unter anderem auf die Möglichkeiten zur Spezifizierung einer Anfrage, die sich in der Regel auf die Eingabe von Suchwörtern oder -phrasen beschränkt, für die die Suchmaschine relevante Dokumente aus ihren Datenbeständen ermittelt. Eine in den meisten Fällen als „advanced search“ bezeichnete Funktion präsentiert eine komplexere Benutzerschnittstelle und erlaubt die Spezifizierung zusätzlicher Parameter, wozu z. B. das Dateiformat (HTML, PDF etc.), die *domain* (`.edu`, `.com`, `.uni-giessen.de` etc.) und die Sprache, in der ein Dokument verfasst wurde, gehören. An diesem Punkt setzt die Motivation der vorliegenden Arbeit an: Ist es – in Anlehnung an das in der Textlinguistik etablierte Konzept der Textsorten – möglich, im *World Wide Web* unterschiedliche *Hypertextsorten* zu differenzieren? Können Hypertextsorten maschinell identifiziert werden, so dass Suchmaschinen ihren Anwendern eine weitere Ebene zur Spezifizierung von Suchanfragen zur Verfügung stellen können?

Die Suchmaschine *Google* (`http://www.google.com`) präsentiert in ihrer Einstiegsseite die Anzahl der HTML-Dokumente, die sich in ihren Datenbeständen befinden. Derzeit, im August 2005, sind etwa acht Milliarden Dokumente verzeichnet, etwa zwei Jahre zuvor befanden sich ca. drei Milliarden Dokumente in den verteilten Indizes dieses Anbieters. Die Schwelle von einer Milliarde Dokumente wurde nach Angaben der nicht mehr auf dem Markt befindlichen Suchmaschine *Inktomi* im Januar 2000 durchbrochen. Da die verteilten *Crawler* einer Suchmaschine niemals das gesamte WWW durchsuchen können, entstand eine Studie, nach deren Auskunft das „deep Web“, d. h. das *vollständige* Web einschließlich der Teile, die von Suchmaschinen aus technischen, logistischen oder algorithmischen Gründen nicht traversiert werden können bzw. sollen, die kaum vorstellbare Zahl von 500 Milliarden Webseiten umfasst (Bergman, 2000). Nach Schätzungen von Gulli und Signorini (2005) enthält der indexierbare Teil des WWW mittlerweile etwa 11,5 Milliarden Dokumente.

Eine Informationsrecherche im *World Wide Web* erfordert den Einsatz von Hilfsmitteln (vgl. Gudivada et al., 1997, Arasu et al., 2001, und Hu et al., 2001). Für diesen Zweck ist entweder ein hierarchischer Themenkatalog wie z. B. *Yahoo!* (`http://www.yahoo.com`) zu konsultieren oder Schlüsselwörter, die das gewünschte Themengebiet charakterisieren, werden in eine Suchmaschine wie *Google* (Brin und Page, 1998) eingegeben. Beide Methoden sind mit Einschränkungen verbunden: Themenkataloge werden von bezahlten Redakteuren gepflegt, die neue Dokumente in bestehende Kategorien einordnen oder neue Kategorien anlegen.[1] Eine auf diese Weise aufgebaute Sammlung von Websites und HTML-Dokumenten ist bezüglich des gesamten WWW zwangsläufig unvollständig, weshalb zu einem gesuchten Thema oftmals keine korrespondierende Kategorie existiert. Suchmaschinen hingegen traversieren das WWW automatisch, hangeln sich dabei von Hyperlink zu Hyperlink und streben die Integration eines möglichst großen Teils des WWW in ihre Volltextdatenbanken an. Ein wesentliches Problem sind die geradezu spartanischen Möglichkeiten, die Suchmaschinen zur Spezifizierung von Suchanfragen bieten: Die Angabe von Schlüsselwörtern und -phrasen resultiert oftmals in Listen mit hunderten oder gar tausenden Ergebnissen, so dass der Benutzer gezwungen ist, diese sukzessive auf ihre Relevanz für das gesuchte Thema zu überprüfen (vgl. z. B. Brown, 2004).[2] Aufgrund der Tatsache, dass eine Suchmaschine niemals das gesamte WWW abdecken kann (Gulli und Signorini, 2005), wurden Metasuchmaschinen entwickelt (Selberg, 1999). Diese reichen eine Anfrage an eine Reihe von Suchmaschinen weiter und präsentieren anschließend eine Ergebnisliste, die nach verschiedenen Kriterien aus den Treffern der einzelnen Suchmaschinen zusammengestellt wird. Obgleich durch die Verteilung einer Suchanfrage eine höhere Abdeckung erzielt wird, bieten auch Metasuchmaschinen keinerlei Möglichkeiten der weiterführenden Spezifizierung von Anfragen.

Der Einsatz von Katalogen und Suchmaschinen sowie der Erfolg einer Anfrage ist unter anderem von der Erfahrung und der Motivation des Anwenders abhängig (Slone, 2002). Einen weiteren Faktor stellt die Zufriedenheit mit der Qualität der Suchergebnisse und ihrer visuellen Aufbereitung dar (Su, 2003a,b). Benutzerstudien zeigen, dass das in Suchmaschinen verwendete Modell der Informationsrecherche gerade für diejenigen Anwender, die zwar mit Computern, nicht jedoch mit dem WWW vertraut sind, zu Missverständnissen und fehlerhaften Interpretationen führt. Es fällt Benutzern z. B. häufig schwer, für einen Informationsbedarf eine Liste präziser Suchwörter zu konstruieren (Pollock und Hockley, 1997). Analysen der Protokolldateien einer Suchmaschine bestätigen dies (Spink et al., 2001): Anfragen sind in den meisten Fällen sehr kurz (vgl. Jansen und Pooch, 2001) und werden kaum iterativ modifiziert oder reformuliert, zudem wird die „advanced search“ nur selten eingesetzt, obwohl sie von Experten gewinnbringend angewendet werden kann (Lucas und Topi, 2002, Dennis et al., 2002). Ford et al. (2001) interpretieren ähnliche Befunde als sehr zurückhaltenden und vorsichtigen Einsatz von Suchmaschinen, der von deren Betreibern durch Funktionen wie z. B. „I'm feeling lucky!“ bzw. „Auf gut Glück!“ zusätzlich forciert wird.[3] Darüber hinaus

[1] Eine Ausnahme stellt das *Yahoo!*-ähnliche *Open Directory Project* dar (`http://www.dmoz.org`), das von Freiwilligen nach dem Open-Source-Prinzip (DiBona et al., 1999, und Rehm und Lobin, 2003) gepflegt wird.

[2] Aus Gründen der besseren Lesbarkeit werden in der vorliegenden Arbeit maskuline Formen wie z. B. Benutzer, Leser, Autor, Rezipient und Produzent verwendet; diese implizieren immer auch die weibliche Form.

[3] Diese auf der Einstiegsseite von *Google* verfügbare Funktion überspringt die Seite mit den Suchergebnissen und leitet den Benutzer unmittelbar zum ersten Treffer.

Abbildung 1.1: Restriktion von Suchanfragen auf Instanzen spezifischer Hypertextsorten

tendieren die Anwender dazu, lediglich die ersten zwei Ergebnisseiten zu betrachten (Spink et al., 2001), so dass möglicherweise relevante Dokumente aufgrund einer schlechten Platzierung nicht wahrgenommen werden. Die Studien zeigen, dass ein Bedarf besteht, Suchmaschinen mit weiterführenden Recherchemöglichkeiten auszustatten, die von den Benutzern intuitiv eingesetzt werden können, um bereits in der Liste der ersten zehn oder 20 Treffer möglichst viele für eine Suchanfrage relevante Dokumente präsentieren zu können.

Die vorliegende Arbeit knüpft an diese Problematik an und basiert auf der Hypothese, dass im *World Wide Web* konventionalisierte Typen von Hypertexten existieren, die im Folgenden unter Rückgriff auf die in der Textlinguistik etablierte Terminologie als *Hypertextsorten* bezeichnet werden. Ihre Entwicklung unterliegt einem zyklischen Prozess, der mit der steigenden Zahl von Anwendern und einem immer umfangreicheren Einsatz des WWW in Forschung, Wirtschaft sowie bei privaten Benutzern einhergeht. Sollten Hypertextsorten distinktive Merkmale besitzen, die eine maschinelle Identifizierung erlauben, wären sie für einen Einsatz im Rahmen einer Suchmaschine ideal geeignet. Abbildung 1.1 veranschaulicht eine derartige Anwendung mit einer fiktiven Suchmaschine, die die Benutzer in die Lage versetzt, neben der Angabe von Stichwörtern auch die gewünschten Hypertextsorten der zu findenden HTML-Dokumente spezifizieren zu können. Mittels einer Filterkomponente könnte die Suchmaschine somit lediglich diejenigen Dokumente zurückliefern, die sowohl die angegebenen Stichwörter enthalten als auch den markierten Hypertextsorten entsprechen. Im Beispiel wird nach Dokumenten gesucht, die die Stichwörter *Hypertext*, *Textsorten*, *WWW* und *Hierarchie* enthalten und den Hypertextsorten *Lehrveranstaltung*, *Homepage* oder *wissenschaftlicher Artikel* zugehörig sind.[4] Nur Dokumente, die *beiden* Kriterien genügen, werden dem Benutzer als Ergebnis der Anfrage präsentiert, um die unter Umständen sehr große Anzahl von Treffern auf eine Teilmenge der relevanten Dokumente einzuschränken.

[4] Es handelt sich um eine fiktive Benutzerschnittstelle, die lediglich der Illustration dient. Die in Abbildung 1.1 präsentierten Hypertextsorten sind ebenfalls als intuitiv zugängliche Beispiele zu verstehen.

Zum Thema der maschinellen Identifizierung von Textsorten liegen verschiedene Arbeiten vor, die sich auf den in der angelsächsischen Literatur etablierten Genre-Begriff beziehen. Beispielsweise schwebt Karlgren und Cutting (1994) eine Sammlung von verschiedenartigen und unterschiedlich mächtigen Werkzeugen zur Filterung vor, die im Rahmen eines *Information Retrieval*-Systems zur Einschränkung des Suchraums eingesetzt werden können:

> In an envisioned application, a user will employ a cascade of filters starting with filtering by topic, and continuing with filters by genre or text type, and ending by filters for text quality, or other tentative finer-grained qualifications. (Karlgren und Cutting, 1994, S. 1073)

Kessler et al. (1997) führen ebenfalls verschiedene grundlegende Experimente zur Klassifikation von Texten in sechs traditionelle Genres durch und betonen die Problematik der Anwendung und Übertragbarkeit derartiger Verfahren in Bezug auf das *World Wide Web*:

> To a large extent, the problems of genre classification don't become salient until we are confronted with large and heterogeneous search domains like the World-Wide Web. (Kessler et al., 1997, S. 32)

Die Heterogenität und der extreme Umfang des *World Wide Web* werden Kessler et al. zufolge zahlreiche Probleme verursachen, jedoch stellt das WWW zugleich eine sehr intuitive Perspektive für eine gänzlich neue sprachtechnologische Anwendung dar, die über die etablierten Methoden hinausgeht (insbesondere die thematische Kategorisierung von Texten, Informationsrecherche, Informationsextraktion und automatisches Textzusammenfassen). Die maschinelle Identifizierung der Hypertextsorten gegebener HTML-Dokumente ist als Komplement zur inhaltlichen Kategorisierung zu verstehen.

Während das Konzept der Hypertextsorte im deutschsprachigen Raum bislang nahezu keine Beachtung gefunden hat, existiert eine Vielzahl von Arbeiten, die sich mit „digital genres" beschäftigen, jedoch nur punktuelle Aspekte untersuchen. In vielen dieser Arbeiten wird das Potenzial einer „genre-enabled search engine" hervorgehoben (vgl. z. B. Roussinov et al., 2001, und Rosso, 2005). Crowston und Kwasnik (2004) zufolge kann die Berechnung der Textsorten von Dokumenten, die sich in der Datenbank eines IR-Systems befinden (vgl. auch Renear, 1997), mehrere Prozesse gewinnbringend unterstützen. Hierzu zählt die Formulierung einer Suchanfrage, der Abgleich der Anfrage mit den internen Dokumentrepräsentationen und die Präsentation der Ergebnisse. Haas und Grams (2000) verwenden statt der Begriffe „digital genre" oder „Web genre" den Terminus „page type" und erläutern:

> We can consider a page type as an additional dimension along which to constrain a search. A single topic such as "celluar phones" could receive many different treatments; allowing the user to additionally specify that he or she is looking for documentation, a picture, or an order form can make the search more successful (or at least less frustrating). (Haas und Grams, 2000, S. 186)

An anderer Stelle konkretisieren Haas und Grams (2000, S. 190) die Schwierigkeit eines derartigen Vorhabens: Beim WWW handelt es sich um ein „quickly moving target", das auf Technologien basiert, die einer kontinuierlichen Weiterentwicklung unterliegen, so dass die

Analyse von Genres und Gestaltungskonventionen als „tricky“ eingeschätzt wird (ebd.). Dass zahlreiche Konventionen existieren, belegt die Ratgeberliteratur zu den Themen Webdesign und Web-Usability, die darauf hinweist, Websites so zu gestalten, dass sie den Erwartungshaltungen der Nutzer entsprechen und ihre Bedürfnisse optimal unterstützen. Die Voraussetzung für die Existenz spezifischer Erwartungshaltungen ist jedoch, dass z. B. bezüglich der Platzierung von Logografiken, der Beschriftung von Hyperlinks oder der Gestaltungsweise unterschiedlicher Dokumenttypen Konventionen existieren, deren Bruch von den Anwendern wahrgenommen wird. Im WWW herrschen jedoch keine expliziten Standards, d. h. die Produzenten unterliegen bezüglich der Anfertigung von Webauftritten lediglich technischen Restriktionen, die etwa die Möglichkeiten von HTML und ihre Umsetzung in Browsern betreffen. Crowston und Williams (1997) formulieren diesen Umstand wie folgt:

> [T]here is no explicit management or enforcement of genres [...]. Instead, individual Web site developers individually choose how to present their information, drawing on their understanding as members of a community, what Orlikowski et al. called implicit structuring. (Crowston und Williams, 1997, S. 32)

Hypertextsorten sind also nicht als normierende Standards aufzufassen. Stattdessen unterliegen sie einem zyklischen Entwicklungsprozess, der auf der Beobachtung von Webangeboten, der Identifizierung von Regeln und Konventionen und der Anfertigung und Pflege eigener Webangebote unter Einbeziehung eben dieser Regeln und Konventionen beruht. Auf diese Weise entstehen innerhalb bestimmter Diskursgemeinschaften konventionalisierte Strukturen, die als Hypertextsorten konzeptualisiert werden können.

1.2 Forschungsstand und beteiligte Disziplinen

In die vorliegende Untersuchung des Bereiches Hypertextsorten sind die vier Themengebiete (i) *Hypertext*, (ii) *World Wide Web*, (iii) *Linguistik* und (iv) *Sprachtechnologie* involviert, deren Bezüge im Folgenden dargestellt werden (vgl. Abbildung 1.2).

Hypertext Bei dem Themengebiet Hypertext handelt es sich um keine klar abgegrenzte wissenschaftliche Disziplin, sondern um einen Bereich, der unter anderem Berührungspunkte zur Linguistik, Informatik, Sprachtechnologie und *Computers and the Humanities*, d. h. dem Einsatz von Rechnern für Fragestellungen in geisteswissenschaftlichen Disziplinen aufweist. Hypertext wird an dieser Stelle als eigenes Gebiet aufgeführt, da es als etablierter Forschungsgegenstand aufgefasst werden kann (Kuhlen, 1997, S. 366). Abstrakt formuliert besteht ein Hypertext aus Knoten, die textuelle oder multimediale Inhalte umfassen. Der Autor eines Hypertextes legt Verknüpfungen an, über die ein Rezipient unmittelbar zu anderen Knoten gelangen kann. Derartige Hyperlinks werden typischerweise visuell hervorgehoben, so dass sie leicht identifizierbar sind. Ein besonderes Merkmal sehr großer Hypertexte sind die vielfältigen Verknüpfungen, die von Knoten zu Knoten führen. Hierdurch können Informationseinheiten auf multilineare Weise organisiert werden. Im Gegensatz zur vornehmlich linearen Sequenzierung, die z. B. in Büchern und Zeitungsartikeln vorliegt, ermöglicht es die multilineare Organisation, innerhalb eines Knotens durch das Aktivieren eines Hyperlinks

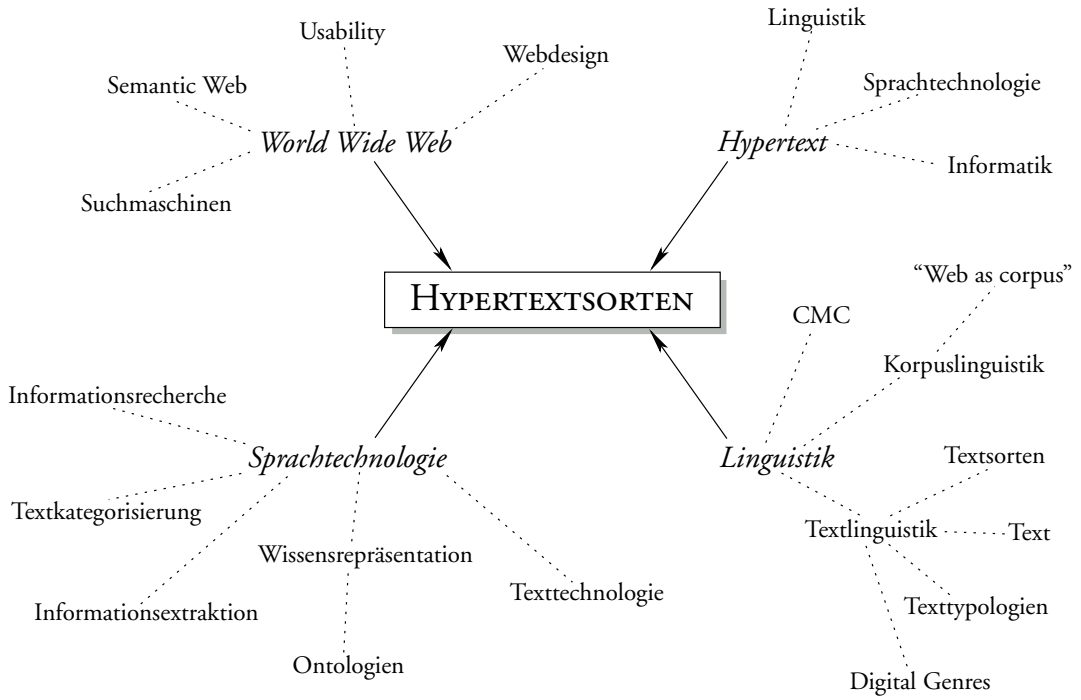


Abbildung 1.2: Die zentralen Themengebiete der vorliegenden Arbeit

z. B. weitere Details zu einem spezifischen Terminus, ein erläuterndes Beispiel, ein verwandtes Thema oder ein Glossar aufzurufen, die sich wiederum in weiteren Knoten befinden. Es existiert eine Vielzahl von Hypertextsystemen; die vorliegende Arbeit bezieht sich auf die mit Abstand meistgenutzte Implementierung – das *World Wide Web*.

Das World Wide Web Das WWW (Berners-Lee et al., 1992, 1994) stellt im Hinblick auf die von der Hypertextforschung vorgeschlagenen Konzepte zwar ein Hypertextsystem mit eingeschränkter Funktionalität dar, gleichwohl handelt es sich um das mit Abstand meistgenutzte System. Das WWW ist, ebenso wie z. B. FTP, IRC oder E-Mail ein Internet-Dienst, der auf dem Client-Server-Paradigma basiert: Ein entfernter Rechner, der Webserver, stellt Dokumente zur Verfügung, die über das Netzwerk von beliebigen Clients angefordert werden können. Die Dokumente werden mit Hilfe der *Hypertext Markup Language* ausgezeichnet und von Browsern nach der Übermittlung dargestellt; die grafische und typografische Aufbereitung hängt vom verwendeten Programm bzw. Endgerät ab. Jede auf einem Webserver befindliche Datei ist über eine eindeutige Adresse referenzierbar (URL) und eben diese URLs werden in HTML-Dokumenten zur Realisierung von Hyperlinks eingesetzt.

Linguistik Es liegt eine Vielzahl von Arbeiten vor, die sich dem Thema Hypertext aus linguistischer Perspektive widmen. Hypertext wird oftmals aufgrund der multilinearen Organisation als eine Erweiterung von Text, als eine besondere Form von Text mit spezifischen Charakteristika aufgefasst. Interessanterweise finden sich nur in den wenigsten Arbeiten ex-

plizite Bezüge zu Hypertextsorten (vgl. Jakobs, 2003). Vielmehr wird das Konzept Hypertext selbst in vielen Fällen als Textsorte bezeichnet (vgl. z. B. Weingarten, 1997b, S. 217, Schmitz, 1997, S. 146, Runkehl et al., 1998, S. 159, Wagner, 1998, S. 195, Schmitz, 2000, S. 265, und Heinemann, 2000d, S. 507). Im Gegensatz zu dieser Auffassung finden sich in einigen weiteren Studien implizite Indikatoren, die durchaus auf die Existenz unterschiedlicher Textsorten in Hypertextsystemen hinweisen, ohne diese jedoch genauer zu thematisieren.

Zu den Internet-Diensten E-Mail, *Internet Relay Chat* (IRC) und Usenet liegen zahlreiche linguistische Arbeiten vor, die dem Forschungsgebiet *Computer-Mediated Communication* (CMC) zugeordnet werden. Eine deutlich geringere Zahl von Studien beschäftigt sich mit dem *World Wide Web* und fokussiert dabei eher abstrakte Aspekte. Ausnahmen stellen beispielsweise die Arbeiten von Bittner (2003), der sich mit privaten Homepages beschäftigt, und Schütte (2004a) dar, die die Einstiegsseiten der Webauftritte russischer und deutscher Unternehmen analysiert und vergleicht. Zu den Ursachen für die eher geringe Beachtung des *World Wide Web* zählt zunächst die Problematik der Datenerhebung: HTML-Dokumente werden häufig aktualisiert, so dass es notwendig ist, die Webseiten zu fixieren. Hierfür bietet sich ein Ausdruck an, wodurch jedoch zahlreiche interaktive und multimediale Merkmale verloren gehen. Die zweite Möglichkeit besteht in der Speicherung der Dokumente auf einem lokalen Rechner, wofür spezialisierte Werkzeuge einzusetzen sind. Eine weitere Problematik betrifft die Vielschichtigkeit der Daten: Neben Abschnitten von Fließtext existieren Listen, isolierte Textfragmente, eine Vielzahl von Bildtypen (Logos, Werbebanner, Fotos etc.), Überschriften, Navigationsleisten und interaktive Menüs, die in die linguistische Analyse eines HTML-Dokuments einzubeziehen sind. Durch Hyperlinks und die Bezugnahme auf weitere Dokumente einer Website verschärft sich diese Problematik zusätzlich.

Eine der zentralen Hypothesen, von denen die vorliegende Arbeit ausgeht, besagt, dass im *World Wide Web* Hypertextsorten existieren. Wenn z. B. in den Definitionen von Hartmann die Termini „Text“ durch „Hypertext“ und „Textsorte“ durch „Hypertextsorte“ ersetzt werden, können Hypertextsorten charakterisiert werden als „Mengen von *Hyper*texten mit bestimmten gemeinsamen Eigenschaften“ (Hartmann, 1964, S. 23) oder, präziser, als „Teilmengen von *Hyper*texten, die sich durch bestimmte relevante gemeinsame Merkmale beschreiben und von anderen Teilmengen abgrenzen lassen“ (Hartmann, 1971, S. 22). Die ursprünglichen Definitionen Hartmanns können nur als erste Näherung an das Konzept der Textsorte verstanden werden. „Im Grunde handelt es sich“, so Heinemann (2000c, S. 524), „um eine komplexe Problematik mit kognitiven, linguistischen und sozialen Aspekten.“ Folglich verwirft de Beaugrande (1997, S. 9) den einschränkenden Terminus „Textlinguistik“ und präferiert stattdessen „Textwissenschaft“, denn „[z]u viele der wichtigsten Fragen – auch bezüglich der Textstrukturen und Textformulierungen – sind nicht nur ›linguistisch‹ in dem Sinne, wie dieser Terminus allgemein ausgelegt wird.“

Von einer solchen übergeordneten Perspektive gehen mehrere Arbeiten aus, die seit etwa 1997 vornehmlich aus der US-amerikanischen und skandinavischen Forschung hervorgegangen sind und sich mit „digital genres“ beschäftigen. Bei diesen Studien handelt es sich primär um kurze Konferenzbeiträge, die spezifische Aspekte von „digital documents“, „digital genres“ und „Web genres“ betrachten und sich dabei auf die *North American Genre Theory* berufen (vgl. Miller, 1984, Swales, 1990, Yates und Orlikowski, 1992, Orlikowski und Yates, 1994, Bazerman, 1994). Diese fasst Genres als primär funktional markierte Mus-

ter der Kommunikation auf, die der Realisierung spezifischer Ziele innerhalb rekurrenter Kommunikationssituationen dienen, welche sich auf bestimmte Diskursgemeinschaften beziehen. Obgleich in diesen Beiträgen zahlreiche Merkmale von Hypertextsorten herausgearbeitet wurden, müssen sie doch letzten Endes als isolierte Einzelarbeiten eingestuft werden, die noch keine einheitliche Forschungslinie hervorgebracht haben. Ein wesentliches Manko sämtlicher Arbeiten ist ihr mangelnder Theoriebezug: Ein Web-Genre wird als Tripel definiert, das aus „content", „form" und „function" besteht – Binnendifferenzierungen oder unterschiedliche Ausprägungen der Ebenen werden, in sehr abstrakter Form, nur in wenigen Fällen vorgenommen. Zudem beschränkt sich der Geltungsbereich nahezu ausschließlich auf das Einzeldokument, d. h. ein Web-Genre wird durch genau ein HTML-Dokument realisiert.

Sprachtechnologie Bezüglich der maschinellen Identifizierung und Repräsentation von Hypertextsorten betrifft der Bereich der Sprachtechnologie insbesondere Methoden aus der Computerlinguistik, der Texttechnologie und der KI-Forschung. Die Ansätze von Karlgren und Cutting (1994) und Kessler et al. (1997) zur maschinellen Kategorisierung von Texten in ihre Genres wurden bereits in Abschnitt 1.1 angesprochenen. Diese basieren unter anderem auf Frequenzangaben von Interpunktionszeichen und *Part-of-Speech*-Informationen. Im Hinblick auf das *World Wide Web* wurden verschiedene Arbeiten zur „automatic genre detection of Web documents" (Lim et al., 2005a,b) vorgelegt, die zusätzliche Merkmale integrieren, die sich z. B. auf HTML-Elemente, Schlüsselwörter und eingebettete Bilder beziehen. Sowohl die von Lim et al. (2005a,b) präsentierten Experimente als auch die verwandten Arbeiten (vgl. z. B. Karlgren et al., 1998, Matsuda und Fukushima, 1999, und Lee und Myaeng, 2002, 2004) zeigen bezüglich des Ziels der Realisierung einer Suchmaschine, die die Identifizierung von Hypertextsorten unterstützt, zwar sehr Erfolg versprechende Ergebnisse auf, jedoch verwenden sie Inventare von Web-Genres, die in ihrem Umfang sehr eingeschränkt sind. Zudem mangelt es auch ihnen an einer theoretischen Fundierung, so dass die realen Gegebenheiten, die im *World Wide Web* existieren, nicht reflektiert werden.

Die maschinelle Detektion von Hypertextsorten bezieht sich auf mehrere Teilgebiete der Sprachtechnologie, wobei zwischen der Repräsentation von Hypertextsorten und ihrer Identifizierung unterschieden werden kann. Webseiten werden mit Hilfe der *Hypertext Markup Language* ausgezeichnet (Raggett et al., 1999). Bei HTML handelt es sich um eine Anwendung der Metasprache *Standard Generalized Markup Language* (ISO 8879), die die Spezifizierung beliebiger Auszeichnungssprachen erlaubt, deren Definitionen auch als Dokumentgrammatiken (DTD, *Document Type Definition*) bezeichnet werden. Diese stellen die konzeptionelle Basis der Texttechnologie dar (vgl. Lobin und Lemnitzer, 2004), die „die linguistisch motivierte Informationsanreicherung und Verarbeitung digital verfügbarer Texte mit standardisierten Auszeichnungssprachen fokussiert." (Rehm, 2004e, S. 138). Eine DTD modelliert oftmals die erlaubten Textstrukturmuster der Exemplare einer spezifischen Textsorte (ein `Buch` besteht aus mehreren Komponenten vom Typ `Kapitel`, die sich wiederum aus einem `Titel` und mehreren `Absatz`-Elementen zusammensetzen). In diesem Sinne ist HTML eine untypische Auszeichnungssprache, da sie sehr heterogene Elemente aufweist, die nicht unmittelbar den Konstituenten einer Textsorte entsprechen. Es ist jedoch möglich, HTML-Dokumente mit weiteren Auszeichnungsebenen anzureichern, so dass die implizit vorhandenen Strukturen expliziert werden können. Weitere Berührungspunkte der maschinellen Iden-

tifizierung von Hypertextsorten beziehen sich auf die Bereiche Informationsextraktion und Textkategorisierung. Diese beschäftigen sich mit der Ermittlung spezifischer Informationen aus Texten und der Einordnung von Texten in vorgegebene Kategorienschemata.

1.3 Zielsetzungen und Forschungsfragen

In Anlehnung an die in der Textlinguistik vorgenommene Differenzierung zwischen Text und Hypertext kann im Zusammenhang der Textsortendiskussion die Existenz von Hypertextsorten im *World Wide Web* postuliert werden. Falls Hypertextsorten distinktive Merkmale besitzen, können sie mit sprachtechnologischen Verfahren identifiziert und in *Information Retrieval*-Systemen nutzbar gemacht werden. Die vorliegende Arbeit verfolgt hinsichtlich des Themenkomplexes Hypertextsorten mehrere Zielsetzungen, die sich der Problematik aus textlinguistischer Perspektive nähern, um den Einsatz abstrakter Repräsentationen von Hypertextsorten in computerlinguistischen und texttechnologischen Anwendungen und sprachtechnologischen Produktionssystemen zu untersuchen.

Eines der Ziele betrifft den Prozess der maschinellen Identifizierung von Hypertextsorten. Diesbezüglich liegen, wie bereits im vorangegangenen Abschnitt angesprochen, erste Arbeiten vor. Diese gehen primär von den eingesetzten Methoden und der Technologie aus und nehmen Inventare von Web-Genres an, die zwischen zwei und 15 Kategorien umfassen und sich jeweils auf die Ebene des einzelnen HTML-Dokuments beziehen. Für die Kategorisierung werden spezifische Merkmalsausprägungen eingesetzt, die in manuell kategorisierten Trainingsdaten beobachtet werden, um mit Hilfe maschineller Lernverfahren Klassifikatoren zu generieren, die durch Merkmalsvergleiche unbekannte Dokumente in ihre Web-Genres sortieren. Sämtliche dieser Arbeiten weisen zwei problematische Aspekte auf, die eine unmittelbare Übertragbarkeit ihrer Ergebnisse auf das Ziel der Realisierung einer Suchmaschine, die in der Lage ist, Hypertextsorten maschinell zu identifizieren, verhindern. Hierzu zählen die sehr eingeschränkten Inventare von Web-Genres, die mit beispielsweise „home page“, „FAQ“, „link collection“ und „research reports“ nur eine sehr geringe Zahl von Kategorien enthalten, und die mangelnde theoretische Fundierung. In der vorliegenden Arbeit wird ein umgekehrter Ansatz verfolgt: Zunächst ist – und zwar vollkommen unabhängig von der Problematik der Implementierung eines Systems zur maschinellen Kategorisierung – die Frage zu klären, welche Charakteristika Hypertextsorten aufweisen, welche Konstituenten sie besitzen und was sie von Textsorten unterscheidet. Es wird also ein textlinguistisch ausgerichtetes Hypertextsortenmodell benötigt, das in der Lage ist, die realen Gegebenheiten im *World Wide Web* adäquat zu beschreiben.[5] Eine weitere Frage, die zwar nicht für einen Forschungsprototypen, sehr wohl jedoch für ein Produktionssystem von essenzieller Bedeutung ist, betrifft die Frage nach der Anzahl von Hypertextsorten im WWW. Wenn eine Suchmaschine in der Lage sein soll, die Hypertextsorten gegebener Dokumente identifizieren zu

[5] Der von Kessler et al. (1997) vorgelegte Ansatz zur maschinellen Kategorisierung bezieht sich ausschließlich auf traditionelle Textsorten. Die Verfasser betonen ebenfalls den zentralen Stellenwert einer theoretischen Fundierung: „In order to do systematic work on automatic genre classification, [...] we require the answers to some basic theoretical and methodological questions.“ (ebd., S. 32). In Bezug auf Hypertextsorten liegen auf diese Fragen bislang kaum Antworten vor.

können, ist es notwendig, zunächst eine Bestandsaufnahme in Form eines Katalogs von Hypertextsorten vorzunehmen. Je umfangreicher ein derartiges Inventar ist, desto unpräziser operieren Ansätze zur maschinellen Kategorisierung. Es ist demnach zu hinterfragen, wie die realen Gegebenheiten im *World Wide Web* beschaffen sind, mit wie vielen Hypertextsorten ein Produktionssystem (versus den bislang präsentierten prototypischen Systemen) also umzugehen in der Lage sein muss. Die vorliegende Arbeit vertritt den Standpunkt, dass zunächst eine textwissenschaftliche Untersuchung des Phänomens Hypertextsorte vorzunehmen ist. Es sind genauere Erkenntnisse über Hypertextsorten, ihre Merkmale, Konstituenten und Geltungsbereiche zu ermitteln, woraufhin empirische Verfahren einzusetzen sind, um gegebene HTML-Dokumente bezüglich ihrer Hypertextsorten zu analysieren. Antworten auf die angesprochenen Fragenkomplexe werden zwingend benötigt, um ein Erkennungssystem konzeptionieren zu können. Ein weiteres Ziel bezieht sich auf die Typologisierung von Hypertextsorten, die eine enge Verbindung zu der bereits angesprochenen Bestandsaufnahme aufweist: Wenn eine Sammlung von Hypertextsorten vorliegt, stellt sich die Frage, ob und wie diese typologisiert werden können. Die Typologisierung hebt abermals den textlinguistischen Bezug hervor und ist für die maschinelle Repräsentation von Hypertextsorten mit texttechnologischen Methoden von Bedeutung: Eine derartige maschinenlesbare Repräsentation von Hypertextsorten ist als zentrale Ressource in einem Erkennungssystem vorzusehen. Abstrakt formuliert wird die Untersuchung von Hypertextsorten in der vorliegenden Arbeit zwischen den Polen der textlinguistischen Beschreibung und ihrer computerlinguistischen Anwendbarkeit aufgespannt.

1.4 Die Arbeit im Überblick

Die vorliegende Arbeit besteht aus vier Teilen und insgesamt 15 Kapiteln, deren Strukturierung und Inhalte im Folgenden vorgestellt werden. Abbildung 1.3 vermittelt diesbezüglich einen grafischen Überblick.

Teil I *Hintergrund und theoretische Grundlagen* Die Kapitel des ersten Teils stellen kritische Bestandsaufnahmen zu den Bereichen Textlinguistik, Hypertext und Digital Genres dar und knüpfen mit Erweiterungen an verschiedene Aspekte an.

Kapitel 2 *Text und Textsorten* Das zweite Kapitel geht auf linguistische Beschreibungsebenen von Texten und die wesentlichen textlinguistischen Grundbegriffe der Textsorten und Texttypen ein und diskutiert die Typologisierung von Textklassen. Darüber hinaus umfasst dieses Kapitel eine Kurzeinführung in die *North American Genre Theory*.

Kapitel 3 *Hypertext und das World Wide Web: Die linguistische Perspektive* Nach einem kurzen historischen Überblick über die Entwicklung des Konzepts Hypertext stellt dieses Kapitel die wesentlichen theoretischen und technologischen Merkmale von Hypertext anhand des Hypertextsystems *World Wide Web* dar. Anschließend werden – analog zu Kapitel 2 – die in der Literatur vorgeschlagenen linguistischen Beschreibungsebenen von Hypertexten thematisiert. Es wurde bereits angesprochen, dass Hypertext in textlinguistischen Arbeiten oftmals als Textsorte bezeichnet wird. Nicht nur Publikationen

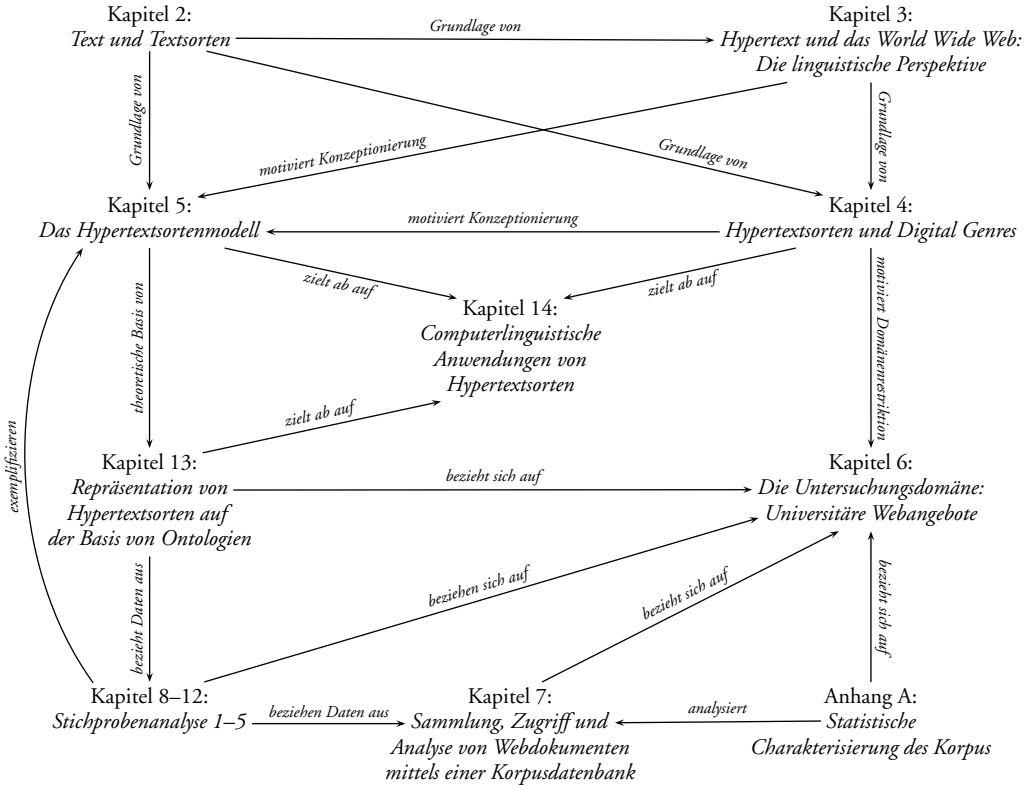


Abbildung 1.3: Die Bezüge zwischen den Kapiteln im Überblick

aus dem textlinguistischen Bereich enthalten Hinweise, die die Existenz von Hypertextsorten nahe legen. Das Kapitel schließt mit einer Sammlung derartiger Indikatoren, aus denen zwei rudimentäre Typologien abgeleitet werden, die sich auf die Ebenen der Hypertextsorten und der Hypertextknotensorten beziehen.

Kapitel 4 *Hypertextsorten und Digital Genres* Das vierte Kapitel setzt an dem Bereich der Digital Genres an und greift die Frage nach der Differenzierung zwischen Medium und Textsorte auf. Daraufhin werden zwei zyklische Modelle eingeführt, die die Entwicklung von Hypertextsorten erklären und sich auf die manuelle und maschinelle Erstellung von HTML-Dokumenten beziehen. Anschließend werden die Methodologien und Ergebnisse von Studien präsentiert, in denen die Sammlung von Web-Genres angestrebt wurde. Eine Schlussfolgerung der Diskussion dieser Arbeiten, in denen vornehmlich Stichproben von HTML-Dokumenten untersucht wurden, betrifft die Notwendigkeit der Einschränkung der Untersuchungsdomäne. Nach einer Erörterung der Kerneigenschaften von Hypertextsorten werden die Ergebnisse derjenigen verwandten Arbeiten vorgestellt, in denen spezifische Hypertextsorten detailliert untersucht werden. Diese Ausführungen stellen eine der Grundlagen von Teil III dar.

Teil II *Das Rahmenmodell und die Methodologie* Die Kapitel des zweiten Teils stellen das Hypertextsortenmodell, die Untersuchungsdomäne sowie das für die vorliegende Arbeit aufgebaute Korpus und eine Korpusdatenbank vor.

Kapitel 5 *Das Hypertextsortenmodell* Dieses Kapitel präsentiert das Hypertextsortenmodell der vorliegenden Arbeit. Neben der textlinguistischen Ausrichtung ist es auf einen gewinnbringenden Einsatz in computerlinguistischen und texttechnologischen Anwendungen ausgelegt. Das Hypertextsortenmodell differenziert zwischen den drei Ebenen Hypertexttyp bzw. -sorte, Hypertextknotentyp bzw. -sorte und Hypertextsortenmodul. Es stellt die zentrale theoretische Grundlage der Teile III und IV dar und bezieht sich nicht nur auf die in Kapitel 6 dargestellte Untersuchungsdomäne, sondern auf beliebige Gebrauchshypertextsorten.

Kapitel 6 *Die Untersuchungsdomäne: Universitäre Webangebote* Die vorliegenden Analysen zur Sammlung von Hypertextsorten beziehen sich nahezu ausschließlich auf die Untersuchung von Stichproben, die den Datenbeständen von Suchmaschinen entnommen wurden und somit *beliebige* Webseiten umfassen, weshalb die Ergebnisse als sehr heterogen und teilweise widersprüchlich einzustufen sind. Die in Teil III vorgestellten empirischen Analysen basieren hingegen auf Stichproben, die aus der Untersuchungsdomäne der deutschsprachigen HTML-Dokumente der Webauftritte deutscher Hochschulen und Universitäten stammen. Die Vorgehensweise ist in der Hypothese begründet, dass diese Restriktion homogenere und detailliertere Resultate ermöglicht. Die Charakteristika der Untersuchungsdomäne werden in diesem Kapitel erläutert.

Kapitel 7 *Sammlung, Zugriff und Analyse von Webdokumenten mittels einer Korpusdatenbank* Für die vorliegende Arbeit wurde ein Korpus aufgebaut, das aus ca. 4 000 000 HTML-Dokumenten besteht, die aus den Webauftritten von 100 deutschen Universitäten und Hochschulen stammen. Um sie zu fixieren, wurden die Dokumente auf einen lokal verfügbaren Webserver übertragen. Neben der Durchführung der Datensammlung erläutert dieses Kapitel die Funktionsweise einer Korpusdatenbank, die verschiedene Zugriffsmöglichkeiten erlaubt. Der indirekte Zugriff erfolgt durch autarke Analysemodule, der direkte Zugriff erfolgt durch eine Web-Oberfläche, die unter anderem die Exploration des Korpus und die interaktive Generierung und Analyse von Stichproben gestattet. Diese Zugriffs- und Analysemethoden stellen die Untersuchungsverfahren dar, die im nachfolgenden Teil III eingesetzt werden.

Teil III *Analysen und Sammlungen von Hypertextsorten* Der dritte Teil umfasst fünf Stichprobenanalysen, die zufällig zusammengestellte Dokumente beinhalten und spezifische Aspekte von Hypertextsorten, Hypertextknotensorten und Hypertextsortenmodulen untersuchen. Die Analysen fungieren als exemplarische Anwendungen des Hypertextsortenmodells (Kapitel 5) und beziehen sich auf die Domäne der universitären Webangebote (Kapitel 6). Alle untersuchten Dokumente stammen aus der Korpusdatenbank und wurden mit Hilfe der Web-Oberfläche bzw. mittels der indirekten Zugriffsmöglichkeiten analysiert (Kapitel 7).

Kapitel 8 *Analyse 1: Quantitative Auswertung persönlicher Homepages* Die von der Literatur bislang am umfangreichsten untersuchten Hypertextsorten sind persönliche bzw. private Homepages, weshalb sich die ersten drei Analysen diesem Thema widmen: Die erste Analyse präsentiert zunächst eine vergleichende Untersuchung, in der die Frequenzen maschinell identifizierbarer Merkmale für konzeptionelle Mündlichkeit einerseits in privaten Homepages von Studierenden und andererseits in persönlichen Homepages von Wissenschaftlern erhoben werden. Die ermittelten Häufigkeiten werden mit einer dritten Stichprobe verglichen.

Kapitel 9 *Analyse 2: Die private Homepage eines Studierenden* Innerhalb der zweiten Analyse werden 15 private Homepages von Studierenden einer Inhalts- und Makrostrukturanalyse unterzogen, um die beteiligten Hypertextsortenmodule und ihre Frequenzen zu ermitteln. Auf diese Weise kann der Grad der Konventionalisierung dieser Hypertextsorte bestimmt werden. Von besonderer Bedeutung sind dabei metadiskursive Äußerungen, d. h. von den jeweiligen Produzenten in ihre Homepages integrierte Anmerkungen, die sich unmittelbar auf die konventionalisierten Inhalte, Strukturen und Funktionen der untersuchten Hypertextsorte beziehen und als indirekter Beleg für das zyklische Modell der Entwicklung von Hypertextsorten aufgefasst werden (Kapitel 4).

Kapitel 10 *Analyse 3: Die persönliche Homepage eines Wissenschaftlers* In der dritten Analyse werden 100 persönliche Homepages von Wissenschaftlern – analog zu Kapitel 9 – hinsichtlich ihrer Funktionen, Inhalte und Makrostrukturkomponenten untersucht. Ebenso wie die zweite Analyse mündet auch diese Untersuchung in einem Hypertextsortenprofil, das die empirisch ermittelten Hypertextsortenmodule und ihre individuellen Merkmalsausprägungen beinhaltet. Das Kapitel schließt mit einer Typologie des Hypertexttyps Homepage einer Person.

Kapitel 11 *Analyse 4: Die Einstiegsseite des Webauftritts einer Universität* Die vierte Analyse umfasst zwei Phasen: Zunächst werden die – als institutionelle Homepages konzeptualisierbaren – Einstiegsseiten der Webauftritte von 35 Universitäten mit dem Ziel untersucht, sie mit kommerziell ausgerichteten Homepages zu vergleichen und ein Profil dieser Hypertextknotensorte zu erstellen. Die zweite Phase der Analyse fokussiert diejenigen Dokumente, die durch Hyperlinks von den 35 Einstiegsseiten erreichbar sind und basiert auf der Hypothese, dass die Sammlung und Identifizierung der korrespondierenden Hypertextknotensorten und zugehörigen Hypertextsorten ein geeignetes Werkzeug zur Konstruktion der „oberen“ Ebenen einer Hypertextsortentypologie für die Untersuchungsdomäne darstellt.

Kapitel 12 *Analyse 5: Untersuchung 750 zufällig ausgewählter Dokumente* Die fünfte Analyse komplementiert Kapitel 11, indem für 750 zufällig ausgewählte und „tief“ eingebettete Dokumente die korrespondierenden Hypertextknotensorten, übergeordneten Hypertextsorten und publizierenden Organisationseinheiten ermittelt werden, um somit die „unteren“ Ebenen der Typologie von Hypertextsorten genauer bestimmen zu können. Zum Abschluss des Kapitels wird die Problematik der Typologisierung von Hypertextsorten in detaillierter Form diskutiert.

Teil IV *Technologische Umsetzung* Der vierte und letzte Teil umfasst zwei Kapitel, die die technologische Umsetzung des Hypertextsortenmodells erläutern und auf die maschinelle Identifizierung von Hypertextsorten eingehen.

Kapitel 13 *Repräsentation von Hypertextsorten auf der Basis von Ontologien* Eines der Kernziele der vorliegenden Arbeit betrifft die Erstellung einer Typologie von Hypertextsorten für die Untersuchungsdomäne der universitären Webangebote. Dieses Kapitel stellt ein Repräsentationsformat für Hypertextsorten dar, das auf dem texttechnologischen Standardformalismus *Web Ontology Language* (OWL) beruht und geeignet ist, die beteiligten Konstituentenebenen zu erfassen. Für diesen Zweck werden die abstrakten Strukturen, die das Hypertextsortenmodell vorgibt (Kapitel 5), in Form einer Ontologie modelliert, in die die Ergebnisse aus Teil III eingearbeitet werden. Innerhalb der Hypertextsortenontologie werden zwei weitere Ontologien eingesetzt. Hierbei handelt es sich um eine Ontologie wissenschaftlicher Themen und Fachgebiete und um ein Domänenmodell, das den Aufbau einer generischen Hochschule beschreibt. Das Kapitel schließt mit einer Diskussion möglicher Anwendungsszenarien der Hypertextsortenontologie in sprachtechnologischen Systemen.

Kapitel 14 *Computerlinguistische Anwendungen von Hypertextsorten* Dieses Kapitel geht von den bislang vorliegenden Ansätzen zur Identifizierung von Web-Genres aus und verdeutlicht, weshalb sie nicht in der Lage sind, die im WWW existenten realen Gegebenheiten zu erfassen. Daraufhin wird die Architektur eines Systems zur Identifizierung von Hypertextsorten vorgestellt, die unter anderem auf der Hypertextsortenontologie basiert. Eine zentrale Problematik der maschinellen Identifizierung betrifft den Umstand, dass Hypertextsortenmodule in sehr flexibler Weise realisiert werden, so können z. B. einerseits *mehrere* Hypertextsortenmodule in *einem* HTML-Dokument aggregiert werden, andererseits können sie auch in jeweils eigenständige Knoten separiert werden. Aus diesem Grund ist sowohl die übergreifende Ebene des gesamten Hypertextes als auch die interne Struktur der Knoten einzubeziehen. Das Kapitel stellt einen Ansatz zur Analyse der Binnenstrukturierung eines HTML-Dokuments vor, um die enthaltenen Hypertextsortenmodule zu ermitteln. Daraufhin werden Methoden diskutiert, die für die vollständige und robuste Implementierung der vorgeschlagenen Architektur benötigt werden. Abschließend werden unterschiedliche Einsatzszenarien von Hypertextsorten in sprach- und informationstechnologischen Anwendungen dargestellt.

Kapitel 15 beendet die Arbeit mit Schlussfolgerungen und einem Ausblick. Der Anhang beinhaltet eine statistische Charakterisierung der im Korpus enthaltenen HTML- und XML-Dokumente (Anhang A), weiterführende Informationen zur Korpusdatenbank (Anhang B) und ein Verzeichnis häufig verwendeter Abkürzungen (Anhang C).

Des Weiteren liegt der Arbeit eine CD ROM bei. Sie umfasst Kerndaten der im Korpus enthaltenen universitären Webauftritte (Anhang D), Aufstellungen der in den Kapiteln 8, 11 und 12 analysierten Stichproben (Anhang E), zwei Bildschirmvideos zur Demonstration der Oberfläche des Textparsers (Anhang F) und eine PDF-Version von Kapitel 13, so dass die teils sehr detailliert visualisierten Ausschnitte der Ontologien dieses Kapitels in hochauflösender Form betrachtet werden können.

Teil I

Hintergrund und theoretische Grundlagen

Überblick

Die Einleitung hat verdeutlicht, dass an einer Untersuchung des Gegenstands der vorliegenden Arbeit mehrere Disziplinen beteiligt sind. Teil I widmet sich zunächst der Textlinguistik, insbesondere werden unterschiedliche linguistische Beschreibungsebenen von Texten sowie textlinguistische Termini erläutert und Ansätze zur Modellierung von Textsorten und zur Konstruktion von Texttypologien diskutiert. Diese Aspekte sind für eine textlinguistische Herangehensweise an Hyper*textsorten* von besonderer Relevanz. Kapitel 2 schließt mit einem Überblick über die *North American Genre Theory*, auf die sich nahezu alle verwandten Arbeiten beziehen. Kapitel 3 geht auf den ersten Bestandteil des Kompositums *Hypertext*sorten ein und beginnt mit einem historischen Abriss der Entwicklung des Konzepts Hypertext. Im Anschluss werden seine theoretischen und technologischen Merkmale anhand des Hypertextsystems *World Wide Web* erläutert, woraufhin textlinguistische Beschreibungsebenen und hypertextspezifische Eigenschaften thematisiert werden. Hypertextsorten finden gerade in der textlinguistisch ausgerichteten Hypertextliteratur nahezu keine Beachtung. Obwohl Hypertext in vielen Fällen selbst als Textsorte bezeichnet wird, können doch oftmals Indikatoren identifiziert werden, die die Existenz von Hypertextsorten nahe legen, von den jeweiligen Verfassern aber nicht näher thematisiert werden. Den Abschluss von Kapitel 3 stellt eine Sammlung derartiger Indikatoren aus unterschiedlichen Bereichen dar (unter anderem Textlinguistik, Webdesign, Informationsarchitektur und Metadatenschemata), aus denen daraufhin zwei initiale und rudimentäre Typologien von Hypertextsorten und Hypertextknotensorten abgeleitet werden können. Kapitel 4 geht zunächst auf unterschiedliche Ausprägungen des Konzepts Digital Genre ein. Daraufhin wird die Etikettierung sowie die Entstehung von Hypertextsorten diskutiert, für die zwei zyklische Modelle eingeführt werden, die sich auf die manuelle und die maschinelle Anfertigung HTML-basierter Hypertexte beziehen. Im Anschluss werden die Methodologien und Ergebnisse verschiedener Studien diskutiert, in denen die Sammlung von Web-Genres mittels der Analyse zufällig zusammengestellter Stichproben von Dokumenten angestrebt wurde. Nach einer Diskussion mehrerer Kerneigenschaften von Hypertextsorten schließt Kapitel 4 mit einer ausführlichen Darstellung der bislang durchgeführten Charakterisierungen von Hypertextsorten bzw. Web-Genres.

2
Text und Textsorten

2.1 Einleitung

Textsorten gehören zweifelsohne zu den zentralen Untersuchungsgegenständen der Textlinguistik. Eines der Kernziele der vorliegenden Arbeit ist die Entwicklung eines Hypertextsortenmodells am Beispiel des verteilten Hypertextsystems *World Wide Web*, weshalb die Frage, inwiefern textlinguistische Theorien zur Charakterisierung von Hypertextsorten eingesetzt werden können, von unmittelbarer Relevanz ist (vgl. auch Rada, 1991, S. 1).

Obgleich die Textlinguistik seit den sechziger Jahren insbesondere im deutschsprachigen Raum eine Vielzahl von Arbeiten zur Sammlung, Charakterisierung und Differenzierung von Textsorten hervorgebracht hat (vgl. zusammenfassend Gülich und Raible, 1972, 1977, Adamzik, 1995, und Brinker, 2001; Beispiele befinden sich z. B. in Göpferich, 1995, und Graefen, 1997), kann auch im neuen Jahrtausend bezüglich der Textsortendiskussion noch nicht von einem Theoriekonsens gesprochen werden:

> Die zahlreichen Einführungen [in die Textlinguistik, G. R.] belegen [eine] immer wieder zum Ausdruck gebrachte Unübersichtlichkeit des Forschungsfeldes, in dem eine große Anzahl von Analyseansätzen entwickelt wurde, die teilweise etwas verbindungslos nebeneinander stehen, mitunter auch nahezu inkompatiblen theoretisch-methodischen Grundsätzen verpflichtet sind [...]. (Adamzik, 2004, S. vii)

Heinemann und Viehweger (1991, S. 274) verdeutlichen, wie gering die Gemeinsamkeiten der unterschiedlichen Analyseansätze tatsächlich sind, denn Textlinguistik „ist keine einheitlich orientierte Wissenschaftsdisziplin, sondern eine Vielzahl älterer, teils auch neuerer Modellvorschläge, die in erster Linie durch das gemeinsame »Leitmotiv« Text zusammengehalten werden, nicht aber durch ein stringentes theoretisches oder methodologisches Programm.“ Antos und Tietz (1997b) konzentrieren sich in der mit „Quo vadis, Textlinguistik?“ (vgl. Abschnitt 1.3 in Heinemann und Viehweger, 1991, S. 83) überschriebenen Einleitung des Bandes *Die Zukunft der Textlinguistik* auf den Status Quo der Textsortenforschung und

kommen zu dem Schluss: „Selbst ein Herzstück der Textlinguistik – nämlich die Textsorten- bzw. Textmusterproblematik [. . .] – scheint heute eher auf der Stelle zu treten." (ebd., S. viii).

Wenn wir diese konstatierte Stagnation auf den Kontext der vorliegenden Arbeit beziehen, d. h. die textlinguistisch orientierte Untersuchung digitaler Medien, so ergibt sich unmittelbar ein weiteres Problemfeld, denn in diesem Bereich ist ein deutlicher Mangel an Vorarbeiten zu verzeichnen. Zwar liegen zu diesem Themenkomplex diverse Arbeiten vor, die sich mit unterschiedlichen Aspekten der computervermittelten Kommunikation (CMC) beschäftigen, doch beziehen sich diese zumeist auf synchrone oder asynchrone Kommunikationsmedien wie E-Mail oder IRC (*Internet Relay Chat*; vgl. etwa Runkehl et al., 1998, Beißwenger, 2001, Ziegler und Dürscheid, 2002). Linguistische oder textlinguistische Analysen, die sich mit dem ebenfalls sehr populären und mittlerweile vielleicht wichtigsten Internet-Dienst *World Wide Web* beschäftigen, können – zumindest im deutschsprachigen Raum – in quantitativer Hinsicht durchaus als Randerscheinung bezeichnet werden (vgl. etwa Storrer, 1999b, 2001b, Dürscheid, 2000, Rehm, 2002a, Huber, 2002, Bittner, 2003, Schütte, 2004a). Als Hypertextsystem ist das WWW eine mögliche Ausprägung des Konzepts *Hypertext*, das sich zwar insbesondere bezüglich des Aspekts kohärenzstiftender Mittel in der textlinguistischen Literatur niedergeschlagen hat, dort aber ebenfalls nur ein Nischendasein führt (vgl. etwa Fritz, 1999, Sager, 2000, Huber, 2002, Storrer, 1999a, 2003, 2004b). Vorarbeiten, die sich explizit in der Tradition der Textlinguistik mit dem Themenkomplex *Hypertextsorten* beschäftigen, existieren bis auf wenige Ausnahmen nicht.

Die von Skepsis gezeichnete Einstellung der Textlinguistik gegenüber dem Hypertextkonzept, und somit auch gegenüber dem WWW, manifestiert sich jedoch nicht nur in einem Mangel an Vorarbeiten, sondern wird auch implizit deutlich, beispielsweise in dem Beitrag „Textsorten in den Massenmedien" (Burger, 2000), der im bislang umfangreichsten Kompendium[1] der textlinguistischen Forschung (Brinker et al., 2000) erschienen ist:

> Der Umfang des Objektbereichs wird noch unbestimmter dadurch, daß wir mit einem neuen Typ von massenmedialen Texten rechnen müssen, dem Internet mit seinen „Hypertext"-Strukturen, das auf dem Wege ist, ein neues Massenmedium zu werden. Das Internet bringt neue Formen von „Text" hervor, es wird alte Textsorten transformieren und neue kreieren. (Burger, 2000, S. 615)

Natürlich ist Burger prinzipiell zuzustimmen, die Setzung der Begriffe Hypertext und Text in Anführungszeichen kann jedoch durchaus als Indiz einer gewissen Zurückhaltung gegenüber den Kommunikations- und Informationsdiensten des Internet interpretiert werden (vgl. hierzu auch Schmitz, 2000, S. 253–256).[2] Jakobs (2003, S. 232) präsentiert einen der wenigen Beiträge, in denen das Konstrukt der Hypertextsorte fokussiert wird. Sie betont ebenfalls den Mangel an Vorarbeiten und führt verschiedene Gründe für diesen Umstand an:

[1] Ein weiteres Indiz für die angesprochene Skepsis ist die Tatsache, dass der insgesamt etwa 1 800 Seiten umfassende Doppelband *Text- und Gesprächslinguistik* aus der Reihe *Handbücher zur Sprach- und Kommunikationswissenschaft* lediglich einen Artikel (Sager, 2000: „Hypertext und Hypermedia") enthält, der sich dem Thema Hypertext widmet (und auf das WWW *nicht* eingeht; stattdessen werden Aspekte hypermedialer CD ROMs diskutiert). Darüber hinaus weist der Index lediglich neun weitere Einträge für diesen Begriff auf.

[2] Storrer (2000b) bezeichnet die Annäherung der Textlinguistik an Hypertexte als „zaghaft und von Sorge um den Gegenstandsbereich geprägt" (S. 223, vgl. auch Hess-Lüttich, 1997, S. 127).

> Probleme der theoretischen Auseinandersetzung mit Hypertextsorten hängen u. a. mit der dazu notwendigen begrifflichen Klärung der Basisbegriffe Textsorte, Hypertext, Hypertextsorte zusammen. Obwohl sich [...] inzwischen so etwas wie eine Textsortenlinguistik herausgebildet hat, *wird nach wie vor zum Teil sehr kontrovers diskutiert, was eine Textsorte ist und wie sie beschrieben werden kann. Ähnliches gilt für den Begriff Hypertext. In Bezug auf Hypertextsorten verschärft sich das Beschreibungsproblem.* Unter anderem ist zu klären, ob sie mit gängigen Textsortenmodellen erfasst werden können oder eigener Ansätze bedürfen [...] sowie ob Hypertextsorten als mediale Schwestern bereits vorhandener Textsorten zu behandeln sind oder ob sie etwas genuin Neues bilden [...] und wie dieses Neue von der Sprachgemeinschaft reflektiert wird [...]. (Jakobs, 2003, S. 233; Hervorhebung hinzugefügt, G. R.)

Dieses Kapitel geht auf die textlinguistischen Grundlagen ein, die in den nachfolgenden Kapiteln zur Charakterisierung von Hypertextsorten und zur Entwicklung eines Hypertextsortenmodells verwendet werden und die die Grenzen der bislang vorgelegten Arbeiten in Bezug auf diesen Gegenstand verdeutlichen. Es ist erneut zu betonen, dass im deutschsprachigen Bereich keine substanziellen Vorarbeiten zum Themenkomplex Hypertextsorten vorliegen, die sich dem Gegenstand aus textlinguistischer Perspektive nähern.[3] Zunächst ist es notwendig, auf der Ebene der textlinguistischen Modelle und Theorien anzusetzen, d. h. eine aktuelle und breit gefächerte Darstellung der Textlinguistik vorzunehmen, um die potenziellen Anknüpfungspunkte zu Hypertextsorten zu diskutieren und eine konsistente Basis bezüglich der in der Literatur häufig inkonsistent verwendeten Termini zu schaffen.

Der sich anschließende Abschnitt 2.2 thematisiert unterschiedliche Beschreibungsebenen von Texten, woraufhin Abschnitt 2.3 auf Beschreibungsaspekte von Textsorten und den Aufbau von Texttypologien eingeht. Von zentraler Bedeutung sind hierbei hierarchische Beziehungen zwischen Textklassen unterschiedlicher Abstraktionsstufen, distinktive Merkmale von Textsorten, die Klassifizierung von Texten in Textsorten, mehrdimensionale Texttypologien und die Prototypikalität von Textexemplaren und Textsorten.

2.2 Beschreibungsebenen von Texten

Die Textlinguistik beschäftigt sich auf unterschiedlichen Beschreibungsebenen mit dem Phänomen Text (vgl. Brinker, 2001, S. 149). Dieser Abschnitt stellt auf den wesentlichen Dimensionen die relevantesten Charakterisierungs- und Analyseansätze vor und legt definitorische Grundlagen für die sich anschließenden Ausführungen.[4] Der Ebenenbegriff bezieht sich dabei nur partiell – insbesondere hinsichtlich der grammatischen, semantischen und pragmatisch-kommunikativen Ebenen – auf eine hierarchische Staffelung. Wie umfangreich die Rahmenfaktoren für textlinguistische Analysen sind, veranschaulichen die Kapitel *Textkonstitution I–IV* in Brinker et al. (2000), die in 23 Beiträgen auf spezifische Aspekte eingehen:

[3] Dürscheid (2004, S. 155) fordert, dass sich die Linguistik neuen Aspekten Internet-basierter Kommunikationsformen zuwenden sollte: „Statt weiter allgemeine Beobachtungen zur Sprache im Internet anzustellen, sollte künftig der Schwerpunkt auf die Analyse einzelner Text- und Diskursarten im Internet gelegt werden."

[4] Einige der vorgestellten Modelle stehen mittlerweile nicht mehr im Zentrum des Interesses, wurden jedoch aufgrund ihres bedeutenden Stellenwerts für die Ausrichtung der heutigen Textlinguistik aufgenommen. Dieser Abschnitt orientiert sich insbesondere an Heinemann und Viehweger (1991), Vater (1994, 2001), Gansel und Jürgens (2002), Heinemann und Heinemann (2002), Adamzik (2004) sowie Brinker et al. (2000).

Textkonstitution I: Voraussetzungen – beispielsweise „Logisch-semantische Voraussetzungen", „Situative Voraussetzungen" und „Kognitive Voraussetzungen"

Textkonstitution II: Grammatische Aspekte – unter anderem „Kohärenz und Kohäsion", „Anapher im Text" und „Textdeixis"

Textkonstitution III: Thematische und pragmatische Aspekte – z. B. „Textuelle Grundfunktionen", „Thema, Themenentfaltung, Makrostruktur", „Handlungsstrukturen von Texten" und die Darstellung unterschiedlicher Vertextungsmuster

Textkonstitution IV: Textproduktion – Textgestaltung – Textrezeption – „Phasen und Verfahren der Produktion schriftlicher Texte", „Text und Stil", „Der Zusammenhang von Text und Bild", „Das Verstehen schriftlicher Texte als Prozeß"

Aufgrund dieser Fülle potenziell beteiligter Gebiete und Disziplinen (neben der Linguistik bzw. Textlinguistik insbesondere die kognitive Psychologie, die Sprachphilosophie und die Sozialwissenschaften) wird deutlich, dass die nachfolgenden Darstellungen der unterschiedlichen Beschreibungsebenen keinen Anspruch auf Vollständigkeit erheben.

2.2.1 Die grammatische Ebene

Die Anfangsphase der Textlinguistik wurde bestimmt von der Auffassung von Texten als transphrastische Ganzheiten. Dem Text wurde eine Stellung zuteil, die auf den Modellen aufbauen, die bis in die späten sechziger Jahre hinein bei der – vornehmlich syntaktischen – Untersuchung des Satzes gewonnen worden waren (vgl. Adamzik, 2004, S. 17 ff.). Die Grundidee war, dass die Bedeutung und Struktur eines Textes durch die enthaltenen Sätze konstituiert wird. Diese Oberflächenstruktur wurde auf der grammatischen Ebene an vorhandene Analyseverfahren angebunden, was zur Entwicklung von Textgrammatiken als Fortführung von Satzgrammatiken geführt hat (vgl. Heinemann und Heinemann, 2002, S. 64).

Die Erweiterung von Einzelsatz- zu Mehrsatzgrammatiken erfolgt durch die Einführung eines neuen Symbols T. Mit Hilfe einer solchen Textregel, wie z. B. $T \rightarrow S_1 S_2 S_3$ expandieren einzelne Sätze zu einem Text (vgl. Isenberg, 1968, van Dijk, 1972, Lang, 1973). Als Hauptkriterium für Textualität wird dabei die Satzverknüpfungshypothese angesehen (Thümmel, 1979). Diese besagt, dass Texte als Satzfolgen charakterisierbar sind, zwischen denen Kohäsionsbeziehungen bestehen, denn nicht alle arbiträren Satzsequenzen können als Text bezeichnet werden. Von wesentlicher Bedeutung sind unterschiedliche Typen der Satzverknüpfung (kausal, modal, temporal etc.), die auch als Vertextungstypen bezeichnet werden (Isenberg, 1968). Diese werden realisiert durch Vertextungsmittel, die Kohäsion herstellen (z. B. Konjunktionen, Pronomina oder Artikel und Tempus). Heinemann und Viehweger (1991, S. 36) charakterisieren die Ursachen, die zum Scheitern des textgrammatischen Ansatzes geführt haben (vgl. auch Lang, 1973, und Thümmel, 1978):

> Das Ziel, eine Text-Erzeugungsgrammatik zu entwickeln, erwies sich bald als illusionär, da Texte nicht auf grammatische Erscheinungen reduzierbar sind und zudem das Explikandum Text […] zum Ausgangspunkt deduktiver Ableitungsprozesse gemacht wurde. […E]ine solche umfassende Textgrammatik [wäre] zwangsläufig überfrachtet und daher praktisch kaum noch handhabbar. (Heinemann und Viehweger, 1991, S. 36)

Im Abschlusskapitel ihrer Einführung machen Heinemann und Viehweger (1991, S. 277) den „naiven Zustand der 70er Jahre“, wie er sich in der Textgrammatik manifestiert hat, für das „methodologische Dilemma“ (ebd.) der Textlinguistik verantwortlich. Heinemann und Heinemann (2002, S. 68) charakterisieren die Beschränkungen eines textgrammatischen Ansatzes, der ihrer Ansicht nach nur „als Ausgangspunkt für Textkennzeichnungen aller Art angesehen werden [soll].“ Die Restriktionen umfassen unter anderem die einer Textgrammatik inhärente Auffassung von Texten als statisch strukturierte Einheiten, keine Verbindungen zu den beteiligten Interaktionspartnern und dass ein „Zugang [...] zur Erklärung des Funktionierens von Texten auf diese Weise nicht hergestellt werden [kann]“ (ebd.).

2.2.2 Die semantische Ebene

Die Textgrammatik geht davon aus, dass Texte Satzsequenzen sind, zwischen denen Kohäsionsbeziehungen herrschen. Die Kohäsion ist jedoch nur für die Vertextung auf der Oberfläche verantwortlich; bei ausschließlicher Betrachung der kohäsiven Elemente wird von der Kohärenz eines Textes abstrahiert. In der Tat spielt die semantische Ebene eine sehr viel wichtigere Rolle für das Textverstehen als die grammatische Ebene. Texte können durchaus Kohärenz aufweisen, ohne kohäsive Mittel zu enthalten. Die Textlinguistik hat unterschiedliche Ansätze zur semantischen Analyse von Texten erarbeitet.

Funktionale Satzperspektive – Thematische Progression

Die funktionale Satzperspektive beschreibt die Verteilung von Informationen in einzelnen Sätzen (vgl. Daneš, 1974c). Als Thema bezeichnet man dabei das, worüber etwas gesagt wird, Rhema umfasst, was über das Thema mitgeteilt werden soll. Daneš (1974a) wendet das Prinzip auf Texte an und postuliert thematische Progressionstypen, die einen Text als Abfolge von Themen betrachten. Entscheidend ist, dass sich nachfolgende Thema-Rhema-Einheiten in gewisser Weise auf die vorangehende Einheit beziehen, so dass die Analyse den Textfortschritt reflektiert. Daneš unterscheidet drei Progressionstypen: Bei der linearen thematischen Progression fungiert das Rhema der jeweils vorangehenden Thema-Rhema-Struktur als Thema der nachfolgenden Thema-Rhema-Sequenz. Zu Beginn eines jeden Textes befindet sich – unabhängig vom Progressionstyp – üblicherweise ein thematischer Satz. Liegt eine Progression mit durchlaufendem Thema vor, so besitzen alle Thema-Rhema-Strukturen eines Textes ein identisches Thema; das globale Thema wird immer wieder aufgegriffen und um rhematische Informationen ergänzt. Bei der Progression mit abgeleitetem Thema liegt keine Beziehung zwischen benachbarten Thema-Rhema-Strukturen vor, da sie von einem globalen Textthema abgeleitet werden. Die Analyse längerer Texte, in denen verschiedene Typen vorliegen, wird mit Hilfe von Kombinationen dieser Grundtypen durchgeführt.

Adamzik (2004, S. 119) zufolge ist „es nie gelungen [...], klare Kriterien zur Abgrenzung von Thema und Rhema zu entwickeln“. Heinemann und Viehweger (1991, S. 34) kritisieren, dass nur „Teilaspekte der Textstrukturierung“ erfasst werden. Weiterhin sei eine „Sequenzgrammatik“ (ebd.) nicht in der Lage, die Beziehungen zwischen Thema-Rhema-Strukturen und der semantischen Basisstruktur eines Textes zu erklären (vgl. Dressler, 1974). Heinemann und Heinemann (2002, S. 72) sind der Ansicht, dass die „Grundtypen nur selten in ›reiner‹ Form auf[treten]; sie sind im Grunde immer nur auf Teiltextstrukturen anwendbar“.

Propositionskomplexe und Makrostruktur

Das Propositionsmodell beschreibt Texte als „Folge von Propositionen, die durch interpropositionale Relationen miteinander verknüpft sind“ (Heinemann und Viehweger, 1991, S. 44). Diese Relationen gelten nicht nur zwischen Propositionen, sondern auch zwischen abstrakteren Einheiten wie z. B. Textteilen.[5] Die Makrostruktur wird über Regeln aus der Mikrostruktur abgeleitet, die aus semantischen Satzstrukturen besteht (van Dijk, 1980). Die Makrostruktur bezeichnet, wie Vater (1994, S. 87) anmerkt, „einen relativen Begriff, eine globale Struktur in bezug auf speziellere Strukturen“.[6] Neben der semantischen Makrostruktur existiert die pragmatische Makrostruktur, die aus den in einem Text enthaltenen Sprechakten besteht (Searle, 1969). van Dijk nimmt auch Superstrukturen an, die das Textthema, die Textbedeutung und die Textsortengebundenheit kennzeichnen. Als Beispiele werden „Narrative“, „Arguments“, „Scholarly Papers“ und „Newspaper Articles“ angeführt, d. h. Texttypen und -sorten besitzen individuelle Superstrukturen (vgl. Heinemann, 2000c, S. 530).

Nach van Dijk (1980) enthält ein Text verschiedene makrostrukturelle Ebenen. An oberster Stelle der semantischen Makrostruktur drückt eine Makroproposition das Textthema aus; die pragmatische Makrostruktur wird dominiert von einem Makrosprechakt, der die Illokution eines Textes reflektiert. Makrostrukturen werden als Bäume aufgefasst: Sätze besitzen Mikrostrukturen, diese bilden für einzelne Textteile Makrostrukturen, die wiederum die Makrostruktur eines Textes bestimmen. Dabei muss nach Vater (2001, S. 70) „jede Makrostruktur dieselben Bedingungen für semantische Konnexion erfüllen wie die Mikrostruktur-Ebenen“. Aufgrund dieser Bedingungen nimmt van Dijk vier Makroregeln an, die zur Abstraktion der Mikroebene eingesetzt werden: Auslassen, Selektieren, Generalisieren und Konstituieren bzw. Integrieren. Nach der ersten Regel können für die weitere Interpretation nicht relevante Propositionen in der aufzubauenden Makrostruktur ausgelassen werden. Die zweite Regel besagt, dass Propositionen, die Prämissen oder Folgen anderer Propositionen sind, ignoriert werden können, da nur die für das Verständnis des Textes wichtigere Proposition selektiert wird. Die dritte Regel drückt aus, dass „Konzepte, die merkmalkonstituierende Kennzeichen von Referenten enthalten, weggelassen und durch ein ›Superkonzept‹ ersetzt“ werden können (Vater, 2001, S. 71). Die vierte Regel betrifft die Bildung der Makroproposition eines Textes.

Vater (1994, S. 93) billigt diesen „Kürzungsvorgängen“ zwar eine „gewisse psychologische Realität“ (ebd.) zu, hält sie aber in kognitiver Hinsicht für unzureichend. Das Modell könne nicht erklären, weshalb z. B. in der Nacherzählung eines Märchens ein Zauberer auftaucht, obwohl dieser nicht Teil der Geschichte war. Vater sieht die Ursachen in kognitiven Schemata, die Texten bzw. Textsorten zugrunde liegen. Nach Adamzik (2004, S. 130), die sich dieser Kritik anschließt, „bleibt es meist bei der Anerkennung, der intuitiven Plausibilität [...] des Modells, dessen praktische Umsetzbarkeit jedoch eher bezweifelt wird.“ (vgl. Gansel und Jürgens, 2002, S. 43). Heinemann und Viehweger (1991, S. 45) zufolge stellt das Modell „eine solide methodische Basis für die Kennzeichnung von semantischen Textstrukturen dar.“

[5] Heinemann und Heinemann (2002, S. 74) charakterisieren den Begriff der Proposition als „semantische Basiseinheit, bestehend aus einem semantischen Prädikat und einer bestimmten Anzahl von Argumenten, da in Termen von Propositionen sowohl die Inhalte von Einzelsätzen als auch die Verknüpfung und Integration dieser Einzelsätze zu komplexeren Ganzheiten beschrieben werden können.“

[6] Obwohl sich die funktionale Satzperspektive in vielen Aspekten wesentlich von van Dijks Ansatz unterscheidet, versuchen beide Modelle letzten Endes, Mikro- und Makrostrukturen abzuleiten (vgl. Vater, 2001, S. 78).

2.2.3 Die pragmatisch-kommunikative Ebene

Die Charakterisierung der Textsemantik hat den Aufbau von Propositionskomplexen zum Ziel, so dass die Textfunktion unberücksichtigt bleibt. Der handlungstheoretische Ansatz (nach Heinemann und Viehweger, 1991, S. 54 ff., sowie Heinemann und Heinemann, 2002, S. 82 ff.) baut auf der Sprechakttheorie auf (Austin, 1962/1979, Searle, 1969): Eine Äußerung wird in Akte aufgeteilt, die auf mehreren Ebenen operieren. Der Äußerungsakt bezeichnet die Tatsache, dass etwas gesagt wird. Der propositionale Akt (unterteilt in Referenz- und Prädikationsakt) beschreibt die Ebenen der Grammatik, der Lexik und der Semantik. Der illokutive Akt gibt die vom Sprecher intendierte Funktion einer Äußerung an, wobei Searle fünf Illokutionsklassen unterscheidet: Repräsentativa (Darstellung eines Sachverhalts, z. B. *Feststellung*, *Behauptung*, *Beschreibung*), Direktiva (der Hörer soll dazu bewegt werden, etwas zu tun, z. B. *Anordnung*, *Befehl*, *Bitte*, *Empfehlung*), Kommissiva (der Sprecher verpflichtet sich zu einer zukünftigen Handlung, z. B. *Versprechen*, *Wette*, *Vertrag*), Expressiva (Ausdruck einer Einstellung des Sprechers zu dem propositional gekennzeichneten Sachverhalt, z. B. *Dank*, *Entschuldigung*, *Glückwunsch*) und Deklarationen (ritualisierte und institutionelle Statusänderung, die bei Vollzug zur Übereinstimmung zwischen dem propositionalem Gehalt und der Wirklichkeit führt, z. B. *Ernennung*, *Trauung*, *Schenkung*, *Kündigung*, *Entlassung*). Der perlokutive Akt bezeichnet die Wirkung der Äußerung auf den Hörer.

Da die Sprechakttheorie nur an Einzelsätzen exemplifiziert worden war, kam die Frage nach ihrer Übertragbarkeit auf Texte auf. Mehrere Arbeiten liefern „eine Präzisierung sprachhandlungs-theoretischer Grundbegriffe und den Konsens, daß Texte als Instrumente kommunikativen Handelns zu bestimmen sind, als komplexe Handlungen [. . .], die sich aus Teilhandlungen zusammensetzen“ (Heinemann und Viehweger, 1991, S. 57). Brinker (2001, S. 102 ff.) setzt auf der Ebene der Illokutionsklassen an und definiert fünf textuelle Grundfunktionen:

1. Informationsfunktion – Der Emittent gibt dem Rezipienten zu verstehen, dass er ihm Wissen vermitteln, ihn über etwas informieren will. Beispiele: *Bericht*, *Gutachten*, *Leserbrief*, *Nachricht*, *Reportage*, *Rezension*, *Vorlesung*.
2. Appellfunktion – Der Emittent gibt dem Rezipienten zu verstehen, dass er ihn dazu bewegen will, eine spezifische Einstellung gegenüber einem Sachverhalt einzunehmen (Meinungsbeeinflussung) oder eine spezifische Handlung durchzuführen (Verhaltensbeeinflussung). Beispiele: *Antrag*, *Arbeitsanleitung*, *Bittschrift*, *Gebrauchsanleitung*, *Predigt*, *Rezept*, *Wahlkampf*, *Werbung*, *Zeitungskommentar*.
3. Obligationsfunktion – Der Emittent gibt dem Rezipienten zu verstehen, dass er sich dazu verpflichtet, eine spezifische Handlung zu vollziehen. Beispiele: *Angebot*, *Garantieschein*, *Gelübde*, *Vereinbarung*, *Versprechen*, *Vertrag*.
4. Kontaktfunktion – Der Emittent gibt dem Rezipienten zu verstehen, dass es ihm um die personale Beziehung zum Rezipienten geht (Herstellung oder Erhaltung des Kontakts). Beispiele: *Ansichtskarte*, *Gratulationsbrief*, *Liebesbrief*, *Trauerkarte*.
5. Deklarationsfunktion – Der Emittent gibt dem Rezipienten zu verstehen, dass der Text eine neue Realität schafft. Beispiele: *Bescheinigung*, *Schuldspruch*, *Testament*.

In einem Text existiert mindestens eine dominierende Illokution, die die Gesamtintention angibt und von subsidiären Illokutionen unterstützt wird (z. B. eine Bitte, die durch

eine Begründung gestützt wird, vgl. Motsch und Viehweger, 1981). Die Illokutionsstruktur kann durch logische und pragmatische Verknüpfungen ermittelt werden, die Beziehungen zwischen Äußerungen betreffen. Das Gesamtziel eines Textes wird demnach über Teilziele realisiert, die Unterstützungsfunktionen besitzen. Teilziele werden durch illokutive Blöcke konstitutiert, die Makroeinheiten der Handlungsstruktur darstellen. Diese Illokutionsstruktur ist mit der propositionalen Struktur eng verknüpft (vgl. van Dijk, 1980).

Die Kritik an handlungstheoretischen Ansätzen zur Charakterisierung der Funktion von Texten betrifft unter anderem die Reduktion der Handlungsziele auf eine kleine Zahl von Handlungstypen, die satzweise Zuordnung von Illokutionen und ihrer aszendenten Integration ohne eindeutige Bezugnahme auf das Textganze und die unzureichende Spezifizierung der pragmatischen Verknüpfung von Illokutionen. Heinemann und Heinemann (2002, S. 85 f.) fassen ihre Kritik plakativ zusammen: „Dass bei einer so weit ausgreifenden und die Gesellschaft in ihrer Komplexität involvierenden Thematik bei weitem noch nicht alle Probleme in zureichender Weise geklärt werden konnten, versteht sich von selbst“.

2.2.4 Die kommunikativ-interaktionale Ebene

Diese Ebene betrifft Aspekte, die bei einer Charakterisierung des Textbegriffs unter Berücksichtigung genereller Faktoren verbaler Interaktion von essenzieller Bedeutung sind, für die vorliegende Arbeit jedoch nur eine untergeordnete Rolle spielen. Die Rahmenkonstituenten umfassen „gesellschaftlich-soziale und psychische Aspekte ebenso [...] wie linguistisch explizierbare sprachlich geformte Strukturen und Prozesse“ (Heinemann und Heinemann, 2002, S. 59). Texte sind demnach in „übergeordnete Interaktionszusammenhänge“ involviert, sie sind als „(Teil-)Prozesse und (Teil)-Resultate“ (ebd.) dieser globalen Zusammenhänge aufzufassen. Eine Explikation von Texten sollte nach Heinemann und Heinemann nur unter Einbeziehung ihrer Rahmenfaktoren stattfinden, wozu unter anderem soziale Konstellationen, die Kooperationsbereitschaft, die Einlassung der Partner auf ein gemeinsames Thema, kognitive Auswahl- und Entscheidungsprozesse von Interagierenden und das Reagieren der Partner auf den Text gehören. Diese Liste von Aktivitäten zeigt, dass Texte das „materiell greifbare (Teil)-Resultat“ (ebd., S. 60) von komplexen, interaktiven Handlungen und Ereignissen sind. Zusätzlich sind Texte die Bezugspunkte, möglicherweise sogar die Auslöser sich anschließender Interaktionen, d. h. Texte „fungieren als *Instrumente des kommunikativen Handelns* zur Durchsetzung bestimmter Ziele der Interagierenden“ (ebd.).

2.2.5 Die argumentativ-rhetorische Ebene

Zwischen Textsegmenten existieren Diskursrelationen, die zur Kohärenz eines Textes beitragen. Zur Beschreibung dieser Strukturen wurde die *Rhetorical Structure Theory* entwickelt (vgl. Mann und Thompson, 1987, 1988, Mann et al., 1989), die ursprünglich in einem System zur Textgenerierung eingesetzt werden sollte. Mann et al. (1989) gehen von verschiedenen Annahmen aus: (i) Texte bestehen aus elementaren Einheiten, aus denen größere Teile aufgebaut werden; (ii) für jede hierarchische Ebene eines Textes existiert eine kleine Menge struktureller Muster (Schemata); (iii) Relationen verknüpfen Textsegmente und konstituieren größere Textteile, zwischen denen ebenfalls Relationen gelten; (iv) die meisten Textstruktu-

rierungsrelationen sind asymmetrisch und werden als Nukleus-Satellit-Relationen bezeichnet, denn „one member of a pair of text spans is more central (the *nucleus*) and one more peripheral (the *satellite*)“ (ebd., S. 8); (v) Textstrukturierungsrelationen sind funktional; sie drücken die Intentionen des Autors und seine Annahmen über den Rezipienten („it is in this sense that an RST structure is ‘rhetorical’.“, ebd.). Es wird ein Test zur Nuklearität angegeben: Wird der Satellit getilgt, sollte der Nukleus noch immer zu verstehen sein. Wird jedoch der Nukleus getilgt, wird der Satellit nicht – oder nur noch sehr schwer – verständlich sein.

Rhetorische Relationen gelten zwischen zwei adjazenten Textsegmenten (üblicherweise ein Nukleus und ein Satellit). Sie werden definiert über die Restriktionen einer Relation in Bezug auf die beteiligten Textsegmente sowie den vom Autor intendierten Effekt der Relation auf den Leser. Mann et al. (1989) schlagen ca. 25 Relationen vor, z. B. „evidence“, „circumstance“, „elaboration“ und „contrast“. Jeder Definition liegt eines von fünf Schemata zugrunde, die meisten Relationen besitzen das Schema vom Typ Satellit/Nukleus (z. B. „evidence“ und „concession“). Die verbleibenden Typen stellen andere Konstellationen dar, z. B. eine Sequenz mehrerer Nuklei. Im Rahmen einer Analyse besitzen diese Schemata die Aufgabe, den hierarchisch angeordneten rhetorischen Strukturbaumes eines Textes zu repräsentieren. Eine RST-Analyse ist abhängig von dem oder den Analysierenden, da „plausibility judgements“ (Mann und Thompson, 1988, S. 246) zu treffen sind; ob eine bestimmte Relation zwischen zwei Textsegmenten gilt, ist also prinzipiell eine subjektive Entscheidung. Einer RST-Analyse liegt eine Aufteilung des Textes in Sätze bzw. Nebensätze zugrunde, die entweder einen Status als Nukleus oder als Satellit besitzen. Relationen, die für größere Texteinheiten gelten, umfassen weitere Relationen. Innerhalb der RST wird angenommen, dass die Struktur eines kohärenten Textes als ein einziger rhetorischer Strukturbaum beschrieben werden kann.

Moore und Pollack (1992) kritisieren, dass die RST nur einen Relationstyp vorsieht, der den Unterschied zwischen der intentionalen und informationalen Struktur eines Diskurses nicht erfassen kann. Ähnliche Kritik äußern Carberry et al. (1993), die für eine Erweiterung der RST hinsichtlich der Erfassung der intentionalen Strukturen von Dialogen plädieren.[7] Die Entwickler der RST betonen, dass die ca. 25 Relationen nicht als vollständig zu betrachten sind. Je nach Textsorte oder Analyseschwerpunkt regen sie zur Definition weiterer Relationen an (vgl. Rösner und Stede, 1993, S. 20). Hovy (1990) führt eine Untersuchung relationaler Theorien zur Diskursmodellierung durch und fasst die insgesamt etwa 350 Diskursrelationen in einer Hierarchie zusammen, wobei er 16 Kernrelationen annimmt, die den Kategorien „Elaboration“, „Enhancement“ und „Extension“ zugeordnet werden. Die verbleibenden Relationen lassen sich nach Hovy wiederum den 16 Kernrelationen zuordnen.

2.2.6 Die kognitive Ebene

Auf der kognitiven Betrachtungsebene nehmen psychische Prozesse (Erinnern, Wahrnehmen, Schlussfolgern etc.) eine zentrale Rolle bei der Produktion und Rezeption sprachlicher Äußerungen ein. Ein Text wird als Resultat mentaler Prozesse betrachtet (vgl. Figge, 2000). Die folgenden Ansätze orientieren sich an kognitiven Operationen und Wissensorganisationen

[7] Die RST stellt einen Versuch dar, die rhetorische Struktur eines Textes so zu abstrahieren, dass sie im Rahmen eines computerlinguistischen Systems traktabel wird. Somit kann die RST von einem ausschließlich textlinguistisch ausgerichteten Standpunkt zwangsläufig als zu starke Vereinfachung kritisiert werden.

(vgl. Scherner, 2000). Es handelt sich um die Script-Theorie (Schank und Abelson, 1977), das Modell von Kintsch und van Dijk (1978) sowie die „story grammar“ (Rumelhart, 1975). Zunächst werden jedoch die kognitionspsychologischen Grundlagen dargestellt.

Wissen, Sprache und Gedächtnis

Der Umgang mit sprachlichen Äußerungen setzt kommunikative Kompetenz voraus, die auf im Gedächtnis gespeicherten, abstrakten, kulturspezifischen und individuell unterschiedlich ausgeprägten Bewusstseinsinhalten und kognitiven Fähigkeiten basiert (vgl. Strohner, 2000). Die Kognitionswissenschaft hat Hypothesen erarbeitet, die die Prinzipien der Organisation des semantischen Gedächtnisses erklären sollen: Seine Basiseinheiten sind im Langzeitgedächtnis hinterlegte Konzepte, die als Repräsentanten von Begriffen fungieren und distinktive Merkmale besitzen. Die Ablage[8] von Konzepten erfolgt in vernetzter Form, weshalb unterschiedliche Distanzen zwischen Konzepten auszumachen sind (z. B. stehen sich *Tisch* und *Stuhl* näher als *Tisch* und *Uhr*). Durch diese Vernetzung bilden sich komplexere kognitive Einheiten, die gelegentlich als Geschehenstypen bezeichnet werden und ebenfalls in Form eines Netzes miteinander verknüpft sind (z. B. *Schreiben*, *Lesen*, *Autofahren*, *Einkaufen*). Einen besonderen Stellenwert besitzen Konzepthierarchien (vgl. Abschnitt 13.2).

Eine andere Betrachtungsweise geht von einer *top-down*-Sicht auf mentale Prozesse aus und beschäftigt sich mit komplexeren kognitiven Einheiten: Bartlett (1932) zeigt, dass neue Informationen auf der Basis von im Gedächtnis vorhandenen Strukturen, z. B. Weltwissen oder Wissen über spezifische Ereignisse, organisiert und erinnert wird. Hierfür hat Bartlett den Terminus Schema geprägt, der in den siebziger Jahren die Entwicklung der KI-orientierten Wissensrepräsentation beeinflusst hat. Schemata[9] sind hierarchisch strukturiert und konstituieren sich aus typischen Zusammenhängen und Spezifika eines begrenzten Bereiches. Sie können Handlungsabläufe, Ereignisse, Dinge und abstrakte Konzepte wie den typischen Ablauf oder die beteiligten Charaktere eines Märchens enthalten.

Neben dem semantischen Gedächtnis (Weltwissen, enzyklopädisches Wissen) existieren weitere Typen von Kenntnissystemen, von denen das sprachliche und das episodische Wissen für die Produktion und Rezeption von Texten relevant sind. Das sprachliche Wissen setzt sich aus dem lexikalischen Wissen (das die Beziehung zum Weltwissen herstellt) und dem grammatischen Wissen zusammen. Das episodische Gedächtnis beinhaltet kognitive Modelle, die das praktische oder kommunikative Handeln in bestimmten Situationen spezifizieren.

Scripts als Repräsentationen standardisierter Ereignisse

Die von Schank und Abelson (1977) entwickelte Script-Theorie besagt, dass standardisierte Ereignisse und Handlungen sowie die Erwartungshaltungen der beteiligten Individuen formal repräsentierbar sind. Ein Restaurantbesuch findet z. B. meist nach einem festgelegten Ablauf statt: Ein Gast betritt das Restaurant, er wird zu einem Tisch geführt, nach der

[8] Die Ablage von Konzepten kann als Lernen bezeichnet werden. Das Abrufen von Entitäten (Erkennen, Erinnern) erfolgt z. B. durch assoziativer Zündung oder durch logische Verknüpfungen (Vergleichs- und Inferenzprozesse), beispielsweise beim Einbringen von Vorwissen in das Verstehen eines Textes.

[9] Dieser Terminus ist ein Oberbegriff, der sich auf unterschiedliche kognitive Einheiten beziehen kann, z. B. Frames, mentale Modelle, Scripts, semantische Netze und Konzepthierarchien.

Durchsicht der Karte wählt er eine oder mehrere Speisen aus, diese werden zubereitet und anschließend serviert. Nach dem Essen bezahlt der Gast die Rechnung und verlässt das Lokal. Scripts stellen für solche ritualisierten Handlungen formale Repräsentationen dar: Neben den Rollen der Akteure (Gast, Ober, Koch etc.) und den Gegenständen (Tisch, Speisekarte, Essen etc.) werden die Abläufe und ihre kausalen Beziehungen in einem Script repräsentiert. Die Repräsentation der Ereignisse erfolgt mit Hilfe semantischer Primitive und unterschiedlicher Relationen, die ihren Ursprung in der *Conceptual Dependency*-Theorie haben (Schank, 1975). Auf der Basis der Script-Theorie wurden Modelle entwickelt, die weitere Einflussfaktoren (Situationswissen, Weltwissen) berücksichtigen (vgl. Johnson-Laird, 1983).

Luger und Stubblefield (1993, S. 386) betrachten Scripts als unflexibel und schlagen vor, Modularisierungen und Generalisierungen vorzunehmen, um Script-Ontologien aufzubauen. Der Script-Theorie ist inhärent, dass sie mit plötzlich auftretenden Ereignissen (z. B. dem Einsturz der Decke des Restaurants) nicht umgehen kann, da nur prototypisches Wissen in Scripts kodiert und antizipiert wird, wobei Sprünge in andere Scripts nicht vorgesehen sind. Dennoch zeigen Experimente (vgl. Görz, 1995, S. 334 ff.), dass Scripts durchaus eine gewisse kognitive Plausibilität besitzen. Werden z. B. Probanden Texte vorgelesen, so nennen diese in einer Nacherzählung oftmals Ereignisse, die nicht explizit in den Texten erwähnt werden, da sie offenbar feste Bestandteile von Script- bzw. Schema-Wissen sind.

Kognitive Verarbeitung von Texten

Der Ansatz von Kintsch und van Dijk (1978) gilt als das erste prozedurale Modell der kognitiven Verarbeitung von Texten. Es beinhaltet interagierende Prozesse, die parallel kohärente Mikrostrukturen aus Propositionen (mit Konzepten als Termen) aufbauen und die Mikrostrukturen zu kohärenten Makrostrukturen reduzieren. Ein weiterer Prozess ist für die Textproduktion zuständig, die auf Makrostrukturen arbeitet. Das Textverstehen wird als zyklischer Prozess modelliert, der sukzessive kurze Sequenzen von Propositionen verarbeitet, was in Abhängigkeit vom Kurzzeitgedächtnis und den Zielen des Textrezipienten geschieht. Im Zuge dieses Prozesses entstehen Makrostrukturen als komprimierte Propositionslisten, die durch die Anwendung von Makroregeln gebildet werden. Die Tilgungsregel erlaubt das Entfernen einer unwichtigen Proposition, die Verallgemeinerungsregel ersetzt eine Sequenz von Propositionen durch eine allgemeinere Proposition und die Konstruktionsregel ersetzt eine Folge von Propositionen durch eine abstrakte Frame- bzw. Schema-ähnliche Proposition. Van Dijk und Kintsch (1983) ergänzen den Ansatz durch ein Situationsmodell, das eine Verbindung zwischen Textverstehen und Weltwissen etabliert.

Story-Grammars – Geschichtengrammatiken

Rumelhart (1975) schlägt mit der „story grammar“ syntaktische Ersetzungs- und semantische Interpretationsregeln vor, die er als Basis für die Analyse der Wohlgeformtheit und der Produktion einfacher Geschichten ansieht (vgl. auch Baddeley, 1990, S. 337–343).[10] Das generative System operiert unabhängig vom Inhalt einer Geschichte und konzentriert sich auf die

[10] Durch die Verwendung kontextfreier Regeln entsteht eine konzeptionelle Ähnlichkeit zu Textgrammatiken. Die „story grammar“ kann unmittelbar in eine Dokumentgrammatik überführt werden (vgl. Rehm, 1998).

Handlungsstruktur, weshalb es als „allgemeines Erzählungs-Schema“ (Figge, 2000, S. 101) betrachtet werden kann. Als grammatische Kategorien fungieren abstrakte semantische Einheiten wie „setting“, „episode“ und „event“ sowie propositionsähnliche Einheiten wie „state“ und „change-of-state“. Gegen diesen Ansatz wurde von vielen Seiten Kritik vorgebracht, die an den nur vage definierten Kategorien und dem Fehlen von Prozeduren zur Zuordnung von Textteilen zu diesen Kategorien ansetzt (Johnson-Laird, 1983, S. 362 ff.). Zwischen den Vertretern des Script-Ansatzes und den Vertretern der Story-Grammar hat sich in den siebziger Jahren eine „weitgehend fruchtlose Debatte“ (Görz, 1995, S. 331) entwickelt (vgl. Black und Wilensky, 1979, Rumelhart, 1980, Mandler und Johnson, 1980, Habel, 1986).

2.2.7 Die mediale und konzeptionelle Ebene

Die Ebene des Mediums bezeichnet die Realisierung einer sprachlichen Äußerung in schriftlicher oder mündlicher Form, wohingegen die Ebene der Konzeption ihren Duktus betrifft, der häufig durch Begriffspaare wie Schriftsprache/Umgangssprache oder formell/informell bestimmt wird (Koch und Oesterreicher, 1994). Obwohl die zwei Ebenen eng miteinander verknüpft sind, besteht dennoch eine „prinzipielle Unabhängigkeit“ (ebd., S. 587).

Nach Koch und Oesterreicher ist die mediale Ebene dichotomischer Natur, d. h. eine Äußerung wird entweder schriftlich oder mündlich realisiert. Adamzik (2004, S. 75 ff.) geht auf Eigenschaften ein, die schriftliche Realisierungen besitzen können.[11] Hierzu zählt der Umstand, dass schriftliche im Gegensatz zu mündlichen Texten situationsentbunden sind. Ein Schrifttext kann auch multimedial gestaltet sein und mehrere Übertragungskanäle besitzen. Zusätzliche Eigenschaften betreffen die Anfertigung der Äußerung mittels Handschrift oder Computer, Satzzeichen, unterschiedliche Schrifttypen und -größen, typografische Merkmale, spezifische Möglichkeiten des Trägermediums (z. B. farbige Gestaltung) und die räumliche Anordnung der beschrifteten Fläche (Layout) – unter Umständen in Kombination mit grafischen Elementen, z. B. Tabellen, Abbildungen, Fotos und Illustrationen.

Die Konzeption stellt ein Kontinuum dar und bezieht sich auf „den Duktus, die Modalität der Äußerung“ (Storrer, 2000a, S. 153). Koch und Oesterreicher verdeutlichen diese Differenzierung[12] mit Beispielen: Ein familiäres Gespräch ist medial und konzeptionell mündlich, auch ein Vorstellungsgespräch ist medial mündlich, wird jedoch etwa in der Mitte des Kontinuums situiert. Ein wissenschaftlicher Vortrag befindet sich nahe der konzeptionellen Schriftlichkeit. Bei medial schriftlich realisierten Äußerungen wird der Privatbrief als konzeptionell mündlich eingestuft, das Zeitungsinterview rangiert zwischen konzeptioneller Mündlichkeit und Schriftlichkeit und der Gesetzestext ist ausschließlich konzeptionell schriftlicher Natur. Koch und Oesterreicher zeigen, dass sich die Pole dieses Kontinuums durch Parameter wie raum-zeitliche Nähe (tendiert zur Mündlichkeit) bzw. Distanz (tendiert zur Schriftlichkeit) beschreiben lassen, wobei Merkmale wie Vertrautheit der Kommunikationspartner, Situationseinbindung, Spontaneität, Dialog/Monolog und Themenfixierung Einfluss darauf haben, auf welcher Position des Kontinuums eine sprachliche Äußerung anzusiedeln ist.

[11] Buhofer (2000) diskutiert diese Thematik ausführlich und geht auch auf die Gebärdensprache ein, die der von Koch und Oesterreicher (1994) aufgestellten Differenzierung nicht ohne Weiteres zugeordnet werden kann.

[12] Die oben angesprochene Unabhängigkeit der beiden Ebenen wird insbesondere durch Untersuchungen zur computervermittelten Kommunikation deutlich (vgl. die Kapitel 4, 8 und 9).

2.2.8 Die Ebene der Textualität

Die Textualität betrifft die abstrakte Frage, welche Merkmale die „Texthaftigkeit“ einer Abfolge von Wörtern und Sätzen ausmachen, was einen Text von einem „Nicht-Text“ unterscheidet. De Beaugrande und Dressler postulieren „sieben Kriterien der Textualität“, die nahezu alle der bereits angesprochenen Ebenen berühren:

1. *Kohäsion* – Dieses Kriterium bezeichnet die Verbindung der Wörter eines Textes, es reflektiert die Zusammengehörigkeit von Oberflächeneinheiten und basiert auf grammatischen Verknüpfungen und Abhängigkeiten. Kohäsion wird hervorgerufen durch explizite Formen der Wiederaufnahme als Ausdruck thematischer Kontinuität (Rekurrenz, Anaphern etc.), Konjunktionen als satzverknüpfende Elemente, textkommentierende und textdeiktische Ausdrücke („siehe Abschnitt . . .“) und durch die Markierung der Grenzen von Texten oder Teilen von Texten, z. B. durch Überschriften oder Gliederungssignale.
2. *Kohärenz* – Der semantische Sinnzusammenhang eines Textes, der der Kohäsion zugrunde liegt, seine Sinnkontinuität, wird als Kohärenz bezeichnet. Sie entsteht durch die Verknüpfung des in einem Text enthaltenen Wissens, das de Beaugrande und Dressler als „Textwelt“ bezeichnen (ebd., S. 8), mit dem „gespeicherten Weltwissen“ (ebd.) des Rezipienten. Die Textwelt besteht aus Konzepten und Relationen. Nach de Beaugrande und Dressler ist Kohärenz also nicht nur ein Merkmal von Texten, „sondern vielmehr das Ergebnis kognitiver Prozesse der Textverwender“ (ebd., S. 7).

Die ersten beiden Kriterien sind textzentrisch (textintern): Sie gehen unmittelbar aus einem Text hervor bzw. sind unmittelbar am Text selbst festzumachen. Die weiteren fünf Kriterien sind pragmatischer Natur und werden als „verwenderzentrierte“ (textexterne) Kategorien bezeichnet, die „die Aktivität der Text-Kommunikation betreffen, sowohl hinsichtlich des Produzenten als auch des Rezipienten von Texten“ (ebd., S. 8).

3. *Intentionalität* – Die Einstellung des Produzenten, einen sowohl kohäsiven als auch kohärenten Text zu generieren, um gewisse Absichten zu erfüllen, d. h. „Wissen zu verbreiten oder ein in einem Plan angegebenes Ziel zu erreichen“ (ebd., S. 8).
4. *Akzeptabilität* – Die Einstellung des Rezipienten, „einen kohäsiven und kohärenten Text zu erwarten, der für ihn nützlich oder relevant ist, z. B. um Wissen zu erwerben [. . .]. Diese Einstellung spricht auf Faktoren an wie Textsorte, sozialen oder kulturellen Kontext und Wünschbarkeit von Zielen“ (ebd., S. 9).
5. *Informativität* – „[D]as Ausmaß der Erwartetheit bzw. Unerwartetheit oder Bekanntheit bzw. Unbekanntheit/Ungewissheit der dargebotenen Textelemente“ (ebd., S. 10). Obwohl jeder Text eine minimale propositionale Textbasis enthält, kann der Grad der Informativität das Interesse des Rezipienten steuern. Zu geringe Informativität kann Langeweile erzeugen, zu hohe Informativität kann abschreckend wirken: „Die Verarbeitung von hochgradig informativen Nachrichten ist anstrengender als von weniger informativen, ist dafür aber auch dementsprechend interessanter“ (ebd., S. 11).
6. *Situationalität* – Die „Faktoren, die einen Text für eine kommunikative Situation relevant machen.“ (ebd., S. 12). Dieses Kriterium bezieht sich auf den Kontext, der

z. B. bei der Disambiguierung potenziell mehrdeutiger Aussagen unterstützend wirken kann. „Ohne Situationsbezogenheit gibt es [...] keinen Text, denn Bedeutung und Gebrauch eines Textes wird eben über die Situation bestimmt“ (Heinemann und Viehweger, 1991, S. 77).

7. *Intertextualität* – Dieses Kriterium „betrifft die Faktoren, welche die Verwendung eines Textes von der Kenntnis eines oder mehrerer vorher aufgenommener Texte abhängig macht. [...] Intertextualität ist, ganz allgemein, für die Entwicklung von Textsorten als Klassen von Texten mit typischen Mustern von Eigenschaften verantwortlich“ (de Beaugrande und Dressler, 1981, S. 12 f.).

Als exemplarische Antwort auf die Frage nach den konstitutiven Merkmalen von Texten wird nachfolgend die Definition[13] von de Beaugrande und Dressler aufgeführt:

> Wie definieren einen Text als eine kommunikative Okkurrenz [...], die sieben Kriterien der Textualität erfüllt. Wenn irgendeines dieser Kriterien als nicht erfüllt betrachtet wird, so gilt der Text nicht als kommunikativ. Daher werden nicht-kommunikative Texte als Nicht-Texte behandelt. (de Beaugrande und Dressler, 1981, S. 3)

Vater (2001, S. 52–54) zeigt in einer äußerst kritischen Betrachtung der Textualitätskriterien, dass keinesfalls alle Faktoren für jeden Text gelten müssen, vielmehr stelle offenbar lediglich Kohärenz das entscheidende Mittel zur Schaffung eines Textzusammenhangs dar.[14] Gansel und Jürgens (2002, S. 31) äußern ähnliche Kritik: „Die These [...], dass Texte nicht kommunikativ sind, wenn eines der Kriterien nicht erfüllt wird und daher auch als Nicht-Texte zu behandeln sind, ist [...] in ihrer Absolutheit nicht haltbar.“ Adamzik (2004, S. 51) erwidert, dass die Definition von de Beaugrande und Dressler lediglich „irreführend formuliert“ und nur ein „Tribut an die formale Grammatik“ sei, da sich die Verfasser „von Anfang an selbst sehr bewusst“ darüber gewesen seien, „dass es sich [bei den Textualitätskriterien, G. R.] um Eigenschaften handelt, die mehr oder weniger ausgeprägt vorliegen können.“ Aus den genannten Gründen nimmt Adamzik (2004, S. 53) an, dass es sich bei den Kriterien um Beschreibungsdimensionen für wesentliche Eigenschaften prototypischer Texte handelt. Zu einem ähnlichen Schluss kommen letzten Endes auch Gansel und Jürgens:

[13] Adamzik (2004, S. 38 f.) geht auf weitere Definitionen ein. Heinemann und Heinemann (2002, S. 64) fassen die Problematik zusammen: „Auf die ›Gretchen-Frage‹ der Linguistik nach dem Wesen von Textualität wurden [...] in nahezu tausend Textdefinitionen teils erheblich voneinander abweichende Antworten gegeben. Auffallend war, dass die größte Zahl der Arbeiten jeweils nur einen Teilaspekt von Texten aus unterschiedlicher [...] Sicht erfassten und adäquat darstellten, dass dagegen das komplexe Phänomen ›Text‹ als eine in der Interaktion funktionierende Ganzheit nur gelegentlich apostrophiert wurde.“ (vgl. auch Tietz, 1997).

[14] Eine kritische und an übergreifenden Forschungsperspektiven ausgerichtete Einschätzung liefert Feilke (2000, S. 76), der die Textualitätskriterien als ein „Spiegelkabinett texttheoretischer Begriffe“ bezeichnet: „[Die Verfasser] definieren »Textualität« über ein Ensemble von sieben Kriterien [...] – die völlig heterogenen Theorietraditionen verpflichtet sind: So verweist [...] *Kohäsion* auf die Textgrammatik, [...] *Kohärenz* ist den Zielen der Textsemantik verpflichtet, *Intentionalität* betont vor allem die sprecherseitigen (Sprechakttheorie) Voraussetzungen und die *Akzeptabilität* sowie *Informativität* beziehen sich auf hörerseitige Konditionen des Textverstehens. Schließlich rekurriert *Situationalität* auf die kontextuelle Einbindung und das Kriterium der *Intertextualität* betont die diachrone Dimension einer Textsortentypik.“ Feilke (2000, S. 72 ff.) schlägt neue „Leitorientierungen“ vor, um der Heterogenität und Undifferenziertheit der Textualitätskriterien entgegen zu wirken: Generativität, Universalität, Kontextualität, Prozessualität, Handeln/Intentionalität und Dialogizität, vgl. auch Heinemann und Heinemann (2002, S. 101 f.) und Adamzik (2004, S. 40 f.).

> Die Kategorie „Text“ entzieht sich einer eindeutigen, auf alle potentiellen Textexemplare zutreffenden Auflistung von Merkmalen. Schon gar nicht lässt sich auf diesem Wege trennscharf zwischen Texten und Nicht-Texten unterscheiden. [...] Um der Kategorie „Text“ in ihrer ganzen Komplexität und Vielschichtigkeit gerecht werden zu können, eigenen sich am besten Definitionen, die relativ vorsichtig und weit gefasst sind. (Gansel und Jürgens, 2002, S. 31)

Adamzik konkretisiert diese Vorgehensweise:

> [S]eit den 1980er Jahren [ist] der Streit um *eine* einheitliche und klare Definition von *Text* denn auch in den Hintergrund getreten und man bemüht sich um eine Überwindung der (teilweise nur vermeintlichen) Gegensätze durch integrative Ansätze, bei denen es weniger um eine Definition geht als um die Zusammenstellung von Aspekten, die sich in der Diskussion als wesentlich für die Charakterisierung und Beschreibung des Phänomens herausgestellt haben. (Adamzik, 2004, S. 39 f.)

Abschnitt 3.5.3 diskutiert Eigenschaften der Textualitätskriterien bei einer Anwendung auf Hypertexte. Abschnitt 5.3.4 geht in detaillierter Form auf die Problematik der Entwicklung einer einheitlichen Textdefinition ein, die für *alle* Hypertexte im WWW Gültigkeit besitzt.

2.2.9 Die prototypische Ebene

Textdefinitionen sind immer wieder Gegenstand kontrovers geführter Diskussionen, in denen unter anderem verschiedene Extrempunkte vertreten werden – neben hochgradig komplexen[15] Definitionen wird z. B. ebenfalls die Frage angesprochen, „ob es überhaupt möglich und sinnvoll ist, einen allgemeinen Textbegriff zu entwickeln, der es erlauben soll, zu bestimmen, was immer und überall als Text zu gelten hat“ (Brinker, 1973, S. 9).

Anfang der neunziger Jahre etablierte sich eine differenziertere Annahme (Sandig, 2000), nach der Text nicht mehr über eine Liste notwendiger und hinreichender Merkmale definiert wird, sondern auf der Grundlage der aus der Kognitionspsychologie stammenden Prototypentheorie[16] (vgl. etwa Rosch, 1977, 1978), deren Kern durch Thesen dargestellt werden kann (nach Mangasser-Wahl, 2000b, S. 15, und Sandig, 2000, S. 93 f.):

1. Kategorien werden nicht immer durch die Verbindung von notwendigen und hinreichenden Merkmalen definiert.
2. Kategorien verfügen nicht immer über klar definierte Grenzen.
3. Kategorien haben Merkmale und sind über sie beschreibbar. Es müssen aber nicht immer alle Merkmale vorhanden sein.
4. Merkmale sind nicht grundsätzlich binär, d. h. sie treffen nicht immer „entweder-oder“ zu, sondern manchmal auch „mehr-oder-weniger“.

[15] Die auf S. 30 reproduzierte Textdefinition von de Beaugrande und Dressler bezieht sich explizit auf die sieben Kriterien der Textualität und kann, insbesondere im Vergleich mit verschiedenen „komplexen Verbaldefinitionen“ (Heinemann und Heinemann, 2002, S. 110), durchaus als knapp bezeichnet werden. Es ist jedoch zu berücksichtigen, dass de Beaugrande und Dressler (1981) jedes der sieben Kriterien durch ein eigenes Kapitel erläutern. Somit kann durchaus der gesamte Band als „komplexe Verbaldefinition“ betrachtet werden.

[16] Adamzik (2004, S. 47) weist darauf hin, dass die „Prototypentheorie kein Gegenkonzept zur Merkmalbeschreibung“ ist (vgl. Abschnitt 2.3.5, S. 47 ff.), sondern den Merkmalen lediglich einen anderen Status zuweist.

5. Merkmale sind gewichtet, d. h. mehr oder weniger wichtig oder zentral.
6. Nicht alle Mitglieder einer Kategorie verfügen über den gleichen Stellenwert. Es existieren bessere und schlechtere Vertreter. Die besten Vertreter sind die Prototypen.
7. Prototypische Vertreter weisen Merkmalbündel auf. Sie haben mit anderen Mitgliedern der Kategorie die meisten Merkmale gemeinsam und möglichst wenige mit anderen Kategorien. Aufgrund übereinstimmender, aber auch verschiedener Merkmale besteht „Familienähnlichkeit“ zwischen den Vertretern einer Kategorie.
8. Mit der Basisebene existiert eine ausgezeichnete Abstraktionsebene bei der Kategorisierung. Kategorien der übergeordneten Abstraktionsebene sind ärmer an Information als die der Basisebene. Kategorien der untergeordneten Ebene enthalten nur minimal spezifischere Informationen.

Der zentrale Aspekt der Prototypentheorie betrifft den Umstand, dass bessere und schlechtere Vertreter einer Kategorie existieren (kulturelle Spezifika spielen hierbei eine entscheidende Rolle). Die besten Vertreter, die als Zentren der mentalen Organisation von Kategorien fungieren, werden als Prototypen bezeichnet und schneller als Vertreter einer Kategorie erkannt. Sandig (2000) untersucht zentrale Merkmale von Text (Kohäsion, Kohärenz, Intentionalität, Situationalität und Thema) auf Basis der Prototypentheorie, um Widersprüche in Textdefinitionen aufzudecken und unterschiedliche Definitionen einander anzunähern:

> Texte als in der Regel komplexe Einheiten werden in Situationen (Situationalität) verwendet, *um* in der Gesellschaft Aufgaben *zu* lösen (Intentionalität/Textfunktion), die auf Sachverhalte (Thema; Kohärenz) bezogen sind. Kohäsion sorgt lokal für die Integration. Das wichtigste dieser zentralen Merkmale ist die Textfunktion. In der Regel ist sie anhand des Textmaterials (Textoberfläche, Aussehen des Textes …) interpretierbar; andernfalls wird die Situation zur Interpretation herangezogen. Ist die Textfunktion wie bei literarischen Texten undeutlich, bekommt das Thema größeres Gewicht. Es gibt also Zusammenhänge zwischen den Merkmalen. (Sandig, 2000, S. 99)

Abbildung 2.1 zeigt, dass weitere „Textmerkmale relevant gemacht werden können; in der Regel bleiben sie aber mehr im Hintergrund“ (ebd., S. 108). Sandig identifiziert in den analysierten Definitionen implizit gesetzte und schwerpunktbedingte Auffassungen, die unmittelbar auf die Prototypentheorie abbildbar sind. Die Gemeinsamkeiten der Definitionen werden auf zwei Prototypen komprimiert: (i) Themadominierte literarische vs. Gebrauchstexte und (ii) themadominierte vs. funktionsdominierte Gebrauchstexte. Als „prototypischen Kern der Prototypen“ bestimmt Sandig (2000, S. 101) die Eigenschaft, dass ein Text eine monologisch geschriebene, sprachliche Äußerung in Form mehrerer Sätze ist, die einen Zusammenhang besitzen. Von diesem prototypischen Kern[17] unterscheiden sich z. B. Ein-Satz-Texte, inkohärente Texte und Texte ohne Thema als weniger typische Vertreter.

Heinemann und Heinemann (2002, S. 104) argumentieren bezüglich der Anwendung der Prototypentheorie auf Texte, „dass ein Kernbereich von Textphänomenen existiert, für den das Texthafte in vollem Umfang zutrifft; daneben aber gibt es [. . .] eine Randzone, für die die Merkmale der Textualität nur in begrenztem Umfang gegeben sein müssen“. Durch die Prototypentheorie können insbesondere diejenigen Texte bzw. Textsorten (als Beispiele werden

[17] Nicht zu verwechseln mit Abbildung 2.1, die eine Abstufung von Textmerkmalen darstellt.

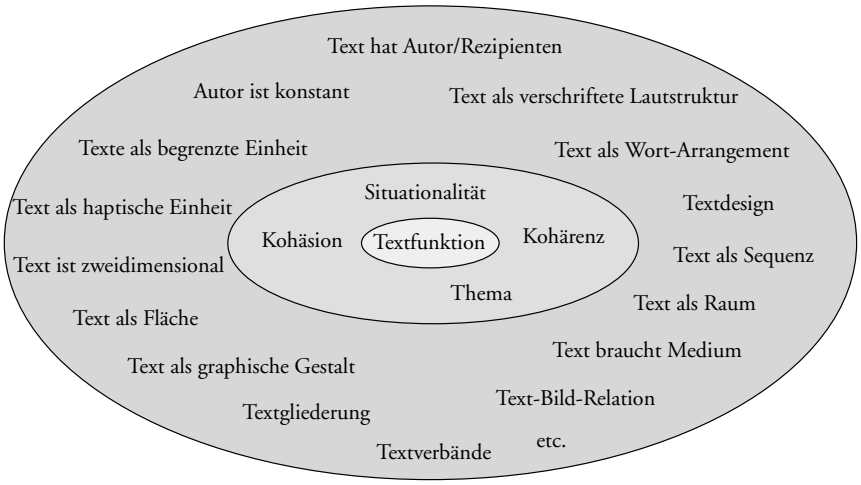


Abbildung 2.1: Abstufungen von Textmerkmalen (nach Sandig, 2000, S. 108)

Busfahrscheine und der *Personalausweis* angeführt), die von der traditionellen Textlinguistik als problematische Grenzfälle betrachtet werden, in bestehende Theorien integriert werden, ohne sie als Nicht-Texte zu verwerfen, weil ein spezifisches Merkmal (z. B. Kohäsion) nicht ausgeprägt ist.[18] Zusätzlich weisen Heinemann und Heinemann darauf hin,

> dass sich Prototypen als Kernbereiche auf allen Abstraktionsebenen von weniger typischen kategorialen Vertretern abheben lassen: Prototypen der Schrift-Texte und der Sprech-Texte, Prototypen als Repräsentanten von Textmustern (Traueranzeigen, Kochrezepten, Nachrichten, [...]), wie auch von Textmuster-Varianten (Bildzeitungs-Nachrichten, Ablehnungsbescheiden usw.). (Heinemann und Heinemann, 2002, S. 104)

Ausgehend von diesen Beispielen sollte es möglich sein, unter den Vertretern von Textsorten bessere und schlechtere Vertreter und somit auch Prototypen zu identifizieren.[19]

2.2.10 Fazit

Die meisten der in den Abschnitten 2.2.4 bis 2.2.9 diskutierten Beschreibungsebenen von Texten berühren in ihrer Grundsätzlichkeit das gesamte Spektrum zwischenmenschlicher Kommunikation. In der vorliegenden Arbeit werden nur ausgewählte Aspekte thematisiert, die insbesondere bei der Konzeptionierung einer allumfassend ausgerichteten Theorie der Texte zu beachten sind. Zugleich zeichnen die Abschnitte die zentralen Phasen nach, die die Textlinguistik in den vergangenen 30 Jahren geprägt haben. Nach der Auffassung von

[18] Bereits Heinemann und Viehweger (1991, S. 142) weisen auf diesen Umstand hin: „Wissen über globale Textstrukturen scheint [...] ein prototypisches Wissen zu sein, in dem ein Prototyp durch eine Menge stereotypischer Eigenschaften repräsentiert wird, die ungleichgewichtig sind, von denen jedoch einige in Abhängigkeit vom Klassifizierungsanlaß prominent und damit klassenbildend sein können."

[19] Sandig (2000, S. 103) geht in einer zweiten Untersuchung, die sich auf das Grußwort bezieht, ebenfalls auf diese Fragestellung ein und kommt zu dem Schluss: „Bezogen auf ein einziges Textmuster [d. h., eine einzige Textsorte] gibt es prototypische Exemplare, die die Kriterien der Beschreibung (fast) vollständig erfüllen, es gibt Exemplare, die weniger Kriterien erfüllen, und schließlich gibt es Randexemplare." (vgl. Abschnitt 2.3.6).

Texten als Satzsequenzen, wurde die semantische Ebene berücksichtigt, wobei sich auch Erkenntnisse aus der Pragmatik und der kognitiven Psychologie niedergeschlagen haben. Die Beschreibungsebenen zeigen auch den der Textlinguistik inhärenten Grad der Komplexität und Interdisziplinarität auf. Hierbei geht es jedoch zunächst „nur“ um traditionelle textlinguistische Modelle und Theorien, die die Einflussfaktoren der digitalen Informations- und Kommunikationsmedien unberücksichtigt lassen. Gerade Hypertext bietet neuartige Möglichkeiten der Anordnung einzelner Textteile, Multimedia-Anwendungen bringen zusätzliche Medien ein (Grafiken, Video und Ton). Das WWW bildet das gemeinsame Dach als unkontrolliertes, verteiltes, multilinguales Hypertextsystem. Die Komplexität wird durch diese Faktoren noch gesteigert, wenn textlinguistische Beschreibungsverfahren für Hypertexte und Hypertextsorten verwendet werden sollen.[20] Die nachfolgenden Kapitel werden verdeutlichen, dass in der vorliegenden Arbeit nicht sämtliche Beschreibungsebenen berücksichtigt werden können. Stattdessen wird der Versuch unternommen, ein generelles Rahmenkonzept für die übergeordnete Ebene der Hypertextsorten zu erarbeiten, das als Ausgangsbasis für Untersuchungen dienen kann, die sich spezifischen Beschreibungsebenen widmen.

2.3 Textsorten und Texttypologien

In Abschnitt 2.2 wurden textlinguistische Beschreibungsebenen vorgestellt. Es wurde deutlich, dass selbst bei der Definition des grundlegenden Terminus Text noch kein einheitlicher und übergreifender Konsens existiert. Ähnlich verhält es sich mit dem Terminus Textsorte, der in der Textlinguistik eine zentrale Stellung[21] einnimmt: „Es ist davon auszugehen, dass der Begriff der Textsorte bisher theoretisch nicht eindeutig definiert ist“ (Gansel und Jürgens, 2002, S. 49), was – zumindest in dieser Frage herrscht ein gewisser Konsens[22] – eine Folge der Vielfalt unterschiedlicher Beschreibungsebenen und Analysedimensionen ist, die sich in gleichem Maße auf die Klassifikation von Texten auswirken:

> Aufgrund der Multidimensionalität der Kategorie „Text“ ist es praktisch nicht möglich, alle potentiellen Texte entsprechend einer einzigen verbindlichen Klassifikation einzuordnen. Es erweist sich als sinnvoller, eine Menge von homogenen Typologien zu akzeptieren, die jeweils nur bestimmte Dimensionen [...] erfassen und einander deshalb nicht etwa ausschließen, sondern ergänzen. (Gansel und Jürgens, 2002, S. 53)

[20] Siehe auch das in der Einleitung (Abschnitt 2.1) wiedergegebene Zitat von Jakobs (2003, S. 233).

[21] Die zentrale Position der Textsorten ist z. B. erkennbar an der Tatsache, dass der Band *Textlinguistik* (Brinker et al., 2000) 80 Artikel enthält, von denen sich sechs mit theoretischen Grundlagen der Textsorten beschäftigen, woraufhin 12 weitere Beiträge auf spezifische Gruppen von Textsorten eingehen, z. B. „Textsorten des Alltags“, „Textsorten in den Massenmedien“ und „Textsorten der Verwaltung“.

[22] Die Vielzahl involvierter Dimensionen resultiert in heterogenen Ansätzen: „Versucht man [...], aus der Vielzahl textsortenspezifischer Einzeldarstellungen [...] ein Resümee abzuleiten, dann zeigt sich, daß sich das Maß an Gemeinsamkeiten [...] eher bescheiden ausnimmt.“ (Heinemann, 2000d, S. 509). Heinemann (2000e, S. 10) identifiziert vier Defizite: „die [...] das ganze Spektrum einer Textsortenlinguistik umfassen“: (i) Aufgrund der zahlreichen und heterogenen Untersuchungsansätze muss noch immer von einem prätheoretischen Gebrauch des Terminus Textsorten gesprochen werden. (ii) Es existiert kein Konsens bezüglich der Methoden, auf deren Basis Textsorten erfasst werden können. (iii) Es ist unklar, nach welchen Kriterien Textsorten untereinander differenziert werden können. (iv) Es liegen nur wenige Versuche vor, die das Relationsgefüge zwischen Textsorten zu erfassen versuchen, um generellere, systemhafte Klassifikationen entwickeln zu können.

Als erste Annäherung an die Multidimensionalität geht Heinemann (2000d, S. 508 f.) auf das Problemfeld der „Textsortenlinguistik“ ein, das unter anderem „von konkreten linguistischen Befunden über textuelle Strukturtypen, soziale, situative und funktionale Spezifika bis hin zu kognitiven Prozeduren der Individuen“ reicht. Es umfasst weiterhin „heterogene Texteinheiten von unterschiedlichem Umfang, unterschiedlicher Frequenz und Spezifizierung“, „unterschiedlichem Abstraktionsgrad“ und „unterschiedlichem Geltungsgrad“, basiert auf „unterschiedlichen lexikalischen und grammatischen Belegungen als Ausdruck der Variabilität und grundsätzlichen Offenheit der Textgestaltung“ und manifestiert sich „in vielschichtiger hierarchischer Abstufung“ (ebd.).

Ein textlinguistisch motiviertes Hypertextsortenmodell setzt eine Diskussion der Begriffe Textsorte und Texttypologie voraus. Vater (2001, S. 157) und Heinemann (2000d, S. 508) reduzieren in diesem Zusammenhang die Aufgabe der Textlinguistik auf drei Fragestellungen: (i) Wie lassen sich Textsorten genauer definieren und ihre zentralen Konstituenten herausarbeiten? (ii) Welche Textsorten existieren? Bestehen systematische Relationen zwischen diesen Textsorten? (iii) Falls solche (regelhaften) Relationen existieren, kann man sie erfassen und auf ihrer Grundlage Typologien von Textsorten aufstellen? Dieser Abschnitt versucht, den aktuellen Kenntnisstand[23] der Textlinguistik in Bezug auf diese und weitere Fragen darzustellen. Hierbei nimmt die Frage nach den „texttypologischen Regeln“ einen wichtigen Stellenwert ein, da das Verhältnis von Arbeiten, in denen Texttypologien erstellt werden, zu denjenigen, die sich der umfassenden Beschreibung einer oder mehrerer (meist verwandter) Textsorten widmen, in der Vergangenheit zu terminologischen Differenzen geführt hat (vgl. Isenberg, 1978, Adamzik, 1995, Heinemann, 2000c, 2000d).

Im Folgenden werden zunächst Textmuster als kognitive Basiseinheiten von Textsorten aufgefasst (Abschnitt 2.3.1), woraufhin eine Differenzierung der Termini Textklasse, Texttyp und Textsorte dargestellt wird (Abschnitt 2.3.2). Abschnitt 2.3.3 geht auf Texttypologien ein und erläutert den Zusammenhang zu Textsorten. Insbesondere neuere Arbeiten betonen die Relevanz der Textfunktion. Abschnitt 2.3.4 thematisiert diesen Aspekt und zusätzliche Typologisierungsebenen. Verfahren zur Binnendifferenzierung von Textsorten, werden in Abschnitt 2.3.5 vorgestellt. Daraufhin thematisiert Abschnitt 2.3.6 die Auffassung von Textsorten als Prototypen. Abschnitt 2.3.7 geht auf die *North American Genre Theory* ein.

2.3.1 Textmuster als kognitive Basiseinheiten von Textsorten

Während der kognitiven Verarbeitung stehen kommunizierenden Individuen in bestimmten Situationen spezifische Schemata zur Verfügung (Abschnitt 2.2.6), die auf das Produzieren und das Rezipieren von Texten ausgerichtet sind (Heinemann, 2000e, S. 22). Diese werden oftmals als Textmuster bezeichnet[24] und stellen, so Heinemann und Heinemann (2002, S. 130), „in erster Annäherung Rahmenmodelle dar für den Ablauf spezifischer Kommunikationsereignisse, die den Handelnden ein schnelles verbales Agieren und Re-agieren [sic] in bestimmten häufig wiederkehrenden Situationen erlauben, indem sie diese Rahmen durch

[23] Die Darstellung erfolgt primär anhand der aktuellsten Überblicksdarstellungen (Gansel und Jürgens, 2002, Heinemann und Heinemann, 2002, Adamzik, 2004) sowie Teil IV aus Brinker et al. (2000).

[24] Beispielsweise von Heinemann und Heinemann (2002, S. 130 ff.); Sandig (2000) versteht unter diesem Begriff die Abstraktionsebene der Textsorten (vgl. den nachfolgenden Abschnitt 2.3.2).

partiell wiederum ›vorgefertigte‹ Äußerungseinheiten und -strukturen ›auffüllen‹.“ Indizien deuten die Existenz derartiger globaler Textmuster an, so sind Menschen z. B. üblicherweise im Stande, in unterschiedlichen Situationen angemessen zu handeln. In spezifischen kommunikativen Situationen (z. B. beim Erzählen eines Märchens) werden immer wieder sehr ähnliche Textstrukturen produziert, wobei häufig spezifische Formulierungen („Es war einmal …“ etc.) eingesetzt werden. Bei der Rezeption eines Textes sind Menschen in der Lage, einen Text einer spezifischen Textklasse zuzuordnen und diese auch zu benennen, obwohl die Klassenzugehörigkeit nicht explizit signalisiert wird.

Heinemann und Heinemann (2002, S. 133 f.) definieren Textmuster als „kognitive Rahmeneinheiten und Operationsfolgen der Individuen zur Lösung von – auf Textganzheiten bezogenen – kommunikativen Aufgaben, d. h. auf erfolgreiche kommunikative Erfahrungen zurückgehende Orientierungsmuster für die Produktion und das Rezipieren von Texten; sie prägen die Erwartungshaltungen der Interagierenden“. Weiterhin werden sie als „Teilmengen des Interaktionswissens“, „konventionell geprägt“, „vage“, „flexibel“ und „mehrdimensional“ bezeichnet, denn „[w]ie die konkreten Texte selbst, so sind auch Textmuster nicht nur durch *eine* Dimension/Ebene zureichend erfassbar; vielmehr müssen für die Konstitution von Textmustern immer Merkmale mehrerer Ebenen zusammenwirken“ (ebd., S. 134).[25]

Bezüglich der Mehrdimensionalität haben die Abschnitte 2.2.8 und 2.2.9 verdeutlicht, dass verschiedene generelle Eigenschaften von Texten identifizierbar sind und dass kein Konsens über ihren jeweiligen Status bei der Definition des Konstrukts Text besteht, was in gleichem Maße auch für die Merkmale von Textmustern (und Textsorten sowie Texttypologien) gilt. Heinemann und Heinemann orientieren sich an Sandig (2000) und bezeichnen die folgenden Dimensionen als „grundlegend für Texte und Textmuster“ (2002, S. 134 f.): Funktionalität (Textmuster „als Modelle zur Lösung spezifischer kommunikativer Aufgaben“, ebd.), Situativität („die situative, interaktionale und diskursive Einbettung eines Textes“, ebd.), Thematizität („die Text-Thema-Geprägtheit“, ebd.), Formulierungsadäquatheit („das Wissen um spezifische Formulierungsmaximen und Formulierungsspezifika“, ebd.) und prototypische Gewichtung. Diese Eigenschaft bezieht sich auf die Überlegungen von Sandig (2000, vgl. Abschnitt 2.2.9), denn ein Textmuster enthält „nicht alle Charakteristika von potenziellen Textexemplaren einer bestimmten Textsorte […], sondern nur jene, die für die Textkonstitution relevant […] sind“ (Heinemann und Heinemann, 2002, S. 135).

Heinemann und Heinemann (2002, S. 138) gehen davon aus, dass die konzeptionellen Einheiten Textmuster und Textsorte unterschiedlich sind. Textmuster können demnach „als kognitive Operationen zur Konstitution und zum Verstehen von Texten angesehen werden […], die einem bestimmten Textsortenrahmen zuordenbar sind“. Es wird betont, dass zwar jede Textsorte ein konventionalisiertes Textmuster besitzt, zugleich aber nicht jedes Textmuster auf eine spezifische Textsorte bezogen ist:

> Die Textsortenspezifik ergibt sich erst sekundär aus dem jeweiligen allgemeineren Musterrahmen, d. h. aus einem solchen Rahmen (*Initialteil*, *Textkern*, *Terminalteil*) können zahlreiche Textsortenrealisierungen hervorgehen: *Privatbrief*, *Geschäftsbrief*, *Antrag*, *Eingabe*, *Widerspruch*, *mündliche Bitte* … (Heinemann und Heinemann, 2002, S. 139)

[25] Heinemann (2000d, S. 517, 2000e, S. 22) geht auf die Korrespondenz des Konzepts Textmuster und den Theorien der Kognitionspsychologie ein, d. h. Frames, Schemata, Scripts und mentale Modelle.

Als Beispiele für Textmuster, die „von nahezu allen Erwachsenen [...] aktiv und passiv beherrscht werden“ (ebd.) geben die Verfasser *Entschuldigungen*, *Privatbriefe* und *Alltagsgespräche* an, wohingegen die Textmuster *Leitartikel* oder *Sportreportage* aktiv nur von Journalisten beherrscht werden. Es existiert kein statisches textsortenbezogenes Textmustersystem, sondern die aktive und passive Kenntnis eines spezifischen Textmusters ist abhängig vom individuellen Entwicklungsstand, von der Bildung und vom Erfahrungsumfeld (vgl. Heinemann, 2000e, S. 22 ff.).[26] Als Differenzierungskriterium wird angeführt, dass Textmuster „allgemeine kognitive Rahmen-/Verfahrensvorgaben, also *kognitive Prozesse* zur Generierung und zum Verstehen/Verarbeiten konkreter Textexemplare [darstellen], während Textsorten *Ergebnisse kognitiver Operationen* – bezogen auf konkrete Textexemplare und deren Merkmale – [...] darstellen“ (ebd., S. 140; zweite Hervorhebung hinzugefügt, G. R.).[27]

2.3.2 Textklassen – Texttypen – Textsorten

Die übereinstimmenden charakteristischen Eigenschaften von Textexemplaren erlauben es, mehrere Texte zu einer Textsorte wie z. B. *Kochrezept*, *Wetterbericht* oder *Kontaktanzeige* zusammenzufassen und von anderen, nicht zugehörigen Textexemplaren abzugrenzen. Die Texte selbst sind als Instanzen der Textsorten aufzufassen. Nach Heinemann und Viehweger (1991, S. 144), die nicht zwischen Textsorte und Textmuster differenzieren, stellen Textsorten „ein bestimmtes Reservoir an Kenntnissen dar, auf die die Mitglieder einer menschlichen Gemeinschaft in ihren sprachlichen Tätigkeiten zurückgreifen“. Der Begriff Textsorte referiert „auf Alltagsklassifikationen, die innerhalb einer menschlichen Gemeinschaft erreicht und mit Lexikonzeichen belegt wurden, die das Wissen über eine bestimmte Textsorte »kondensieren«.“ (ebd.). Heinemann und Heinemann (2002, S. 140) betonen den Stellenwert der kognitiven Ebene (vgl. die Abschnitte 2.2.6 und 2.3.1), denn „der Begriff ›Textsorte‹ [erweist sich] letztlich als ein kognitives Phänomen, als ein auf einer bestimmten Menge von übereinstimmenden Merkmalen basierender Operator für Zuordnungsoperationen der Individuen; und als Ergebnis dieser kognitiven Operationen ergibt sich dann die Zusammenfassung einer bestimmten Menge konkreter Textexemplare zu einer (Text-)Klasse.“

Textsorten im Alltag

Häufig wird der Bezug von Textsorten zum Alltagswissen (Techtmeier, 2000, Heinemann, 2000d), zur Alltagssprache und -kommunikation (Heinemann, 2000e, Brinker, 2001) und zum kommunikativen Haushalt (Jakobs, 2003) herausgestellt, zu denen eine Fülle von Text-

[26] Heinemann (2000d, S. 516) fügt hinzu: „Im Alltagsverständnis sind Textsorten stets an konkrete Textexemplare gebunden, die [...] auch atypische Merkmale aufweisen können. Der Ausdruck *Textmuster* dagegen wird [...] als etwas Idealtypisches verstanden, als abstraktes Modell, in dem atypische Elemente keinen Platz haben.“

[27] An anderer Stelle charakterisieren Heinemann und Heinemann (2002, S. 155) den Zusammenhang zwischen Textmustern und Textsorten: „Das Textmusterwissen darf als Wissen über kognitive Operationen, vor allem zur Herstellung und Darstellung neuer Texte charakterisiert werden; das Textsortenwissen dagegen ist ein Wissen über Zuordnungsoperationen zur Klassenbildung auf der Basis von Relationen zwischen Texten und prototypischen Merkmalen von Textexemplaren. Von Textexemplaren als Repräsentanten einer bestimmten Textsorte kann man auf übergreifende Textmuster schließen; andererseits sind die Textmuster wiederum Voraussetzung für Zuordnungsoperationen bei der Bildung von Textklassen auf niederer Abstraktionsstufe.“

sortenbezeichnungen gehört.[28] Dimter (1981) demonstriert dies anhand einer Liste von mehr als 1 600 Textsortennamen, die aus dem Rechtschreibduden von 1973 zusammengetragen wurden. Etwa 500 dieser Bezeichnungen werden als „grundlegend" angesehen (z. B. *Bericht*), die anderen sind „abgeleitet" (z. B. *Wetterbericht*, *Segelflugwetterbericht*). Eine vergleichbare Vorgehensweise zur Sammlung von Hypertextsortennamen erscheint – trotz der Existenz der Lexeme „Homepage" und „Webseite" im aktuellen Duden, Band 01 (2004) – nicht sonderlich vielversprechend, da sich noch keine einheitliche Terminologie gebildet hat.

Somit ist zu unterscheiden zwischen der textlinguistischen Herangehensweise zur Charakterisierung von Textsorten und dem Umgang mit Textsorten, wie er von Individuen im Alltag verwendet wird, um mit etablierten Etiketten auf ausgezeichnete Typen von Texten zu referieren. Sammlungen von Textsortenbezeichnungen werden gelegentlich als „prätheoretisch" oder „vorwissenschaftlich" bezeichnet (z. B. von Linke et al., 2001, S. 248, und Gansel und Jürgens, 2002, S. 57). Schon Sandig (1972, S. 113) leitet ihren Beitrag mit dem Hinweis ein: „Ziel der Analyse ist die Beschreibung derjenigen Textsorten, für die die natürliche Sprache großenteils Lexeme besitzt." Seit der Arbeit von Dimter (1981) wird der Stellenwert dieser Etiketten für die Textlinguistik immer wieder betont und kontrovers diskutiert:

> Eine Charakterisierung der Benennungen für vorkommende Textsorten als (vorwissenschaftliches) Alltagsvokabular verkennt die Tatsache, dass [...] die „alltagsweltlichen Konzepte", die sich mit den Benennungen verbinden, nicht nur Alltagswissen über Texte, Textsorten, Textproduktion und Textrezeption darstellen, sondern auch über situative Einordnungen. (Gansel und Jürgens, 2002, S. 57)

Adamzik (1995, S. 12 ff.) kommt zu dem Schluss, „daß die Natürlichsprachlichkeit einer Textsortenbezeichnung als Argument gegen ihre Tauglichkeit nichts hergibt", denn alltagssprachlichen Klassifikationen fehlten die Merkmale wissenschaftlicher Klassifikationen, d. h. „Systematik, Explizitheit und Geschlossenheit" (ebd., S. 24). Des Weiteren hält Adamzik „die Anlehnung an alltagssprachliche Ausdrücke für sekundär; das Vorliegen solcher Ausdrücke stellt keinen hinreichenden [...] Grund für die Untersuchung dar." (ebd., S. 26). Heinemann (2000c, S. 537) hebt den wesentlichsten Aspekt in der Diskussion des alltagssprachlichen Textsortenbegriffs hervor: „Jede Typologie [...] sollte an das konventionelle Alltagswissen der Kommunizierenden über Textsorten anknüpfen, mit ihm kompatibel, zumindest darauf beziehbar sein".[29]

Textsorten und Texttypen als Textklassen

Die Begriffe Textsorte, Textklasse, Texttyp, Textart, Textmuster, Textform, Äußerungssorte und Kommunikationssorte (Heinemann, 2000d, S. 515, zählt einige weitere auf) werden häufig mit unterschiedlichen Definitionen, gelegentlich aber auch synonym verwendet

[28] Nach Heinemann (2000e, S. 15) sind Textsorten „keine Erfindungen von [...] Linguisten, sondern kommunikationspraktische Gegebenheiten [...] – auch in der Alltagskommunikation." Wissen über Textsorten wird in Bildungseinrichtungen (Schule, Hochschule etc.) vermittelt. Darüber hinaus bezieht man es sukzessive aus den eigenen Erfahrungen im privaten und beruflichen Alltag (Heinemann und Heinemann, 2002, S. 141).

[29] Dieser Aspekt wird auch von Miller (1984, S. 27) diskutiert (vgl. Abschnitt 2.3.7): „The genre classification I am advocating [...] seeks to explicate the knowledge that practice creates. This approach insists that '*de facto*' genres, the types we have names for [...], tell us something theoretically important about discourse."

(z. B. von Heinemann und Viehweger, 1991, S. 144, und Brinker, 2001, S. 129, vgl. auch Schoenke, 2000, S. 127).[30] Zur Beibehaltung einer definitorischen Konsistenz folgen wir in terminologischer Hinsicht Heinemann und Heinemann (2002, S. 142 ff.), die den Begriff *Textklasse* „allgemein und unspezifisch" auffassen als „Gesamtheit von potenziellen Textmengen/-Klassen überhaupt" (ebd.). Als Beispiele werden die „Gesamtmenge aller Presse-Texte", die „Menge aller Schrifttexte" oder „die Menge aller Wetterberichte oder aller Reisewetterberichte" angeführt. Die hierarchische Stufung, die durch diese Folge von Beispielen bereits angedeutet wird, ist auf den Geltungsbereich einer Textklasse zurückzuführen: Diejenigen Textklassen mit einem sehr umfangreichen Geltungsbereich – Heinemann und Heinemann sprechen von „Großklassen" – umfassen nur wenige Merkmale, die zwingend ausgeprägt sein müssen (z. B. dass ein Text in schriftlicher Form vorliegt bei der „Gesamtmenge aller Schrifttexte") und besitzen dementsprechend einen hohen Grad der Abstraktion. Textklassen mit einem niedrigeren Abstraktionsgrad benötigen ein umfangreicheres Inventar distinktiver Merkmale und besitzen somit einen geringeren Geltungsbereich. Derartige Basisklassen werden im Folgenden als *Textsorten*[31] bezeichnet, wohingegen Textklassen mit einem großen Geltungsbereich, d. h. Großklassen, *Texttypen* genannt werden. Zwischen diesen beiden Polen einer Textklassenhierarchie können nach Heinemann und Heinemann verschiedene Zwischenstufen existieren, die als *Textsortenklassen* bezeichnet werden.[32] Schließlich können noch unterhalb der Textsorten „vor allem *inhaltlich* geprägte Subklassen" (ebd., S. 143; Hervorhebung hinzugefügt, G. R.) existieren – *Textsortenvarianten* (vgl. Abbildung 2.2).[33]

Heinemann und Heinemann (2002, S. 144) betonen, dass diese Hierarchisierung keinesfalls als Absolutum aufzufassen ist, da Ausprägung und Anzahl der Ebenen und die Positionierung einer spezifischen Klasse in der Hierarchie „vom Anliegen des Klassifikators" abhängig sind, so könnte mit Bezug auf die linke Beispielhierarchie in Abbildung 2.2 auch *Zeitungstext* auf der Ebene des Texttyps angenommen werden, wenn eine detaillierte Analyse der Textklassenhierarchie von Zeitungstexten das Ziel der Untersuchung ist. Auch die in der

[30] Bereits Gülich und Raible (1972, S. 2) weisen auf dieses Problem hin. Gansel und Jürgens (2002, S. 53) ergänzen: „Die Linguistik ist von einem terminologischen Konsens, den Textsortenbegriff und die Textklassifikationen betreffend, nach wie vor weit entfernt." (vgl. Adamzik, 1995, S. 11 ff.). Nach Raible (1996, S. 59) existieren sogar „allergische Reaktionen auf den Terminus ›Textsorte‹". Neben der Terminologie ist auch die Methodologie betroffen: „Es ist [...] bisher nicht gelungen, eine einheitliche ›gültige‹ *Textsortenklassifikation* zu erstellen und es besteht auch noch kein [...] Konsens darüber, nach welchen Verfahren die Zuordnung eines Textes zu einer Textsorte genau erfolgen müsste." (Linke et al., 2001, S. 248).

[31] Der Begriff der Textsorte hat sich als Bezeichnung für Gebrauchstexte etabliert. Er grenzt sich (in der deutschsprachigen Forschung) von den literaturwissenschaftlichen Termini der Gattung und des Genres ab (vgl. Dammann, 2000, Gansel und Jürgens, 2002, S. 51, und Adamzik, 2004, S. 98).

[32] Gansel und Jürgens (2002, S. 56) kritisieren den Ausdruck, denn es werden „zwei Hierarchiestufen miteinander verschmolzen [...]. Dies geht u. E. nicht.". Diese Kritik ist ungerechtfertigt, wenn der Begriff Textklasse „allgemein und unspezifisch" als „Gesamtheit von potenziellen Textmengen/-Klassen überhaupt" (Heinemann und Heinemann, 2002, S. 142), d. h. als Terminus für beliebige Mengen von Texten aufgefasst wird.

[33] Adamzik (2004, S. 101) kritisiert die Ansätze, in denen eine „spezifische Lesart" (Adamzik, 1995) des Begriffs Textsorte präferiert wird. Diese hätten „zu einer terminologischen Verwirrung beigetragen, und ich sehe keine Anhaltspunkte dafür, dass irgendeiner dieser Vorschläge Chancen hätte, sich allgemein durchzusetzen." Bei der Wahl einer unspezifischen Lesart „könnten wir unter den Begriff *Textsorte* alle Ausdrücke für irgendwelche Mengen von Texten mit gemeinsamen Merkmalen fallen lassen, gleichgültig sogar, ob es sich dabei nun um geläufige alltagssprachliche Ausdrücke (*Briefe*, *Wetterberichte*), Syntagmen (*blaue Briefe*, *literarische Texte*) oder Fachausdrücke bzw. Kunstbegriffe (*Nämlichkeitsbescheinigungen*, *Paratexte*) handelt." (Adamzik, 2004, S. 101).

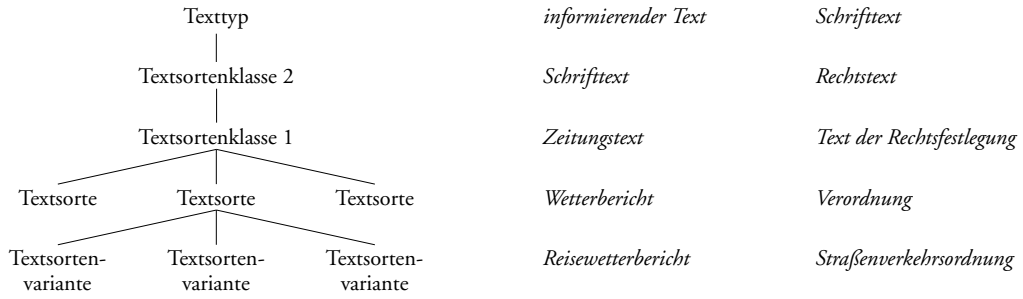


Abbildung 2.2: Hierarchische Abstufungen von Textklassen (nach Heinemann und Heinemann, 2002, S. 143, Heinemann, 2000d, S. 514, 2000e, S. 17)

Abbildung dargestellte Baumstruktur muss nicht zwangsläufig beibehalten werden, da „gerade die Zwischenstufen, die Textsortenklassen, mit anderen Textklassen-Repräsentationen derselben Hierarchiestufe vernetzt" sind (ebd.), wodurch eine Graphstruktur entsteht.

2.3.3 Textsorten und Texttypologien

Der Begriff Texttyp bezeichnet die oberste Ebene einer Hierarchie von Textklassen. Im Folgenden wird die Problematik des Aufbaus von Texttypologien thematisiert, die sich auf eine höhere Abstraktionsebene beziehen als die Binnendifferenzierung einer oder mehrerer Textsorten (vgl. Abbildung 2.2). Texttypologien verfolgen den Zweck, das „Textuniversum" (Dimter, 1981, S. 8), d. h. „die unendliche Vielfalt realer Texte auf eine überschaubare Menge von Grundtypen zu reduzieren" (Heinemann und Viehweger, 1991, S. 145). Besonders prägnant ist die in Bußmann (2002, S. 693) enthaltene Definition[34] des Forschungsbereichs Texttypologie, die „das Ziel einer konsistenten, theoretisch begründeten Klassifizierung von Texten [verfolgt], insbesondere einer systematischen Rekonstruktion des Begriffs Textsorte. Die Texttypologie soll einen Rahmen für das interaktionale und sprachliche Wissen liefern, das sich in der Benennung, Verwendung und Identifizierung von Textsorten zeigt."

Nach Heinemann und Heinemann (2002, S. 160 ff., die die Differenzierung von Adamzik, 1995, erweitern) sind fünf Grundtypen von Textordnungen[35] zu unterscheiden: (i) *Additive Aufzählungen und Reihungen von Textklassen* werden gerade in der Alltagskommunikation verwendet, denn „[solche] einfachen lockeren Bündelungen von Textklassen machen [...] bestimmte Teilbereiche der ›Textwelt‹ transparenter und damit leichter überschau- und beherrschbar." (ebd., S. 161). (ii) Als *Textsorten-Reihungen* werden obligatorische Sequenzen

[34] Die Definition orientiert sich an der Auffassung, dass es möglich sein sollte, *alle* Texte in *einer* Typologie zu erfassen, was im Lexikoneintrag für „Textsorte" deutlich wird: „Die (noch weitgehend ungelöste) Aufgabe einer Texttypologie ist es, die [...] alltagssprachlichen Textsorten-Konzepte mit textlinguistischen Kriterien zu beschreiben und sie [...] in einen theoretischen Rahmen zu integrieren." (Bußmann, 2002, S. 691).

[35] Heinemann (2000c, S. 540) stellt ebenfalls eine Reihe von Haupttypen vor, geht bei der Zusammenstellung jedoch nach forschungsbezogenen Schwerpunkten vor: (i) Klassifikation auf der Basis formaler Merkmale wie z. B. Texte mit dominant elliptischen Konstruktionen; (ii) Zuordnung auf der Grundlage der Textstrukturierung/Sequenzierung, z. B. von Texten mit spezifischen Initialteilen; (iii) Klassifikation auf der Basis von Inhalten, etwa Textsorten der Wirtschaft oder der Hochschule (vgl. Abschnitt 6.2, S. 300 ff.); (iv) Zuordnung auf der Grundlage situativer Spezifika (z. B. Face-to-Face-Kommunikation); (v) Funktionsorientierte Zuordnungen (vgl. z. B. Große, 1976); (vi) Mehrebenenmodelle, die auf multiplen Einordnungsinstanzen operieren.

der Anwendung von Textsorten bezeichnet. Als Beispiel dient der Ablauf zur Verabschiedung eines Gesetzes, der sich in der Anwendung spezifischer Textsorten widerspiegelt (*Vorschlag* bzw. *Antrag*, *Gesetzesentwurf*, *parlamentarische Beratung* unter Einbeziehung von *Gutachten*, dreimalige *Beratung* im Bundestag, *Abstimmung*, *Beschlussfassung* im Bundesrat, *Ratifizierung* durch den Bundespräsidenten, *Publikation* im Bundesgesetzblatt). Zu diesem Grundtyp können auch „Textsortengruppierungen" (Heinemann, 2000c, S. 540) gezählt werden, für die der Antrag auf Baugenehmigung mit vollständigen Unterlagen als Beispiel angeführt wird: *Lageplan*, *Baubeschreibung*, *Bauzeichnungen*, *Finanzierungsplan*, *Kaufvertrag* etc. (iii) *Lockere Zuordnungen von Textexemplaren zu Text-Großklassen* werden vorgenommen, um grundlegende Eckpunkte eines einzelnen Texttyps zu markieren, z. B. können *Nachrichten*, *Leitartikel*, *Todesanzeigen*, *Kultur-*, *Wirtschafts-* und *Sportberichte* als *Zeitungstexte* bezeichnet werden.

Eindimensionale Typisierungen

(iv) *Eindimensionale hierarchische Typisierungen* gehen von einem Texttyp aus, der als gemeinsame Einordnungsinstanz fungiert (Heinemann, 2000c, S. 539 f.). Die Zuordnung findet auf der Basis *eines* spezifischen Aspekts statt (z. B. Textfunktion, Situativität oder Inhalt, vgl. Abschnitt 2.2.3), wobei auch hierarchische Zwischenstufen gebildet werden: „Diese Subklassen repräsentieren [...] immer nur diese eine Grundeigenschaft von Textklassen, und die Klassifikatoren begnügen sich [...] mit wenigen, auf relativ hoher Abstraktionsstufe anzusetzenden, nicht distinktiven Sub-Einheiten, die folglich nicht bis zu den kommunikativen Basiseinheiten, den Textsorten, hinunterreichen" (Heinemann und Heinemann, 2002, S. 162).[36] Das Modell von Große (1976)[37] ist ein Beispiel einer derartigen Funktionstypologie (vgl. Tabelle 2.1). Schon anhand dieses sehr abstrakten Modells wird deutlich, dass ein einzelnes Basiskriterium für die Aufstellung einer umfassenden Typologie nicht ausreichend ist. Isenberg (1978)[38] stellt daher Anforderungen an eine theoretisch fundierte Texttypologie: *Homogenität* (einheitliche Typologisierungsbasis), *Monotypie* (keine Mehrfachzuordnung von Texten zu Textklassen), *Striktheit* (Ausschluss der Möglichkeit typologisch mehrdeutiger Texte im Geltungsbereich eines Texttyps) und *Exhaustivität* (die Typologie gilt für alle in ihrem Bereich auftretenden Texte). Nach Heinemann und Heinemann (2002, S. 158) ist jedoch „die Gesamtheit dieser Aspekte nicht gleichzeitig erfüllbar und in einer Typologie realisierbar".[39]

[36] Adamzik (1995, S. 32) zeigt, dass hierbei meist auf der Grundlage einer spezifischen Definition von Text vorgegangen wird (z. B. „kohärente Folge von Sätzen"; „abgeschlossene sprachliche Äußerung mit erkennbarer kommunikativer Funktion", ebd.), von der unterschiedliche Ausprägungen als Basis der Subtypen fungieren. Heinemann und Viehweger (1991, S. 143) zählen weitere Faktoren auf: „So gehen zahlreiche Klassifikationen von außerlinguistischen Faktoren wie Tätigkeitsbereichen, Situationen u. a. aus, andere wiederum sehen in den Zielen, Funktionen, Intentionen u. a. relevante Klassifizierungskriterien."

[37] Zitiert nach Heinemann und Heinemann (2002, S. 157).

[38] Zitiert nach Heinemann und Heinemann (2002, S. 167 f.).

[39] Kritik an diesem Ansatz übt auch Adamzik (1995, S. 21): „Und als wäre es der Probleme noch nicht genug, kommt Isenberg [1978] am Ende mit seiner These vom ›typologischen Dilemma‹ auch noch zu dem Schluß, daß die von ihm als notwendig angesehenen Anforderungen [...] nicht alle gleichzeitig erfüllbar sind, daß also die von ihm als einzig akzeptabel geforderte Art von Typologie ein Ding der Unmöglichkeit ist." Das „typologische Dilemma" wird als „methodologische Fehlkonstruktion" bezeichnet, „die dadurch zustandekommt, daß wissenschaftliche Analysekategorien [...] mit ›Sortierschubladen‹ einer technisch-praktischen Klassifikation verwechselt werden [...]." (ebd., S. 41).

Textklasse	Textfunktion	Beispiele
Sachinformierende Texte	Informationstransfer	*Nachricht*
Auffordernde Texte	Aufforderungen	*Gesuch*, *Werbung*
Selbstdarstellende Texte	Kundgabe	*Tagebuch*, *Biografie*
Kontakttexte	Kontaktfunktion	*Glückwunsch*
Normative Texte	Normative Funktion	*Gesetze*, *Vertrag*
Gruppenindizierende Texte	Gruppenindizierende Funktion	*Gruppenlieder*
Poetische Texte	Poetische Funktion	*Roman*, *Komödie*
Übergangsklasse	Zwei Funktionen dominieren gleichermaßen	*Gesetze*

Tabelle 2.1: Eindimensionale Funktionstypologie von Texten (nach Große, 1976)

Mehrdimensionale Typisierungen

(v) *Mehrdimensionale hierarchische Typisierungen* sind auf praktische Anwendbarkeit ausgelegt. Prinzipiell bauen sie auf eindimensionalen Modellen auf, es werden jedoch mehrere Dimensionen aufeinander bezogen, um Texte und Textsorten eines bestimmten Texttyps in einem hierarchischen System auf mehreren Abstraktionsebenen anzuordnen (vgl. Göpferich, 1995). Auf der obersten Hierarchiestufe wird ein Basiskriterium angenommen, woraufhin sukzessive weitere Ebenen in die Hierarchie eingezogen werden, die auf anderen Kriterien basieren (vgl. Abbildung 2.3). Hierdurch kann ein systematischer „Nachteil“ (Heinemann und Heinemann, 2002, S. 163) entstehen, da „ein und dieselbe Dimension mehrfach als Differenzierungskriterium herangezogen werden muss, je tiefer sie hierarchisch angesetzt ist.“ (ebd.). Aus diesem Grund kommt es auch häufig zu Mehrfachzuordnungen von Textsorten.

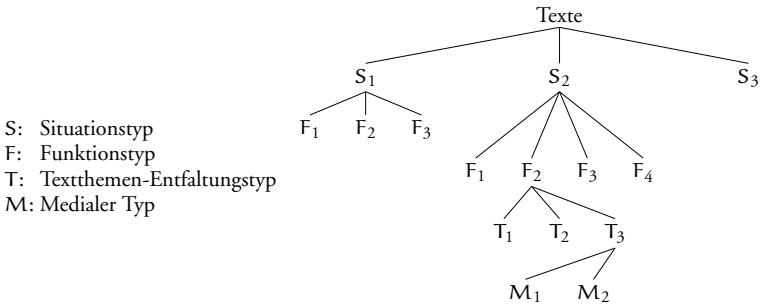


Abbildung 2.3: Mehrdimensionale hierarchische Texttypologie (nach Adamzik, 1995, S. 35)

Adamzik (1995, S. 30) geht ausführlich auf das Verhältnis von Texttypologien und Beschreibungen von Textsorten ein und verwendet eine „scharfe Abgrenzung“: Der Texttypologie geht es um die „systematische Klassifizierung von Texten mittels universell anwendbarer wissenschaftlicher Kategorien“ (ebd.), wohingegen sich die Textsortenforschung „auf die Beschreibung einzelsprachspezifischer kommunikativer Routinen richtet“ (ebd.), so dass die Typologisierung nur eine untergeordnete Rolle spielt. Dennoch stellt sich die Frage nach dem Verhältnis der Konzepte Textsorte und Texttyp. Nach Adamzik ist es zunächst „aussichtslos, irgendwo eine Grenze zu ziehen oder festlegen zu wollen, was einer standardisierten Textsorte und was einem systematisch abgrenzbaren Texttyp entspricht“, weshalb sie „die-

se unterschiedlichen Gegenstände als essentiell durch das Forschungsinteresse konstitutierte“ betrachtet (ebd., S. 31). Adamzik liefert eine Einschätzung unterschiedlicher Typologien:

> [Eindimensionale Texttypologien] stellen [...] nicht mehr und nicht weniger bereit als die Kategorien, mit deren Hilfe Texteigenschaften analysiert werden. Das Ziel dieser Modelle liegt gerade in der Erarbeitung solcher Kategorien – die selbstverständlich theorieabhängig sind, nämlich entsprechend der jeweiligen Auffassung darüber etabliert werden, *was eine (für bestimmte Fragestellungen relevante) Texteigenschaft ist, welcher Aspekt von Texten also (entsprechend welchem texttheoretischen Ansatz) systematisiert werden soll.* (Adamzik, 1995, S. 37; Hervorhebung hinzugefügt, G. R.)

Aus diesem Grund betrachtet Adamzik (1995, S. 38) eindimensionale Typologien als Verfahren „zur Erarbeitung von Analysekategorien“, schließlich gehöre „Typologisierung im Sinne von Sortierung nicht zum eigentlichen Interesse dieser Ansätze“ (ebd.). Bei mehrdimensionalen Typologien handelt es sich dagegen „tatsächlich um Typologisierungen“ im Sinne der Sortierung gegebener Mengen von Texten (ebd., S. 39). Als Einsatzzweck wird die Gewinnung eines ersten „Einblick[s] in die Struktur(ierungsmöglichkeit) des jeweiligen Gegenstandskomplexes“ genannt (ebd.).[40] Abschließend ist zu betonen, dass starre, statische Texttypologien nicht das Ziel der Textlinguistik sein sollten:

> Auch für die Sortierung von Texten und Textsorten gibt es kein mehr oder minder verbindliches, allgemeines und in sich geschlossenes System, sondern eine außerordentlich große Flexibilität und Variabilität der Klassenbildung und des Miteinander-in-Beziehung-Setzens von Texten [...]. Das allgemeinste Ziel von Text-Klassifikationen besteht daher nicht in der Aufstellung eines stringenten und abgeschlossenen/unveränderlichen Systems, sondern darin, eine bestimmte Teilmenge von Texten – immer mit dem Blick auf bestimmte Zwecke und die Bezogenheit auf andere Textmengen – überschaubarer zu machen [...]. (Heinemann und Heinemann, 2002, S. 159)

Für das Ziel der Erstellung einer Typologie von Hypertextsorten ist dieser Aspekt von zentraler Bedeutung, wie die Kapitel 11 bis 13 zeigen werden.

2.3.4 Ebenen der Texttypologisierung

Im Rahmen ihres pragmatisch-kommunikativen mehrdimensionalen Modells gehen Heinemann und Viehweger (1991, S. 148 ff.) auf fünf Typologisierungsebenen ein, die nachfolgend dargestellt und um eine zusätzliche Ebene ergänzt werden. Die fünf Ebenen bauen, beginnend mit den Funktionstypen, aufeinander auf, d. h. das Modell ist ein „Rahmenansatz über die Ziel- und Intentionskomponente“ (Heinemann und Viehweger, 1991, S. 150).[41]

[40] In späteren Arbeiten distanziert sich Adamzik von hierarchischen Modellen, denn es seien „Modelle vorzuziehen, die nicht eine Hierarchisierung der Typologisierungsbasen [...], sondern voneinander unabhängige Zuordnungen zu bestimmten Funktions-, Situations- und Themen(behandlungs-)typen vorsehen. Sie erlauben allgemein ein flexibleres Vorgehen [...].“ (Adamzik, 2004, S. 101).

[41] Bei einer Anwendung „kommt diesen Ebenen [...] unterschiedliches Gewicht zu: Bei einem Antrag/einer Bitte dominiert der funktionale Aspekt; Telegramme/Briefe/Fernsehsendungen sind eher primär situativ geprägte Konzepte, und bei bestimmten Texten vor allem des pädagogischen Bereichs kann [...] das Beherrschen bestimmter Textdarstellungs- bzw. Herstellungsverfahren oder charakteristischer Textstrukturen im Fokus des Interesses [...] stehen.“ (Heinemann und Viehweger, 1991, S. 171).

Funktionstypen

Die Textfunktion umfasst nicht nur eine produzentenseitige Intention,[42] deren Realisierung mit einem Text angestrebt wird. Ein Text ist vielmehr in übergreifende soziale und kommunikative Zusammenhänge und Interaktionen eingebettet. Die Auswahl einer begrenzten Menge von Funktionstypen als Basis einer Texttypologisierung ist prinzipiell arbiträr, da Texte unterschiedliche Funktionen einnehmen können.[43] Heinemann und Viehweger (1991, S. 149) gehen von vier „Primärfunktionen des Kommunizierens" aus, die reflektieren, „was Texte in Interaktionsakten generell bewirken können" (ebd.): (i) „sich ausdrücken", (ii) „kontaktieren" (Begrüßungen, Pausengespräche, Grußpostkarten), (iii) „informieren" (umfasst informationsermittelnde und informationsvermittelnde Texte) und (iv) „steuern" (Handlungsanweisungen, Befehle, Gesuche, Ratschläge, Bittschriften, Studienprogramme, Absprachen etc.). Die Grundtypen stehen untereinander in einem Inklusionsverhältnis, d. h. steuernde Texte, vermitteln z. B. immer auch Informationen und basieren auf einem Kontakt der Kommunikationspartner; „sich ausdrücken" ist daher die allgemeinste Grundfunktion. Es können keine trennscharfen Grenzen zwischen den Typen gezogen werden, so kann ein Kontakttext z. B. in einen Informationstext münden. Häufig liegt jedoch eine dominante kommunikative Funktion vor. Heinemann und Viehweger nennen eine weitere Funktion, die sich auf das Erzielen einer ästhetischen Wirkung bezieht.[44] Sie betrifft vornehmlich literarische Texte und überlagert typischerweise die anderen vier Funktionen.

Situationstypen

Texte hängen auch von der Kommunikationssituation ab, so wird z. B. eine Bitte an einen sozial gleichgestellten Partner in anderer Weise formuliert als eine Bitte identischen Inhalts an den Vorgesetzten (vgl. Heinemann und Viehweger, 1991, S. 153 f.).[45] Heinemann und Viehweger (1991, S. 154) verwenden einen weit gefassten Situationsbegriff und geben zu bedenken, dass der „skizzierte Ansatz [...] nur als Versuch verstanden werden [kann], aus dem Kontinuum der Gesamtheit situativer Faktoren einzelne [...] Aspekte herauszuheben."

[42] Nach Heinemann (2000e, S. 14) besitzen funktionsorientierte Ansätze zwar eine gewisse Plausibilität, da „Texte ja immer nur dann produziert werden, wenn Sprecher bei Partnern etwas bewirken wollen", zugleich dürfe der Begriff nicht missverstanden werden: Einerseits bezieht er sich auf die Absicht, das Ziel, die Intention des Produzenten bzw. auf den Zweck der kommunikativen Handlung, andererseits jedoch auch auf die „generelle Empfänger-Orientiertheit der Textgestaltung" (ebd.). Auch Klein (2000, S. 33) äußert sich kritisch: „Üblicherweise wird Texten und Textsorten unproblematisch unterstellt, dass ihnen eindeutig eine und nur eine (dominierende) kommunikative Funktion zugesprochen werden könne. Das muss [...] bezweifelt werden."

[43] Heinemann (2000c, S. 533 ff.) geht auf Texttypologien ein, die sich am Funktionsbegriff orientieren und stellt die „Heterogenität der zahllosen Funktionsmodelle" durch eine auflistende und explizit als „unvollständig" markierte Zusammenfassung eines Teils der in der Literatur berichteten Funktionen von Texten deutlich heraus (ebd., S. 534): „informieren", „mitteilen", „sachinformierend handeln", „behaupten", „assertieren", „darstellen", „referentiell handeln", „feststellen", „festlegen" und 41 weitere (vgl. Abschnitt 2.3.5).

[44] Vater (2001, S. 178) ist zuzustimmen, wenn er anmerkt: „Die Schaffung [...] einer fiktionalen Welt scheint mir dabei eine weit größere Rolle zu spielen als die Auslösung emotionaler Bewusstseinsprozesse, da diese auch durch einen informierenden Text (z. B. über die Krankheit eines dem Produzenten und/oder Adressaten nahe stehenden Menschen) oder einen steuernden Text (z. B. die Bitte um einen Kuss) angestrebt werden können."

[45] Die Situativität betrifft eine „abstraktere Frage [...], nämlich die nach der Welt, in der die Texte angesiedelt sind bzw. in der die Interaktanten sie situieren" (Adamzik, 2004, S. 61, vgl. Heinemann, 2000c, S. 531 ff.).

(ebd.). Die Basis der Grundtypen wird durch die interaktionalen Rahmentypen gebildet, die sich in eigenständige kommunikative Tätigkeiten und Tätigkeiten im Dienste übergeordneter nicht-kommunikativer Tätigkeiten unterteilen lassen. Auf der zweiten Ebene können Interaktionsereignisse nach der sozialen Organisation differenziert werden, denn die „meisten Interaktionsereignisse sind institutionell geprägt" (ebd., S. 155), wobei z. B. Handel und Dienstleistungen, Gesundheitswesen oder Wissenschaft als Kommunikationsbereiche fungieren können, in denen Kommunizierende spezifische soziale Rollen besitzen (ebd., S. 156). Die weiteren Ebenen beziehen sich auf eine Klassifizierung nach der Anzahl der Kommunikationspartner, ihren sozialen Rollen (symmetrische versus asymmetrische Verteilung) und der Umgebungssituation (z. B. in Bezug auf das Kommunikationsmedium).[46]

Verfahrenstypen

Die dritte Ebene ist sehr heterogener Natur. Unter dem Sammelbegriff der Verfahren verstehen Heinemann und Viehweger (1991, S. 158) „in erster Annäherung Vorgehensweisen von Handelnden zur effektiven Lösung vorgefaßter oder sich aus bestimmten Situationen ergebender Ziele", wobei textproduzierende und textrezipierende Verfahren unterschieden werden, die sich z. B. auf die Entscheidung für ein bestimmtes Textthema und das Verfahren zur Realisierung des sprecherseitigen Anliegens beziehen. Zwei Prozeduren sind insbesondere für die Textstrukturierung von Bedeutung: Textentfaltungsprozesse betreffen unter anderem die zu berücksichtigende Informationsmenge und die Entfaltung von Subthemen. Die strategische Prozedur umfasst die Frage, ob ein Thema z. B. mit narrativen, deskriptiven oder argumentativen Verfahren möglichst erfolgversprechend zu realisieren ist.

Textstrukturierungstypen

Heinemann und Viehweger (1991, S. 161) vertreten den Standpunkt, dass es nicht möglich ist, „feste Strukturierungsmuster für jede einzelne Textklasse aufzustellen". Daher gehen sie von vier „grundlegenden Strukturierungstypen" aus (ebd.), denen die Strukturen von Texten zugeordnet werden können. Hierzu zählt die Aufgliederung eines Textes in Teiltexteinheiten, die Ausgliederung eines Initial- oder Terminalteils, die thematische Fixierung des Textkerns und die Abfolge der Teiltexte. Zur internen Strukturierung von Teiltextkomplexen sind Sequenzierungs- und Konnexionsprozesse relevant. Hierbei gelten die bereits angesprochenen vier Typen, sie beziehen sich jedoch auf eine spezifischere Textebene.

Formulierungsmuster

Diese Ebene betrifft „prototypisches Wissen über Formulierungsmerkmale bestimmter Textklassen" (ebd., S. 165). Zur Begründung führen Heinemann und Viehweger an, dass der Produzent keine vollständige Freiheit bei der Wahl von Textformulierungen besitzt und dass Rezipienten in der Lage sind, Textexemplare mit unterschiedlichen Formulierungen ein und

[46] Auf der Grundlage dieser fünf Ebenen charakterisieren Heinemann und Viehweger (1991, S. 158) den situativen Rahmen eines Briefes mit der Bitte um Verlegung eines Telefonanschlusses als (i) „gerichtet (vor allem) auf gegenständlich-praktische Tätigkeit des Rezipienten", (ii) „institutionelle" und (iii) „dyadische Kommunikation", (iv) „asymmetrisch" sowie (v) „Aufzeichnungskommunikation".

derselben Textklasse zuordnen zu können. Es werden mehrere Bereiche unterschieden: Die „textklassenspezifischen Kommunikationsmaximen“ sind nach Heinemann und Viehweger „generelle Ordnungs- und Formulierungsprinzipien“, die sich aus „textklassenspezifischen Gestaltungsprinzipien“ ergeben und den Spielraum des Produzenten zur individuellen Ausgestaltung eines Textes einschränken; diese Prinzipien werden auch als „Stilzüge“ bezeichnet, die mit intuitiven Schlagwörtern wie z. B. „knapp“, „prägnant“ und „höflich“ charakterisiert werden.[47] Der zweite Bereich betrifft die eigentlichen Formulierungsmuster, unter denen Wörter und Konstruktionen verstanden werden, die sich bei einer Kommunikationsaufgabe bewährt haben (ebd., S. 166). Heinemann und Viehweger unterscheiden zwischen (i) Einzellexemen (z. B. „Haft“, „Urteil“ und „Plädoyer“ aus dem Bereich des Rechtswesens), (ii) Kollokationen (z. B. „Lehrer und Erzieher“ und „leistungsschwache Schüler fördern“ aus dem Bereich des Bildungswesens oder unmittelbare Indikatoren für Textsorten, z. B. „in tiefer Trauer“, „im Namen des Volkes“ und „es war einmal“), (iii) stereotypen Textkonstitutiven (z. B. „Guten Tag“, „Grüß Dich!“ und „Hallo!“ als Begrüßungsformeln oder „Womit kann ich dienen?“ und „Was darf's sein?“ zu Beginn eines Verkaufsgesprächs) und (iv) Gliederungssignalen. Heinemann und Viehweger (1991, S. 168) weisen darauf hin, dass Formulierungsmuster zwar als Hinweise auf bestimmte Textsorten fungieren können, eine isolierte Betrachtung der Formulierungsebene erweise sich aber als unzureichend.

Thema und Thementypen

Heinemann und Viehweger (1991) widmen dem Textthema keine eigene Typologisierungsebene, bei den Verfahrenstypen wird es lediglich angerissen.[48] Bei der Analyse eines umfassenden Korpus inhaltlich heterogener Textexemplare ist es notwendig, eine Unterscheidung auf der Basis des Themas/Inhalts vornehmen zu können.[49] Der Terminus Thema ist mehrdeutig und bezeichnet in der Alltagssprache Brinker (2001, S. 55) zufolge den „kommunikativen Hauptgegenstand“, d. h. „den dominierenden Referenzträger“ eines Textes. Das Thema stellt den Kern des Textinhalts, seinen Hauptgedankengang dar. Es ist entweder explizit in einem spezifischen Textteil realisiert oder muss durch die verkürzende Paraphrase aus dem Textinhalt abstrahiert werden (ebd., S. 56, vgl. auch Abschnitt 2.2.2).

Adamzik (2004) nimmt eine weitere Differenzierung vor und unterscheidet das Thema (i) als zentrales Referenzobjekt bzw. fokussierten Gegenstand, (ii) als Informationskern und (iii) „als Problemstellung bzw. als das Fragliche, die Quaestio oder auch Strittige“ (Adamzik, 2004, S. 120). Diese Lesarten korrelieren mit (i) deskriptiven, (ii) narrativen und (iii) argumentativen Texten. Diese drei Texttypen sind wiederum eng mit verschiedenen Themenentfaltungstypen bzw. Vertextungsmustern verbunden (Deskription, Narration, Argumentation, siehe Brinker, 2001, S. 55–82, Brinker et al., 2000, sowie Abschnitt 2.2.2). Adamzik (2004,

[47] Heinemann und Viehweger (1991, S. 165) erläutern, dass „Anzahl und Klassifikation der Stilzüge heute noch als weitgehend ungelöste Probleme angesehen werden [müssen], ebenso die Frage der »Bindung« von Stilelementen an solche allgemeinen Textgestaltungsprinzipien“.

[48] Nach Ansicht von Adamzik (2004, S. 118) besteht bei der Analyse inhaltlicher Aspekte „weniger Einheitlichkeit und Klarheit über dabei zu verwendende Kategorien“ als bei funktionalen Aspekten und „die Frage [...], wovon denn in Texten überhaupt die Rede sein kann, [wird] als solche kaum einmal aufgeworfen“.

[49] Die maschinelle Kategorisierung von Texten bezieht sich nahezu immer auf thematische Kategorien wie z. B. „Politik“, „Sport“, „Wissenschaft“, „Wirtschaft“ und „Unterhaltung“ (vgl. Abschnitt 14.2.1).

S. 125) strebt den Aufbau eines abstrakten Kategorieninventars von Textthemen an.[50] Auf der Basis von Brinker (2001) wird eine Liste konstruiert: *Vorgang*, *Ereignis*, *Geschehen*, *Sachverhalt*, *Lebewesen*, *Gegenstand*, *Ding*, *Zustand*, *Prozess*, *These*, *Behauptung*, *Proposition* und *Aussage*. Daraufhin stellt Adamzik drei Gruppen auf, die als Thementypen (jedoch nicht als Taxonomie) fungieren: (i) Statische Objekte (unbelebte Dinge, Gegenstände, Lebewesen, Zustände; die Objekte besitzen Eigenschaften und sind situierbar), (ii) dynamische Objekte (einmalige, typische oder wiederholbare Ereignisse, Vorgänge, Handlungen; sie erfordern einen Agens und sind oftmals mit statischen Objekten verbunden; Handlungen können Bewertungen, Motiven und Zwecken unterliegen) und (iii) kognitive Objekte (Begriffe, Kategorien, Propositionen, Theorien). Adamzik spezifiziert für die Typen verschiedene Einschränkungen, die verdeutlichen, „dass bei einem gegebenen Thementyp die inhaltliche Ausführung auf einem sehr abstrakten Niveau voraussehbar ist“ (Adamzik, 2004, S. 125). Von zentraler Bedeutung ist nun die Frage, ob bei einer gegebenen Textsorte ein korrespondierer Thementyp vorhersehbar sein kann, welche Wechselwirkungen also zwischen Textsorte und Textthema (sowie Textfunktion und Textsituation) bestehen.[51] Adamzik (2004, S. 128) kommt zu dem Schluss, „dass die aus dem Alltag bekannten Textsorten(bezeichnungen) großen Aufschluss [...] über die behandelten Themen geben.“ Der „große Aufschluss“ bezieht sich jedoch in der Regel nur auf eine grobe „Vororientierung“ (ebd.) bezüglich der Einschränkung der Gesamtmenge aller potenziellen Themen auf eine Gruppe möglicher Themen und eine weitere Gruppe ausgeschlossener Themen. Weiterhin erlauben die meisten Textsortennamen Aussagen über Merkmale des Thementyps und potenziell beteiligte Teilthemen: „Bereits ein inhaltlich spezifisches Stichwort reicht aus, um kognitive Schemata zu aktivieren, die miteinander zusammenhängende Konzepte und Verbindungen zwischen ihnen aufzurufen.“ (ebd.).

2.3.5 Differenzierung und Modellierung von Textsorten

Texte können auf der Grundlage gemeinsamer Eigenschaften verschiedenen Textklassen zugewiesen werden (vgl. Abschnitt 2.3.2), die unterschiedliche Abstraktionsgrade und Geltungsbereiche besitzen, welche wiederum von der Anzahl und Komplexität der Merkmale abhängig sind. Die Auswahl und Gewichtung der Merkmale, die im Rahmen einer Textsortenanalyse oder -beschreibung verwendet werden, ist arbiträr, weil Texte von beliebigen Perspektiven betrachtet werden können. Daher ergeben sich die Fragen nach der Annahme einer oder mehrerer Basisebenen, auf die spezifische Merkmale bezogen werden, der Auswahl der Merkmale und auf welche Weise Merkmalsausprägungen auf bestimmten Basisebenen schließlich eine Textsorte ergeben. Heinemann und Heinemann (2002, S. 145) kritisieren

[50] Es handelt sich also um eine *top-down*-Ansicht, d. h. es sollen möglichst präzise Aussagen über Mengen heterogener Textexemplare gemacht werden. Die *bottom-up*-Ansicht entspräche der Analyse der thematischen Struktur eines gegebenen Einzeltextes (vgl. Abschnitt 2.2.2).

[51] Dimter (1981, S. 116 ff.) stellt anhand einer Stichprobe von 40 „grundlegenden“ und 40 „abgeleiteten“ Textsortenbezeichnungen fest, dass Textsortenbezeichnungen Hinweise auf diese drei Ebenen geben. Informationen über die Kommunikationssituation geben 84,2% der Bezeichnungen, 80,3% der Namen implizieren die Textfunktion, und 75% geben Hinweise auf den Inhalt. Informationen über alle drei Kategorien werden in 57,9% der Bezeichnungen gegeben, wohingegen „ausschließlich inhaltlich mit 0% (!) [...] eine weniger wichtige Rolle [spielt].“ (ebd.). Aus diesem Grund spekuliert Dimter, „daß beim Reden über Texte als Instanzen von Textklassen zunächst die Kommunikationssituation und die Textfunktion von Interesse sind.“ (ebd., S. 117).

in diesem Zusammenhang, dass die Beschreibung einer Textsorte nicht nur die Aufzählung auffälliger Merkmale umfassen sollte, stattdessen sollte eine systematische Charakterisierung stattfinden, die bereits auf der Ebene der Merkmalseruierung ansetzt.

Heinemann (2000d, S. 509 ff.) unterscheidet vier Ansätze zur Differenzierung von Textsorten: Die erste Gruppe basiert auf Textgrammatiken (vgl. Abschnitt 2.2.1), die Merkmalbündel liefern, welche zur Differenzierung von Textsorten verwendet werden können (z. B. Sandig, 1972). Die zweite Gruppe beschäftigt sich mit Bedeutungskomplexen (z. B. van Dijk, 1980). Situationsmodelle fokussieren die Umgebungs- und Kommunikationssituation (Medium, Handlungsbereich etc.). Funktionsmodelle betrachten primär das kommunikative Funktionieren von Texten. Nachfolgend werden einige dieser Ansätze diskutiert.

Die Überblicksdarstellung von Linke, Nussbaumer und Portmann (2001)

In ihrem *Studienbuch Linguistik* behandeln Linke et al. (2001) das Thema Textsorten auf etwa sieben Seiten in stark kondensierter, theorieneutraler Form (ebd., S. 248–255), weshalb es für einen initialen Überblick besonders geeignet erscheint. Die Rede ist zunächst nicht von Merkmalen, sondern von heterogenen, „auf den verschiedensten Analyseebenen von Sprache [vgl. Abschnitt 2.2, G. R.] anzusiedelnden [...] Kriterien“ (ebd., S. 248), die es erlauben, einen Text einer bestimmten Textsorte zuzuordnen, wobei zwischen textinternen und textexternen Kriterien unterschieden wird (vgl. Abschnitt 2.2.8).[52] Es wird explizit darauf hingewiesen, dass die Liste der dargestellten Kriterien nicht vollständig ist.

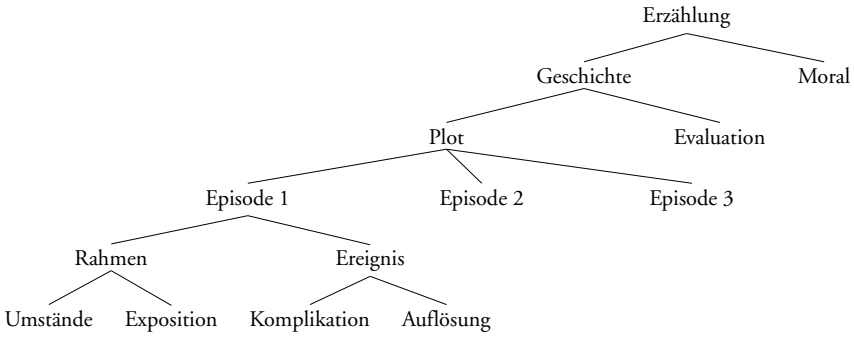


Abbildung 2.4: Beispiel eines Textstrukturmusters (nach Linke et al., 2001, S. 249)

Zu den textinternen Kriterien zählt die grafische Ebene (Hand- oder Maschinenschrift, Druck, grafische Textgestaltung), die Wortwahl (Wortschatz, charakteristische Schlüsselwörter), Art und Häufigkeit von Satzbaumustern, die Themenbindung und der Themenverlauf, das Thema[53] selbst und Textstrukturmuster (charakteristische Gliederungs- oder Baumstruktur). Für das zuletzt genannte Kriterium geben Linke, Nussbaumer und Portmann das in

[52] Bei der Diskussion textinterner und textexterner Kriterien wird häufig auf de Beaugrande und Dressler (1981) verwiesen. Eine derart eindeutige Zuordnung kann jedoch nicht immer vorgenommen werden, da textexterne Merkmale „selbstverständlich auch im Text thematisiert sein können“ (Gülich und Raible, 1972, S. 3).

[53] Vermutlich in Anlehnung an Dimter (1981) meinen Linke et al. (2001, S. 249): „Dies wird z. T. bereits in den Benennungen von Textsorten deutlich, so z. B. wenn wir *Liebesromane* von *Abenteuerromanen* unterscheiden oder von *Kriegsberichten*, *Todeslyrik* oder *Geburtsanzeige* sprechen.“

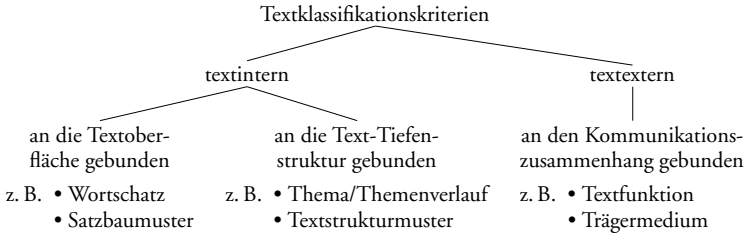


Abbildung 2.5: Hierarchie von Klassifikationskriterien (nach Linke et al., 2001, S. 251)

Abbildung 2.4 reproduzierte Beispiel an.[54] Zu den textexternen Kriterien zählen Linke et al. die Textfunktion, das Kommunikationsmedium und die Kommunikationssituation.

Zusammenfassend stellen Linke et al. (2001, S. 251) die Kriterien in Form von Abbildung 2.5 dar: „[Es] sind nicht alle [...] Klassifikationskriterien gleich gut operationalisierbar. Die Kriterien, die sich an der Textoberfläche ›materiell‹ festmachen lassen, sind in jedem Fall praktikabler als diejenigen, die nur indirekt im Text selbst manifestiert sind“. Der Aspekt der Operationalisierbarkeit wirkt sich unmittelbar auf computerlinguistische Verfahren zur maschinellen Detektion von Textsorten aus, die in Abschnitt 14.2 diskutiert werden.

Merkmalbasierte Differenzierung von Textsorten nach Sandig (1972)

Texttypologien haben sich, wie Heinemann und Viehweger (1991, S. 135) betonen, zu Beginn der siebziger Jahre vor allem an der Komponentialitätsthese orientiert, „nach der sprachliche Entitäten [...] aus elementaren, diskreten Bausteinen konstituiert werden. Eine Textsorte wurde demzufolge als eine Kombinatorik [...] von Merkmalen verstanden, [...] die jeweils spezifische Aspekte einer Textsorte reflektieren.“ (ebd.). In ihrem „fragmentarischen Vorschlag zur Differenzierung von Textsorten mittels Merkmalsoppositionen“ berücksichtigt Sandig (1972, S. 113) sowohl die „Gegebenheiten der Kommunikation anhand des Kommunikationsmodells“ als auch „einige sprachliche Eigenschaften von Textsorten“ (ebd.).[55] Auf unterschiedliche Typen von Rezipienten geht Sandig dabei nicht ein: „Eine Abstraktion von den Benutzergruppen und den typischen Verwendungssituationen ist nur vorläufig zum Zweck genauerer Untersuchung der Textsorten gerechtfertigt.“ (ebd.).

Den Kern des Modells bilden 20 Merkmale, die sich auf allgemeine Kommunikations- und Handlungsbedingungen, grammatische Merkmale und Präsignale beziehen. Es können drei unterschiedliche Werte (+: Merkmal ausgeprägt; −: Merkmal nicht ausgeprägt; ±: Ausprägung möglich, aber nicht notwendig) vorliegen. Das Merkmal [+ gesp] bedeutet z. B. „gesprochen“, [− gesp] meint „geschrieben“. Weitere Beispiele sind „spontan“ [+ spon], „mo-

[54] Es handelt sich um eine Darstellung von Gülich und Raible (1977, S. 267), die wiederum auf einer Abbildung von van Dijk beruht. Besonders deutlich ist der Einfluss des Story-Grammar-Ansatzes (vgl. Abschnitt 2.2.6).

[55] Der Beitrag entstammt dem Tagungsband des ersten speziell den Textsorten gewidmeten Kolloquiums (Gülich und Raible, 1972). Nach Adamzik (1995, S. 11) hat von diesen Beiträgen „eigentlich nur der von Sandig [...] in der späteren Forschung eine größere Rolle gespielt. [... Es] kann nicht übersehen werden, daß dieser erste Entwurf einer systematischen Abgrenzung von Gebrauchstexten außerordentlich stimulierend wirkte.“

	gesprochen	spontan	monologisch	dialogische Textform	räumlicher Kontakt	zeitlicher Kontakt	akustischer Kontakt	Form des Textanfangs	Form des Textendes	festgelegter Textaufbau	Thema festgelegt	1. Person	2. Person	3. Person	Imperativformen	Tempusformen	ökonomische Formen	Redundanz	Ausschließlich Sprachliches	Gleichber. Kommunikationsp.
Interview	+	±	−	−	±	+	+	±	±	−	+	+	+	+	±	±	±	±	+	−
Brief	−	±	±	−	−	−	−	+	+	−	±	+	+	+	±	±	±	±	+	±
Telefongespräch	+	±	−	−	−	+	+	+	+	−	±	+	+	+	±	±	±	±	±	±
Gesetzestext	−	−	+	−	−	−	−	+	+	−	+	−	−	+	−	−	−	−	+	−
Arztrezept	−	−	+	−	−	−	−	+	+	+	+	−	−	−	−	−	+	−	+	−
Kochrezept	±	−	+	−	±	±	±	+	−	+	+	−	−	+	±	−	±	−	+	−
Wetterbericht	±	−	+	−	−	+	±	+	−	+	+	−	−	+	−	−	±	−	+	−
Traueranzeige	−	−	+	−	−	−	−	+	+	+	+	±	−	+	−	−	±	−	±	±
Vorlesung(stunde)	+	±	+	−	+	+	+	+	±	−	+	±	±	+	±	±	−	±	±	−
Vorlesungsmitschrift	−	−	+	−	−	−	−	±	−	−	+	−	−	+	−	−	+	−	±	+
Reklame	±	±	±	±	±	±	±	±	±	−	±	±	±	±	±	±	±	±	±	−
Stelleninserat	−	−	+	−	−	−	−	+	+	+	+	±	±	+	±	−	±	−	+	−
Rundfunknachrichten	+	−	+	−	−	+	+	+	+	−	−	−	−	+	−	+	−	±	+	−
Zeitungsnachricht	−	−	+	−	−	−	−	+	−	−	+	−	−	+	−	+	−	−	+	−
Telegramm	−	−	+	−	−	−	−	+	+	−	+	±	±	+	±	−	+	−	+	±
Gebrauchsanweisung	−	−	+	−	−	−	−	±	−	−	+	−	±	+	±	−	±	±	±	−
Diskussion	+	±	−	−	±	+	+	+	+	−	+	+	+	+	±	+	−	±	+	±
familiäres Gespräch	+	+	−	±	+	+	+	±	−	−	−	+	+	+	±	±	±	+	±	+

Tabelle 2.2: Merkmalbasierte Klassifikation von Gebrauchstextsorten (nach Sandig, 1972, S. 118; Spaltenköpfe teilweise nach Heinemann und Viehweger, 1991, S. 136)

nologisch“ [+ mono] und „dialogisch“ [− mono].[56] Tabelle 2.2 zeigt eine Anwendung der Merkmale.[57] Sandig bezeichnet ihren Ansatz explizit als „fragmentarischen Vorschlag“:

> [Diese] Charakterisierungen von Textsorten sind aber viel zu grob [...]. Bei der großen Zahl der Merkmale, die notwendig wären, um eine möglichst genaue Abgrenzung aller gebrauchssprachlichen Textsorten zu erreichen, ist es die Frage, ob es sinnvoll erscheint, solche Merkmalskombinationen überhaupt zu erstellen. (Sandig, 1972, S. 119)

Eine „möglichst genaue Abgrenzung“ kann also nicht angestrebt werden, weshalb die Fragestellung der jeweiligen Untersuchung im Vordergrund stehen muss (versus dem vermeintlichen Idealkonstrukt einer universalen, allgemein gültigen Texttypologie), die somit zwangsläufig einer hohen Abstraktionsebene unterliegt (bei Sandig ist dies eine Unterscheidung der „groben Texteigenschaften“ von 18 Gebrauchstextsorten). Nach Vater (2001, S. 164) sind die Merkmale „recht brauchbar, müßten jedoch an größeren Textkorpora überprüft werden“. Ähnlicher Meinung sind Heinemann und Viehweger (1991), die den Ansatz als „einen

[56] Die weiteren Merkmale bedeuten: [tdia] (monologische Kommunikation, obwohl der Text eine Dialogform besitzt), [rkon] („räumlicher Kontakt von Sender und Empfänger“, Sandig, 1972, S. 116), [zkon] („zeitliche Kontinuität der Kommunikation“, ebd.), [akon] (es besteht akustischer Kontakt), [anfa] und [ende] (Formelhaftigkeit, d. h. besondere sprachliche Formen von Textanfang und Textende), [aufb] (durch Konvention weitgehend festgelegter Textaufbau), [them] (markiert „[ob] das Thema ziemlich genau festgelegt ist“, ebd., S. 117), [1per], [2per], [3per], [impe] (die „Weise der Interaktion der Kommunikationspartner“, ebd., d. h. , erste, zweite oder dritte Person oder Imperativ), [temp] (restringierter Gebrauch der Tempora), [ökon] („ob eine Textsorte ökonomische Formen enthält“, ebd., z. B. grafische Kürzungen, Ellipsen oder Kurzsätze), [redu] (markiert, ob eine Textsorte „sprachliche Redundanz“ aufweist, ebd.), [nspr] (ausschließlicher Einsatz sprachlicher Mittel) und [part] (Gleichberechtigung der Kommunikationspartner).

[57] Bei Sandig (1972) befinden sich in der Tabelle sechs Fußnoten, die in Tabelle 2.2 nicht berücksichtigt wurden.

der am besten ausgearbeiteten Textsortenklassifikationsvorschläge“ (ebd., S. 135) bezeichnen, obwohl sie doch grundsätzliche Probleme sehen: „So haben sich die [merkmalbasierten] Analysen offenbar nicht ernsthaft die Frage vorgelegt, wie die einzelnen Merkmale zu gewinnen sind, welchen Status sie besitzen und welche linguistischen Eigenschaften sie abbilden.“ (ebd.). Es werden weitere Kritikpunkte angeführt, die jedoch die Tatsache unbeachtet lassen, dass es sich nur um einen fragmentarischen Vorschlag handelt: „Mit Merkmalen können nur grobe Textcharakteristika angegeben werden, nicht die internen Strukturen von Textsorten.“ (Sandig, 1972, S. 122). Nach Heinemann (2000c, S. 527) nimmt der Ansatz „Grundannahmen sowohl von pragmatischen als auch von Mehrebenen-Modellen“ vorweg, da einige textexterne Merkmale berücksichtigt werden, die jedoch nur „willkürlich“ zusammengestellt und „quantitativ unterrepräsentiert“ seien, zudem „fehlen so wichtige Merkmale wie kommunikative Funktion bzw. Intention“ (ebd., S. 528).

Das Textmustermodell von Sandig (1997)

Das Textmustermodell von Sandig (1997) beruht nach eigenen Angaben auf Überlegungen aus der Text- und Textsortenlinguistik, Stilistik sowie der Konversationsanalyse und weist kaum noch Gemeinsamkeiten mit dem früheren Vorschlag auf (Sandig, 1972). War dieser noch explizit als „fragmentarisch“ gekennzeichnet, wird nun eine „ganzheitliche, »holistische« Beschreibung von Textmustern“ (Sandig, 1997, S. 26) angestrebt, womit „komplexe sinnhafte Ganzheiten“ (ebd.) gemeint sind, die aus sehr unterschiedlichen Eigenschaften bestehen:

> Ein Textmuster ist ein standardisiertes (konventionelles) Mittel zur Lösung in einer Gesellschaft auftretender Standardprobleme. Der gesellschaftlichen Relevanz entsprechend gibt es in der Gesellschaft (mindestens) eine Benennung dafür. [...] Ein Textmuster kann beschrieben werden als Zusammenhang von (nicht sprachlichem) Handlungstyp und (sprachlicher) Textsorte. (Sandig, 1997, S. 26)

Abbildung 2.6 zeigt den Zusammenhang zwischen Handlungstyp und Textsorte: „Der Handlungstyp steuert die konventionelle Erwartung bezüglich der Textsorte; die Textsorteneigenschaften »kontextualisieren« den Handlungstyp [...]; sie zeigen ihn an“ (ebd., S. 27 f.). Das Modell wird anhand linguistischer Artikel exemplifiziert.

Der Handlungstyp besteht nach Sandig (1997, S. 28) in einem „sozialen Sinn“, der „durch die standardmäßige Lösungsart des Standardproblems“ vermittelt wird (ebd.), es liegt also eine rekurrente Problemsituation vor. Der soziale Sinn wissenschaftlicher Arbeiten besteht z. B. darin, neue Erkenntnisse mitzuteilen; die Problemlösung erfolgt in einem Artikel oder einer Monografie. Ein Handlungstyp wird über einen oder mehrere spezifische Kanäle realisiert, der oder die die Textsorte beeinflussen. Der Kanal wird durch die Wahl eines bestimmten Mediums spezifiziert (Buch, Zeitschrift, Overheadfolien, Poster- und Computerpräsentation). Ein weiterer Aspekt betrifft die Situationsbeteiligten, in der Beispieldomäne handelt es sich um Wissenschaftler. Texte werden als „Mittel des Handelns mit spezifischem sozialem Sinn“ (ebd., S. 29) verstanden und sind nach prototypischen Textsorten gestaltet, sie besitzen Merkmale, die ihren sozialen Sinn markieren (vgl. die Abschnitte 2.2.9 und 2.3.6 sowie Sandig, 2000). Textsorten weisen Eigenschaften auf, die in Abbildung 2.6 dargestellt sind.

Textmuster(wissen) Benennung(en) in der Sprache	
Handlungstyp	Handlungsmittel: Textsorte prototypische Textsorteneigenschaften
Gesellschaftlicher Zweck • sozialer Sinn • Art der Problemlösung	Handlungshierarchie • konstitutive und fakultative Teilhandlungen • generelle Textherstellungshandlungen, die genutzt werden • eingelagerte Themenstruktur
Situationseigenschaften • Problemsituation • Institution/Handlungsbereich • Kanal • Medium	Sequenzmuster • textmusterspezifisch • allgemeine Sequenzmuster, die nutzbar sind
Situationsbeteiligte (Rollen) • Sprecherschreiber • Adressaten • Beziehungsart	Formulierungsmuster • nach Heinemann/Viehweger • allgemeine Textherstellungsmuster • global (Fachsprachen, Stilebenen, Sprachökonomie . . .) • auf (Teil-)Themen bezogen: Frames • auf Teilhandlungstypen (auch: Sequenzpositionen) bezogen • allgemeine Darstellungsmuster: Dialogisieren, Kontrastieren, Erzählen . . .
	Materielle Textgestalt Graphische (+ bildliche)/prosodische . . . Gestalt
	Durchschnittsumfang (Länge, Dauer)

Abbildung 2.6: Das Textmustermodell von Sandig (1997, S. 27)

Die Handlungshierarchie wird nach Sandig durch den sozialen Sinn determiniert. Der Hierarchiebegriff deutet an, dass eine wichtigste Handlung sowie untergeordnete Handlungen existieren können, die obligatorischen oder optionalen Charakter besitzen können. Als wichtige Handlungstypen beim wissenschaftlichen Aufsatz werden das „Präsentieren“ von Neuem und das „Aufbauen“ auf Bekanntem angegeben, was üblicherweise durch „Argumentieren“ geschieht, d. h. das Thema eines Textes wird in die Handlungsstruktur „eingelagert“ (ebd.). Untergeordnete Handlungstypen sind „Verweisen“ auf und „Zitieren“ aus der Literatur und das „Einbetten“ von Zitiertem in den Kontext. Die Handlungshierarchie bezieht sich ebenfalls auf generelle Textherstellungshandlungen wie Reformulieren, das Markieren von Absätzen oder „metakommunikativ die Aktivität [zu] verdeutlichen“ (ebd., S. 30).

Der Aspekt der Sequenzmuster betrifft allgemeinere Sequenzmuster und textmusterspezifische Eigenschaften, wobei für bestimmte Positionen Formulierungsvorgaben vorliegen können. Als allgemeines Sequenzmuster sieht Sandig die „Anfang-Mitte-Ende-Struktur“ an, die sich in wissenschaftlichen Texten als „Einleitung, Hauptteil, Schluss“ manifestiert (ebd., S. 31). Auch die Auszeichnung des Textanfangs durch eine Überschrift und die Nennung des Autorennamens gehören zu dieser Kategorie. Als „[t]extmusterspezifisch für [. . .] Wissenschaftstexte“ bezeichnet Sandig die finale Position des Literaturverzeichnisses und unterschiedliche Möglichkeiten, Anmerkungen zu integrieren (z. B. Fuß- oder Endnoten).

Konventionelle Vorgaben eines Handlungsbereiches, der Kanal und unter Umständen auch das Medium determinieren konventionelle Formulierungsweisen einer Textsorte. Die

Beispieldomäne Wissenschaft als Handlungsbereich umfasst nach Sandig globale Formulierungsvorgaben wie z. B. „Hoch- und Schriftsprache“, „allgemeine fächerübergreifende Wissenschaftssprache bzw. Wissenschaftsstil“, „fachsprachliche Elemente wie Textmuster, Formulierungsmuster“ und „mehr oder weniger klar strukturierte Frames“ für das Thema (ebd., S. 31 f.). Fachsprachliche Elemente hängen mit ihren Formulierungen von Wissenschaftsparadigmen und Schulen ab. Ein weiteres Charakteristikum sieht Sandig in textsortenspezifischen, in unterschiedlichen Graden festgelegten Formulierungsmustern, z. B. Einzelelemente (in linguistischen Artikeln z. B. „Äußerung“, „Formulierung“), Kollokationen („ein Textmuster realisieren“, „sprachlicher Ausdruck“), Phraseologismen („generative Grammatik“) und „feste syntagmatische Verbindungen, die z. T. auch für Handlungsbereiche oder für mehrere Textmuster charakteristisch sind“ (ebd.), wobei die Verfasserin auf das Zitieren eingeht, das häufig unter Nennung des Autorennamens, der Jahreszahl und der Seitenzahl (in dieser Reihenfolge) geschieht. Ein weiteres prototypisches Formulierungsmuster sind Gliederungssignale, die „wichtige Scharniere darstellen“ (Sandig, 1997, S. 34). Als globale Gliederungssignale werden Kapitelüberschriften angesehen, wohingegen „erstens“, „zweitens“, „zunächst“, „weiterhin“ und „schließlich“ als lokale Gliederungssignale fungieren. Zu dieser Gruppe gehören im Bereich des Wissenschaftsstils deiktische Ausdrücke, die sich unmittelbar auf den Text selbst (z. B. „in dieser Arbeit wird gezeigt, dass …“), auf andere Positionen zu einem spezifischen Thema („während X der Meinung ist, dass …“), auf Listenstrukturen (katadeiktisch: „folgendes“, anadeiktisch: „auf diese Weise“, „so“) oder bestimmte Positionen des Textes („hier“, „jetzt“, „bereits“, „oben“ etc.) beziehen.

Der Textsorte zugehörig ist auch die materielle Textgestalt, die vom Handlungsbereich, dem Kanal und dem Medium beeinflusst wird. Für das Beispiel der linguistischen Aufsätze nennt Sandig (1997, S. 36 f.) den Einsatz von Kursivschrift für objektsprachliche Ausdrücke, typografisch abgesetzte längere Zitate, kurze Literaturverweise im Text, längere Bezugnahmen und relevanzrückgestufte Kommentare als Fußnoten, die Gliederung in Absätze, die optionale Voranstellung eines Abstracts und die optionale, optisch abgesetzte Integrierung eines Mottos als „stimulierende Einleitung in das Thema“ (ebd., S. 36).[58] In Bezug auf „das Textbild als Ganzes“ können auch unterschiedliche Konventionen herrschen, so wird dieses bei einigen Zeitschriften oder Handbüchern in zwei oder mehr Spalten gesetzt.

Der Durchschnittsumfang bezieht sich auf die Problemsituation und die Art der Problemlösung. Er „kann individuell oder durch Zwänge im Handlungsbereich verändert werden (Zeitdauer der Präsentation, Kosten der Publikation)“ (ebd., S. 37).

Der Ansatz von Brinker (2001)

Brinker (2001, S. 136 ff.) schlägt verschiedene Kriterien zur Differenzierung von Textsorten vor, die an dem Basiskriterium der Textfunktion ansetzen, so dass zunächst die „noch recht umfangreich[en]“ (ebd., S. 137) Textklassen Informations-, Appell-, Obligations-, Kontakt- und Deklarationstexte unterschieden werden (vgl. Abschnitt 2.2.3). Der Subdifferenzierung dienen (i) kontextuelle, d. h. situative Kriterien und (ii) strukturelle bzw. thematische Merkmale. Die erste Gruppe bezieht sich vornehmlich auf die Kommunikationssituation, wo-

[58] Fast alle der von Sandig aufgeführten Spezifika gelten nicht nur für linguistische Aufsätze, sondern für wissenschaftliche Artikel beliebiger Disziplinen (vgl. z. B. Göpferich, 1995).

bei zwischen sechs Kommunikationsformen (Gespräch, Telefongespräch, Rundfunksendung, Fernsehsendung, Brief und Zeitungsartikel/Buch) und den Rollenverhältnissen der Partner im privaten, offiziellem und öffentlichen Handlungsbereich unterschieden wird. Kommunikationsformen sind nicht mit Textsorten zu verwechseln, da sie „allein durch situative bzw. mediale Merkmale definiert, in kommunikativ-funktionaler Hinsicht also nicht festgelegt sind." (ebd., S. 139). Die Rollenverhältnisse stellen ebenfalls ein differenzierendes Merkmal dar: Im privaten Bereich kommunizieren Privatpersonen (*Ansichtskarte*, *Privatbrief*), wohingegen im offiziellen Bereich Träger beruflicher Rollen bzw. Funktionen kommunizieren, wodurch die Verbindlichkeit steigt (*Weisung*, *Anordnung*, *Gesetz*); den öffentlichen Bereich bezieht Brinker auf Presse, Funk und Fernsehen. Die Gruppe der strukturellen Kriterien subsumiert das Textthema und die Themenentfaltung. Für das Thema sind Brinker (2001, S. 142) zufolge „gewisse thematische Restriktionen anzugeben" wie z. B. „Thema = Emittent" für die *Werbeanzeige* oder „Thema = Rezipient" für die *Stellenanzeige*. Bei der Themenentfaltung wird zwischen deskriptiven, narrativen, explikativen und argumentativen Verfahren unterschieden. Diesbezüglich schlägt Brinker weitere Differenzierungen vor, z. B. „sachbetonte" und „meinungsbetonte" Realisierung der deskriptiven sowie „persuasiv-überredende" und „rational-überzeugende" Realisierung der argumentativen Themenentfaltung (ebd., S. 143).

Mehrebenenklassifikationen nach Heinemann und Heinemann (2002)

Heinemann und Heinemann (2002, S. 144) gehen von vier Basisebenen aus, auf die Merkmale bezogen werden können: (1) Funktionalität, (2) Situativität, (3) Thematizität und Strukturiertheit sowie (4) Formulierungsadäquatheit. Weitere Aspekte können als zusätzliche Bezugsdimensionen eingeführt werden. Der Ansatz basiert auf drei Schritten: Zunächst werden die für eine Textsorte relevanten Merkmalsausprägungen auf den Basisebenen eruiert. Daraufhin erfolgt die gewichtete „Integration der Einzelmerkmale zur charakteristischen Ganzheit der jeweiligen Textsorte" (ebd., S. 145).[59] In einem letzten Schritt kann die Merkmalsbelegung als natürlichsprachliche Beschreibung der Textsorte paraphrasiert werden.

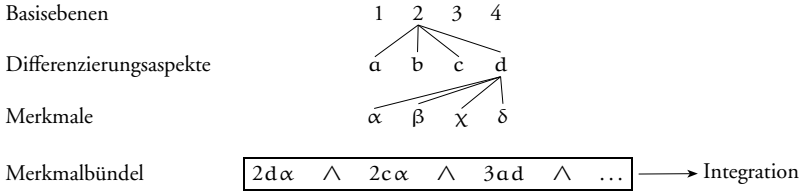


Abbildung 2.7: Beschreibungsaspekte für die Kennzeichnung von Textsorten (nach Heinemann und Heinemann, 2002, S. 146)

Abbildung 2.7 zeigt die Beschreibungsaspekte schematisch, wobei a, b, c und d Differenzierungsaspekte und α, β, χ und δ Merkmalsausprägungen symbolisieren. Jede Ebene wird in Subkategorien aufgeteilt, es ist jedoch schwierig, „aus der Fülle von Beschreibungsrastern

[59] Die Integration betrifft die Ermittlung der distinktiven Merkmale: „[W]ir könnten im Alltag nicht von ›Textsorten‹ sprechen, wenn es nicht [...] übereinstimmende Konstanten in Textexemplaren gäbe." (ebd., S. 146).

in der Spezialliteratur die jeweils relevanten Aspekte herauszufiltern, da dabei insbesondere der Grad der Theoriebezogenheit der jeweiligen Darstellung zum Tragen kommt.“ (ebd., S. 147). Für die Basisebenen werden Subkategorien angenommen, „die für den größten Teil der Textsorten von Relevanz sind“ (ebd.):[60]

1. Funktionalität – Hauptfunktionen: (a) Sich Ausdrücken; (b) Kontaktieren; (c) Informieren; (d) Steuern (α: Aufforderung zum praktischen Handeln; β: Aufforderung zu einer Sprachhandlung/Anworthandlung; χ: Aufforderung zu Bewertungen und Handlungen; δ: Aufforderung zur kognitiven Verarbeitung); (e) Ästhetisch Wirken
2. Situationalität – Situationsklassen: (a) Tätigkeitssituationen; (b) Soziale Organisation der Tätigkeiten in Kommunikationsbereichen; (c) Kanal/Medium; (d) Anzahl der Partner (α: dyadische Kommunikation, z. B. *Alltagsgespräch*, *Privatbrief* ; β: Gruppenkommunikation, z. B. *Gruppengespräch*, *Rundschreiben*; χ: Massenkommunikation, z. B. *Leserkreis einer Tageszeitung*); (e) Soziale Rollen der Partner; (f) Umgebungssituation
3. Thematizität und Strukturiertheit – (a) Thematische Geprägtheit; (b) Text-Thema-Entfaltungen (einschließlich Vertextungsmuster); (c) Textstrukturierung
4. Formulierungsadäquatheit – (a) Kommunikationsmaximen; (b) Textsortenspezifische Formulierungsmuster (α: syntaktische Spezifika, z. B. dominierender Satztyp, Verhältnis Haupt- zu Nebensätze, Komplexitätsgrad, Verhältnis Satzlänge zu Kürze der Sätze; β: lexikalische Spezifika, z. B. Indikatoren der Textfunktion, textsortenspezifischer Wortschatz, Fremd-/Fach-/Kurzwörter etc.); (c) Stilistische Besonderheiten

Es wird betont, dass die „entscheidenden Schritte für die Textsortenkennzeichnung [...] vom Textinterpreten vorgenommen werden [müssen]: die Aussonderung aller nichttypischen Merkmale, die Selektion und Wichtung aller für die jeweilige Textsorte relevanten und konstitutiven Parameter [...] und schließlich die Bündelung und Integration der konstitutiven Merkmale zu der je spezifischen Merkmalkomplexion der jeweiligen Textsorte“ (ebd., S. 149). Heinemann (2000e, S. 18 f.) merkt an, dass sich eine derartig detaillierte Herangehensweise nur auf *standardisierte* Textsorten beziehen kann, „also Textmengen, die in der Alltags- und in der institutionellen Kommunikation usuell sind und immer wieder in – annäherend – gleicher Weise fast stereotyp konstituiert werden.“ Heinemann und Heinemann (2002) erläutern ihre Mehrebenenklassifikation anhand der Textsorten *Arztrezept* und *Telegramm*.

2.3.6 Textsorten und prototypische Textexemplare

Das Konzept Text besitzt nach Sandig (2000) einen prototypischen Kern, um den sich typische und weniger typische Textualitätsmerkmale gruppieren (vgl. Abbildung 2.1, S. 33). Zusätzlich diskutiert Sandig das Verhältnis von Textexemplaren zu Textsorten (von ihr als Textmuster bezeichnet) und geht „auf ein komplexes Merkmal [ein], das die genannten Merkmale als informationsreiche Bündel kookurrierender Merkmale in je unterschiedlicher Gewichtung zusammenfaßt: die Musterbezogenheit von Texten.“ (ebd., S. 101).

Dass Textsorten variabel sind, wurde wiederholt angesprochen. Sandig verdeutlicht zentrale Eigenschaften sehr vieler Charakterisierungen einzelner Textsorten: „Mit Textmuster*beschreibungen* werden jeweils *prototypische Exemplare* beschrieben: die besten Exemplare, die

[60] Der angesprochene „größte Teil der Textsorten“ wird dabei nicht genauer charakterisiert.

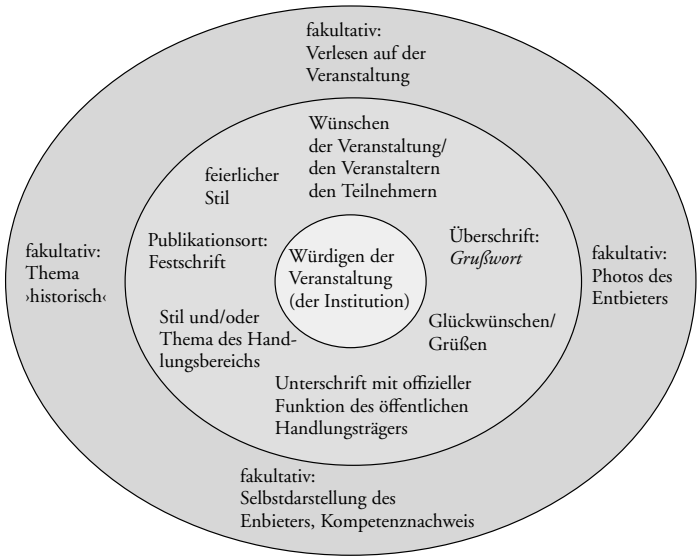


Abbildung 2.8: Der Prototyp der Textsorte *Grußwort* (nach Sandig, 2000, S. 104)

klar als solche erkennbar sind." (ebd., S. 103; Hervorhebungen hinzugefügt, G. R.).[61] Analysierende Darstellungen beziehen sich somit meist auf prototypische Vertreter einer Textsorte (vgl. Sandig, 1997, und de Beaugrande und Dressler, 1981, S. 188 f.). Prototypische Merkmale sind aber nicht als obligatorisch anzusehen: „Es geht nicht um Klassen von Texten [...] mit gleicher Struktur, sondern die Realisierungen sind variabel je nach individueller Situation und individuell damit verfolgtem Zweck [...]: im Textmuster sind bereits Grade der Variabilität vorgesehen." (Sandig, 2000, S. 103). Sandig exemplifiziert dies anhand eines typischen und eines weniger typischen Vertreters des Textmusters *Grußwort* (vgl. Abbildung 2.8).

2.3.7 Die North American Genre Theory

In der deutschsprachigen Literatur, die sich mit dem Thema Hypertext aus textlinguistischer Perspektive beschäftigt, wird das Konzept der Hypertextsorte beinahe vollständig ignoriert. Kapitel 4 behandelt vornehmlich aus den USA stammende Arbeiten, die unterschiedliche Facetten von „digital documents" und „digital genres" untersuchen. Diese Beiträge berufen sich fast ausnahmslos auf die soziolinguistisch ausgerichtete *North American Genre Theory*, wobei

[61] Heinemann und Viehweger (1991, S. 170) bezeichnen Textsorten als Prototypen: „Textsorten stellen sich daher in einer Typologie als idealtypische/prototypische Phänomene dar, als Verallgemeinerungen, die auf Durchschnittserfahrungen [...] basieren". Bei der Diskussion der Textsorte Telegramm wechseln sie die Ebene: „Textexemplare, die alle hier genannten Merkmale aufweisen, dürfen als Prototypen der Textsorte Telegramm angesehen werden; sie repräsentieren [...] das »Idealmuster« dieser Textsorte." (ebd., S. 172). Weiterhin wird angemerkt, „daß es für bestimmte Textsorten offenkundig gar keine prototypische Strukturierung gibt" (ebd.), z. B. Werbetexte, „die nicht durch eine spezifische Textstruktur, sondern durch die implizite Aufforderung zum Kauf von Waren determiniert sind" (ebd.). Werbetexte sind jedoch aufgrund des sehr großen Geltungsbereichs dieses Etiketts als Texttyp und nicht als Textsorte zu charakterisieren, wobei durchaus Textsorten mit spezifischen Strukturierungen existieren, z. B. *gewerbliche Kleinanzeigen* und *Flyer* (Androutsopoulos, 2000).

insbesondere auf Miller (1984), Bazerman (1988) und Swales (1990) sowie Yates und Orlikowski (1992) und Orlikowski und Yates (1994) Bezug genommen wird. Die *North American Genre Theory* (auch: „situated genre theory", Erickson, 2000) ist – obwohl sie die tatsächlich linguistischen Charakteristika vernachlässigt – in ihren Grundprinzipien mit den neueren pragmatischen Texttheorien vereinbar, die im Rahmen der Textlinguistik im deutschsprachigen Raum vorgelegt wurden. Nachfolgend werden zunächst die Grundlagen thematisiert (vgl. auch Breure, 2001, und Boudourides und Peticca, 2001), woraufhin der Einsatz einer „genre lens" (Yates und Orlikowski, 1992) zur Analyse von Prozessen der Organisationskommunikation erläutert wird. Nach der Darstellung eines Beispiels wird abschließend eine Genre-Taxonomie zur Analyse von Kommunikationsprozessen diskutiert.

Eine adäquate Übersetzung des Begriffs „Genre" kann nur durch die generische Bezeichnung „Textklasse" erfolgen.[62] Eine Unterscheidung von Texttypen und Textsorten (vgl. Abschnitt 2.3.2) existiert in der *North American Genre Theory* nicht, der Terminus Genre bezieht sich gleichermaßen auf abstrakte wie spezifische Klassen von Texten. Miller (1984, S. 36) identifiziert diesen Umstand als „one of the most important problems raised by recent genre theory" und schlägt für den Geltungsbereich des Konzepts eine Konzentration auf konventionalisierte Diskursstrukturen vor, die sich in einer Gesellschaft als Mittel des miteinander Agierens etabliert haben.[63] Genres werden als Konventionen bezüglich Inhalt, Form und kommunikativer Funktion einer sprachlichen Handlung definiert. Miller präzisiert ihre Auffassung des Genrebegriffs als sprachliche Handlungen in situationsgebundenen und sozialen Kontexten, die der Realisierung bestimmter Ziele dienen und als Reaktion auf eine spezifische Situation zu interpretieren sind. Die seit der Antike dem Begriff inhärente Konzentration auf taxonomische Anordnungen von Genres solle in den Hintergrund treten, da Genres einer fortwährenden Evolution ausgesetzt seien; durch naive Klassifikationen könnten Miller (1984) zufolge „reductionism" und „formalism" entstehen. Die Benutzung von Genres unterliegt abstrakten Regeln, die rekurrente und einander ähnelnde Situationen auf spezifische Merkmale der Dimensionen Inhalt und Form abbilden. Nach Bazerman (1988) sind Genres auf Seiten der Produzenten und Rezipienten mit individuellen Erwartungshaltungen verbunden: Durch die Erkennung formaler Hinweise wissen Leser, was von einem Textexemplar zu erwarten ist und auf welche Kommunikationssituation ein Text reagiert. Genres können dabei derart ausgeprägte Formen der Etabliertheit annehmen, dass Rezipienten überrascht oder unkooperativ sein könnten, falls eine bestimmte Situation nicht von der qua Konvention zu erwartenden sprachlichen Handlung begleitet wird. Bazerman (1994, S. 97) beschäftigt sich mit Patentanträgen und führt den Begriff der „systems of genre" ein, der als Erweite-

[62] Aus diesem Grund wird im Folgenden ausschließlich die Bezeichnung „Genre" verwendet und auf die zu abstrakt erscheinende Übersetzung verzichtet.

[63] Siehe auch Bazerman (1994, S. 81): „A textual form which is not recognized as being of a type, having a particular force, would have no status nor social value as a genre. A genre exists only in the recognitions and attributions of the users." Während Miller die arbiträre Extension als Problem auffasst, sehen Yates und Orlikowski (1992, S. 303) hierin einen immensen Vorteil: „[T]his flexible approach seems more useful in dealing with the vast range of communication in organizations; that is, the business letter and the recommendation letter, the meeting and the personnel committee meeting may all be designated as genres of organizational communication if there can be identified for a recurrent situation, a common subject [...], and common formal features." Crowston und Williams (2000, S. 202) schließen sich an: „Rather than argue about the proper level of analysis for a genre, we believe it is most useful to [...] consider genres at any of these different levels."

rung von „genre set“ anzusehen ist: Ein „genre set“ repräsentiert den vollständigen Bestand von Texten, die vom Träger einer bestimmten beruflichen Rolle erstellt werden. Dieser Terminus bezieht sich also nur auf *eine* Seite komplexer sprachlicher Interaktionen – „systems of genre“ hingegen bezeichnet die vollständigen Bestände von Genres, die von *allen* an solchen Interaktionen beteiligten Partnern sukzessive instanziiert werden (z. B. die Genres, die schließlich zur Ausstellung eines Patentes führen).[64] Swales (1990) untersucht verschiedene Genres aus Wissenschaft und Forschung zur Konzipierung didaktischer Grundlagen für das wissenschaftliche Schreiben. Nach Swales (1990, S. 58) fungieren Genres als institutionalisierte und instrumentalisierte Vermittler zwischen Individuen und Diskursgemeinschaften. Genres sind also funktional markierte Kommunikationsmuster und dienen der Realisierung spezifischer Ziele innerhalb einer Diskursgemeinschaft. Das mit ihrer Instanziierung vefolgte Ziel legt den Aufbau eines Diskurses fest und beeinflusst seinen Inhalt. Exemplare besitzen mehrere Gemeinsamkeiten hinsichtlich Struktur, Stil, Inhalt und Zielgruppe. Wenn von einem Textexemplar alle Erwartungen eines bestimmten Genres realisiert werden, kann es von der Diskursgemeinschaft als prototypisch angesehen werden.

Genres in der Organisationskommunikation

Yates und Orlikowski (1992) wenden die Genre-Theorie zur Untersuchung von Kommunikationspraktiken in Organisationen an. Genres werden aufgefasst als „typified communicative actions characterized by similar substance and form and taken in response to recurrent situations.“ (ebd., S. 299).[65] Der Begriff der „substance“ bezieht sich auf soziale Motive sowie Themen und Inhalte, die durch eine Sprachhandlung ausgedrückt werden sollen, während „form“ die physikalische Manifestation der Handlung und ihre sprachlichen Eigenschaften bezeichnet. Yates und Orlikowski (1992, S. 301 f.) zufolge existieren mindestens drei Formaspekte: Strukturelle Eigenschaften (z. B. Abschnitte, Listen, typografische Auszeichnungen), ein Kommunikationsmedium und die Art der sprachlichen Realisierung, die z. B. umgangs- oder fachsprachlich erfolgen kann. Die Verfasserinnen gehen (wie Miller, 1984) davon aus, dass Genre-Regeln bestimmte Merkmale von „form“ und „substance“ mit rekurrenten Situationen assoziieren. Derartige Regeln können stillschweigend existieren oder einer standardisierenden Kodifizierung unterliegen. Genres können sich verändern, indem einzelne Aspekte

[64] Bazerman (1994, S. 97) definiert „systems of genre“ als „interrelated genres that interact with each other in specific settings.“ Und: „The system of genres would be the full set of genres that instantiate the participation of all the parties [...]. This would be the full interaction, the full event, the set of social relations as it has been enacted.“ (ebd., S. 99). Die Reihenfolge der Instanziierung ist also signifikant (Yoshioka et al., 2001, S. 437), was durch das Beispiel der Bewerbung um eine Anstellung deutlich wird, denn „the job ad, job letter and resume, and rejection letter (or invitation to interview, interview, and job offer) form a genre system. Such systems are composed of a well-coordinated set of communicative moves that together accomplish an interaction [...].“ (Yates et al., 1997, S. 51). Bergquist und Ljungberg (1999) stellen ein Genre-System zur Dokumentation der Projekte eines schwedischen IT-Unternehmens vor. Saunders und Chiasson (2005) beschäftigen sich mit dem „case file“, in dem Anwälte sämtliche Geschäftsvorgänge eines Falls pflegen.

[65] In einer späteren Arbeit betonen Orlikowski und Yates (1994, S. 543) den Einfluss der Diskursgemeinschaft und definieren Genre als „a distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form [...]. The communicative purpose of a genre is not rooted in a single individual's motive for communicating, but in a purpose that is constructed, recognized, and reinforced within a community“. Eine zusätzliche Definition fasst Genres als kommunikative Handlungen auf, „that are habitually enacted by members of a community to realize particular social purposes“ (ebd., S. 542).

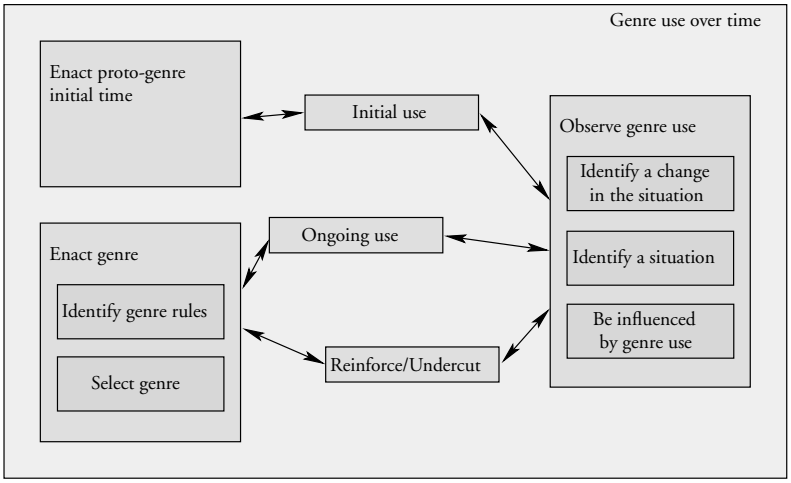


Abbildung 2.9: Genre-Verwendung und -Evolution (nach Yoshioka et al., 2001, S. 446)

der Genre-Regeln ignoriert oder alternativ realisiert werden (vgl. Abbildung 2.9).[66] Yates und Orlikowski (1992, S. 311–318) exemplifizieren diesen Vorgang anhand der Entwicklung des Genres *Memorandum*, das auf den Geschäftsbrief des 19. Jahrhunderts zurückgeht.[67]

Orlikowski und Yates (1994, S. 546) führen den Begriff Genre-Repertoire ein, worunter „the set of genres routinely enacted by a particular community" verstanden wird. Durch das Anfertigen von Exemplaren solcher Genres wird die Mitgliedschaft in einer Diskursgemeinschaft signalisiert und neue Mitglieder von Diskursgemeinschaften verwenden diejenigen Genres, die sie schon in anderen Gemeinschaften benutzt haben (ebd., S. 547).[68]

[66] Abbildung 2.9 stellt diese Prozesse in sehr abstrakter Form dar (vgl. Yoshioka et al., 2001, S. 439, S. 446 f.): Zu einem gewissen Zeitpunkt, der meist nur retrospektiv und näherungsweise bestimmt werden kann, findet in einer spezifischen Kommunikationssituation die erstmalige Verwendung eines Protogenres statt, das sich in einem zyklischen Prozess (Beobachtung und Benutzung eines Genres) innerhalb einer Diskursgemeinschaft zu einem etablierten Genre entwickelt. Die Mitglieder reproduzieren ein Genre in der Form, die ihnen aus vergangenen sprachlichen Handlungen bekannt ist. Falls jedoch eine abweichende Kommunikationssituation vorliegt, werden – bewusst oder unbewusst – Modifikationen durchgeführt. Die Rezipienten identifizieren das Genre bzw. die Variante in Bezug auf ihre eigenen Erfahrungen in vergleichbaren Kommunikationssituationen.

[67] Orlikowski und Yates (1994, S. 545) präzisieren den Prozess: „While members typically reinforce established genres through their communicative actions, they can and, on occasion, do challenge and modify these genres, both deliberately and inadvertently. When changes [...] are repeatedly enacted and become widely adopted within the community, genre variants or even new genres may emerge, either alongside existing genres or to replace those that have lost currency." Ihre Benutzung stellt einen zyklischen Prozess dar, der aus „enacting a genre" und „observing genre use" besteht (Yoshioka et al., 2001, S. 439).

[68] Orlikowski und Yates (1994, S. 550–570) untersuchen ca. 2 000 Nachrichten einer Mailingliste, die zwischen 1981 und 1983 von den Mitgliedern eines Projekts zur Entwicklung der Programmiersprache Common LISP verwendet wurde. Es werden die Genres „Memo", „Dialogue", „Proposal" und das „Ballot genre system" („ballot questionaire", „ballot response", „ballot results") identifiziert. Bezüglich der Untersuchung digitaler Kommunikationsmedien kommen die Autorinnen zu folgendem Schluss: „Understanding organizing processes mediated by new technologies becomes increasingly important as more and more organizational work becomes a matter of electronic symbol manipulation and information exchange. The genres through which information is shaped and shared for particular purposes (reports, spreadsheets, meetings, or teleconferences) are no longer merely an aspect of organizational work; rather, they *are* the organizational work." (ebd., S. 572).

Ein Beispiel – Das Genre „Lebenslauf"

Erickson (2000) stellt eine Beispielanalyse des Genres „résumé" dar.[69] Der in Deutschland geläufige tabellarische Lebenslauf (vgl. Heiber, 2001, S. 4 f.) unterscheidet sich nur in Details von seinem US-amerikanischen Pendant, so dass die Darstellung für beide Textsorten Gültigkeit besitzt (vgl. auch Abschnitt 4.6.3).

Der kommunikative Zweck des Genres *résumé* besteht in der Vermittlung von Informationen, die für eine berufliche Anstellung relevant sind; das Ziel des Produzenten ist die Realisierung der angestrebten Anstellung. In Bezug auf Form und Inhalt existieren verschiedene Konventionen: Textexemplare sind kurz, strukturiert und enthalten Informationen über ausgeübte Berufe und Kontaktinformationen. Diese Konventionen sind Reaktionen auf die rekurrente Verwendungssituation. Der Inhalt wurde geformt durch diejenigen Informationen, die für eine spezifische Branche als adäquat betrachtet werden. Gleichzeitig existieren auf Seiten des Produzenten Annahmen darüber, auf welchem Wege der Rezipient antworten wird, wodurch bestimmt wird, welche Kontaktinformationen (E-Mail- und/oder Postadresse, Telefonnummer etc.) einzufügen sind. Die Strukturierung erlaubt es dem Rezipienten (Personalchef, Abteilungsleiter etc.), das Textexemplar schnell zu überfliegen und als Referenz in einem Bewerbungsgespräch zu verwenden. Die Form wird auch von technischen Aspekten beeinflusst, so haben moderne Textverarbeitungssysteme den Einsatz typografischer Auszeichnungen wie Fett- und Kursivdruck erhöht und die bei der Benutzung von Schreibmaschinen übliche Verwendung von Unterstreichungen oder Großbuchstaben reduziert. Die Konventionen des Genres sind als Reaktionen auf technische, soziale und institutionelle Faktoren zu verstehen. Seine Diskursgemeinschaft besteht aus Personen, die Lebensläufe produzieren und rezipieren, sowie aus denjenigen Branchen, die Personen bei der Berufssuche und Bewerbungsprozessen unterstützen. Basierend auf diesem Beispiel stellt Erickson (2000) seine Definition von Genre dar, die als Synthese vorhandener Definitionen aufgefasst wird: „A genre is a patterning of communication created by a combination of the individual, social and technical forces implicit in a recurring communicative situation. A genre structures communication by creating shared expectations about the form and content of the interaction, thus easing the burden of production and interpretation."

Die Genre-Taxonomie von Yoshioka, Herman, Yates und Orlikowski (2001)

Yoshioka, Herman, Yates und Orlikowski (2001) stellen eine „genre taxonomy" als „knowledge repository" vor, das Manager, Unternehmensberater und Software-Entwickler bei der Konzeptionierung und Optimierung von Kommunikationsprozessen unterstützen soll. Die Taxonomie ist in der Lage, sowohl einzelne Genres als auch Genre-Systeme innerhalb ihrer Verwendungskontexte zu repräsentieren, wobei zur Operationalisierung und Abstraktion derart komplexer Konzepte das von Orlikowski und Yates (1998) vorgeschlagene 5W1H-Modell

[69] Ihlström und Lundberg (2003, S. 2) geben mit der Titelseite einer Tageszeitung ein weiteres Beispiel, das das Zusammenspiel der zentralen Begriffe „content", „form" und „purpose" (bzw. „function") erläutert: „[O]ne purpose of the first page of a printed newspaper is normally to inform the reader about what news item is considered as having the highest news value. These news items (content) are placed in the top position. Thus the layout of the page (form) affects the interpretation of the content, since the same article switching positions with other articles (content) on the first page, would affect the news value (purpose) of both articles."

(„why, what, who, when, where, how“) verwendet wird, das eine Charakterisierung von Genres hinsichtlich der Dimensionen „purpose“, „content“, „participants“, „timing“, „location“ und „form“ erlaubt. Die Bestimmung dieser Charakteristika für spezifische Genres erfolgt nach Yates und Orlikowski (1992) und Orlikowski und Yates (1994).[70] Die Ergebnisse dieser beiden Studien sind als Teil des Bestandes über „specific genres“ in die Taxonomie eingeflossen (Yoshioka et al., 2001, S. 443). Der Prototyp der Taxonomie enthält zusätzlich ein „open set of widely recognized genres“ wie z. B. „business letter“, „memo“, „expense form“ und „report“.[71] Nachfolgend werden die sechs Dimensionen genauer thematisiert.

Der Zweck („why“) eines Genres oder Genre-Systems wird in verschiedenen Kategorien notiert, die auf den illokutionären Akten der Sprechakttheorie, den Vorarbeiten von Orlikowski und Yates (1994), einem Thesaurus und WordNet (Fellbaum, 1998) basieren: „inform“, „request“, „express (emotion)“, „decide“, „propose“, „respond“, „record“ und „other“ (Yoshioka et al., 2001, S. 435). Da mit Genres unterschiedliche Zwecke verfolgt werden können, erlaubt die Taxonomie eine Unterscheidung von „primary and secondary purposes“ (ebd.). Innerhalb der Taxonomie werden die Kategorien in hierarchischer Form repräsentiert. Der Inhalt („what“) von Genres, die zugleich Teil von Genre-Systemen sein können, wird unterhalb der „relevant purpose categories“ (ebd., S. 437) in dem „description field of the activity for representing the content of a genre“ abgelegt (ebd., S. 444).[72] Der Inhalt eines Genre-Systems besteht aus einzelnen Genres, die ebenfalls in hierarchischer Form repräsentiert werden. Die Information, wer an einem Genre beteiligt ist („who“), bezieht sich auf die jeweilige Diskursgemeinschaft. Darüber hinaus können Genres auch mit Informationen über Rollen wie „sender“ und „receiver“ versehen werden. Zeitliche Beschränkungen („when“) liegen z. B. bei den Genres „thank you note“ oder „daily morning meeting“ vor.[73] Diese Angabe wird ebenfalls im „description field“ hinterlegt. Der Ort („where“) eines Genres wird als Zeichenkette wie „Japan“ oder „northeastern United States“ innerhalb des Feldes „location“ hinterlegt (ebd., S. 446). Besonders deutlich wird die Notwendigkeit der Verortung von Genres, wenn Internet-basierte Kommunikationsformen modelliert werden sollen, denen die Loslösung von physikalischen Orten inhärent ist. Die Form eines Genres („how“) bezieht sich auf die von Yates und Orlikowski (1992) eingeführten Merkmale, die in der Taxonomie „along with purposes for identifying a genre“ abgelegt werden (Yoshioka et al., 2001, S. 438). Als Beispiel wird das Genre „Electronic Traditional Memo in Japan“ angeführt, das eine „Kanji signature“ und „no embedded message“ als strukturelle Merkmale besitzt. Als

[70] Es wird explizit darauf hingewiesen, dass diese sechs Dimensionen lediglich einen Ausschnitt des Wissens über ein Genre modellieren können: „We do not intend these six dimensions to be exhaustive or definitive, but rather offer them as a grounded starting point for classifying characteristics of genres and genre systems based on empirical evidence in organizations.“ (Yoshioka et al., 2001, S. 434).

[71] Die Autoren gehen nicht explizit darauf ein, mit welchen Technologien die Taxonomie gepflegt wird oder mit welchen Verfahren Benutzer auf sie zugreifen können. Der Prototyp basiert auf dem „Process Handbook“ (Yoshioka et al., 2001, S. 444), einem betriebswirtschaftlichen Projekt zur Modellierung von Geschäftsprozessen (vgl. `http://ccs.mit.edu`). Das Handbuch ist als kommerzielles Produkt erhältlich und wird von einer Datenbank und Zugriffswerkzeugen flankiert.

[72] Es wird nicht deutlich, was hiermit gemeint ist. Vermutlich handelt es sich um ein Datenbankfeld zur Eingabe freien Textes oder der zu erwartende Inhalt wird allein durch den Titel eines Genres repräsentiert.

[73] Eine „thank you note“ sollte innerhalb eines gewissen Zeitraums nach Erhalt eines Geschenks verschickt werden, falls „some appreciation for the gift or activity bestowed by another“ zu verzeichnen ist (Yoshioka et al., 2001, S. 437). Beim „daily morning meeting“ liegen Erwartungen vor, wann dieses Treffen stattfindet.

Medium wird „Usenet news group adjusted to Japanese environment“ angegeben und die linguistischen Eigenschaften umfassen eine „Kanji subject line“ und „no dialect“ (ebd.).

Abschließend demonstrieren Yoshioka et al. (2001, S. 448 ff.) den Einsatz der Genre-Taxonomie anhand einer Analyse der Arbeitsprozesse und Genres, die am Bewerbungsprozess für einen spezifischen Studiengang beteiligt sind. Dieser Bewerbungsprozess wurde im Jahr 2001 von ausschließlich papierbasierten Dokumenten auf digitale Medien umgestellt. Die Taxonomie ermöglicht eine Charakterisierung der unterschiedlichen Facetten der beteiligten Genres und ist in der Lage, für spezifische Prozesse, z. B. die Auswahl von Bewerbern, geeignete Genres vorzuschlagen: Der Zweck eines Auswahlprozesses ist „to decide. The process serialization hierarchy suggests two alternative ideas for the current review system: the ballot genre system and the bidding genre system“ (ebd., S. 452). Die Genre-Taxonomie kann behilflich sein, vorhandene Prozesse und Genres zu analysieren, innerhalb der Hierarchie zu verorten und Alternativen vorzuschlagen. Da die Taxonomie derzeit erst 15 „generally accepted genres and several kinds of specific genres used in particular organizations“ enthält (ebd., S. 454), weisen die Autoren darauf hin, dass die eigentlichen Vorteile des Prototypen deutlicher werden, sobald zusätzliche Genres in das System integriert worden sind.

Zusammenfassung und Fazit

Breure (2001) fasst die Kernmerkmale der *North American Genre Theory* zusammen: Ein Genre strukturiert Kommunikationsprozesse durch geteilte Erwartungshaltungen bezüglich Form und Inhalt, wodurch Produktions- und Interpretationsaufwand reduziert werden. Solche Kommunikationsmuster sind nicht auf einzelne Texte bzw. Genres beschränkt, sondern existieren auch auf abstrakteren Ebenen (Genre-Repertoires, Genre-Systeme). Genres sind sprachliche Handlungen, die untrennbar mit Kommunikationssituationen verbunden sind. Als institutionalisierte Antworten auf rekurrente Situationen reflektieren sie Normen, Ideologien und Gewohnheiten von Diskursgemeinschaften in Bezug auf Kommunikationssituationen. Ein weiterer Aspekt betrifft die Entwicklungsdynamik von Genres, die durch statische Kriterien nicht adäquat zu beschreiben sind. Zusätzlich zum Inhalt und zur Form hat sich mittlerweile der Aspekt der Funktion eines Genres als dritter Bestandteil etabliert.

Die Darstellung der primär soziolinguistisch ausgerichteten *North American Genre Theory* verdeutlicht verschiedene Unterschiede zu den Konzeptualisierungen von Textsorten und Texttypen, wie sie von der germanistischen Textlinguistik vertreten werden. Die wesentlichste Differenz betrifft die Abstraktionsebene, insbesondere die Loslösung von der sprachlichen Oberfläche, und die Fokussierung eher sozialer als tatsächlich linguistischer Aspekte. Innerhalb der *North American Genre Theory* hat sich – vermutlich als Folge des hohen Abstraktionsgrades – ein relativ klar umrissenes Inventar von Begriffen und Definitionen etabliert, während in der germanistischen Textsortenlinguistik immer wieder das Fehlen einer etablierten Terminologie beklagt wird (vgl. Abschnitt 2.1).[74] Für die maschinelle Identifizierung von Textsorten ist ein derart hoher Abstraktionsgrad durchaus mit Vorteilen bezüglich der Operationalisierbarkeit verbunden (vgl. Kapitel 14).

[74] Die *North American Genre Theory* wird von der deutschsprachigen Textlinguistik de facto nicht wahrgenommen. In Brinker et al. (2000) existiert z. B. diesbezüglich lediglich ein knapper Hinweis (auf Swales, 1990) innerhalb des Beitrags „Textlinguistik im englischsprachigen Raum“ (Thiele, 2000, S. 134 f.).

2.4 Zusammenfassung

Dieses Kapitel geht auf die Eckpfeiler zur linguistischen Beschreibung von Texten, Textsorten und Texttypologien ein. Nachfolgend werden die zentralen Aspekte, die in eine nähere Bestimmung von Textsorten und Texttypologien einfließen müssen, hervorgehoben. Die Ausführungen orientieren sich an Heinemann (2000c,d,e) und stellen für die nachfolgenden Kapitel eine Art programmatische Leitorientierung dar: Textsorten sind Abstraktionen einer begrenzten Menge von Textexemplaren, die sich Gemeinsamkeiten teilen. Sie können aufgrund ihrer Variabilität und Vagheit nach Anzahl und Umfang nicht exakt festgelegt werden. Die Gemeinsamkeiten beziehen sich auf unterschiedliche Ebenen, wie z. B. die Textgestalt, die Struktur, Formulierungsmuster, inhaltlich-thematische Aspekte, situative Bedingungen und kommunikative Funktionen; diesen Ebenen können Merkmale zugeordnet werden. Die textsortenspezifischen Besonderheiten dieser Dimensionen „sind aufeinander bezogen und bedingen sich wechselseitig“ (Heinemann, 2000d, S. 513). Versuche zur Differenzierung von Textsorten müssen auf Kernbereiche beschränkt bleiben. Textsorten sind an Konventionen innerhalb von Sprachgemeinschaften gebunden, die auf Erfahrungen der Kommunizierenden beruhen, weshalb Textsorten prototypische Kerne besitzen, die einen breiten Spielraum für ihre Realisierung bieten. Texttypen unterscheiden sich voneinander auf der Grundlage sehr weniger Kriterien. Textsorten sind Textklassen auf einer niedrigen Abstraktionsstufe und umfassen umfangreichere Inventare distinktiver Merkmale. Zusätzliche (meist inhaltlich-thematische) Merkmale unterscheiden Textsorten von Textsortenvarianten. Ansätze zur Beschreibung von Textsorten können unterschiedliche Geltungsbereiche besitzen, so bezieht sich z. B. Sandig (1972) explizit auf 18 Textsorten aus dem Bereich der Gebrauchstexte; literarische Texte werden in den meisten Arbeiten vollständig ausgeschlossen. Texttypologien reflektieren „die vielfältigen Möglichkeiten des In-Beziehungs-Setzens von Textsorten (und Textklassen anderer Abstraktionsstufen), wobei Einzelaspekte oder komplexe Merkmalbündel als Basis für Zuordnungsprozesse [...] fungieren.“ (Heinemann, 2000c, S. 543). Weiterhin vertritt Heinemann den Standpunkt, dass „die pragmatische Effizienz von Klassifikationen“ für die Konstruktion von Texttypologien relevant sein sollte. Dabei geht es um die Frage, „ob und in welchem Grade eine Typologie/Teiltypologie potentielle reale Aspekte kognitiver und kommunikativer Prozesse und Zusammenhänge abbildet oder nicht. Der Versuch einer exhaustiven, universellen und absoluten Erfassung irgendeines ›Gesamtsystems‹ von Textsorten erscheint daher weder sinnvoll noch erreichbar.“ (ebd.).

2.5 Fazit

Abschnitt 2.4 hat die zentralen Aspekte aufgeführt, die in die Beschreibung von Textsorten und den Aufbau von Texttypologien einzubeziehen sind und die das in Kapitel 5 dargestellte Hypertextsortenmodell maßgeblich beeinflusst haben. Adamzik (1995) fasst einen Teil der angesprochenen Aspekte in Form von Leitfragen zusammen:

> Auf welchem Abstraktionsniveau werden Textklassen unterschieden? Werden bei der Klassenbildung mehrere (Abstraktions-)Ebenen einbezogen [...]? Welchen Status haben die [...] verwendeten Kategorien? Wieviele und welche Merkmale/Eigenschaften von

> Texten werden bei der Klassifizierung von Texten berücksichtigt? Wenn mehrere Differenzierungskriterien berücksichtigt werden: Stehen diese Kriterien in einem hierarchischen oder in einem additiven Verhältnis zueinander? Und schließlich die [...] entscheidende Frage: Welches ist das leitende Erkenntnisinteresse bei der Arbeit mit Textklassen? (Adamzik, 1995, S. 14)

Der entscheidende Faktor bei der Zusammenstellung von Merkmalen oder Kriterien zur Textsortendifferenzierung und bei der Wahl einer oder mehrerer Einordnungsinstanzen zum Aufbau einer Texttypologie ist somit das Erkenntnisinteresse, d. h. die übergreifende Fragestellung, unter der die textlinguistische Analyse durchgeführt wird. Diesbezüglich besitzen alle Klassifikationen Heinemann und Viehweger zufolge zumindest *ein* gemeinsames Ziel:

> Jede Textklassifikation setzt sich das Ziel, die unendliche Vielfalt realer Texte auf eine überschaubare Menge von Grundtypen zu reduzieren, um auf diese Weise die kommunikative Praxis und letztlich auch gesellschaftliche Beziehungen und Strukturen durchschaubarer zu machen. (Heinemann und Viehweger, 1991, S. 145)

Heinemann (2000c, S. 539) geht auf weitere Zwecke von Textklassifikationen und Texttypologien ein und nennt Dokumentationen, technisch-praktische Aufgaben, z. B. zur Erstellung von Bibliotheksklassifikationen, und didaktische Zielsetzungen. Heinemann stellt seine Überlegungen zu unterschiedlichen Einsatzgebieten texttypologischer Untersuchungen unter ein Leitmotto: „Aufgabe von Textklassifikationen aber sollten weniger Theorie- oder System-Explikationen sein, sondern vielmehr Modellierungen realer Gegebenheiten des Kommunizierens." (ebd., vgl. auch Miller, 1984, S. 27 und S. 36).

In diesem Kapitel konnten verschiedene Aspekte nur partiell angerissen werden, z. B. Vertextungsmuster und Themenentfaltungstypen. Diese Aspekte sollten, bei einer ausschließlich textlinguistisch ausgerichteten Perspektive, durchaus in die Untersuchung von Hypertextsorten einfließen; das Erkenntnisinteresse der vorliegenden Arbeit betrifft jedoch unter anderem die Erstellung eines textlinguistisch fundierten Hypertextsortenmodells, das in computerlinguistischen Systemen – z. B. zur automatischen Identifizierung von Hypertextsorten – eingesetzt werden kann. Das in Kapitel 5 dargestellte Modell orientiert sich an dem Leitmotiv von Heinemann (2000c), so dass Teil III die „realen Gegebenheiten" der Untersuchungsdomäne – deutschsprachige Webseiten deutscher Universitäten – modelliert, weshalb zwangsläufig verschiedene Kernbereiche fokussiert werden. Hierbei geht es vornehmlich um diejenigen Aspekte, deren Operationalisierbarkeit durch computerlinguistische Verfahren gewährleistet ist. Das Hypertextsortenmodell ist explizit als Ausgangsbasis für weitere Forschungen zu verstehen. Es kann verwendet werden, um in zukünftigen Untersuchungen spezifische textlinguistische Aspekte innerhalb dieses globalen Rahmenmodells zu verorten.

3

Hypertext und das World Wide Web: Die linguistische Perspektive

3.1 Einleitung

Nachdem im vorangegangenen Kapitel ausschließlich textlinguistische Beschreibungsmethoden des Konzepts Text diskutiert wurden, widmen sich die folgenden Ausführungen dem Thema Hypertext. Insbesondere ist die Frage zu beantworten, ob im *World Wide Web* tatsächlich unterschiedliche Hypertextsorten voneinander differenziert werden können oder Hypertext selbst als eine Textsorte aufzufassen ist.[1] Eine Annäherung an Hypertext und die damit verbundenen Konzepte kann aus verschiedenen Perspektiven erfolgen. In informatischer bzw. informationswissenschaftlicher Hinsicht können z. B. unterschiedliche Hypertextsysteme als spezifische Software-Anwendungen mit individueller Funktionalität differenziert werden. Ebenfalls der Informatik zugehörig ist der Bereich des *Information Retrieval*, der in Bezug auf Hypertext die Frage thematisiert, mit welchen Methoden der Anwender für ihn relevante Informationseinheiten bzw. Teiltexte in komplexen Hypertexten auffinden kann (Stichwortsuche, Visualisierung etc., vgl. Agosti und Smeaton, 1996). In literaturwissenschaftlicher und auch künstlerischer Hinsicht existiert die Tradition der Hyperfiction-Erzählungen,

[1] Diese Ansicht wird z. B. deutlich bei Heinemann (2000d, S. 507): „Bezeichnenderweise haben Sprecher (vor allem Jugendliche) nach einer Lernphase auch kaum Probleme im Umgang mit *neuen Textsorten* (E-Mail, Hypertext).“ (ähnlich bei Heinemann und Heinemann, 2002, S. 141). Vergleichbarer Auffassung sind Weingarten (1997b, S. 217), Wagner (1998, S. 195), Naumann et al. (2003, S. 73) sowie Runkehl et al. (1998, S. 159): „[…] Besonderheiten von Hypertexten bringen es mit sich, daß dem Leser *dieser Textsorte* neue Herausforderungen bei der Rezeption von Texten entgegentreten […].“ (Hervorhebung hinzugefügt, G. R.). Amitay (1997, S. 57) zufolge ist Hypertext „*a new genre* with its own, already defined, linguistic patterns and conventions“ (Hervorhebung hinzugefügt, G. R.). Nach Schmitz (1997, S. 146) handelt es sich beim WWW um eine „Multimedia-Sorte“, in einer späteren Arbeit bezeichnet Schmitz (2000, S. 265) Hypertext und Hypermedia als „neue Textsorten“. Eichhoff-Cyrus (2000, S. 61) ist hingegen der Ansicht, dass der „Computer und das Internet […] eine Fülle von neuen Textsorten mit sich [brachten …]; diese Entwicklung wird sich fortsetzen.“

in denen die Leser – ähnlich den in den achtziger Jahren beliebten, rein textbasierten *Adventure*-Spielen – durch die individuelle Auswahl von Hyperlinks unterschiedlichen, vom Autor vorgegebenen Strängen einer Geschichte folgen können (vgl. Runkehl et al., 1998, S. 157 ff., und Landow, 1997). Das Ziel dieses Kapitels ist sowohl die Etablierung einer einheitlichen Terminologie, insbesondere unter Bezugnahme auf das *World Wide Web*, als auch die Darstellung der relevanten Konzepte, die in Arbeiten zum Thema Hypertext aus primär linguistischer bzw. textlinguistischer Perspektive vorgelegt wurden. Der Einsatz von Hypertextsorten in spezifischen Systemen – z. B. zum Zwecke des *Information Retrieval* – tritt dabei zunächst in den Hintergrund und wird in Teil IV der Arbeit diskutiert.

Es sind drei Beschreibungsebenen zu unterscheiden. Zunächst hat sich seit den siebziger Jahren eine eigenständige Terminologie etabliert, mit der unterschiedliche Merkmale des Konzepts Hypertext bezeichnet werden. Die zweite Ebene betrifft die technische Realisierung dieser abstrakten Konzepte in einem implementierten Hypertextsystem. Die dritte Ebene bezieht sich schließlich auf linguistische Beschreibungen konkreter Hypertextexemplare in spezifischen Hypertextsystemen. Die beiden ersten Ebenen werden im Folgenden ausführlich diskutiert (Abschnitt 3.3). Da sich diese Arbeit mit Hypertextsorten am Beispiel des WWW beschäftigt, werden diejenigen Möglichkeiten, die zwar in der Hypertext-Forschung als etabliert gelten, im WWW jedoch aus technischen Gründen nicht bzw. noch nicht realisierbar sind, nur am Rande diskutiert. Abschnitt 3.4 thematisiert verschiedene kritische Aspekte, die bei einer Diskussion von Hypertext häufig unbeachtet bleiben. Abschnitt 3.5 schwenkt daraufhin – analog zu Kapitel 2 – über zu verschiedenen linguistischen bzw. textlinguistischen Beschreibungsebenen von Hypertexten, woraufhin Abschnitt 3.6 Indikatoren für die Existenz von Hypertexttypen und Hypertextsorten aus mehreren Bereichen darstellt. Auf der Basis dieser Indikatoren werden zwei initiale Typologien entwickelt, die sich auf die Ebenen des Hypertexttyps und des Hypertextknotentyps beziehen. Zunächst erfolgt jedoch in Abschnitt 3.2 ein kurzer Abriss über die Entwicklung des Hypertextkonzepts.

3.2 Historische Grundlagen

Das zentrale Prinzip von Hypertext ist – ohne zunächst auf eine konkrete Definition einzugehen – die nichtlineare Verknüpfung von Informationseinheiten, üblicherweise beliebigen Texten und Teiltexten sowie Grafiken, Fotos und Tondateien. Die Verknüpfung erfolgt über Hyperlinks, die von jedem Bestandteil einer Informationseinheit zu anderen Bestandteilen anderer Informationseinheiten reichen können.[2]

[2] In diesem historischen Abriss werden die wesentlichen Entwicklungsschritte bis hin zum *World Wide Web* verfolgt. Ausführliche Historien und auch Überblicksdarstellungen verschiedenster Hypertextsysteme finden sich in der umfangreichen Einführungsliteratur, unter anderem bei Conklin (1987), Horn (1989), Jonassen (1989), Kuhlen (1991), McKnight et al. (1991), Bolter (1991), Rada (1991), Nielsen (1995b), Rada (1995), Landow (1997) und Hammwöhner (1997).

In der schon 1994 „längst nicht mehr überblickbar[en]“ Literatur zum Thema Hypertext (Freisler, 1994, S. 21, vgl. auch Meyrowitz, 1989, sowie Huber, 2002, S. 10 ff.) befinden sich immer wieder Darstellungen über die „neuzeitliche Mediengeschichte“ (Runkehl et al., 1998, S. 155 ff., auch bei Hammwöhner, 1997, und Jucker, 2000), die die Entwicklung von Schrift und Buchdruck, den Wandel in den Aufgaben von Textproduzenten und die Möglichkeiten, die ihnen mit modernen Textverarbeitungssystemen zur Verfügung stehen, thematisieren. Diese Aspekte werden in der vorliegenden Arbeit nur am Rande angesprochen.

Der Artikel „As We May Think“ von Vannevar Bush (1945a) gilt als Ausgangspunkt der Hypertextidee, da er eben dieses Prinzip etwa 20 Jahre vor seiner erstmaligen Realisierung vorwegnimmt.[3] Bush argumentiert, dass die Informationsmenge immer mehr zunimmt und dass Wissenschaftler kaum über effiziente Möglichkeiten der Informationsrecherche verfügen, weshalb er verschiedene Geräte und Technologien skizziert, die sowohl Forschern als auch Privatpersonen die Arbeit erleichtern könnten. Als ein wesentliches Problem sieht Bush die traditionellen hierarchischen Klassifikationen an, auf deren Basis Bibliotheken ihre Bestände sortieren: „When data of any sort are placed in storage, they are filed alphabetically or numerically, and information is found (when it is), by tracing it down from subclass to subclass.“ (ebd., S. 106). Dabei kann der Suchende auf Sackgassen stoßen; in jedem Falle benötige er aufwändige Regeln, um die richtigen Suchpfade zu verfolgen:

> The human mind does not work that way. It operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain. It has other characteristics, of course; trails that are not frequently followed are prone to fade, items are not fully permanent, memory is transitory. Yet the speed of action, the intricacy of trails, the details of mental pictures, is awe-inspiring beyond all else in nature. (Bush, 1945a, S. 106)

Diese Beobachtungen über die Wirkungsweise des Gehirns (vgl. Abschnitt 2.2.6) münden in der Idee des Memex (*Memory extender*) – „a sort of mechanized private file and library“ (ebd.). Der Memex, nach Bush eine Art Schreibtisch mit Tastatur, mehreren Sichtfenstern und diversen Hebeln, sollte die automatisierte Ablage von Büchern, Akten, Dokumenten, Briefwechseln und Notizen in einem Mikrofilm-ähnlichen Format ermöglichen, so dass sie zu einem späteren Zeitpunkt wieder effizient aufgefunden werden können. Da ein Memex potenziell sehr viele Materialien unterschiedlichster Art speichern könnte, sieht Bush insbesondere den effektiven Zugriff als wichtig an:

> It affords an immediate step, however, to associative indexing, the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another. This is the essentiell feature of the memex. The process of tying two items together is the important thing. (Bush, 1945a, S. 107)

Bush skizziert einen typischen Arbeitsablauf mit dem Memex, in dem sich der Anwender über die Ursprünge von Pfeil und Bogen informieren möchte. Der Benutzer hat Dutzende von Büchern und Artikeln in seinem Memex gespeichert und beginnt mit der Recherche in einer Enzyklopädie. Der Eintrag ist jedoch nur oberflächlich, weshalb ein geschichtliches Werk konsultiert und mit dem Lexikonartikel verknüpft wird. Weitere Beiträge zu spezifischen Subthemen werden aufgerufen, der Benutzer fertigt hierzu unmittelbar im Memex Notizen an und speichert den Recherchepfad anschließend unter einem Code-Wort ab: „Thus

[3] Der Beitrag ist im Juli 1945 in der Zeitschrift *Atlantic Monthly* erschienen, eine minimal revidierte und mit Abbildungen versehene Version wurde im September 1945 in *Life* veröffentlicht. Der Band *From Memex to Hypertext* (Nyce und Kahn, 1991a) enthält die ursprüngliche Fassung des Textes, in der die Änderungen, die in der späteren Version durchgeführt wurden, hervorgehoben sind (vgl. Bush, 1945b). Zusätzlich versammeln Nyce und Kahn weitere Arbeiten von Bush, die sich sowohl auf Vorstufen des Memex beziehen (Bush, 1941) als auch Revisionen des Konzepts (Bush, 1959, 1967) darstellen. Die Herausgeber gehen ausführlich auf die ursprünglichen (Nyce und Kahn, 1991b) und revidierten Überlegungen ein (Kahn und Nyce, 1991).

he builds a trail of his interest through the maze of materials available to him." (ebd.).[4] Prinzipiell kann diese Darstellung unmittelbar auf eine Recherche im *World Wide Web* übertragen werden: Der Benutzer geht von einem Interesse aus, identifiziert – mit Hilfe verschiedener Suchmaschinen – unterschiedliche Quellen und kann sich auf diese Weise zu einem bestimmten Thema informieren. Einige der von Bush dargestellten Möglichkeiten sind im WWW aus technischen Gründen nicht ohne Weiteres möglich bzw. haben sich noch nicht als Funktionen in den Browsern durchsetzen können. Hierzu gehören das Annotieren von Dokumenten mit (handschriftlichen) Notizen und das Abspeichern einer Recherche.[5]

Im Februar 1965 – 20 Jahre nach der Darstellung des Memex – hat Theodor Nelson in einem Vortrag am Vassar College (Poughkeepsie, New York) den Begriff Hypertext geprägt.[6] Ein im gleichen Jahr publizierter Konferenzbeitrag Nelsons enthält die erste Definition:

> Let me introduce the word "hypertext" to mean a body of written or pictorial material interconnected in such a complex way that it could not conveniently be presented or represented on paper. It may contain summaries, or maps of its contents and their interrelations; it may contain annotations, additions and footnotes from scholars who have examined it. Let me suggest that such an object and system, properly designed and administered, could have great potential for education [...]. Such a system could grow indefinitely, gradually including more and more of the world's written knowledge. (Nelson, 1965, S. 96)

Nelson geht mit „written or pictorial material" sowie „hyperfilm" (ebd.) explizit auf unterschiedliche Medien ein, die von dem Sammelbegriff „hypermedia" (ebd.) subsumiert werden. Allerdings wird das zentrale Konzept des Hyperlinks von Nelson *nicht* erwähnt (vgl. Wardrip-Fruin, 2004, S. 127). In einer späteren Veröffentlichung definiert Nelson (1987) Hypertext als „non-sequential writing – text that branches and allows choices to the reader, best read at an interactive screen".[7] Nelson publizierte verschiedene Konzeptionierungsphasen des bis

[4] Meyrowitz (1989, S. 298) vergleicht die von Bush (1945a) skizzierten Geräte mit den 1989 aktuellen Technologien und kommt zu dem Schluss: „A lot of the technology is just around the corner for putting this all [Bush's vision of the Memex, G. R.] together." McKnight et al. (1991, S. 8) betrachten die Überlegungen von Bush retrospektiv: „The memex was a problem in search of a solution, and the solution was the computer."

[5] Das Annotieren (vgl. Kuhlen, 1991, S. 114 ff.) von Webdokumenten, z. B. auf der Basis von Annotations-Servern, die die Anmerkungen von den Dokumenten isoliert speichern, wird seit etwa 1993 untersucht (vgl. z. B. Ovsiannikov et al., 1999, Vasudevan und Palmer, 1999, und Iorio und Vitali, 2005). Bereits die ersten Versionen des Browsers *Mosaic* enthielten derartige Mechanismen, die – nach dem Vorbild des NNTP-basierten Usenet – auf dezentralen Annotations-Servern operieren sollten; diese Funktion wurde von den Entwicklern folgender Browser nicht übernommen. Neben einigen kommerziellen Produkten verfügt der vom W3C entwickelte, experimentelle Browser *Amaya* über die Möglichkeit, Webdokumente zu annotieren.

[6] Nelsons Vortrag diskutierte die Grundideen von PRIDE (*Personalized Retrieval Indexing and Documentary Evolution*), einem System zur Organisierung verschiedenster Materialien. Ein in der College-Zeitung erschienener Artikel berichtet: „In this system passages of material would be [...] filed [...] in any sequence. [... T]he machine would print out any sequence the writer wished to try, freeing him from the necessity of keeping the ideas in his head. Mr. Nelson pointed out that we often do not think in linear sequences but rather in "swirls" and in footnotes. He introduced the concept of hyper-text, which would be a more flexible, more generalized, non-linear presentation of material on a particular subject." (Wedeles, 1965, S. 4).

[7] Nelson hat *Literary Machines* (1987) als Hypertext in Buchform konzipiert (ähnlich wie Horn, 1989, und Jonassen, 1989) und auch in einer digitalen Version (für das Hypertextsystem GUIDE) vertrieben, weshalb dieser Band keine Seitenzahlen im üblichen Sinne enthält.

heute nicht in vollem Umfang implementierten Hypertextsystems *Xanadu*[8] (vgl. z. B. Nelson, 1987, sowie Nelson, 1972, für einen Vergleich mit dem Memex), das verschiedene Funktionen umfasst, die im WWW aus technischen Gründen nicht realisierbar sind, z. B. bidirektionale Verknüpfungen (im WWW nur unidirektional), Versionskontrolle (im WWW abhängig vom verwendeten Editor bzw. *Content Management System*) oder die systeminhärente Möglichkeit zur Bezahlung von Inhalten (im WWW nur durch aufwändige Zusatzfunktionen realisierbar, die nicht Bestandteil des Protokolls HTTP sind). Nelson macht kein Geheimnis aus seiner Auffassung, dass das *World Wide Web* in konzeptioneller Hinsicht, z. B. in Bezug auf *Xanadu*, ein deutlicher Rückschritt sei. In „Embedded Markup Considered Harmful" bezeichnet er HTML als „one of the worst mistakes of the current software world" (Nelson, 1997, S. 129), da Probleme hinsichtlich des Editierens von Dokumenten bestehen, das Einfügen fremder Textteile zum Zwecke des Zitierens mit integrierter *micropayment*-Funktion („Transclusion", „Transpublishing") nicht möglich ist und die hierarchische Strukturierung von HTML-Dokumenten für viele Daten nicht geeignet erscheint.[9]

Das erste funktionsfähige Hypertextsystem wurde 1968 von Douglas Engelbart, dem „Edison of the personal computer" (Horn, 1989, S. 256) entwickelt (vgl. Engelbart und English, 1968, die ebenfalls die erste Computer-Maus vorstellen).[10] Das System NLS[11] (*oNLine System*, 1977 umbenannt in *Augment*, vgl. auch McKnight et al., 1991, S. 9) sollte primär in verteilten Arbeitsgruppen eingesetzt werden, so dass die Mitarbeiter in einem zentral administrierten, virtuellen Raum Dokumente und Informationen austauschen können.

Tim Berners-Lee entwickelte das *World Wide Web* in den Jahren 1989/1990 während eines Aufenthalts am europäischen Kernforschungszentrum CERN. Zehn Jahre zuvor implementierte er ein Hypertextsystem namens Enquire, das auf den folgenden Überlegungen basierte:

[8] Raskin (1987) kritisiert, dass sich Nelson ausschließlich auf sehr abstrakte theoretische Konzepte konzentriert und die für eine breite Akzeptanz des Systems zentrale Frage der Benutzerschnittstelle außer Acht lässt.

[9] Nelson (1997, S. 133) kommt zu der Schlussfolgerung: „Today's popular but trivially-structured Web hypertext has excused people from seeing the real hypertext issues, or being able to create and publish deep complexes of thought." In verschiedenen Punkten ist Nelson durchaus zuzustimmen – insbesondere hinsichtlich der Anforderungen an die Baumstrukturierung von SGML/XML-Instanzen – doch übersieht er, dass sich HTML, das in gewisser Weise durchaus „inhospitable" ist und „unsatisfied problems" (ebd., S. 134) mit sich führt, langfristig durchgesetzt hat, weshalb der geforderte Paradigmenwechsel als unrealistisch eingestuft werden muss. Auch abseits der hypertexttheoretischen Konzepte existieren Probleme: „HTML represents the worst of two worlds. We could have taken a formatting language and added hypertext anchors, so that users had beautifully designed documents on their desktops. We could have developed a powerful document structure language, so that browsers could automatically do intelligent things with Web documents. What we actually *have* with HTML is a hybrid: ugly documents without formatting *or* structural information." (Greenspun, 1995; vgl. auch Greenspun, 1999, S. 129 ff.). Aus Sicht des Webdesigns kritisiert Siegel (1999b, S. 4) die Tatsache, dass die Browser HTML-Code unterschiedlich interpretieren und darstellen: „Das Ganze ist so, als würde man einem Künstler erzählen wollen, wie er seinen Pinsel zu halten hat!"

[10] Engelbart bezieht sich, ebenso wie Nelson, unmittelbar auf Bush (1945a), was durch einen Brief deutlich wird, den Engelbart (1962) an Bush geschrieben hat. Die Aussage, Engelbart hätte das Konzept Hypertext unabhängig von Bush erfunden (vgl. Hofmann und Simon, 1995, S. 2), ist somit nicht korrekt.

[11] Das beinahe fertige System NLS wurde am 9.12.1968 auf der *Fall Joint Computer Conference* in einer Demonstration vorgeführt, die mittlerweile als „Mother of All Demos" gilt, weil erstmalig ein immenser technischer Aufwand betrieben wurde (unter anderem entfernter Zugriff auf das System über eine eigens angemietete Telefonleitung und großflächige Datenprojektion auf eine Leinwand). Weiterhin fand in dieser Demonstration der erste öffentliche Einsatz einer Computer-Maus statt. Ein Video der 90-minütigen Veranstaltung ist online verfügbar unter `http://sloan.stanford.edu/mousesite/1968Demo.html`.

> *Suppose all the information stored on computers everywhere were linked*, I thought. *Suppose I could program my computer to create a space in which anything could be linked to anything.* [...] There would be a single, global information space. [...] By being able to reference anything with equal ease, a computer could represent associations between things that might seem unrelated but somehow did, in fact, share a relationship. A web of information would form. (Berners-Lee, 1999, S. 4)

Diese grundlegenden Konzepte erarbeitete Berners-Lee nach eigenen Angaben unabhängig von Bush, Nelson und Engelbart: „Unbeknownst to me at that early stage in my thinking, several people had hit upon similar concepts“ (ebd., S. 5). Eine gleichermaßen ausführliche wie kurzweilige Darstellung der Ereignisse und strategischen Entscheidungen, die die technischen Grundlagen des WWW (u. a. HTML als Auszeichnungssprache und URLs als Adressierungsschema) geprägt haben, befindet sich in Berners-Lee (1999).

Für den globalen Erfolg des WWW sind mehrere Faktoren verantwortlich: Seine Architektur basiert auf dem Client-Server-Paradigma, d. h. die Server können als Datenspeicher in beliebiger Weise im Netzwerk verteilt sein, und unabhängig davon können Browser als Clients auf die Server zugreifen. Hypertext wurde somit erstmals auch physikalisch über die Grenzen eines Einzelplatzrechners hinaus vernetzt, potenziell zugreifbar aus dem gesamten Internet, dessen wachsende Popularität – Ende der achtziger Jahre vornehmlich bezogen auf wissenschaftliche Einrichtungen – ebenfalls ein wichtiger Faktor war.[12] Ein zentrales Merkmal des Internet ist wiederum, dass es auf offenen Standards wie TCP/IP basiert, d. h. es ist nicht auf spezifische Betriebssysteme oder Rechnerplattformen beschränkt.[13] Gleiches gilt in den meisten Fällen auch für das Client-Server-Paradigma, denn es existieren häufig unterschiedliche Implementierungen eines Servertyps (Mailserver, Newsserver etc.) und auch korrespondierender Clients, so dass die Anwender individuell präferierte Werkzeuge benutzen können. Berners-Lee und seine Gruppe haben in diesem Zusammenhang sowohl einen Browser für die Benutzeroberfläche der NeXT-Rechner als auch einen Client für textbasierte UNIX-Terminals entwickelt, um unterschiedlichen Benutzergruppen einen Zugang zum WWW zu ermöglichen (Berners-Lee et al., 1992, 1994).[14] Diese Software wurde nicht auf dem üblichen Wege gegen Zahlung einer Lizenzgebühr vertrieben, sondern konnte frei, d. h. kostenlos, per FTP von einem Server im CERN heruntergeladen werden. Zuzüglich war das initiale Software-Paket – der „line-mode“ Browser für UNIX-Terminals, ein rudimentärer Webserver (vgl. Abschnitt A.3.3) und eine Bibliothek mit gemeinsamen Funktionen – aber auch in dem Sinne frei, dass Modifikationen an der Code-Basis erlaubt und sogar erwünscht waren. Somit konnten Interessierte die Quelltexte als Grundlage einsetzen, um die Software auf neue Plattformen zu portieren, bestehende Funktionen zu erweitern oder Programm-

[12] Im WWW existiert kein zentrales Verzeichnis: „So long as I didn't introduce some central link database, everything would scale nicely. [...] Hypertext would be most powerful if it could conceivably point to absolutely anything. Every node [...] would have an address by which it could be referenced. They would all exist together in the same space – the information space.“ (Berners-Lee, 1999, S. 16).

[13] Eine detaillierte Betrachtung der technischen Grundlagen des Internet würde den Rahmen dieser Arbeit sprengen. Einen Überblick, insbesondere in Bezug auf Internet-Dienste, die vom WWW verdrängt wurden und mittlerweile kaum noch verwendet werden, bietet Krol (1994). Wilde (1999) geht auf das WWW ein.

[14] Dieser erste Browser namens *WorldWideWeb* war sowohl Browser als auch Editor. Der ursprünglich von Berners-Lee als wichtig erachtete Aspekt des unmittelbaren Editierens von Dokumenten ist bis zur Veröffentlichung des kombinierten Browsers und Editors *Netscape Communicator* in den Hintergrund getreten.

fehler zu beheben.[15] Eine solche Vorgehensweise war in der UNIX-Szene keine Seltenheit. Mit dem Erfolg des Betriebssystems *Linux* wurde das Open-Source-Prinzip weltweit bekannt (vgl. DiBona et al., 1999, Rehm und Lobin, 2003, und Hars und Ou, 2001). Neben der Software waren auch das zugrunde liegende *Hypertext Transfer Protocol*, die *Hypertext Markup Language* und das Adressierungsschema URL frei zugänglich, auch sie wurden nicht als proprietäre Technologien betrachtet, sondern Standardisierungsprozessen unterzogen (vgl. Abschnitt A.4.8).[16] Die von Berners-Lee konzipierte und einfach zu erlernende Auszeichnungssprache HTML basiert auf der Metasprache SGML (ISO 8879, Goldfarb, 1990), weshalb Anwender beliebige ASCII-Editoren für die Erstellung von Dokumenten verwenden können. Ein weiterer Grund für den unmittelbaren Erfolg war die Intuitivität, mit der die ersten grafischen Browser bedient werden konnten sowie die Tatsache, dass sich das WWW nicht auf den Transfer von Dateien (im Regelfall HTML-Dokumenten) beschränkt, sondern in der Lage war, weitere Internet-Dienste durch die Angabe ihrer Protokollbezeichner in einer URL benutzen zu können. Zuvor kompliziert auf der Kommandozeile zu bedienende Dienste wie WAIS (RFC 1625) oder FTP (RFC 0959) waren nun verkapselt in einer übergreifenden Oberfläche, dem Webbrowser, wodurch das WWW gerade bei eher unerfahrenen Benutzern schnell eine große Beliebtheit erreicht hat – Hammwöhner (1997, S. 23) nennt das WWW daher auch ein „Metainformationssystem für Mehrwertdienste“.[17] Durch die Möglichkeit, Fotos und Grafiken darstellen zu können, hob sich das WWW spürbar von Gopher (RFC 1436) ab, das als sehr eingeschränktes, ursprünglich ebenfalls ausschließlich auf der Kommandozeile zu verwendendes Hypertextsystem bezeichnet werden kann. Das WWW hat eine derart große Popularität erreicht, dass die Begriffe Webseite und HTML-Dokument häufig synonym mit Hypertext verwendet werden. Es handelt sich um „den globalen Hypertext schlechthin“ (Jucker, 2000, S. 12), „um einen einzigen großen Hypertext [...], der sich in permanenter Veränderung, in ständiger Bewegung befindet.“ (Sandbothe, 1997, S. 73).

Die zugrunde liegende Technologie des WWW realisiert nur einen Bruchteil der etablierten Hypertext- und Hypermediakonzepte, die im Laufe der vergangenen Jahrzehnte entwickelt worden sind.[18] Furuta und Marshall (1996) fassen diese offensichtliche Diskrepanz – Heyer und Wolff (1999, S. 101) sprechen von einer „Theorie-Praxis-Kluft“ – wie folgt zusammen (vgl. auch Smith et al., 1997, und ausführlich Bieber et al., 1997):

[15] Das CERN hat am 30.04.1993 alle Rechte an dieser Software abgetreten: „CERN's intention in this is to further [...] standards in networking and computer supported collaboration. [...] CERN relinquishes all intellectual property rights to this code [...] and permission is granted for anyone to use, duplicate, modify and redistribute it.“ (vgl. `http://info.web.cern.ch/info/Announcements/CERN/2003/04-30TenYearsWWW/`). Die URL deutet an, dass dieses Datum vom CERN als die eigentliche Geburtsstunde des WWW angesehen wird.

[16] Zur weiteren Entwicklung der Basistechnologien des WWW wurde im Oktober 1994 das *World Wide Web Consortium* (W3C) als *non-profit*-Organisation gegründet (vgl. `http://www.w3.org`). Neue Technologien durchlaufen einen komplexen Standardisierungsprozess, der schließlich in einer *Recommendation* mündet.

[17] Das WWW wird häufig als „Killer-Applikation“ für das Internet bezeichnet (vgl. z. B. Vora und Helander, 1997, S. 899). Nielsen (1995b, S. 183) fokussiert den ersten populären grafischen Browser: „The improved user interface made all the difference and Mosaic became the “killer app” of the Internet in the sense that many people started using the net just to use this one application.“ (vgl. auch Andreessen und Bina, 1994).

[18] Der Begriff Hypertext bezog sich ursprünglich auf textuelle Informationseinheiten, wohingegen der Terminus Multimedia eher zur Charakterisierung von Systemen verwendet wurde, die mehrere Medien unterstützen (Waterworth und Chignell, 1997, S. 916). In der vorliegenden Arbeit werden die Begriffe Hypertext und Hypermedia synonym verwendet.

> The hypermedia research community often views the Web with a combination of awe and frustration. Awe because of its meteoric ascendance (coupled with the voyeuristic sentiment that perhaps an equivalently dramatic extinction may be on the horizon). Frustration because of the sensation that hard lessons learned with hypermedia are not being reflected. (Furuta und Marshall, 1996, S. 193)

Fand die ursprüngliche Konzeptionierung und Implementierung der Basistechnologien des WWW noch unabhängig von der Hypertext-Forschung statt, so hat spätestens die Standardisierung von XML einen Konvergenzprozess initiiert, der die ehemals gespaltenen Lager einander annähert.[19] XML (Bray et al., 2004b) und flankierende Standards wie die XML *Linking Language* (DeRose et al., 2001) ermöglichen nun die Entwicklung von Hypertexten mit einem deutlich erweiterten Funktionsumfang. Derzeit können solche Realisierungen jedoch nur von Experten vorgenommen werden, die mit den texttechnologischen Grundlagen und Software-Systemen vertraut sind. Hinzu kommt, dass die Veröffentlichung derartiger Hypertexte letzten Endes meist im WWW erfolgt, weshalb zwangsläufig eine Konvertierung in die strukturarme *Hypertext Markup Language* nötig ist, da viele Anwender ältere Browser verwenden, die XML-Dokumente nicht darstellen können. Doch auch bei den neuesten Browsern kann derzeit nur von einer rudimentären Unterstützung der zahlreichen XML-Standards gesprochen werden. Die Entwicklung theoretischer Konzepte (vgl. die umfangreiche Liste von Standards auf dem Webserver des W3C) ist der Technologie, die dem Endanwender zur Verfügung steht, um mehrere Jahre voraus; die „Theorie-Praxis-Kluft“ besteht daher noch immer, nun allerdings in gewissermaßen entgegengesetzter Weise.

3.3 Hypertext: Theoretische und technologische Konzepte

Nachfolgend werden die relevanten theoretischen und technologischen Hypertextkonzepte erläutert. Ihre Menge und Komplexität ist mittlerweile kaum noch überschaubar, weshalb nur die zentralen Aspekte thematisiert werden, die im *World Wide Web* technisch realisierbar sind. Die Realisierbarkeit bezieht sich auf die Basistechnologien des WWW (d. h. HTTP und HTML). Mit *Java*-Applets oder proprietären *Plug-ins* können zwar weiterführende Funktionen realisiert werden, diese stellen jedoch – insbesondere in Bezug auf das Korpus (vgl. Abschnitt A.4.5) – Ausnahmen dar. In gewisser Weise folgt die Darstellung Heinemann (2000c, S. 539) und versucht, die technologischen und konzeptionellen Komponenten der „realen Gegebenheiten des Kommunizierens“ im WWW zu beschreiben (vgl. Abschnitt 2.5).

Bezüglich der Untersuchung von Hypertextsorten kommt ein weiterer Aspekt hinzu: Die meisten im WWW publizierten HTML-Dokumente unterliegen einer sehr einfachen Strukturierung: Vermutlich nur ein Bruchteil der Autoren, die Webseiten veröffentlichen, ist mit den abstrakten und größtenteils sehr komplexen hypertexttheoretischen Konzepten, sprachlichen Prämissen und kognitiven Grundlagen vertraut, wie sie z. B. von Kuhlen (1991) und

[19] Siehe hierzu auch Kuhlen (1997, S. 366): „Sicherlich ist die Hypertextmethodologie insgesamt weiter entwickelt als das für viele nur einen kleinen gemeinsamen Nenner verwirklichende WWW, aber mit WWW ist Hypertext zum umfassenden Vorbild von nicht-linearer Wissensdarstellung und Informationserarbeitung [sic] geworden. Die Universalisierung des Hypertextprinzips [...] mag das Ende einer Spezialdisziplin »Hypertext« bedeuten, keinesfalls aber das Ende der Innovationsfähigkeit der Idee.“

Hammwöhner (1997) detailliert vorgestellt werden (vgl. auch Eckkrammer, 2001, S. 63). Hinzu kommt, dass HTML eine einfach zu erlernende Auszeichnungssprache ist, Anfänger können erste Dokumente schon nach einer kurzen Einarbeitungszeit realisieren.[20] Auch die Einführungsliteratur zu den Themen HTML und *Web-Publishing* reflektiert nur selten die „hard lessons" (Furuta und Marshall, 1996), die über Jahrzehnte hinweg von der Hypertextforschung gemacht worden sind (vgl. Ratner et al., 1996, und Cunliffe, 2000). Stattdessen werden die Syntax von HTML und die Darstellungsspezifika bestimmter Elemente in unterschiedlichen Browsern, also die Realisierung primär visueller Spezialeffekte besprochen (vgl. z. B. Grigoleit, 1995, S. 251 f.).[21] Autoren von Webseiten orientieren sich also nicht an Hypertextkonzepten, sondern gehen einerseits von technischen Aspekten und dem Funktionsumfang von HTML aus (vgl. Lutz, 1995, S. 155 f., und Foltz, 1996, S. 110), andererseits basieren sie die eigenen Dokumente bewusst oder unbewusst auf bereits rezipierten HTML-Dokumenten (vgl. Eckkrammer, 2001, S. 62 f.). Ein besonders gelungener Seitenaufbau, ein ansprechender Farbraum oder eine intuitiv sinnvoll erscheinende Auswahl von Informationseinheiten in einem anderen – zufällig gefundenen oder explizit gesuchten – Dokument wird bei der Anfertigung der eigenen Webseiten übernommen, d. h. Produzenten orientieren sich insbesondere an anderen Dokumenten sowie ihnen bekannten Textsorten (vgl. Abschnitt 4.3.2).[22] In der Konsequenz bedeutet dies, dass hochgradig vernetzte und im eigentlichen Sinne nichtlineare, d. h. prototypische[23] bzw. idealtypische Hypertexte, die einen sehr großen Produktionsaufwand mit sich führen, im *World Wide Web* nur den Status punktueller Randerscheinungen besitzen.

[20] Der Funktionsumfang von HTML/HTTP bleibt weit hinter den hypertexttheoretischen Konzepten zurück (vgl. Hammwöhner, 1997, S. 110, und Nürnberg und Ashman, 1999). Eine Extremposition vertritt Pang (1998), der die von Bolter (1991) und Landow (1992) vorgeschlagenen Konzepte einer kritischen Analyse unterzieht und zu folgendem Schluss kommt: „[H]ypertext [...] doesn't actually exist, though hypertext theory is based on the assumption that it does. Versions of some of hypertext's features can be found in software programs [...] and on the World Wide Web. But none of these is hypertext as described in the literature. [...] Until these capabilities are popularly (or even universally) available, readers cannot be transformed into writers, and authorial power cannot be diffused; [...] The fact that hypertext theory is writing about a revolutionary impact of a nonexistent technology wouldn't matter if its proponents recognized that fact. But they don't."

[21] Es hat sich nicht etwa „Web-Autor" oder „Web-Texter", sondern „Web-Designer" als Berufsbezeichnung etabliert, weil das Design eines Webauftritts im professionellen Kontext einen hohen Stellenwert besitzt (vgl. Nielsen, 1996, sowie Fußnote 36, S. 78). Storrer (2001b, S. 107) merkt an, dass das technische Potenzial häufig voll ausgeschöpft wird, um Medienkompetenz zu signalisieren (vgl. Rosenfeld und Morville, 1998, S. 5 f., Berker, 2001, S. 218 f., sowie Abschnitt 4.6.3). Der Umstand erklärt die Präzisierung des Autorenbegriffs, die Williams (2002, S. 132) vornimmt: „As authors and their purposes flex, so must readers (or more accurately, "users"). As users and their purposes shift, so must authors (or more accurately, "*designers*")." (Hervorhebung hinzugefügt, G. R.; ähnlich bei Crystal, 2001, S. 197 f.).

[22] Vora und Helander (1997, S. 896) vertreten die Ansicht: „Most hypertext designs are still based on the designer's intuitions and past experience." Die in der Vergangenheit rezipierten Dokumente können auch eine implizite Auswirkung auf die eigenen HTML-Dokumente haben (vgl. Fleming, 1998, S. 87). Rosenfeld und Morville (1998, S. 2) geben den Ratschlag: „You can't really become a proficient web site architect unless you first know what it's like to really use the Web on a regular basis. [... T]he best web site producer is an experienced consumer." (vgl. Taylor et al., 2001). Nielsen (1995a, S. 75) merkt an, dass eine „competetive usability analysis", d. h. eine Analyse der Produkte von Mitbewerbern, eine effiziente Möglichkeit ist, die eigene Website zu verbessern (vgl. Cunliffe, 2000, S. 301, Siegel, 1999a, S. 148, sowie Abschnitt 3.6.7).

[23] In Anlehnung an Sandig (2000, vgl. Abschnitt 2.2.9) und Rouet und Levonen (1996, S. 15: „the prototypical representation of hypertext is a set of text units connected through links, that is, a text network").

3.3.1 Linearität und Nichtlinearität

Eine eindeutige und generell anwendbare Charakterisierung des Terminus Hypertext existiert bislang nicht. Derartige Definitionen sind – ähnlich wie die in Abschnitt 2.2.8 angesprochene Problematik der Definition von Text – vom jeweiligen Forschungsinteresse bzw. fokussierten Beschreibungsaspekt abhängig (vgl. Pfammatter, 1998, S. 47 ff., und Storrer, 2000b, S. 223). Für Conklin (1987, S. 17) ist Hypertext „quite simple: Windows on the screen are associated with objects in a database, and links are provided between these objects, both graphically [...] and in the database [...]." Conklin beschränkt sich auf den Zugriff und die Speicherung von Informationen, was auch erklärt, weshalb nicht zwischen Hypertext und „nonlinear text" (ebd.) differenziert wird. Vora und Helander (1997, S. 878) schränken den Geltungsbereich des Hypertextkonzepts auf drei Aspekte ein: (i) Erstellung, (ii) Ablage und Verwaltung sowie (iii) Präsentation von bzw. Zugriff auf Informationen. Nelson (1987) geht von der Perspektive des Textproduzenten aus. Seine Auffassung von Hypertext als „non-sequential writing" gehört zur ersten Kategorie. Nielsen (1995b, S. 4) betont mit „true hypertext should also make users *feel* that they can move freely through the information" die dritte Ebene. Kuhlen (1991, S. 27) geht von der zweiten und dritten Kategorie aus und definiert „Hypertext als ein *Medium der nichtlinearen Organisation* von Informationseinheiten", dessen Grundidee darin besteht, „daß informationelle Einheiten [...] flexibel über Verknüpfungen manipuliert werden können" (ebd., S. 13). Sager (2000, S. 589) integriert mehrere Perspektiven und definiert Hypertext als „ein kohärenter, nichtlinearer, multimedialer, computerrealisierter, daher interaktiv rezipier- und manipulierbarer Symbolkomplex über einem jederzeit vom Rezipienten unterschiedlich nutzbaren Netz von vorprogrammierten Verknüpfungen." Storrer (2004b, S. 31) liefert eine kommunikationstheoretische Annäherung: Hypertext lässt sich „bestimmen als eine an Computertechnik gebundene neue Kommunikationsform, die Aspekte von Text und Diskurs miteinander verbindet und damit neue Formen der Wissensorganisation und der Wissenserarbeitung ermöglicht [...]."

Werden die Gemeinsamkeiten der zahlreichen Definitionen verglichen, sind die Kerneigenschaften von Hypertext[24] demnach Nichtlinearität und Multimedialität sowie das Merkmal, im Computer realisiert zu sein (vgl. Flender und Christmann, 2000). Die Nichtlinearität bezieht sich auf einen Gegensatz zu gedruckten Texten (vgl. Abbildung 3.1): Bücher, Artikel oder Briefe werden sequenziell rezipiert (vgl. Bucher, 1999, S. 12), wohingegen sich die Leser eines Hypertextes frei in der verfügbaren Hypertextbasis bewegen können.[25]

[24] Häufig werden die Bestandteile des Kompositums Hypertext analysiert, um seine Spezifika herauszustellen, z. B. von Runkehl et al. (1998, S. 157), Storrer (2000b) und Schütte (2004a, S. 27 f.).

[25] Aus diesem Grund betont Nelson (1972, S. 253): „The best current definition of hypertext [...] is "text structure that cannot be conveniently printed." This [...] fits best." Ein Hypertext ist also ein Text, „der sich nicht ohne Wertverlust auf Papier ausdrucken lässt" (Storrer, 2000b, S. 229). Lutz (1995, S. 161) fügt hinzu: „Das Ausdrucken von Hypertexten kann eigentlich nie zur vollen Zufriedenheit gelingen. Bei stärker vernetzten Strukturen ist ein Ausdrucken meist gar nicht mehr sinnvoll [...]." Dies erkennt auch Handler (1997, S. 98): „Je »hyper« der Hypertext ist [...], desto schwieriger wird der Transfer auf Papier." (Aspekte der Linearisierung von Hypertexten und der Entlinearisierung traditioneller Texte werden von Ghaoui et al., 1990, Kuhlen, 1991, S. 162 ff. und Storrer, 1997 thematisiert). Rieffel (1999, S. 5) weist darauf hin, dass das Ausdrucken *keinesfalls* ein kritischer Faktor ist, denn „any static hypertext document can be converted to a paper one which maintains the links by numbering all lines [...], putting a box around the link anchor, and putting a margin note on the side giving the lines the link goes to. [...] The interface might not be as convenient to use as the online one, but

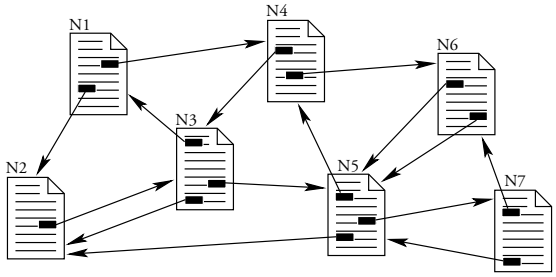


Abbildung 3.1: Vereinfachte Hypertextstruktur (nach Vora und Helander, 1997, S. 878)

> Die nicht-lineare Organisation unterstützt das selektive Lesen und ermöglicht es, Wissen für heterogene Adressatengruppen [...] zu vermitteln. Das Netzwerk [...] wird [...] auf je individuellen Rezeptionspfaden durchschritten, d. h., jeder Rezipient entscheidet nach Vorwissen, Vorlieben und Interessen selbst, welche Module er in welcher Reihenfolge und Zusammenstellung abrufen möchte [...]. (Storrer, 2000b, S. 228)

Die nichtlineare Organisation eines Hypertextes ist dafür verantwortlich, dass „die Leserin aktiv entscheiden muss, wohin sie sich als nächstes wendet. Man kann also sagen, dass der [...] linear rezipierte Text erst durch seine Interaktion mit dem [...] multilinearen Hypertext entsteht." (Jucker, 2000, S. 23; zum Terminus der Multilinearität siehe Fußnote 78, S. 94).[26] Traditionelle Texte, die schon in ihrer „logischen und referentiellen Struktur inhärent nichtlinear" sind (Hammwöhner, 1997, S. 27), verfügen mit Fuß- oder Endnoten über explizite nichtlineare Verknüpfungen, deren Rezeption optional ist, weshalb Hypertext auch gelegentlich als „generalisierte Fußnote" bezeichnet wird (Nielsen, 1995b, S. 2, vgl. auch die Abbildungen bei Horn, 1989, S. 30 f., und Kuhlen, 1991, S. 39, sowie Handler, 1997, und Rieffel, 1999).[27] Nichtlineare Verknüpfungen können auch auf der sprachlichen Ebene rea-

it is still easy to use for reasonable sized documents." Die realen Gegebenheiten im WWW widersprechen den eingangs genannten Auffassungen: Auf vielen Webseiten existieren Hyperlinks, die z. B. mit „Druckversion" beschriftet sind und eine um Grafiken, Navigationselemente und Werbebanner reduzierte Variante des Textes bereitstellen (vgl. auch Nielsen, 1999, S. 94 ff.). Auch die Tatsache, dass jeder Browser eine Druckfunktion besitzt, ist ein Indiz dafür, dass Hypertexte im WWW nicht annähernd die Komplexität prototypischer Hypertexte aufweisen. Da Webseiten häufig aktualisiert werden, kann das Ausdrucken „nur schnappschussartig ihren momentan Zustand einfangen." (Storrer, 1999b, S. 4; vgl. auch Kapitel 7).

[26] In der Literatur finden sich Aussagen wie z. B. „Die Benutzer können die Bausteine in einer nichtlinearen Weise lesen" (Schweiger, 1996, S. 327) oder „hypertext [is] reading being regarded as non-linear" (Askehave und Nielsen, 2005, S. 3; ähnlich bei Géry, 2002b). Die Textrezeption findet jedoch, *unabhängig* von der Organisationsform, *ausschließlich* sequenziell statt: „Die faktischen Pfade, die ein Hypertextleser [...] einschlägt, sind im Sinne des reinen Abfolgearguments natürlich linear." (Kuhlen, 1991, S. 33). Zimmer (1997) formuliert überspitzt: „»Linear« nämlich ist und bleibt Lesen immer – kreisförmig oder um die Ecke oder mehreres simultan kann keiner lesen." De Saint-Georges (1998, S. 70) fügt hinzu: „A first generation of enthusiastic literary theorists saw in hypertext the first physical realization of a mode of writing, which, assuming to be non-linear, would precisely transform the linear reading of texts. [... W]e see that pages on the Internet are still objects having intrinsically a beginning and an end, a top and a bottom and that the reading process is organized as with non-hypertext text." (vgl. auch Jucker, 2004).

[27] Storrer (2001c, S. 196) identifiziert zwei Unterschiede zwischen Hyperlinks und Verweisen in linear organisierten Texten: Hyperlinks sind notwendig und sie sind appellativ, d. h. die Verfolgung eines Links ist automatisiert, der Leser muss das Verweisziel nicht selbst erschließen.

lisiert werden, z. B. anaphorische („siehe oben“) und kataphorische Verweise („im Folgenden ...“ etc.) auf andere Textteile (vgl. Sager, 2000, S. 587). Nach Runkehl et al. (1998, S. 127) hat ein Hypertext „einen (scheinbaren) Anfang und kein bzw. viele Enden“, doch gilt dies in gleichem Maße auch für viele traditionelle Textsorten: Niemand liest ein Lexikon (vgl. zu diesem Beispiel auch Hess-Lüttich, 1997, S. 137, Lobin, 1999a, S. 155, und Jucker, 2000, S. 11) oder ein Telefonbuch vom ersten bis zum letzten Eintrag, und auch bei Fachbüchern und -artikeln kann häufig ein eher nach Schlagwörtern, Definitionen, Literaturverweisen, Formeln, Tabellen (vgl. McKnight et al., 1991, S. 17) oder Abbildungen „scannendes“ Lesen, Selektieren oder Überfliegen beobachtet werden als eine Wort-für-Wort-Rezipierung, wie sie z. B. für Privatbriefe und Erzählungen typisch und auch notwendig ist (vgl. Foltz, 1996, S. 119 f.).[28] Freisler (1994, S. 27) resümiert: „Der Übergang zwischen linearen Texten und Hypertexten ist – abgesehen vom Medium – fließend.“

3.3.2 Hypertextsystem

Als Hypertextsystem wird die Software bezeichnet, die dem Benutzer einen Hypertext präsentiert und die Interaktion mit dem Hypertext gewährleistet.[29] Einige Hypertextsysteme integrieren zusätzlich Funktionen zur Eingabe und Manipulation des Hypertextes. Übersichten über Hypertextsysteme bieten Conklin (1987), Nielsen (1995b) und Hammwöhner (1997, S. 105–113).[30] Im Kontext des WWW bilden Webserver und Webbrowser, die per HTTP (RFC 2616) miteinander kommunizieren, sowie die zur Erstellung der Dokumente verwendeten Editoren und Werkzeuge gleichermaßen das Hypertextsystem.

3.3.3 Knoten

Der Begriff Knoten ist eine abstrakte Bezeichnung für Informationseinheiten in Hypertexten (vgl. Kuhlen, 1991, S. 79 ff.).[31] In multimedialen Hypertext- bzw. Hypermediasystemen können in Knoten unterschiedliche Medientypen miteinander kombiniert werden, z. B. Text, Bild, Ton und Video. Hyperlinks, die im nachfolgenden Abschnitt thematisiert werden, fungieren als Bindeglieder zwischen Knoten (vgl. Abbildung 3.1, S. 75).

[28] Nielsen (1999, S. 104) berichtet von einer Studie, derzufolge 79% der Probanden eine neue Webseite zunächst „scannen“, um sich einen Überblick zu verschaffen (vgl. Morkes und Nielsen, 1998, Bucher, 2000, S. 682, und Storrer, 2001c, S. 181 ff.). In Bezug auf journalistische Printtexte ist Schmitz (2001, S. 218) der Ansicht: „Wer heute Texte liest, tastet nur noch partiell [...] eine linear konstruierte Ganzheit von vorne nach hinten ab; meist wählt er nach komplizierteren, durch Textdesign [...] teilweise vorgeprägten Mustern das, worauf sein Auge fällt oder was ihn interessiert.“ Bucher (1996) bezeichnet diesen Lesertyp als „Anleser“.

[29] Freisler (1994, S. 40 f.) liefert eine abstrakte Definition: „Ein Hypertext-Hypermedia-System ist eine im elektronischen Medium realisierte Kommunikations- und Publikationsmaschine, die es erlaubt, synästhetisierte informationelle Einheiten dialogisch-flexibel in entlinearisierten Strukturen beliebig zu verknüpfen und zu manipulieren.“ (ähnlich bei Storrer, 2000b, S. 229).

[30] Nielsen (1995b, S. 132 ff.) geht auf unterschiedliche Architekturen von Hypertextsystemen ein und diskutiert das Modell von Campbell und Goodman (1988), das auch auf das WWW anwendbar ist.

[31] Diese Terminologie wählt z. B. Rada (1995, S. 21): „Hypertext is nodes of text connected by links.“ (vgl. Whitehead, 2000, der terminologische Aspekte von Hypertextsystemen untersucht). Nach Ansicht von Nielsen (1995b, S. 136), der in Bezug auf Knoten „frame-based systems“ und „window-based systems“ unterscheidet, sind Knoten die „fundamental unit of hypertext“ (Nielsen, 1995b, S. 136).

In Bezug auf das WWW ist eine Definition des Knotenbegriffs mit Problemen verbunden. HTML unterliegt den syntaktischen Beschränkungen der korrespondierenden DTD (Raggett et al., 1999), doch waren schon die ersten Browser in der Lage, auch HTML-Dokumente korrekt darzustellen, die hinsichtlich der DTD syntaktische Fehler aufweisen (vgl. Abschnitt A.4.8). Daher kann der Knotenbegriff nicht auf die syntaktische Korrektheit des HTML-Markups reduziert werden. Stattdessen wird der Begriff im Kontext des WWW in sehr genereller Weise auf Dokumente bezogen, die auf HTML basierende Auszeichnungen enthalten und von den meisten Browsern offenbar[32] korrekt und vollständig dargestellt werden. Damit es sich auch in graphentheoretischer Hinsicht um einen Knoten handeln kann, muss weiterhin mindestens ein Hyperlink enthalten sein, der auf ein anderes Dokument verweist (vgl. Freisler, 1994, und Runkehl et al., 1998). Im umgekehrten Fall, wenn keine abgehenden Hyperlinks existieren, liegt ein Hypertextblatt vor.[33] Synonym zu Hypertextknoten werden die Begriffe HTML-Dokument, Webseite und Webdokument verwendet.

HTML-Dokumente werden in einer speziellen Verzeichnishierarchie auf dem Webserver hinterlegt, so dass sie auf eingehende HTTP-Anfragen hin unmittelbar ausgeliefert oder dynamisch aus Datenbanken generiert werden können. HTML 4.01 (Raggett et al., 1999) spezifiziert 91 Elemente und 119 Attribute, mit denen die Dokumentstruktur markiert werden kann. Ihre Präsentation übernimmt der Browser, dessen *Rendering Engine* über Regeln verfügt, die HTML-Elemente auf grafische Objekte abbilden. Obligatorisch ist die Angabe des Wurzelelements `html`, das wiederum die Elemente `head` (für Metainformationen) und `body` enthalten muss. Innerhalb von `body` wird der Inhalt festgelegt, der im Browser dargestellt wird. Die Möglichkeiten der strukturierten Textauszeichnung von HTML sind begrenzt, so stehen z. B. Elemente für unterschiedliche Überschriften (die Elemente `h1` bis `h6`), Textabschnitte (`p`, `div`, `span`), Zitate (`blockquote`), Adressen (`address`), vorformatierten Text (`pre`), Listen (`ol`, `ul`, `dl`), Tabellen (`table`) und interaktive Formulare (`form`) zur Verfügung. Innerhalb von Textabschnitten können einzelne Zeichenketten gesondert markiert werden (`font`, `em`, `strong`, `code`, `samp`, `kbd`, `var`, `cite`, `dfn`, `acronym`, `abbr`, `b`, `i`, `tt`, `u` etc.), was üblicherweise in einem Wechsel der Schriftart (z. B. `tt` für dicktengleiche Schrift) bzw. des Schriftschnitts (z. B. `em` und `i` für die Kursive) resultiert.[34] Weiterhin können multimediale Objekte integriert werden (`img`[35] für Bilder, `embed` und `object` für weitere Dateitypen). Zusätzlich können dem Benutzer durch den Einsatz von Framesets mehrere HTML-Dokumente gleichzeitig präsentiert werden, indem der Darstellungsbereich des Browsers in rechteckige Flächen aufgeteilt wird, die jeweils den Inhalt eines Dokuments aufnehmen. Trotz dieses umfangreichen Vokabulars beschränken sich Produzenten auf den Einsatz nur weniger Elemente (vgl. Abschnitt A.4.2). Für die maschinelle Auswertung von Webseiten ist es sehr problematisch,

[32] In Zweifelsfällen existiert keine Möglichkeit der Überprüfung, ob das visuelle Ergebnis der *Rendering*-Stufe der vom Autor intendierten Darstellung entspricht (vgl. auch Siegel, 1999b, S. 64).

[33] Auch andere Formate können als Knoten fungieren, z. B. *Java*-Applets oder *Flash*-Animationen. Dieser Aspekt wird nicht berücksichtigt, da es sich um Randerscheinungen handelt. HTML ist das native Format des WWW.

[34] Logisches (`em`, `strong`) wird von physikalischem Markup (`i`, `b`, `tt`) unterschieden (vgl. Abschnitt A.4.2).

[35] Hierbei kann es sich um beliebige visuelle Darstellungen handeln, z. B. Fotos, Strichzeichnungen, Abbildungen, typografisch gestalteten Text (Siegel, 1999b, S. 99, spricht von „Typobildern“) und visuelle Platzhalter, die für Layoutzwecke verwendet werden und im Browser zusätzlichen Zwischenraum erzeugen. Da die Dateitypen GIF, JPEG und PNG keine Metadaten über den Typ des Bildes bereitstellen, kann eine maschinelle Differenzierung nur approximativ stattfinden.

dass viele Autoren nicht auf den vollen Funktionsumfang von HTML zurückgreifen; Zitate werden z. B. nur in den seltensten Fällen mit dem Element `cite` markiert. Weiterhin werden Elemente zweckentfremdet, um einen bestimmten visuellen Effekt zu erzielen (Barnard et al., 1996, bezeichnen dieses Phänomen als „tag abuse"). So sollten etwa Überschriften mit den Elementen `h1` bis `h6` markiert werden, sehr häufig werden jedoch Kombinationen aus `font` (zur Aktivierung einer größeren Schrift) und `strong` bzw. `b` (für Fettdruck) benutzt. Dies sind nur zwei Indizien dafür, dass die Autoren vieler Webseiten primär bestrebt sind, ein möglichst ansprechendes grafisches Erscheinungsbild ihrer Dokumente und nicht eine möglichst präzise und standardkonforme Auszeichnung der Dokumentstruktur zu realisieren.

3.3.4 Hyperlinks als Verbindungen zwischen Knoten

Hyperlinks sind *die* zentrale Eigenschaft, die Hypertexte von auf Papier gedruckten Texten unterscheiden (Nielsen, 1995b, S. 138). Hyperlinks (auch: Links, Verweise, Verknüpfungen) verbinden Knoten untereinander und spannen multilineare, netzartige Strukturen (d. h. gerichtete Graphen) auf, in denen die Anwender[36] navigieren können (vgl. Abbildung 3.1, sowie Ziegler, 2004). Die Navigation ist ein inhärent interaktiver Vorgang, denn nach Sager (2000, S. 589) ist Interaktivität eine „Verknüpfungstechnik [...], die es erlaubt, potentielle, auf dem Bildschirm [...] markierte Verbindungen zu anderen medialen Einheiten anzubieten, auf die der Benutzer unmittelbar durch eine bestimmte Aktion (Tastatureingabe, Mausklick [...]) reagieren kann. Dies wiederum führt zu einer Reaktion des Rechners." In vielen Hypertextsystemen – auch bei den meisten grafischen Browsern – verändert sich der Mauszeiger, wenn er auf einen Hyperlink bewegt wird, der möglicherweise nicht eindeutig als solcher gekennzeichnet oder gar unsichtbar ist.[37]

Während die in HTML eingesetzte Verknüpfungstechnologie einen eher geringen Spielraum bietet, hat die Hypertextforschung diesbezüglich zahlreiche Konzepte entwickelt, die sich jedoch nicht in HTML niedergeschlagen haben. Hierzu zählt z. B. eine Unterscheidung von $1:1$, $n:1$, $1:m$ und $n:m$ Hyperlinks, d. h. Verknüpfungen mit exakt einem Ausgangs- und Zielpunkt, einem Ziel, das aber von mehreren Ausgangspunkten aus angesteuert werden kann, einem Ausgangspunkt, der zu mehreren Zielen führen kann und Hyperlinknetzen, in denen zu einem Punkt mehrere Einheiten führen und von ihm auf mehrere verweisen können (Kuhlen, 1991, S. 113). Einige Systeme unterstützen auch bidirektionale Hyperlinks, so dass in jedem Knoten bekannt ist, welche eingehenden Verweise existieren – wird ein Knoten gelöscht, werden gleichzeitig auch alle Verweise auf diesen Knoten entfernt. Auf diese Weise kann garantiert werden, dass eine Verknüpfung immer ein wohldefiniertes Ziel hat.

[36] Aufgrund der obligatorischen Realisierung von Hypertexten im Rechner und der damit verbundenen interaktiven Komponente werden ihre Rezipienten primär als Benutzer oder Anwender und nur seltener als Leser bezeichnet (vgl. Haas und Grams, 1998a, S. 485, Groß, 2000, Storrer, 2000b, sowie Fußnote 21, S. 73).

[37] Sager (2000, S. 600) bezeichnet den Cursor als „komplexe semiotisch-mediale Maschinerie [...], die ebenfalls charakteristisch für das Hypermedium" ist und erstellt ein Schema von „Cursoraktivitäten" (ebd., S. 602), das seiner Ansicht nach deutlich macht „dass der Umgang mit dem Hypermedium bereits ein recht differenziertes Repertoire von Fähigkeiten erfordert und durchaus als eine neue Kulturtechnik für den Umgang mit einem neuen Medium angesehen werden kann." (ebd.). Doch nicht nur Hypermediasysteme, sondern die grafischen Oberflächen aller modernen Betriebssysteme bauen auf diesem Konzept auf und verwenden unterschiedliche „Cursoraktivitäten". Dass der Mauszeiger nicht inhärent mit Hypertext verbunden ist, zeigen Browser wie *lynx*, die in einem Terminalfenster mit der Tastatur bedient werden (vgl. auch Fußnote 118, S. 111).

Hyperlinks werden mit dem HTML-Element `a` (*anchor*) realisiert (vgl. Abschnitt A.4.3).[38] Das Element benötigt im Attribut `href` (*hypertext reference*) die Adresse (URI, RFC 2396) des Linkziels und umschließt eine Zeichenkette oder Abbildung, die im Browser visuell hervorgehoben wird und als Hyperlinkanzeiger fungiert (vgl. Abschnitt 3.5.5).[39] Ein Beispiel für einen in HTML realisierten Hyperlink lautet `<a href="http://www.uni-giessen.de"> Justus-Liebig-Universität</a>`. Verknüpfungen in HTML sind ausschließlich unidirektional, d. h. am Zielpunkt ist nicht bekannt, dass ein eingehender Link vorliegt. Hyperlinks können relativ oder absolut sein, d. h. sie können sich auf den gleichen Webserver oder einen entfernten Webserver beziehen. Durch die Angabe anderer Protokolle können weitere Internet-Dienste referenziert werden, z. B. `news`, `gopher` und `ftp`. Wenn ein HTML-Dokument referenziert wird, beziehen sich Hyperlinks immer auf das gesamte Dokument. Um einen spezifischen Teilbereich zu verknüpfen, ist es notwendig, das Ziel zunächst zu markieren. Hierzu wird das Element `a` in Verbindung mit dem Attribut `name` im Zielknoten benutzt (z. B. `name="definition4"`). Mit Hilfe einer speziellen Syntax kann dieser Zielbereich referenziert werden (in diesem Beispiel etwa `.../artikel/index.html#definition4`), so dass der Browser die Darstellung des Zielknotens an eben dieser Stelle beginnt.[40] Hyperlinks können weiterhin einen Titel besitzen, der mit Hilfe des Attributs `title` angegeben und vom Browser z. B. als Pop-up-Fenster dargestellt werden kann, wenn der Mauszeiger auf den Hyperlinkanzeiger bewegt wird (Storrer, 2001c, S. 199, spricht von „Link-Etiketten“, vgl. auch Weinreich und Lamersdorf, 2000). In Verbindung mit Framesets wird über das Attribut `target` spezifiziert, auf welchen Frame sich der Link auswirken soll.

Es existieren zahlreiche Untersuchungen zu funktionalen Typisierungen von Hyperlinks.[41] In HTML steht nur ein minimales, nicht erweiterbares Typeninventar zur Verfügung, das nur selten eingesetzt wird. Über die Attribute `rel` (*forward link*) und `rev` (*reverse link*) können die Typen (der Standard spricht von *roles*) *Alternate*, *Stylesheet*, *Start*, *Next*, *Prev*, *Contents*, *Index*, *Glossary*, *Copyright*, *Chapter*, *Section*, *Subsection*, *Appendix*, *Help* und *Bookmark* angegeben werden. Die Attribute `rel` und `rev` werden verwendet, um einen Knoten in einem Hypertext zu verorten. Weiterhin können sie auch in `link` benutzt werden.[42] Dieses Element kann Hyperlinkinformationen über den aktuellen Knoten bereitstellen, es wird im Metadatenteil eines Dokuments hinterlegt. Derartige Verknüpfungen werden vom Browser in einer speziellen Menüzeile eingeblendet, um dem Anwender eine zusätzliche Navigationsebene zur Verfügung zu stellen (vgl. Abschnitt 3.5.5, insbesondere Fußnote 112, S. 109).

[38] Dieser Begriff ist im Vergleich zu alternativen Termini noch recht jung und geht auf Meyrowitz (1989) zurück.

[39] Der Einsatz von *Cascading Style Sheets* (CSS, Bos et al., 1998) erlaubt die Beeinflussung der Formatierung von Hyperlinkanzeigern, was sowohl vom Autor als auch vom Leser – durch entsprechende Browsereinstellungen – durchgeführt werden kann (die Kaskadierung bezieht sich auf den simultanen Einsatz mehrerer Stylesheets). Üblicherweise wird ein Hyperlink in blauer Schrift und unterstrichen dargestellt. Aus Sicht der Benutzerfreundlichkeit sollte die Voreinstellung, noch nicht besuchte Links blau darzustellen, nicht verändert werden, weil sich die Benutzer an diese Kennzeichnung gewöhnt haben, oder, wie es Nielsen (1999, S. 64) ausdrückt: „They just go: blue, boom, click.“ (vgl. Abschnitt 3.6.7).

[40] Diese Verknüpfungsart kann auch innerhalb eines Dokuments verwendet werden. Nielsen (1999, S. 115) spricht von „within-page links“ und rät von ihrer Verwendung ab, weil der Benutzer Gefahr läuft, nicht zu bemerken, dass er sich nach der Aktivierung solcher Hyperlinks noch im gleichen Knoten befindet.

[41] Trigg (1983) unterscheidet z. B. 75 unterschiedliche Hyperlinktypen. Kopak (1999) vergleicht verschiedene Ansätze zur Typisierung von Verknüpfungen (vgl. auch Abschnitt 3.5.5).

[42] Dieses Element wird vornehmlich zur Referenzierung von CSS-Stylesheets eingesetzt (vgl. Abschnitt A.4.2).

3.3.5 Navigation und Browsing

Bei der Interaktion mit dem Webbrowser ist – neben dem Vorgang der Rezeption von Inhalten – zwischen Browsing und Navigation zu differenzieren. Browsing bezieht sich auf das explorative Erkunden der Inhalte und Strukturen eines Hypertextes, es verläuft also nicht zielgerichtet.[43] Gerade dies trifft jedoch für das Navigieren in einem Hypertext zu, der Benutzer verfolgt also ein bestimmtes Ziel (Vora und Helander, 1997, S. 880).[44]

Das Navigieren in einem Hypertext bereitet in vielen Fällen Probleme (McKnight et al., 1991, S. 65 ff., Runkehl et al., 1998, S. 159 ff.). Die Produzenten müssen dafür Sorge tragen, dass sich Benutzer möglichst intuitiv in einem Hypertext bewegen können (vgl. die Abschnitte 3.4 und 3.5). Die verbreiteten Browser – *Mozilla*, *Firefox* und der *Internet Explorer*[45] – implementieren einen kanonischen Funktionsumfang, der aus Vor- und Zurück-Buttons, einer History- und einer Bookmark-Liste besteht. Die Knoten eines Hypertextes werden sukzessive besucht, so dass eine individuell konstruierte, lineare Sequenz entsteht. Es existieren verschiedene Optionen, in einer derartigen Sequenz mittels Vor- und Zurück-Knöpfen zu navigieren; in den verbreiteten Browsern wird ein chronologisches Stack-Modell verwendet (vgl. Nielsen, 1995b, S. 249 ff.). Die History-Liste umfasst eine chronologische Auflistung aller besuchten Webseiten, die mit Zeitstempeln versehen werden. *Mozilla* stellt diese Liste in einer hierarchischen Struktur dar, in der die Seitenaufrufe der jeweils letzten sechs Tage gezielt angesprungen werden können. Adressen, die der Anwender als wichtig oder interessant einschätzt und zu einem späteren Zeitpunkt erneut verwenden möchte, können in der Bookmark-Liste gespeichert und hierarchisch organisiert werden.

3.3.6 Zur Erstellung von Hypertexten für das World Wide Web

Die Erstellung von HTML-Dokumenten kann auf unterschiedlichen Prozessen beruhen. Die verwendete Software determiniert die Produktionsbedingungen und sie ist für die Untersuchung von Hypertextsorten von großer Relevanz, da sie sich auf die Strukturierung und Gestaltung von Dokumenten auswirken kann.

[43] Das Navigieren hat zunächst nichts mit dem Auffinden spezifischer Informationen zu tun. Die Metapher des Surfens im WWW bezeichnet eine Mischung aus Browsing und Navigation, wobei der Aspekt des Zeitvertreibs hinzu kommt (vgl. Sørensen, 1998, Spence, 1999, Storrer, 1999b, S. 4 f., sowie Bucher, 2001, S. 158, der von „flanieren“ spricht). Storrer (2001b, S. 96) bestätigt diese Ansicht: „Für das Herumstöbern [...] ohne feste Zielvorgabe hat sich der Ausdruck »surfen« eingebürgert, in Anlehnung an den amerikanischen Ausdruck »channel surfing«, das spaßgeleitete Herumzappen in Fernsehkanälen.“ (vgl. Fußnote 61, S. 87). Kuhlen (1997, S. 361) ist der Ansicht, dass Browsing heute „Surfen“ genannt wird.

[44] Kuhlen (1991, S. 126 ff.) unterscheidet „gerichtetes Browsing mit Mitnahmeeffekt“, „gerichtetes Browsing mit Serendipity-Effekt“, „ungerichtetes Browsing“ und „assoziatives Browsing“. Bucher (2001, S. 158 ff.) differenziert zwischen „Top-Down-“, „Ressort- oder Quer-“ und „Nabe-Speichen-Navigation“. Bucher zeigt, dass verschiedene Seitentypen (vgl. Indikator 51, S. 146) mit Navigationsstrategien korrelieren: „Ob eine Seite als Hyperlink-Seite genutzt wird, ist allerdings nicht davon abhängig, dass sie als solche gemeint ist, sondern nur davon, dass der Nutzer sie als solche definiert.“ (ebd., S. 161). Neben *browsing*, *navigation* und *scanning* ist die Terminologie von zahlreichen weiteren Metaphern geprägt (vgl. Abschnitt 3.5.4).

[45] Es existieren verschiedene Browser, die innovative Benutzeroberflächen und erweiterte Navigationsfunktionen anbieten, z. B. zur Visualisierung von Hyperlinkstrukturen. Sie werden jedoch im Alltag kaum eingesetzt, da sie meist nicht in der Lage sind, HTML-Dokumente, die aktuelle Auszeichnungstechnologien verwenden, korrekt darzustellen. Bezüglich der Usability bleiben die aktuellen Browser hinter ihren Möglichkeiten: „Actually, current web browsers do a lousy job of navigation support, but they do try.“ (Nielsen, 1999, S. 362).

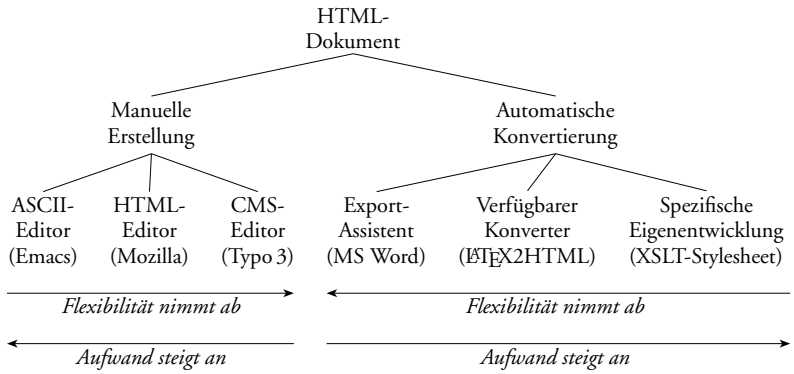


Abbildung 3.2: Möglichkeiten der Herstellung von HTML-Dokumenten

Es können die Kategorien der manuellen Erstellung von HTML-Dokumenten und der automatischen Konvertierung bereits vorhandener Dokumente unterschieden werden (vgl. Abbildung 3.2). Die manuelle Erstellung kann mit beliebigen Editoren erfolgen, die ASCII-Dateien verarbeiten, z. B. *Emacs* oder *Notepad* (vgl. Abschnitt A.4.2). HTML-Editoren erlauben die komfortable Auswahl von Elementen und Attributen aus Menüs. Einige Produkte umfassen vorgefertigtae Schablonen, die als Vorlagen dienen können (z. B. *FrontPage*).[46] HTML-Editoren können zusätzlich unterschieden werden in Werkzeuge, die die Bearbeitung der HTML-Struktur eines Dokuments unterstützen und WYSIWYG-Editoren (vgl. Abschnitt 3.6.4, insbesondere S. 133); Kombinationen sind möglich. Die Arbeit mit einem *Content Management System* erfolgt im Browser, d. h. es werden Textfelder, die im Template der zu erstellenden Seite vorgesehen sind, mit Inhalten gefüllt.[47] Besitzt der Autor mit einem ASCII-Editor jegliche Freiheiten und maximale Flexibilität bei der Gestaltung und Strukturierung eines Dokuments, existieren bei einem zentral administrierten CMS verschiedene Einschränkungen. Farbräume, Logos und das Layout sind meist vorgegeben und können nicht modifiziert werden. Zugleich ist der Produktionsaufwand bei der Anwendung eines ASCII-Editors sehr hoch, er nimmt jedoch schon mit einem HTML-Editor deutlich ab und mit einem CMS sind schließlich neue Dokumente in extrem kurzer Zeit erstellbar und werden meist automatisch auf dem Webserver veröffentlicht und in den Hypertext eingebunden.

Bei der manuellen Erstellung existieren entweder noch keine Inhalte oder es liegen Textbausteine vor, die in das zu erstellende HTML-Dokument integriert werden.[48] Bei der zweiten Kategorie, der automatischen Konvertierung, liegen hingegen vollständige Texte in Form

[46] Bei der Benutzung von ASCII-Editoren und rudimentären HTML-Editoren kann es zu verschiedensten Problemen kommen (z. B. fehlerhafte Verlinkung durch die Angabe nicht existenter Adressen in Hyperlinks).

[47] Zu dieser Kategorie können auch Weblogs (vgl. Abschnitt 4.6.7) und Gästebücher (vgl. Abschnitt 4.6.8) gezählt werden. Eckkrammer (2001, S. 61) merkt an, dass „der Schreiber mit den Strukturen der zugrundeliegenden Hypertexte nicht zwingend etwas zu tun hat, daher vielfach die hypertextuelle Struktur nur annehmen aber nicht modifizieren kann." Iorio und Vitali (2005, S. 3 f.) schlagen eine Klassifikation von vier Prozesskategorien der Herstellung von HTML-Dokumenten vor, orientieren sich jedoch an dem Grad der Überlappung hinsichtlich der Rollen des Produzenten und des Rezipienten, der für ein spezifisches „web authoring scenario" gilt, z. B. „total separation" (HTML-Editor, Konvertierungswerkzeuge) oder „overlap of roles" (Weblog).

[48] Zu den Produktionsbedingungen müssen weitere Faktoren gerechnet werden, diese können jedoch nur durch eine Befragung der Autoren ermittelt werden (vgl. Adamzik, 2000b, S. 110, und Kapitel 4).

spezifischer Dateien vor, die maschinell nach HTML überführt werden können.[49] Der Konvertierungsprozess basiert auf der physikalischen (z. B. bei WYSIWYG-Textverarbeitungen oder Textsatzsprachen wie troff oder TEX) oder der logischen (bei Auszeichnungssprachen wie LATEX oder XML-basierten Markup-Sprachen) Dokumentstruktur (vgl. Rada, 1991, S. 9 ff.). Ein häufig gewählter Weg betrifft den Einsatz von Export-Assistenten. Die Textverarbeitung *Word* erlaubt z. B. die Speicherung eines Dokuments als HTML-Datei, wobei nur wenige Möglichkeiten existieren, den Konvertierungsprozess zu beeinflussen. Kuhlen (1991, S. 162 ff.) differenziert verschiedene Strategien, nach denen Texte in Hypertexte überführt werden können. Der Export einer Textverarbeitungsdatei nach HTML entspricht der ersten „Konversionsform", der „einfachen Übertragung": „Die einfachste Form der Konversion besteht in der 1 : 1 Übertragung eines Textes in eine Datei eines Hypertextsystems. Damit ist natürlich noch keine Hypertextbasis aufgebaut, sondern nur eine gedruckte Version durch eine On-line-Version ersetzt." (ebd., S. 163). Beim Einsatz spezialisierter Werkzeuge ist der Spielraum größer, z. B. bietet *LATEX2HTML* vielfältige Optionen, mit denen die Verlinkung und Aufteilung eines LATEX-Dokuments in eine oder mehrere HTML-Dateien beeinflusst werden kann. Dies entspricht dem zweiten von Kuhlen (1991, S. 163) beschriebenen Konversionstyp der „Segmentierung und Relationierung über formale Texteigenschaften", bei dem „häufig die an formalen Eigenschaften erkennbaren textuellen Makrostrukturen, wie Absätze, Unterabschnitte, etc., ausgenutzt" werden (ebd.). Die dritte Möglichkeit betrifft die Implementierung eines Werkzeugs, um vorhandene Daten (z. B. eine XML-Datei oder eine ASCII-Datei, die durch Kommata getrennte Datensätze enthält) nach HTML zu überführen. Dies erfolgt im Falle von XML-Instanzen meist mit XSLT-Stylesheets, im Falle anderer Daten mit Skriptsprachen wie *Perl*, *awk* oder *sed*. Je nach Komplexität der Quelldaten und den Anforderungen an das zu erzeugende HTML-Dokument kann sich eine derartige Eigenimplementierung als sehr aufwändig erweisen.[50] Zugleich bietet diese dritte Option die größte Flexibilität, denn die weiteren von Kuhlen genannten Konversionformen können – jedenfalls in Teilen – durchaus mit speziellen Werkzeugen und auf der Basis einer entsprechend annotier-

[49] In der Literatur ist von der Konversion verfügbarer Texte die Rede (vgl. Abschnitt 3.6.2). Dieser aufwändige Prozess (vgl. Ford, 1995, und Nielsen, 1995b, S. 329 ff.) ist auf prototypische Hypertexte ausgerichtet und kann bestenfalls semiautomatisch erfolgen: „the conversion of a linear text to a well-written hypertext can involve an extensive amount of modification [...] in order for the text to be more appropriate in a hypertext format." (Foltz, 1996, S. 130). Der Terminus Konvertierung deutet einen vollautomatischen Transformationsprozess an, und diese Form scheint im WWW die mit Abstand häufigste zu sein (vgl. etwa Wagner, 1998).

Sandbothe (1997, S. 74) verneint die Frage, ob das WWW letzten Endes das Buch verdrängen wird, denn „die meisten Texte [...] im Netz [...] sind keine Hypertexte, sondern ganz normale Aufsätze und Bücher, die in HTML-Code konvertiert [...] wurden. Derzeit dient das [WWW] in erster Linie dazu, Bücher und Aufsätze besser und schneller zugänglich zu machen. [... Das] eigentliche, dem neuen Medium angemessene Schreiben und Denken im Hypertextstil [stellt] noch eine anspruchsvolle Zukunftsaufgabe dar. Schulen und Universitäten, Lehrer, Wissenschaftler und Autoren müssen darauf erst noch vorbereitet werden." (vgl. Kendall, 2000). Storrer (2001b, S. 109) vertritt bezüglich des Umgangs mit vorhandenen Texten die Ansicht: „Wer mit der Konversion keine Mehrwerte gegenüber dem gedruckten Text schaffen kann, sollte [...] das WWW in seiner Funktion als Datenverteiler nutzen, d. h. das Dokument als Ganzes zum Ausdruck anbieten." Pfammatter (1998, S. 72 f.) fasst beide Aspekte zusammen: „Der [...] Versuch in Hypertext gedruckte Texte zu imitieren ist nicht sinnvoll. Die Diskussion über die Konkurrenzsituation von Buch und Multimedia ist irrelevant."

[50] Das unmittelbare Publizieren von XML-Dokumentinstanzen, die *on the fly* vom Browser durch referenzierte CSS- oder XSLT-Stylesheets dargestellt werden können, stellt einen Spezialfall dieser Kategorie dar, der in der Untersuchungsdomäne als Ausnahme zu betrachten ist.

ten Textgrundlage realisiert werden. Hierbei handelt es sich um „Segmentierung und Relationierung nach Kohärenzkriterien“ und „intertextuelle Konversion“ (Kuhlen, 1991, S. 164 f.). Erst die Segmentierung nach Kohärenzkriterien „führt zu einer als Netzwerk strukturierten Hypertextbasis“ (Storrer, 1997, S. 125), aber sie „erfordert einen hohen Aufwand für Analyse und Reorganisation der Ausgangstexte, der beim gegenwärtigen Stand wissensbasierter Systeme allenfalls bei standardisierten und einfach strukturierten Textsorten automatisierbar ist.“ (ebd., S. 126). Derartige Implementierungen können nur von Experten vorgenommen werden, die mit den beteiligten Programmierkonzepten vertraut sind, weshalb auf diese Weise erzeugte HTML-Dokumente als Randerscheinung einzustufen sind. Automatisch erzeugte Dokumente können mit ASCII- oder HTML-Editoren nachbearbeitet werden. Viele Werkzeuge hinterlassen einen Fingerabdruck innerhalb des `meta`-Elements (vgl. Abschnitt A.4.7).

3.3.7 Hypertexte, Hypertextnetze und elektronische Texte

Eine im Hinblick auf die Erstellung von HTML-Dokumenten wichtige Unterscheidung trifft Storrer (2003, vgl. 1999a, S. 38 f., 2000b, 2001c, S. 180 f., 2004b, S. 29), die Hypertexte von elektronischen Texten abgrenzt:

> *E-Texte* sind linear organisierte Texte, die [...] durch Links in ein Hypernetz eingebunden sind. Es handelt sich dabei häufig um digitale Kopien gedruckter Texte, [...] die als PDF- oder als HTML-Dateien im WWW publiziert sind. E-Texte können zwar Links zu anderen Dokumenten im WWW enthalten; es fehlt ihnen aber die modulare Aufbereitung von Inhalten im Hinblick auf selektive Rezeptionsformen, die für Hypertextdokumente charakteristisch ist. (Storrer, 2003, S. 284)

An anderer Stelle bemerkt Storrer (2000b, S. 230), dass es sich bei E-Texten häufig um „Parallel- oder Vorversionen von Print-Publikationen“ handelt und dass ihnen „die modulare Aufbereitung von Inhalten im Hinblick auf selektive Rezeptionsformen [fehlt]“ (Storrer, 2003, S. 284). Wenn es sich um „linear organisierte Texte“ (ebd.) handelt, kann sich diese Eigenschaft auf die Ebene des Knotens oder die des Hypertextes beziehen: Ein E-Text ist also entweder in einem Knoten enthalten oder auf mehrere linear organisierte Knoten verteilt (vgl. Abschnitt 3.5.1).[51] Er kann sowohl bei der automatischen als auch bei der manuellen Erstellung entstehen. Viele HTML-basierte E-Texte dürften auf rudimentär durchgeführte Konvertierungen durch Export-Assistenten zurückzuführen sein (vgl. Abbildung 3.2).

Storrer (2000b, S. 236) führt eine Differenzierung zwischen Hypertext und Hypertextnetz ein, die „den Vergleich von Text und Hypertext im Hinblick auf Fragen der Themenstrukturierung, der Kohärenzplanung auf Produktionsseite bzw. der Kohäsionsbildung auf Rezeptionsseite“ erleichtert und eine Grundlage zur Beantwortung der „Frage nach der Konstitution

[51] E-Texte können nach Storrer (2001c, S. 180 f.) Teil umfassenderer (journalistischer) Hypertexte sein: „E-Texte können als Module in Hypertexte eingebaut werden, die in gedruckter Form rezipiert werden, als Lektüre zwischendrin oder im Anschluss an eine WWW-Sitzung. Der Wechsel zwischen Print- und Online-Lektüre ist kein Bruch mit dem Hypertext-Prinzip, sondern erhöht dessen Mehrwertpotenzial.“ Handler (1997, S. 95) unterscheidet in diesem Zusammenhang (in Anlehnung an Storrer, 1997) zwischen „parasitärer“ und „originärer“ Verwendung: Bei der parasitären Nutzung entspricht das WWW „in etwa einem über Distanz funktionierenden Kopiergerät“, d. h. ein bereits vorliegender Text wird „1:1 in die elektronische Distribution“ übernommen. Die originäre Verwendung zeichnet sich dadurch aus, dass das Medium einen Einfluss auf die Produktion bzw. Adaption eines Textes besitzt, es dominiert also die Nutzungsfunktion.

und Delimitation von Hypertexten“ (ebd.) sein kann (vgl. Heinemann und Heinemann, 2002, S. 107). Storrers Definition von „Hypertext“ ist allgemein ausgelegt und wird nachfolgend für das WWW spezifiziert.[52] Hypertextbasis wird in der vorliegenden Arbeit synonym zu Hypertext (im Sinne von Hypertextexemplar) und Website verwendet (vgl. Kuhlen, 1991, S. 17). Der Begriff Website bezeichnet oftmals eine abstraktere funktionale Ganzheit, die mehrere Hypertexte umfasst (vgl. Nielsen, 1999, S. 222 ff.).[53] Ein Hypertext ist demnach

> eine von einem Hypertextsystem verwaltete Menge von [HTML-Dokumenten], die als Resultate von Herstellungshandlungen vor dem Hintergrund einer bestimmten thematischen Gesamtvorstellung und zu einem bestimmten kommunikativen Zweck produziert werden. Textfunktion und Thema fungieren als übergeordnete Einordnungsinstanz und liefern den kontextuellen Rahmen für das Verständnis der einzelnen [HTML-Dokumente]. Sie konstituieren den Hypertext als Ganzheit, deren Bestandteile durch „interne“ Links zusammengehalten werden und durch „externe“ Links [...] in einem [...] Hypertextnetz verknüpft sein können. (Storrer, 2000b, S. 236)

Die von Storrer gelieferte Definition von Hypertextnetz (bzw. Hypernetz, vgl. Storrer, 2003) knüpft mittelbar an das *World Wide Web* an:

> Hypertextnetze verknüpfen Hypertexte [...] durch Hyperlinks. Das WWW kann in diesem Sinne als [...] Hypertextnetzwerk angesehen werden, das aus [...] Teilnetzen besteht, sich in steter Veränderung befindet und in seiner Gesamtheit von niemandem überblickt werden kann. (Storrer, 2000b, S. 236)

Jakobs (2003) merkt an, dass das aus Teilnetzen bestehende WWW als Ganzheit aus der Textsortendiskussion ausgeschlossen werden kann, weil es nicht kommunikativ bestimmt ist.

Storrer (2000b, S. 236) unterscheidet geschlossene von offenen Hypertexten. Weitere Typisierungen können „nach technischen, strukturellen und funktionalen Aspekte[n]“ vorgenommen werden. Geschlossene Hypertexte verfügen über eine „feste Anzahl“ von HTML-Dokumenten: „Auch wenn sie durch externe Links in größere Hypertextnetze, z. B. ins WWW eingebunden sind, sind sie konzipiert als statische Produkte mit stabiler Struktur, auf die spätere Produkte ohne Risiko Bezug nehmen könnten.“ (ebd.). Offene Hypertexte sind eher dem prototypischen Hypertextkonzept zuzurechnen: Sie besitzen keine stabile Struktur, denn Autoren und Benutzer können weitere HTML-Dokumente hinzufügen, diese aktualisieren, ein offener Hypertext kann also „das zugrunde liegende Thema über eine

[52] Storrer (1999a, S. 35) verwendet „in Anlehnung an die Terminologie zum Textdesign“ (vgl. Abschnitt 3.5.6) statt „HTML-Dokument“ den Begriff „Modul“ (in früheren Arbeiten „Hypertexteinheit“ bzw. „Einheit“, vgl. Storrer, 1997) und im Kontext von Online-Zeitungen den Terminus „Info-Modul“ (Storrer, 2001c, S. 180). Bei einer abstrakten Sichtweise auf Hypertext erscheint der Modul-Begriff durchaus plausibel, in der vorliegenden Arbeit wird jedoch die Ansicht vertreten, dass HTML-Dokumente *intern* aus Modulen zusammengesetzt sind (vgl. Abschnitt 5.6). Whitehead (2000) untersucht die Bezeichnungen des Knotenkonzepts in 36 Hypertextsystemen: Vornehmlich werden „node“, „document“, „object“ und „component“ verwendet; der Begriff „module“ wird in *keinem* System benutzt. Whitehead (2000, S. 14) selbst schlägt „work“ vor.

[53] Storrer (1999a, S. 39) fügt hinzu: „Ein als »Site« bezeichnetes Teilnetz ist [...] durch technische und institutionelle Rahmenbedingungen begrenzt: Eine Site wird von [...] Web-Redakteuren kontrolliert“. In einer späteren Arbeit wird erläutert, dass „sich in offenen Hypertexten Zahl und Inhalt der Module stets ändern können, so sind [...] »Sites« meist [...] auf eine sehr langfristige Existenz hin ausgelegt. Diese Ganzheiten bzw. Teile davon können jederzeit [...] archiviert werden, sie werden dadurch zu zitierbaren Produkten mit langfristiger Überlieferungsqualität“ (Storrer, 2000b, S. 238, vgl. auch Nielsen und Tahir, 2002, S. 11).

unbestimmte Zeitspanne hinweg im Gespräch“ halten (ebd.). „Über das Thema kann sich eine Diskussion zwischen Autoren und Nutzern entspinnen, deren Beiträge wiederum wechselseitig kommentiert und diskutiert werden. Es entsteht eine Ganzheit, die nicht einmal durchlaufen, sondern regelmäßig besucht wird“ (ebd., S. 236 f.). Gerade die Kombination geschlossener und offener Hypertexte ergibt eine große Flexibilität (ebd., S. 238), z. B. können abgeschlossene wissenschaftliche Veröffentlichungen mit in offener Weise gepflegten Anmerkungen, sukzessive erscheinenden Rezensionen und Leserbriefen angereichert werden.

3.4 Kritische Anmerkungen

In einem Plenarvortrag auf der Konferenz *Hypertext '89* hinterfragte Meyrowitz (1989) angesichts des rapide zunehmenden *Hypes* um Hypertexttechnologien und der explosionsartigen Zunahme von Publikationen zu diesem Thema, ob Hypertext vielleicht auch in der Lage sei, den Cholesterinspiegel zu senken.[54] Retrospektiv kann konstatiert werden, dass Hypertext die hochgesteckten Erwartungen nicht erfüllen konnte (vgl. Pang, 1998). Rieffel (1999, S. 1) charakterisiert die Ursachen treffend: „The claims that the chunked and linked style of hypertext is new and revolutionary, and that it is the computer that makes non-sequential text possible, or at least removes severe limits imposed by paper, are both patently false.“ (vgl. auch Dalgaard, 2001, S. 176). Die Euphorie um Hypertext hat in den vergangenen Jahren deutlich abgenommen. Mittlerweile ist eine Konzentration auf das Hypertextsystem WWW festzustellen – insbesondere auf die Vision des *Semantic Web* (vgl. Abschnitt 13.2).

Das Medium Hypertext weist in der Ausprägung des WWW gegenüber traditionellen Medien (Büchern, Zeitschriften etc.) durchaus zahlreiche Vorteile auf, z. B. das nahezu alle Themen- und Interessengebiete umfassende Informationsangebot und einen sehr effektiven Zugriff mittels Suchmechanismen. Diese Vorteile sind jedoch mit medieninhärenten Nachteilen verbunden, die sich auf konzeptionelle und technologische Aspekte beziehen.

3.4.1 Die These der kognitiven Plausibilität

Gerade in kaum reflektierenden Arbeiten wird oftmals die schon von Bush (1945a)[55] diskutierte Möglichkeit der nichtlinearen, assoziativen Verknüpfung einzelner Texte und Textteile mit den Verarbeitungsmechanismen des Gehirns gleichgesetzt (Belege liefert Freisler, 1994, S. 43), was als zentraler Vorteil des Hypertextkonzepts angesehen wird. Dillon (1996, S. 28) bezeichnet diese Einstellung als „naturalistic associationism“.[56] Freisler (1994, S. 42) geht ausführlich auf die „These der kognitiven Plausibilität“ von Hypertext ein, die auf zwei

[54] Eine Antwort blieb Meyrowitz (1989, S. 317) nicht schuldig: „There are a couple of answers. One is that sources tell me that yes, in fact, HyperCard Version 2.0 will have a cholesterol reduction feature. I'm not sure how reliable those sources are. Another way to answer the question is to say that just as oat bran is important for reducing cholesterol, so is hypermedia important for reducing information clogging and information glut.“

[55] Bush wollte „a mechanical analog to the brain“ konstruieren: „Bush's intention was to use machines to innovate by improving on parts of the imperfect biological process“ (Nyce und Kahn, 1991b, S. 63).

[56] Im Kontext des Lernens mit digitalen Medien wird Hypertext häufig als natürliche Art des Assoziierens betrachtet, was ihn gegenüber linearen Texten zwangsläufig für das Lernen prädestiniere (vgl. auch Tergan, 1997, S. 130). Perfetti (1996, S. 157) stellt diese Auffassung polarisierend dar: „hypertext research [...] tried to persuade that hyper is good and linear is not.“ (vgl. ausführlich hierzu Dillon, 1996).

Annahmen fußt: „Die Hypertextstruktur *ist* quasi die externalisierte Struktur des semantischen Netzes aus dem Kopf des Schreibers in elektronifizierter Gestalt. Diese Art der Wissens(re)präsentation verbessert die Kohärenzbildung und damit die Rezeption schlechthin." (ebd.).[57] Freisler unterscheidet zwischen natürlichen und künstlichen Netzen zur Repräsentation von Wissen und kommt zu dem Schluss: „In bezug auf die *nur* natürlichen Netzen zuschreibbaren Eigenschaften des *Metawissens*, des *Vergessens*, des *Kontextes* und der *Relevanz* ist es jedoch nun völlig sinnlos und unproduktiv, derartig simple Technologien mit solch komplexen menschlichen Fähigkeiten [...] zu vergleichen." (ebd., S. 45; vgl. auch McKnight et al., 1991, S. 94 ff., Kuhlen, 1991, S. 55 f., und Storrer, 2004b, S. 38). Zudem geht Freisler (1994, S. 42) davon aus, dass die fragmentartige Struktur eines Hypertextes einen Lerner bei der Konstruktion einer kohärenten Repräsentation eher behindert als unterstützt: „Kritiker der Hypertexttechnologie weisen deshalb ganz richtig darauf hin, daß überall dort, wo es den Hypertextschreibern auf eine argumentative Stringenz ankomme, diese nicht anders könnten als auf erprobte und bewährte Mittel der Linearisierung zurückzugreifen."[58] Kuhlen (1997, S. 99 ff.) nähert sich der These der kognitiven Plausibilität durch eine Betrachtung der Möglichkeiten, Notizen und spontane Gedankengänge nachhaltig zu speichern, z. B. mit handschriftlichen Kommentaren in gedruckten Texten, Spickzetteln, Karteikarten, Notizblöcken und Diktiergeräten. Er stellt fest, dass das „für Hypertext konstitutive Verknüpfungskonzept [...] der bislang konsequenteste Versuch [ist], assoziativem Denken Rechnung zu tragen. Wenn [...] auch noch andere Denkformen und andere Zugriffsformen auf Wissen unterstützt werden, z. B. über gezielte, selektive Retrievalfunktionen, dann kann das Hypertext nur umso mehr befördern." (ebd., S. 101). Die These der kognitiven Plausibilität von Hypertext gehört zwar „in den Bereich der Fiktion" (Weingarten, 1997b, S. 216), Hypertexte können das assoziative Denken jedoch durchaus unterstützen. Allerdings können zwei zentrale Probleme Beeinträchtigungen hervorrufen. Diese werden nachfolgend thematisiert.

3.4.2 Orientierungslosigkeit und kognitiver Ballast

Conklin (1987, S. 38) diskutiert unter den Schlagworten „cognitive overhead" und „disorientation" zwei Probleme, „that may in fact ultimately limit the usefulness of hypertext".

Die Orientierungslosigkeit resultiert aus der größeren Freiheit, die Benutzer von Hypertexten haben. Gerade aufgrund dieser Flexibilität verlieren sich Leser häufig im Knotendickicht und wissen nicht, wie sie einen gesuchten Knoten finden oder zu einem bestimmten Knoten zurückkehren können – die Navigation kann „in ein Chaos münden" (Kuhlen, 1991, S. 123). Solche Probleme existieren auch in gedruckten Texten, diese enthalten jedoch häufig ein Inhaltsverzeichnis und einen Index, die das Auffinden von Informationen erleichtern. Die greifbare Materialität eines gedruckten Textes vermittelt dem Leser indirekt seine Grenzen,

[57] Freisler hat eine sehr differenzierte Ansicht kognitiver Plausibilität: „›Kognitiv plausibel‹ ist ein informationsvermittelndes System dann, wenn zwischen diesem Weg von der Repräsentation im System zur Repräsentation im Rezipienten möglichst wenig Umformungsprozesse notwendig sind und der Leser zu einer besseren Informationsaufnahme gelangt, *weil* weniger Umformungsprozesse zwischen der Wissensrepräsentation des Systems und seiner eigenen notwendig sind." (Freisler, 1994, S. 43).

[58] Vora und Helander (1997, S. 894) gehen auf Studien zum Vergleich von Text und Hypertext ein: Lesern wurden Fragen zu Artikeln gestellt, die zuvor entweder als Text oder als Hypertext präsentiert wurden; die Leser der Printversionen konnten die Fragen schneller und präziser beantworten (vgl. auch Schweiger, 1996).

was bei einem Hypertext nur durch (navigierbare) Visualisierungen erreichbar ist.[59] Zur Vermeidung des Problems werden technische Lösungen eingesetzt, z. B. Suchfunktionen oder Navigationshilfen, die die aktuelle Position innerhalb des Hypertextes explizieren.

Das zweite Problem – der kognitive Ballast[60] – bezieht sich auf die Produzenten und Rezipienten von Hypertexten. Die Freiheit der beliebigen Verknüpfung setzt voraus, dass der Autor intuitive und sinnvoll erscheinende Hyperlinks antizipieren kann: „it is difficult to become accustomed to the additional mental overhead required to create, name, and keep track of links." (Conklin, 1987, S. 40). Auf der Rezipientenseite manifestiert sich das Problem in entgegengesetzter Weise: Der Autor präsentiert zahlreiche Wahlmöglichkeiten, und nur der Leser selbst kann entscheiden, welcher Hyperlink verfolgt werden soll: „These choices engender a certain overhead of metalevel decision making" (ebd.; vgl. Naumann et al., 2003). Der Leser ist förmlich gezwungen, kurz innezuhalten, sobald ein Hyperlink die Rezeption unterbricht („cognitive loading", ebd.), was ihn wiederum von der Lektüre ablenkt, um das Aktivieren des Links in Erwägung zu ziehen. Er „kann in seiner Aufnahme- und Verarbeitungsfähigkeit ganz schnell überfordert werden und zum reizüberfütterten Zapper degenerieren" (Schmitz, 2001, S. 209), da zirkulär zwischen der Rezeption und der Navigation gewechselt wird (vgl. Askehave und Nielsen, 2005).[61] Zur Minderung des Problems werden z. B. deskriptive Hyperlinkanzeiger oder erläuternde Pop-up-Fenster eingesetzt.

3.4.3 Probleme im Umgang mit dem World Wide Web

Viele der im WWW existenten Probleme wurden bereits diskutiert. Zunächst ist das WWW nur eine sehr rudimentäre Ausprägung der bislang in alternativen Systemen oder Forschungsplattformen implementierten Konzepte (vgl. Abschnitt 3.3). Die beiden von Conklin auf einer systemneutralen Ebene diskutierten Probleme sind gerade im WWW offensichtlich, was Schmitz (2001, S. 215) treffend charakterisiert: „Angesichts steigender Informationsmengen [...] bewegen sich Anbieter und Leser ständig an der Grenze zu Chaos, Heterogenität, Auseinanderfall, Desorientierung und Verwirrung." Ein weiteres Indiz für die Orientierungslosigkeit vieler Benutzer wird als *serendipity*-Effekt bezeichnet: Der Anwender sucht Informationen zu einem spezifischen Thema, findet dabei jedoch andere Informationen, die ihn

[59] Hinzu kommt, dass der Umgang mit digitalen Dokumenten nicht unmittelbar mit Gebrauchsspuren wie Unterstreichungen, Eselsohren oder handschriftlichen Anmerkungen verbunden ist, die für nachfolgende Leser durchaus hilfreich sein könnten. Derartige Spuren werden in Hypertextsystemen z. B. durch Diskussionsforen oder verteilte Annotationen emuliert. Wexelblat (1999) demonstriert, dass eine maschinelle Auswertung der Zugriffsprotokolle von Webservern als Basis einer „history-based navigation" dienen kann.

[60] Diese Übersetzung von „cognitive overhead" liefert Kuhlen (1991, S. viii), der „disorientation" als „Orientierungsverlust" wiedergibt und sich an die „getting lost in space"-Metapher anlehnt (Conklin, 1987, S. 38).

[61] Wenz (2000, S. 23) ist der Ansicht, dass „Fernseherfahrung, das Zappen durch die Fernsehkanäle" beim Erlernen der „Regeln des Mediums [Hypertext]" hilfreich ist. Bucher (2001, S. 141) äußert sich kritisch: „Die Rezeption von Online-Angeboten kann [...] nicht mit dem Begriff des Zappens [...] erklärt werden: Die Stelle, an der ein Zapper auf das nächste Programm umschaltet, gibt ihm keine Anhaltspunkte, was ihn dort erwartet. Er verfährt gleichsam blind. Deshalb ist das Zappen hinsichtlich des neuen Angebots unmotiviert. Es dominiert der diffuse Wunsch, etwas anderes zu finden. Die Nutzung einer Absprungstelle in einem [Hypertext] zeichnet sich aber dadurch aus, dass begründete *Fortsetzungserwartungen* gebildet werden können." (vgl. Bucher, 2004, S. 141 ff.). Fleming (1998, S. 1) fügt hinzu: „You may be a couch potato while watching TV, which requires [...] the periodic twitching of your finger on the remote, but passive viewing on the Web is a much trickier proposition." (vgl. Schmitz, 1996, Nielsen, 1999, S. 365 ff., und Fußnote 43, S. 80).

so sehr „beschlagnahmen“ (Kuhlen, 1991, S. 129), dass das ursprüngliche Ziel vergessen wird.[62] Technische Gründe sind für weitere Probleme verantwortlich: Die Anforderung entfernter Dokumente kann in Verzögerungen resultieren (z. B. bedingt durch eine Überlastung des Netzwerks oder des entfernten Rechners), Webserver können aus verschiedenen Gründen nicht erreichbar sein, Dokumente können Dateitypen enthalten, die der Browser nicht darstellen kann, und da das WWW einen dezentralen Aufbau besitzt, existiert keinerlei Qualitätskontrolle, d. h. jede Person kann zu jedem nur denkbaren Thema beliebige Meinungen äußern oder Produkte der Fantasie als Fakten präsentieren.[63] Nach Vora und Helander (1997, S. 902 f.) „bevölkern tausende mittelmäßige Webseiten das Netz“. Die konstatierte Mediokrität bezieht sich auf eine Vielzahl von Aspekten: „poor, outdated, and incomplete content [...]; unnecessary and gratuitous use of graphics; [...] poor navigation support; dangling links; and so forth. [...] The links provided on the Web are often cumbersome, irrelevant, or trivial.“ (ebd., vgl. auch Raskin, 1987, S. 329). Auch bezüglich des Inhalts sind bei sehr vielen Webseiten Defizite zu verzeichnen, obwohl sie in puncto Webdesign häufig auf dem neuesten Stand sind, „aber auch ein mit allen Finessen der Web-Sprache HTML aufgeputztes Nichts bleibt ein Nichts, und seine Nichtigkeit wird nur um so deutlicher.“ (Zimmer, 1997).

Ein weiteres Problem betrifft die Informationsrecherche, die entweder mit generellen oder spezialisierten Suchmaschinenen oder mit Webkatalogen durchgeführt wird. Da Suchmaschinen nur ein Bruchteil des WWW indexieren und dieser Prozess unter häufig unklaren Aspekten[64] erfolgt, kann der Anwender nach einer Recherche keine Gewissheit darüber haben, ob wirklich alle relevanten Dokumente gefunden werden konnten (vgl. Abschnitt 1.1).

[62] Kuhlen (1991, S. 57) verwendet den Begriff in neutraler Weise und spricht von „Überraschungseffekten“. Mit *serendipity* war ursprünglich ein positiver Mitnahmeeffekt gemeint: Er „mag zwar auch nützlich und anregend sein, entspricht aber kaum einem effizienten Informationsverhalten, zumal nicht in professionellen Umgebungen – vielleicht akzeptabel oder erwünscht als Infotainment bei der globalen Internet-WWW-Navigation.“ (Kuhlen, 1997, S. 361). Diese Auffassung teilen Haas und Grams (2000, S. 189), denn „serendipity is one of the joys of browsing. This type of reading is more for enjoyment than for finding a specific piece of information.“ Nach Ansicht von Sager (1997, S. 119) besteht „die Gefahr, daß dieses Phänomen Verwirrung, Aggression oder Frustration im Rezipienten auslöst und damit jede Form von Erkenntnis unmöglich macht.“

[63] Die Möglichkeit der freien Meinungsäußerung stellt einen immensen Vorteil dar, der sich jedoch z. B. innerhalb der Wissenschaftskommunikation als eklatanter Nachteil erweisen kann: „Anybody can put up a site, and increasingly, anybody does. As a result, users don't quite know what to make of information retrieved from the Web. It can be the deep truth, or it can be the ramblings of a nut.“ (Nielsen, 1999, S. 92, vgl. auch Adamzik, 2004, S. 91). Walker (2000, S. 112 ff.) gibt ein Beispiel und geht auf den Faktor Glaubwürdigkeit bei der Rezeption persönlicher Homepages ein. Graham und Metaxas (2003) zeigen, dass Studierende für Recherchen sehr häufig Suchmaschinen einsetzen, wobei von Regierungen oder Firmen zu Werbezwecken veröffentlichte Informationen oftmals nicht hinterfragt und für bare Münze genommen werden („Students are also not consistently able to differentiate between advertising and fact.“, ebd., S. 75; vgl. auch Grimes und Boening, 2001, und Perry et al., 1998). Bezüglich der Gestaltung kommerzieller Websites rät Nielsen, mit der Vermittlung von Glaubwürdigkeit bereits auf der Ebene des Webdesigns anzusetzen (vgl. Abschnitt 3.5.3); dies wird von empirischen Studien bestätigt (vgl. Fogg et al., 2001, 2003, und Liang und Lai, 2002). Nielsen und Tahir (2002, S. 46 f.) sind der Ansicht, dass explizite Angaben über die Datenschutzrichtlinien sowie den Betreiber einer Website zusätzliche Mittel sind, Glaubwürdigkeit und Akzeptanz zu erreichen (vgl. Abschnitt 4.6.2).

[64] Schlüsselbegriffe in `meta`-Elementen (vgl. z. B. Siegel, 1999b, S. 176), die Werbetreibende vielfach zur Manipulation von Suchmaschinenergebnissen missbrauchen (Zhang und Dimitroff, 2005a,b), werden z. B. aus eben diesem Grund mittlerweile von vielen Suchmaschinen ignoriert.

3.5 Beschreibungsebenen von Hypertexten

Im Folgenden werden Beschreibungsebenen dargestellt, die – neben den in Kapitel 2 diskutierten generischen Textdimensionen – der linguistischen Analyse und Beschreibung von Hypertexten dienen. Einige dieser Ebenen sind auch auf traditionelle Textsorten anwendbar, werden jedoch von der Textlinguistik eher vernachlässigt. Andere Ebenen sind zu modifizieren, um sie an die medialen Spezifika anzupassen. Nachdem Abschnitt 3.3 in Bezug auf Hypertext primär technologische Konzepte eingeführt hat, wird nachfolgend eine linguistische Verortung vorgenommen, ohne jedoch genauer auf Hypertextsorten einzugehen. Dieser Aspekt wird ausführlich in Abschnitt 3.6 und den sich anschließenden Kapiteln thematisiert.

3.5.1 Nichtlinearität und Sequenziertheitsgrade

Storrer (2000b, S. 239) geht auf das Problem der Nichtlinearität für textlinguistische Untersuchungen ein, da die von Nelson prognostizierte Befreiung der Autoren von der „Bürde der Sequenzierung“ gerade „bei vielen Textwissenschaftlern Skepsis und Befremden“ (ebd.) hervorrufe (vgl. Abschnitt 2.1). Auf der Knotenebene liegt zwar eine lineare Ausrichtung vor, auf der Ebene der Hypertextbasis kann jedoch auf Linearität verzichtet werden, um dem Leser einen multilinearen Zugang zu ermöglichen. Storrer (2000b, S. 239) wählt mit dem *Krimi* und dem *Witz* zwei „drastische“ (ebd.) Beispiele, um zu zeigen, dass nicht alle Textsorten nichtlinear aufbereitet werden können. Es existieren demnach Textsorten, die für eine Umsetzung als multilinearer Hypertext geeigneter sind als andere Textsorten. In Anlehnung an Koch und Oesterreicher (1994, vgl. Abschnitt 2.2.7) führt Storrer eine Differenzierung zwischen konzeptioneller Linearität bzw. Nichtlinearität und linearen bzw. nichtlinearen Medien ein. Bei der Verwendung linearer Medien kann die lineare Rezeption „nicht oder nur schwer unterlaufen“ werden (Storrer, 2000b, S. 240), z. B. bei Filmrollen, Ton- und Videobändern (vgl. Sager, 2000, S. 599). Nichtlineare Medien können Daten in unterschiedlichen Sequenzen wiedergeben. Dieses Kriterium erfüllt auch das Buch, da sich der Leser relativ frei zwischen unterschiedlichen Teilen und Kapiteln bewegen kann (vgl. Rieffel, 1999).

Die konzeptionelle Linearität und Nichtlinearität sind als Eigenschaften anzusehen, die der Produzent bewusst wählt. Ghaoui et al. (1990, S. 112) sprechen von einem „degree of linearity“, der je nach Textsorte unterschiedlich sein kann.[65] Storrer führt den Begriff der Sequenziertheit von Texten ein, der synonym mit konzeptionelle Linearität verwendet und auf drei Strukturierungsformen bezogen wird: Monosequenziertheit, Mehrfachsequenziertheit und Unsequenziertheit. Diese Differenzierung verdeutlicht, dass der Unterschied zwischen Hypertexten und traditionellen Texten gerade nicht in der medialen, sondern in der konzeptionellen Linearität bzw. Nichtlinearität besteht.[66] Für monosequenzierte Texte ist

[65] Siehe auch Dahlström (2002, S. 150): „The hypertextual structuring [...] can take on different forms, in varying degrees of complexity and linearity, ranging from strictly locked sequence, via axial to hierarchical structures, and finally to completely retinal Web forms, where every single fragment is linked to all the others.“ Jonassen (1989, S. 50–54) unterscheidet „unstructured“, „structured“ und „hierarchical hypertext“.

[66] Es wurden zahlreiche Grundtypen der Makrostrukturierung von Hypertexten vorgelegt, die sich meist einer graphentheoretischen Terminologie bedienen (vgl. Gillenson et al., 2000, Schlobinski, 2000b, und Ziegler, 2004). Sager (2000, S. 593 f.) konzentriert sich auf hypermediale CD ROMs und unterscheidet z. B. Kette (mit den Subtypen einfache, erweiterte und verzweigte Kette), Kreis bzw. Ring („selbstablaufende Schlaufen“),

ein thematisch kontinuierlicher Rezeptionsweg charakteristisch, „auf dem sich jedes Textsegment inhaltlich-thematisch auf der Grundlage der bereits rezipierten Textsegmente einordnen lässt." (Storrer, 2000b, S. 240). Sie sind auf eine vollständige Lektüre ausgelegt, d. h. einzelne Textteile können nicht ohne Beeinträchtigung der Textkohärenz getauscht oder entfernt werden. Als Beispiele werden argumentative Texte (*Urteilsbegründung*), narrative Texte (*Märchen*, *Novelle*) und Informationstexte (*wissenschaftliche Monografie*, *Fachartikel*) genannt. Bei mehrfachsequenzierten Texten ist die vom Produzenten intendierte vollständige Rezeption nicht mehr gegeben. Es können unterschiedliche Pfade existieren, von denen der Leser einen wählt. Das globale Textthema muss daher in mehrfachsequenzierten Texten so verschriftlicht werden, dass unterschiedliche zweckgebundene Rezeptionen ermöglicht werden (z. B. beim *Reiseführer* oder *Computerhandbuch* sowie bestimmten *Lehrbüchern* oder *Monografien*). In mehrfachsequenzierten Hypertexten existiert meist eine Einstiegsseite, von der aus mehrere Pfade „durch eine Struktur mit hierarchischem Grundgerüst führen" (ebd., S. 242). Unsequenzierte Texte entsprechen prototypischen Hypertexten, denn nach Storrer wird in dieser Sequenzierungsform auf vorbestimmte Lesepfade vollständig verzichtet, wodurch die Rezeptionsreihenfolge beliebig wird. Die Idee hierbei ist, „das Thema so abzuhandeln, dass zu verschiedenen Typen von Benutzungssituationen gezielt Informationen abgerufen werden können" (ebd., S. 241). Als Beispiele nennt Storrer *Wörterbücher* und *Lexika*. Die Knoten in unsequenzierten Hypertexten sind in der Regel nach funktionalen oder thematischen Kriterien miteinander verknüpft. Der Zugang erfolgt nach Storrer meist auf der Grundlage von Suchfunktionen, die den Anwender bei der Informationsauffindung unterstützen.

In Abschnitt 3.3 wird argumentiert, dass unsequenzierte Hypertexte im WWW den Status von Randerscheinungen besitzen, dass also das Gros der Webseiten eher Printtexten als prototypischen Hypertexten ähnelt (vgl. Grigar, 2002, S. 165). Die Behandlung eines Themas findet meist nicht durch eine große Menge unsequenziert vernetzter Knoten statt. Stattdessen ist die Anzahl der Knoten eines Hypertextes im WWW in der Regel überschaubar, ihre Inhalte besitzen eine heterogene Konsistenz und umfassen Bilder jeglicher Art, Textfragmente (vgl. Dürscheid, 2000, S. 71) und insbesondere Listen von Hyperlinks. Listen sind im WWW derart omnipräsent, dass Crystal (2001, S. 197) der Ansicht ist, „it would seem that list organization is intrinsic to the structure of the Web."[67] Der Beobachtung von Storrer, die Navigation in unsequenzierten Hypertexten erfolge meist über Suchfunktionen, ist prinzipiell zuzustimmen, doch ist gerade die *Abwesenheit* von Suchfunktionen, wie Storrer sie beschreibt, in sehr vielen Websites ein deutliches Indiz dafür, dass *keine* Unsequenziertheit vorliegt.[68] In diesem Zusammenhang wird eine Diskrepanz deutlich, denn die Mehrzahl der

Stern („Typische Formen [...] sind die auch in Hypermedia-Anwendungen immer noch als »Hauptmenü« bezeichneten Überblicks- oder Startscreens, von denen aus man in die [informationellen Einheiten ...] der Anwendung gelangt."), Baum bzw. Hierarchie („treten dort auf, wo eine klare Gliederung in Themen und Subthemen vorliegt") und Netz. Sager weist darauf hin, dass diese Basistypen „in der Regel im Rahmen von konkreten Hypermedia-Anwendungen in vielfältiger Weise miteinander kombiniert" werden (ebd.).

[67] Schütte (2004a, S. 333) ermittelt ebenfalls eine „Dominanz von Listentextstrukturen" und bezeichnet diese als „übereinzelsprachlich homepagetypisches Vertextungsprinzip".

[68] Zwar bieten viele Websites Suchfunktionen an, diese besitzen jedoch nicht den von Storrer gemeinten Status als *charakteristische* Zugangsfunktion zu einem unsequenzierten Hypertext. Vielmehr fungieren sie nur als weiteres Werkzeug zur Informationsauffindung. Die primäre Orientierung erfolgt über Navigationshilfen. Ein Indiz für diesen abgeschwächten Status ist die unauffällige Platzierung von Suchfunktionen (z. B. als kleines Eingabefeld

an der Textlinguistik orientierten Arbeiten beschäftigt sich mit unsequenzierten Hypertexten, d. h. der prototypischen Ausprägung des Hypertextkonzepts (vgl. Abschnitt 3.3).[69]

3.5.2 Hypertext und Paratext

Der Begriff Paratext (Genette, 2001) bezieht sich auf Texte, „die einen Basistext ergänzend oder kommentierend begleiten“ (Bußmann, 2002, S. 497), bei einem Buch z. B. Vor- und Nachwort, Widmung, Motto und Danksagung. Genette nimmt eine Subklassifizierung vor: Peritexte umfassen diejenigen paratextuellen Elemente, die sich im Basistext selbst befinden (z. B. Titelei, Überschriften, Autorname etc.), Epitexte beziehen sich auf einen Basistext (im öffentlichen Bereich z. B. Rezensionen oder Werbetexte, im privaten Bereich ein Briefwechsel mit dem Autor).[70] In Abschnitt 3.3.1 wurde angemerkt, dass traditionelle Texte Verweise enthalten, die einen nichtlinearen Zugang unterstützen (z. B. Fußnoten). Derartige Verweise sind der peritextuellen Ebene zuzuordnen, wenn sie sich nicht implizit auf den Basistext selbst, sondern explizit auf den Peritext beziehen (vgl. Hammwöhner, 1997, S. 27). Daher besteht eine enge Verbindung zwischen Peritext und der typografischen Gestaltung eines Textes (vgl. Sager, 2000), deren wichtigste Merkmale der Textsatz, die für Überschriften und die Grundschrift gewählten Schrifttypen, der Satzspiegel und der Zeilenabstand sind (vgl. auch Tschichold, 1960, S. 11, und S. 16, sowie Abschnitt 3.5.6).

Hammwöhner (1997, S. 26) beobachtet in Bezug auf Paratexte (speziell Peritexte) und ihren Status in Hypertexten eine „Metamorphose“, einen „Formwandel“ (ebd., S. 29). Einem Buch zugeordnete Peritexte werden z. B. in einem Hypertext oft für jeden Knoten ein-

in der oberen rechten Ecke oder als Hyperlink, der zu einem Suchformular führt, vgl. Abschnitt 4.6.2). Ein Gegenbeispiel stellen die Wikipedia-Enzyklopädien dar (z. B. `http://de.wikipedia.org`). Ihre Einstiegsseiten enthalten zwar Gruppen unterschiedlich sortierter Links zu einzelnen Artikeln, die meisten Anwender dürften jedoch von der auch hier unauffällig platzierten Suchfunktion Gebrauch machen, um in der mehr als 200 000 Artikel umfassenden Hypertextbasis Informationen zu finden (vgl. Abschnitt 4.6.5).

[69] Besonders drastisch wird die Auffassung, nur ein unsequenzierter Hypertext sei ein „richtiger“ oder „guter“ Hypertext, von Todesco (1997, S. 113) vertreten: „Natürlich lassen sich auf Hypertextmaschinen auch Texte schreiben, die dem Leser viel weniger Freiheiten lassen als ein Buch [...]. *Hier ist aber von Hypertexten [...] die Rede*.“ (Hervorhebung hinzugefügt, G. R.). Lutz (1995, S. 160) hat eine differenziertere Meinung: „Ein *guter* Hypertext sollte sich [...] nicht an einem [...] linearen Text orientieren [...].“ (vgl. Brown, 1990, und Foltz, 1996). Mit Bezug auf das WWW meint Nielsen (1999, S. 4): „Readers and writers must both adjust to non-linear information spaces, that is, how to *write* in ways that utilize hypertext and how to *read* without the safety of mind that comes from making no decisions beyond turning the page. Nothing but time and plenty of experience and exposure to well-crafted hypertexts will make this change happen. Unfortunately, [...] well-crafted hypertexts will not happen until good writers have become skilled in writing hypertexts.“ Todesco vergleicht Hypertext mit einem gedruckten Buch. Pang (1998) bezeichnet dies als „reductionist attitude toward the products of print culture“. Mit Bezug auf Landow (1992) meint Pang: „the words “text” and “book” are essentially interchangeable, and “books” are taken to be academic monographs or novels. But the book is not the only kind of print technology: newspapers, magazines, brochures, [...], and a million other printed artifacts form a kind of textual white noise background to our lives.“ Darüber hinaus warnt Pang vor inadäquaten Generalisierungen, da zahllose, teils sehr unterschiedliche Ausprägungen des Konzepts „Buch“ existieren.

[70] Klein (2000, S. 36) unterscheidet Vor-, Parallel- und Nach-Textsorten: Exemplare von Vortextsorten sind „modellbildend, subsidiär oder motivierend für die Produktion von Exemplaren der zu beschreibenden Textsorten“. Die Exemplare von Paralleltextsorten werden „unter einem einheitlichen Gesichtspunkt (in etwa) gleichzeitig mit der zu beschreibenden Textsorte produziert“. Nachtextsorten sind Textsorten, „für die die beschriebene Textsorte eine Vor-Textsorte darstellt“. Zusätzlich werden Filtertextsorten eingeführt, die „in komprimierter Reformulierung [...] den Inhalt von Exemplaren der beschriebenen Textsorte“ wiedergeben.

zeln genannt, z. B. die Autorangabe, weshalb „die Vorstellung von der Gesamtautorschaft eines Hypertexts verschwimmt." (ebd., S. 28).[71] Ein weiterer Unterschied besteht in peritexuellen Angaben zur Aktualität. Ein Buch besitzt eine Auflagennummer, die Knoten eines Hypertextes können hingegen zu unterschiedlichen Zeiten aktualisiert werden, was sich nach Hammwöhner in der Angabe des Datums der letzten Änderung auf der Knotenebene widerspiegelt.[72] Einer ähnlichen Metamorphose unterliegt der Index, der in Büchern ein Zugangswerkzeug und in vielen Hypertexten ein „bedeutende[s] Such- und Orientierungsmittel" darstellt (ebd., S. 29). Vor- und Nachwort, deren Positionierung in einem linearen Text schon durch ihre Bezeichnungen ausgedrückt wird, können nur in monosequenzierten Hypertexten auftreten, ihren eigentlichen Funktionen kommt nach Hammwöhner aber gerade in umfassend vernetzten Hypertexten eine besondere Bedeutung zu (z. B. Angaben über die Zielgruppe, Ziel des Hypertexts, Hinweise auf spezifische Lesestrategien). Verlegerische Peritexte schließlich befinden sich bei Büchern in der Regel auf dem Schutzumschlag und dienen der Werbung. Hammwöhner beobachtet auch hier einen Wandel, weil der Umschlag „als Ort derartiger peritextueller Elemente [...] bei Hypertexten [entfällt]." (ebd.). Im unmittelbaren Zusammenhang mit dem Wegfall des Umschlags stellt Hammwöhner die im WWW „geübte Praxis" (ebd.) dar, die wegfallenden Funktionen durch eine Reihe von Gestaltungsmitteln „*im Prinzip auf jeder einzelnen* Hypertextseite wahrzunehmen" (ebd., Hervorhebung hinzugefügt, G. R.). Dies betrifft die Aufnahme eines Logos, das die publizierende Organisation kennzeichnet (vgl. Abschnitt 3.5.6, S. 111 ff.), eine „spezielle Aufbereitung des Seitenhintergrunds erhöht den Wiedererkennungswert", und ein Hyperlink „auf die Startseite(n) ermöglicht dem Leser den Zugriff auf das gesamte Informationsangebot eines Anbieters unabhängig davon, wie er eine bestimmte Seite erreichte." (ebd.).

Huber (2002) ordnet die Bestandteile eines HTML-basierten Hypertextes den von Genette (2001) eingeführten Ebenen zu (vgl. auch Dalgaard, 2001): Zum anwenderspezifischen Peritext gehören die Navigationsfunktionen des Browsers. Dem verlegerischen Peritext entsprechen das Webdesign sowie möglicherweise verwendete Metaphern (vgl. Schönefeld, 2001, sowie Abschnitt 3.5.4).[73] Die Navigationshilfen, die sich im Hypertext selbst befinden, werden dem vom Autor bestimmten Peritext zugerechnet. Die „Gliederung" bzw. „Knoten-Struktur" des Hypertextes sieht Huber als „Peritext im engeren Sinn" an (ebd., S. 84; vgl. Abschnitt 3.5.7). Die Ebene des öffentlichen Epitextes wird in zwei Kategorien gruppiert: Ver-

[71] Hammwöhner stellt diesen Aspekt verkürzt dar. Bei einer Monografie liegt nur ein Autor vor, der zu Beginn genannt wird. Bei einem Sammelband existieren zahlreiche Beiträger, die zu Beginn ihrer Artikel genannt werden. In gleicher Weise kann ein Hypertext von einer oder mehreren Personen geschrieben werden. Erst in letzterem Fall verschwimmen die Grenzen, jedoch nur in Bezug auf den Vergleich von Einzelbeiträgen eines Sammelbandes zu den Knoten eines Hypertextes.

[72] Jucker (2004, S. 19) beschäftigt sich mit der „Flüchtigkeit der Internetdaten" und sieht es als das „größte Problem" an, mit dem die Korpuslinguistik konfrontiert wird, da sie fixierte Daten benötigt. Eine Webseite unterliegt jedoch durchaus einer Fixierung, nur ist es eben auch möglich, Änderungen an ihrem aktuellen Zustand vorzunehmen, so dass diesbezüglich der Vergleich mit den Auflagen eines Buches herangezogen werden kann: Ein bereits publiziertes Buch kann in einer aktualisierten Version erneut publiziert werden, wodurch die veraltete Auflage aus korpuslinguistischer Perspektive aber nicht ihre Gültigkeit verliert. Solange nicht die Änderungsprozesse im Zentrum des Interesses stehen, ist es durchaus adäquat, zu einem gegebenen Zeitpunkt einen Schnappschuss einer Menge von Webseiten anzufertigen und ihn zu analysieren.

[73] Hammwöhner (1997, S. 30) sieht „Informationen über Neuerscheinungen bzw. neue Hypertextknoten" sowie „zu erwartende oder vollzogene Umstellungen der Hypertextstruktur" als verlegerischen Epitext an.

linkte Hypertexte (z. B. Diskussionsbeiträge) werden als Teil des Diskursuniversums und nicht als Bestandteil des Hypertextes aufgefasst, der Hypertextautor empfindet die Texte jedoch als wichtig und bietet in Folge dessen einen Hyperlink an – gerade derartige Verknüpfungen konstituieren Intertexualität (vgl. Abschnitt 3.5.3). Weiterhin können Hypertexte *über* den aktuellen Hypertext existieren, die vom Autor jedoch nicht verlinkt wurden.

3.5.3 Hypertext und Textualität

Bei einem Vergleich der von de Beaugrande und Dressler postulierten Textualitätskriterien (vgl. Abschnitt 2.2.8) in Bezug auf die Charakteristika von Hypertexten[74] muss hinterfragt werden, ob sich die Spezifika dieses Mediums auf die Textualität auswirken, d. h. es ist „vor allem die Frage nach der spezifischen Textualität von Hypertexten zu stellen." (Sager, 2000, S. 587).[75] Sager fasst Hypertexte als „Textkonglomerate aus *mehreren in sich eigenständigen kohärenten Subtexten*" auf, „Hypertexte sind also *nichtlineare und dennoch kohärente* Textgebilde, denen *grundsätzlich* Textualität [...] zugesprochen wird." (Sager, 2000, S. 588, Hervorhebungen hinzugefügt, G. R.).[76] Die Kohärenz stellt dasjenige Kriterium dar, das am deutlichsten von der Multilinearität betroffen ist, die Darstellung von Sager kann daher nur als Annäherung verstanden werden: Wenn ein Hypertext keine fixierten Anfangs- und Endpunkte besitzt, scheint sein semantischer Sinnzusammenhang durchaus gefährdet zu sein, da der Produzent aufgrund des Wegfalls der linearen Gliederung die individuell unterschiedlichen Rezeptionspfade der Leser nur partiell kontrollieren kann, was insbesondere für unsequenzierte, d. h. prototypische Hypertexte gilt. Die Hyperlinks zwischen einzelnen Knoten bestimmt im WWW ausschließlich der Produzent, weshalb in den zahlreichen Arbeiten zur Kohärenz in Hypertexten auch meist von der Perspektive der veränderten Kohärenzbedingungen und Kohärenzbildungshilfen ausgegangen wird, die der Autor berücksichtigen sollte.

Sowohl bei der Analyse eines spezifischen Hypertextes als auch bei einer abstrakten Betrachtung des Mediums zur Eruierung seiner spezifischen Textualitätscharakteristika ist es notwendig, die jeweilige Bezugsgröße zu bestimmen. Diese kann sich auf die gesamte Hypertextbasis, einen oder mehrere Knoten beziehen.[77] Eine zusätzliche Möglichkeit besteht in

[74] Mit diesem Vergleich wird nicht das Ziel verfolgt, den Textbegriff von de Beaugrande und Dressler (1981) einzusetzen, um einen an das Medium Hypertext adaptierten Textbegriff zu (re)konstruieren. Vielmehr dienen die Textualitätskriterien als eine Art Leitfaden zur Erörterung linguistischer Spezifika des Mediums.

[75] Storrer (2004b, S. 14) zufolge bleibt der Einsatz von Hyperlinks „nicht ohne Konsequenzen für die am sequenziell geordneten Printtext entwickelten Vorstellungen von Textualität."

[76] Eine entgegengesetzte Meinung, die sich auf Desktop-Technologien bezieht, vertritt Schmitz (1997, S. 140 f.): „Multimedia-Zeichen weisen [...] nicht die Kohärenz auf, die wir von [...] Texten gewohnt sind. [...] Schriftliche Texte in multimedialen Kontexten und auch diese selbst können zwar [...] kohärent sein [...], sind es typischerweise aber nicht. Vielmehr sind sie inkohärent, flüchtig, beweglich, experimentell und offen." Schmitz bezieht sich auf grafische Oberflächen moderner Betriebssysteme und führt ein Bildschirmfoto mit Standardanwendungen als Beispiel an. Schütte (2004a, S. 104 f.) empfindet dies – ohne Angabe näherer Gründe – als „konstruierte Bildschirmansicht, die in der Praxis eher selten sein dürfte, da mit ihr kaum sinnvoll zu arbeiten ist." Zuzustimmen ist Schütte jedoch bezüglich der Feststellung, dass es nicht das Ziel einer grafischen Oberfläche ist, Kohärenz (im Sinne einer „auf Sinnkontinuität zielenden Rezeption") zu stiften (ebd., S. 105).

[77] Der Umfang des Hypertextes und die Anzahl seiner Autoren spielen eine wesentliche Rolle bei der Bestimmung seiner kommunikativen Funktion: „Ist ein kleiner Hypertext vielleicht noch als ein intendiertes Kommunikat aufzufassen, so ist der ganze Diskurse umfassende Hypertext, wie ihn sich Ted Nelson vorstellt, dieser Betrachtungsweise sicher entzogen." (Hammwöhner, 1997, S. 38; vgl. auch Abschnitt 3.3.7).

Usability-Analysen, also der Beobachtung von Benutzern, die – frei oder vorgegebene Aufgaben lösend – in einem Hypertext navigieren und individuelle Textsequenzen konstruieren (vgl. Bucher, 2001).[78] Hammwöhner (1997, S. 39) fasst die Rezeption eines Hypertextes als „eingeschränkte Form eines Dialogs" auf (ebd., S. 72 ff.). Nach Ipsen (2001, S. 69) liegt „eine Art bedingter Dialogizität" vor. Bucher (2004, S. 141) stellt den zentralen Aspekt heraus, denn „wir haben es aus der Perspektive des Nutzers [...] mit einer dialogischen Situation zu tun: In der Aneignung des digitalen Kommunikationsangebots wird eine dialogische Situation *unterstellt*". Es liegt also eine „*antizipierte Dialogkonstellation*" (ebd.) vor, in der die „Nutzer handeln, *als ob* das Angebot ein Kommunikationspartner wäre" (ebd., S. 145).[79]

Kohäsion

Kohäsion bezeichnet den sprachlichen Zusammenhalt auf der Textoberfläche. Sie wird unter anderem durch Anaphern, Überschriften und Gliederungsebenen hergestellt. In Hypertexten ist ein Rückgang sprachlicher kohäsiver Mittel auszumachen, der durch den vermehrten Einsatz peritextueller Elemente kompensiert wird, so übernehmen Logos, spezifische Layouts und typografische Merkmale Aufgaben der Kohäsionsbildung (vgl. die Abschnitte 3.5.2 und 3.5.6). Hammwöhner (1997, S. 41) weist diesen Elementen, „die mediale Objekte verbinden oder abgrenzen" und „Kombinationen von Schrift und Graphik oder Bild ausdrücklich ein[schließen]" (ebd., S. 42) eine primär kohäsive Wirkung zu.

In Hypertexten kann die Rezeptionsreihenfolge vom Autor nur partiell antizipiert werden, weshalb Anaphern als Einleitung eines Knotens keine Verwendung finden dürften, denn die Kohäsion erfährt aufgrund der Abwesenheit des Koreferenten einen Bruch. Kohäsion existiert also meist nur in Form rekurrierender Ausdrücke (vgl. Jucker, 2000, S. 25, und Horn, 1989, S. 49).[80] Weingarten (1997b, S. 226) analysiert einen Hypertext und zeigt, dass

[78] Aus diesem Grund präferiert Fritz (1999, S. 222, ebenso wie Bucher, 2001, S. 140) den Terminus Multilinearität: „For the user hypertext is [...] multilinear. A sequence produced by travelling through such a network is called a *path*. [... A] path is something like a text [...]." Pfammatter (1998, S. 65) teilt diese Auffassung: „Am Ziel der Realisierung kohärenter Pfade mißt sich sowohl die Intention des Autors bei der Kreation wie auch das Verhalten des Nutzers bei der Rezeption." In der älteren Hypertextliteratur wird der Begriff *path* synonym mit *Guided Tour* benutzt (vgl. Jonassen, 1989, S. 10, und Abschnitt 3.6.5).

[79] Fritz (1999, S. 223 f.) verwendet ebenfalls den Dialogbegriff: „The fact that the user himself chooses from alternatives reminds one of the activities of a speaker in dialogue. Therefore the structure of dialogues obviously provides a useful object of comparison for the interactive aspect of hypertext." Ähnlicher Auffassung ist Storrer (2003, S. 286): „Als leitende Metapher für die Kohärenzplanung [...] eignet sich [...] der Dialog zwischen Nutzer und Hypertextsystem, dessen Ablauf vom Produzenten [...] durch Hypertext-Strukturierung und den Einsatz hypertextspezifischer Navigations- und Orientierungshilfen gesteuert werden kann." Ein Dialog im eigentlichen Sinne liegt bei der Hypertextrezeption nicht vor. Dürscheid (2004, S. 146) merkt an, dass „die dialogische Komponente meist völlig im Hintergrund [steht]." (vgl. auch Michalak und Coney, 1993).

[80] Kuhlen (1991, S. 37) merkt an, dass „[t]raditionelle textuelle kohäsive Strukturen [...] in Hypertexten kaum eine Rolle [spielen]; vielmehr werden sie entweder ganz aufgelöst oder durch die Verknüpfungstechnik explizit gemacht." Kuhlen fordert daher die „kohäsive Geschlossenheit" von Hypertextknoten. Hierunter versteht er, „daß in informationellen Einheiten nicht auf die in Texten üblichen kohäsiven Gestaltungsmittel über die Grenzen von Einheiten hinweg zurückgegriffen werden kann [...]. Informationelle Einheiten müssen in kohäsiver Sicht autonom sein und sollten entsprechend autonom rezipiert werden können." (ebd., S. 87), was auch eine Grundvoraussetzung dafür ist, diese Knoten von anderen Einheiten aus zu referenzieren. Kuhlen (1991) bezieht sich fast ausschließlich auf die knotenübergreifende Kohäsion (Huber, 2002, S. 57), denn erst auf dieser Ebene sind deutliche Unterschiede zwischen Text und Hypertext festzustellen.

„es keine kohäsiven Elemente außer der Rekurrenz einzelner Ausdrücke [...] gibt."[81] Für den Hypertext ergibt sich nach Weingarten „eine größere sprachliche Unabhängigkeit der einzelnen Seiten" (ebd., vgl. Fußnote 80). Durch den Wegfall kohäsiver Mittel nimmt die „sprachliche Integration der Texte" ab, d. h. „Textabschnitte werden eher aggregativ in einem räumlichen Cluster zusammengeführt." (ebd., S. 235; vgl. Abschnitt 3.5.6).

Auch Hyperlinks können als Bestandteile von (paratextueller, vgl. Schmitz, 2003, S. 268) Kohäsion aufgefasst werden.[82] Ihre bloße Existenz ist jedoch, wie Hammwöhner (1997, S. 38) annimmt, noch kein Kriterium für Kohäsion: „Durch das Bestehen von Verknüpfungen kann der Hypertext als kohäsiv angesehen werden."[83] Nach Kuhlen (1991, S. 102 ff.) bilden assoziative referenzielle Hyperlinks lexikalische Kohäsionsstrukturen (ebd., S. 114): Autoren verwenden Hyperlinkanzeiger über direkte Koreferenzierung, d. h. das Vorkommen einer Bezeichnung ist für den Autor ein Anlass, eine Verknüpfung auf einen Knoten zu integrieren, in der die als Anzeiger fungierende Bezeichnung ebenfalls thematisiert wird. Kuhlen nennt dies „hypertextuelle lexikalische Kohäsion" (ebd.).

Hammwöhner (1997, S. 42) schreibt einzelnen Knoten die Funktion von Absätzen in traditionellen Texten zu, denn an dieser Stelle „fällt die Trennlinie zwischen mikro- und makrostrukturellen Kohäsionsphänomenen weitgehend mit den Grenzen eines Hypertextknotens zusammen". Außerhalb von Knoten wird Kohäsion somit durch Hyperlinks hergestellt: Der Hyperlinkanzeiger fungiert als Thema und sein Ziel als Rhema (vgl. Abschnitt 2.2.2). Hyperlinks konstituieren Thema-Rhema-Strukturen und fügen Knoten zu „thematischen Absatzkomplexen" zusammen (ebd., S. 43). Der Leser navigiert daher aktiv und iterativ zwischen Themen und Rhemen (vgl. de Saint-Georges, 1998, Pfammatter, 1998, S. 67, und Jucker, 2000, S. 26). Hammwöhner vertritt den Standpunkt, dass durch „diese explizite textuelle Deixis" der Sequenz der angesteuerten Knoten „eine höhere Kohäsion zuzumessen" ist „als den nicht gelesenen Alternativen." (ebd.).[84] Storrer (2003, S. 276) fasst Kohäsion „als Spezialfall von Kohärenz (als durch grammatische Mittel gestiftete Kohärenz)" auf und subsumiert Kohäsionsmittel unter dem Begriff der Kohärenzbildungshilfen, zu denen Storrer auch layoutbezogene Hilfen zählt.

[81] Ipsen (1999) demonstriert mittels eines *JavaScript*-Programms, das die rezipierten HTML-Dokumente protokolliert, dass die Verwendung anaphorischer Verweise wie z. B. „wie Sie bereits gesehen haben" sehr wohl möglich, jedoch mit einem gewissen technischen Aufwand verbunden ist: Beim Besuch eines Knotens wird getestet, ob ein Knoten, auf den im aktuellen Knoten Bezug genommen wird, bereits rezipiert wurde. In diesem Fall kann der oben wiedergegebene Ausdruck anaphorisch auf ihn verwiesen.

[82] Storrer (2001c, S. 198 f.) weist darauf hin, dass (textuelle) Hyperlinkanzeiger drei Funktionen parallel erfüllen: Sie tragen, wenn sie in den Fließtext integriert sind, zur Textbedeutung bei, sie fungieren als Schaltflächen und sie sollten den Benutzer darüber informieren, „wie das Linkziel beschaffen ist, welche Beziehung hinter der Verknüpfung steckt und was das Linkziel" zum aktuellen Knoten beiträgt (ebd., S. 199). In HTML existieren diesbezüglich nur eingeschränkte Möglichkeiten der Hyperlinktypisierung (vgl. Abschnitt 3.5.5).

[83] Mehler (2001, S. 332) kritisiert zu Recht die Auffassung, „daß Kohäsion eine Eigenschaft sei, die Hypertexten schon aufgrund des Bestehens von Verknüpfungen zukommt, so daß der Kohäsionsgrad eines Hypertexts proportional zur Zahl seiner Verknüpfungen steigt. [...] Nicht die Existenz einer Verknüpfung produziert Kohäsion, sondern erst die Einlösung, des mit der Verknüpfung verbundenen »Versprechens« eines kohäsiven, (sprachlich, semantisch) kontinuierlichen Anschlusses von Textmodulen."

[84] Es handelt sich um eine extreme Ansicht. Kohäsion ist ein textinternes Kriterium und von den Eigenschaften des Textes abhängig, die der *Autor* realisiert hat. Es ist höchstens bei Hypertextsystemen, in denen der Leser schreibenden Zugriff auf die Hypertextbasis besitzt, denkbar, dass ein rezipierter (und dabei möglicherweise implizit oder explizit editierter) Pfad eine höhere Kohäsion besitzt als ein nicht rezipierter Pfad.

Kohärenz

Zahlreiche Arbeiten diskutieren die veränderten Kohärenzbedingungen in Hypertexten aus textlinguistischer Perspektive (z. B. Kuhlen, 1991, Freisler, 1994, Foltz, 1996, Hammwöhner, 1997, Vora und Helander, 1997, Weingarten, 1997b, van Berkel und de Jong, 1999, Storrer, 1999a, Fritz, 1999, Jucker, 2000, Bucher, 2001, Ipsen, 2001, Mehler, 2001, und Storrer, 2003). Nachfolgend werden die zentralen Aspekte thematisiert, wobei festzuhalten ist, dass „Kohärenz unter verschiedenen Aspekten [...] betrachtet werden kann", weshalb „die bislang vorgebrachten Meinungen [...] recht uneinheitlich" sind (Storrer, 1999a, S. 33).

Die für die Kohäsion gemachten Aussagen gelten in ähnlicher Weise für die Kohärenz: „Hypertexte verlangen offenbar eine Kohärenz, die auf Rekurrenz und thematischer Progression beruhen [sic] und auf kohäsive Elemente [...] ganz verzichten." (Jucker, 2000, S. 25). Die wesentliche Bezugsgröße stellen Hyperlinks dar. Nur durch eine dem Rezipienten sinnvoll erscheinende Verlinkung kann der Leser in die Lage versetzt werden, einen semantischen Sinnzusammenhang zu konstruieren. Der individuelle Pfad durch einen Hypertext, dem Fritz (1999) Textstatus zuschreibt, ist dann kohärent, „wenn die Knoten in sich kohärent sind und die Verbindungen so angelegt sind, dass jeder Ausgangsknoten für alle davon erreichbaren Zielknoten einen sinnvollen Kontext abgibt." (Jucker, 2000, S. 26). Nach Storrer (2001c, S. 26) müssen Hypertextknoten „so gestaltet sein, dass sie einerseits für sich verständlich sind [...], dass andererseits der Bezug zu einem übergreifenden Ganzen erkennbar bleibt." (vgl. auch Bucher, 1999, S. 15).[85] Auch Weingarten (1997b, S. 216) geht vom Rezipienten (also von der Kohärenzbildung) aus, der immer bestrebt sein wird, einen Hypertext „zu einem kohärenten Ganzen zu integrieren", ebenso wie generell unterstellt werden kann, „dass der Autor unabhängig vom Medium ein Interesse daran hat, den Leser beim Aufbau kohärenter Wissensstrukturen möglichst gut zu unterstützen." (Kohärenzplanung, Storrer, 1999a, S. 34). Dieser Prozess ist abhängig vom Kohärenzgrad des Hypertextes. Ist dieser gering, steigen die kognitiven Integrationsanforderungen an den Rezipienten, der jedoch gleichzeitig die Freiheit besitzt, die Informationen an die individuellen Interessen und Intentionen anzupassen (Hammwöhner, 1997, S. 235). Bucher (2001, S. 155) stellt den Prozess des Aufrufens von Zielknoten dar: „Die neuen Seiten werden als Kommunikationsbeitrag eines Angebotes verstanden, der in das Sequenzmuster eingepasst wird [...]. Diese kontinuierliche Kohärenzkontrolle manifestiert sich in Kohärenzurteilen, wenn die aufgerufene Seite den Erwartungen nicht entspricht."[86] Auch Storrer (1999a, S. 41) geht von dieser prozessbezogenen Perspektive aus und betrachtet Kohärenz nicht nur als textinterne Eigenschaft, „sondern als einen

[85] Kuhlen (1991, S. 36) vertritt daher die Ansicht, „daß es wenig Sinn macht, von der *Gesamtkohärenz* einer Hypertextbasis zu sprechen. Die eine Hypertextkohärenz kann es nicht geben. [...] Hypertexte sind [...] rezipientenabhängige Informationssysteme. Zwar ist Kohärenz [...] auch in traditionellen Texten nicht nur eine Leistung des Autors, sondern beruht auch auf der Rezeptionskompetenz des Lesers, in Hypertext wird dies aber zum generellen Prinzip gemacht." Im übergreifenden Kontext betont Sager (2000, S. 600): „Universelle Verknüpfbarkeit [... ist ...] ein Indiz dafür, dass das [WWW] als Ganzes keine Textualität und keine Kohärenz besitzt. Die kann sich erst einstellen, wenn Nutzer sich einen [...] Pfad [...] durch das Netz bahnen."

[86] Bucher (2001, S. 157) untermauert diese Ansicht mit empirischen Befunden, die durch Benutzerstudien gewonnen wurden. Es zeigt sich, dass die Nutzer „in Bezug auf die Abfolge von Angebotsseiten bestimmte Sequenzmuster unterstellen: So soll auf eine Strukturseite [...] eine Inhaltsseite folgen und eben [...] keine weitere Strukturseite. Ähnlich wie in der gesprochenen Kommunikation kann nicht auf eine Ankündigung eine weitere Ankündigung folgen."

übergreifenden Sinnzusammenhalt, der sich durch einen Kommunikationsprozess hindurchzieht", weshalb das Vorwissen und die Handlungsziele der Partner für die Kohärenzplanung und -bildung eine wesentliche Rolle spielen (ebd., S. 42).

Es ist zwischen lokaler und globaler Kohärenz zu differenzieren, die ineinander greifen. Lokale Kohärenz bezieht sich auf den semantischen Zusammenhang benachbarter Textteile. Globale Kohärenz bezeichnet den „Gesamtzusammenhang, der die thematische und funktionale Gliederung des ganzen Textes in Textsegmente determiniert." (Storrer, 2003, S. 285, vgl. Bucher, 1999, S. 20 ff.).[87] Nach Storrer (1999a, S. 43) kann in unsequenzierten Hypertexten nur auf der Knotenebene lokale Kohärenz herrschen. Hinzu kommt, dass Rezipienten solcher Hypertexte individuelle Pfade wählen, d. h. der Sinnzusammenhang sukzessive rezipierter Knoten (lokale Kohärenz) und ihr Stellenwert innerhalb des Hypertextes (globale Kohärenz) ist nur in Teilen vom Produzenten planbar; er muss hingegen vom Leser konstruiert werden (ebd., S. 44; vgl. auch Schütte, 2004a, S. 106). Dieser Konstruktionsprozess sollte vom Autor unterstützt werden, so dass einzelne Knoten kontextualisiert werden können:

> Das Ganze des Hypertextes muss erst sichtbar gemacht, der Leser muss die Ausgestaltung des „virtuellen" Textraums überblicken können. Bei Hypertexten, die in größere Hypertextnetze eingebunden sind, muss die Kontextualisierung der einzelnen Module auf verschiedene Größen bezogen sein: Auf das globale Thema und die Textfunktion des Hypertextes; auf die Organisation der Site, in der der Hypertext verortet ist; auf den thematisch-funktionalen Stellenwert im gesamten Hypertextnetz. (Storrer, 1999a, S. 44)

Das „Sichtbarmachen" des Hypertextes erreicht der Produzent durch Gestaltungsmittel, die als Kohärenzbildungshilfen fungieren.[88] Navigationshilfen heben die globale Struktur eines Hypertextes und den Status des aktuellen Knotens hervor. Nach Storrer (2003, S. 287) existieren drei funktionale Typen: (i) Überblickshilfen unterstützen den Rezipienten beim Aufbau eines mentalen Modells der thematischen und funktionalen Struktur (z. B. Sitemaps, vgl. Pilgrim und Leung, 1999, spezielle Visualisierungen, vgl. Vora und Helander, 1997, und Metaphern, vgl. Abschnitt 3.5.4). (ii) Globale Kontextualisierungshilfen kennzeichnen den

[87] Vora und Helander (1997, S. 885 f.) unterscheiden „cohesion or local coherence" und „global coherence" und betonen, dass der Autor auf beiden Ebenen Hinweise („cues") geben sollte: „To improve coherence within the node, one can use the traditional reading models [...] to improve local and global coherence. At the net level, however, to increase coherence, the designers of hypertext should attempt to limit the "fragmentation" so users can understand how the text is distributed over several nodes and what they have to do with each other." (ebd.). Die lokale Kohärenz kann nach Vora und Helander durch die Verwendung von Link-Etiketten verbessert werden. Übersichtsdarstellungen tragen zur Stiftung globaler Kohärenz bei.

[88] Oftmals wird das Problem thematisiert, dass Anwender mit einem Klick unbeabsichtigt auf eine andere Website gelangen können, wodurch ein Kohärenzbruch entstünde, da der Ziel- nicht mit den Ausgangsknoten in Einklang gebracht werden könne. Gerade bei erfahrenen Benutzern existiert dieses Phänomen nicht. Mehrere Indikatoren zeigen an, dass ein Wechsel der Website stattgefunden hat, wozu schon das meist unterschiedliche Webdesign gehört. Die meisten Browser bieten eine prophylaktische Kohärenzbildungshilfe an: Beim Überfahren eines Hyperlinks mit dem Mauszeiger wird in der Statuszeile die URL des Ziels eingeblendet. Dem Benutzer ist die URL des aktuellen Knotens entweder unmittelbar bekannt oder er betrachtet die Adresszeile am oberen Rand, die diese URL enthält. Ein Vergleich der Adressen gibt schnell Auskunft darüber, ob es sich um einen externen oder internen Link handelt und ob das Ziel ein HTML-Dokument, eine Tondatei oder ein Video ist (vgl. auch Bucher, 2001, S. 148, und Storrer, 2001b, S. 103). Aus Sicht der Benutzerfreundlichkeit sollten Anwender jedoch gar nicht erst in die Lage versetzt werden, derartige Hilfsmittel verwenden zu müssen.

funktionalen und thematischen Stellenwert eines Knotens und erleichtern die globale Kohärenzbildung (z. B. Überschriften und thematische Sätze, grafische Strukturübersichten). (iii) Lokale Kontextualisierungshilfen verdeutlichen, welche Zielknoten erreichbar sind und in welcher Relation sie zum aktuellen Knoten stehen. Hierdurch wird die Planung des Rezeptionswegs erleichtert und die lokale Kohärenzbildung unterstützt (z. B. Link-Etiketten und typisierte Hyperlinks, vgl. Abschnitt 3.5.5). Storrer (1999a, S. 49) führt zusätzlich (iv) retrospektive Hilfen ein, die eine Orientierung im bereits durchschrittenen Leseweg erlauben. Zu den globalen Kontextualisierungshilfen zählen auch typografische und gestalterische Aspekte des Peritextes (vgl. Abschnitt 3.5.2, sowie Bittner, 2003, S. 96 f.), z. B. „identitätsstiftende Elemente wie Hintergrundfarbe, Logos oder charakteristische Navigationsleisten“ (Storrer, 2003, S. 288).[89] Abschnitt 3.5.6 geht genauer auf diese Thematik ein.

Intentionalität und Akzeptabilität

Intentionalität bezieht sich nach de Beaugrande und Dressler (1981) auf die Einstellung des Produzenten, einen sowohl kohäsiven als auch kohärenten Text zu erstellen, um eine Absicht zu erfüllen. Akzeptabilität geht vom Rezipienten aus, der einen kohäsiven und kohärenten Text erwartet, der für ihn nützlich oder relevant ist. Der vorangegangene Abschnitt ist bereits auf die Erstellung und Erwartung eines kohäsiven und kohärenten Textes eingegangen: Nach Weingarten (1997b, S. 216) ist der Rezipient eines Hypertextes bestrebt, diesen „zu einem kohärenten Ganzen zu integrieren“, Storrer (1999a, S. 34) zufolge ist der Autor in gleicher Weise bestrebt, den Leser bestmöglich bei der Kohärenzplanung zu unterstützen. Hierbei handelt es sich jedoch nur um den Idealfall, denn ein Text oder Hypertext kann für einen Leser aus verschiedenen Gründen inkohäsiv oder inkohärent sein:

> Eine sprachliche Struktur muß als Text *intendiert* und *akzeptiert* werden, um in der kommunikativen Interaktion verwendet werden zu können. Diese Einstellungen bedingen eine [...] Toleranz gegenüber Kohäsions- oder Kohärenzstörungen, solange die Zweckhaftigkeit der Kommunikation besteht. (de Beaugrande und Dressler, 1981, S. 118)

Der zweite Aspekt der eingangs paraphrasierten Definition von de Beaugrande und Dressler betrifft die Ziele des Autors, die sich auf die Textfunktion auswirken und häufig als Textsorten in komplexe und institutionalisierte Handlungszusammenhänge eingebettet sind (vgl. Storrer, 2004b, S. 18). Storrer unterscheidet zwischen Hypertexten mit erkennbarer Funktion und Hypertextnetzen, die ohne Intention zusammengestellt wurden: „Die Abgrenzung ist [...] wichtig, weil [...] die Trennung von Eigenem und Fremdem und die Verantwortlichkeit für den Textinhalt von hoher Relevanz ist, z. B. im Zusammenhang mit Urheberrechtsfragen oder mit der Verantwortlichkeit i. S. des Presserechts.“ (ebd., S. 18 f., vgl. Fußnote 91, S. 99). Nach Hammwöhner (1997, S. 52) ist in Hypertexten häufig das Fehlen „einer durchgängigen Intention“ die Ursache von Kohäsions- und Kohärenzbrüchen. Die Gründe liegen in zusätzlichen Äußerungsebenen, die gerade im WWW deutlich werden: Auf der ersten Ebene kann ein Autor den Inhalt eines oder mehrerer Knoten verfassen. Auf der zweiten Ebene

[89] Auch die Ratgeberliteratur betont: „An attractive site is distinguished by a cohesive and consistent look that presents a unique identity [...]. [... T]he user doesn't notice the individual images so much as he or she enjoys the overall atmosphere and experience created by the site.“ (Rosenfeld und Morville, 1998, S. 7).

kann die Konvertierung linear organisierter Texte in Hypertexte vom ursprünglichen Autor vorgenommen werden, es ist jedoch gerade bei umfangreichen Hypertexten häufig der Fall, dass die Segmentierung von einer anderen Person oder einem Konverter durchgeführt wird. Die dritte Ebene betrifft die Verknüpfung, wobei Hammwöhner drei Fälle unterscheidet: (i) Die Verknüpfung wird von der Person produziert, die auch Autor dieser Knoten ist; (ii) der Autor der Verknüpfung ist der Produzent von nur einem der beiden Knoten; (iii) die Verknüpfung wird von einer Person durchgeführt, die keinen der Knoten produziert hat (vgl. Nielsen, 1995b, S. 326). Diese „Differenzierung der Autorenrollen" ist dafür verantwortlich, dass es zu „Inkongruenzen der Intentionen und damit der verfolgten Strategien" kommen kann (Hammwöhner, 1997, S. 52), weshalb der oder die Autoren eines Knotens oder einer Verknüpfung genannt werden sollten. Hierdurch können Erwartungshaltungen bestätigt, widerlegt oder modifiziert werden: „Ist die Verknüpfung vom Autor der Texteinheiten vergeben, so werden höhere Anforderungen an Kohäsion und Kohärenz zu stellen sein, als wenn es sich um eine von einem Dritten erstellte Verknüpfung handelt" (ebd.).

Hammwöhner bezieht sich in abstrakter Weise auf das Konzept Hypertext. Auf universitären Websites kann die folgende Situation festgestellt werden: Der Autor eines Hypertextes ist praktisch immer auch der Autor der Verknüpfungen, die von Knoten ausgehen. Ausnahmen stellen Hypertexte dar, die von mehreren Autoren produziert oder gepflegt werden (z. B. die Webdokumente eines Instituts).[90] Das von Hammwöhner dargestellte Problem manifestiert sich im WWW letzten Endes in der Hinsicht, dass der Leser im Falle eines Kohäsions- oder Kohärenzbruches hinterfragt, weshalb ein spezifischer Hyperlink überhaupt existiert, da keine produzentenseitige Äußerungsstrategie erkannt wird. Ein weiteres Problem kommt hinzu: Der Rezipient kann nur anhand von Indizien erschließen, wer der tatsächliche Autor eines Knotens ist, da HTTP keine obligatorische und explizite Nennung des Produzenten vorsieht. Der expliziteste Hinweis ist die – in technischer Hinsicht fakultative – Angabe des Autors im Dokument selbst bzw. innerhalb eines diesem Knoten zugeordneten Impressums.[91] Im Falle der Nennung eines Autors können mehrere Szenarien unterschieden werden: (i) Der Name ist der Name des Autors; (ii) der Name ist ein Pseudonym, Spitz- oder Rufname des Autors; (iii) die Bezeichnung bezieht sich auf eine Rolle bzw. Funktion des Autors (z. B. „Webmaster" oder „Webmistress") oder einer Autorengruppe („Web-Redaktion", „Web-Team"); (iv) der Name ist nicht mit dem Namen des Autors identisch.[92] Gerade bei persönlichen Homepages von Wissenschaftlern kann dieser letzte Fall beobachtet werden: Das Dokument beschreibt

[90] Die Situation ist auf Websites, die wirtschaftliche Interessen verfolgen, sehr viel komplexer, da hier Datenbank-gestützte Websites die Regel sind, die ihre Bestandteile aus Vorlagen beziehen, die von mehreren Autoren oder Autorengruppen erstellt wurden (vom Webdesigner, der z. B. Navigationselemente entwirft, über den Administrator, der einen Link zur Suchmaschine eingepflegt hat, zur Redaktion, die sich um die Inhalte kümmert). Auch Online-Werbung, bei der der Verlinkende für jede Aktivierung eines Werbebanners mit einem kleinen Betrag entlohnt wird, muss berücksichtigt werden (vgl. Fortanet et al., 1998, 1999, und Crijns, 2001).

[91] Laut § 6 des Teledienstegesetzes vom 14.12.2001 ist die Anbieterkennung für die Betreiber „geschäftsmäßiger" Websites obligatorisch (vgl. Brinker und Hoffmann, 2004, S. 71). Ausgenommen sind rein private Websites. In Sonderfällen gelten die Vorschriften aus § 312 c BGB und § 10 des Mediendienste-Staatsvertrags.

[92] Neben der oder den unmittelbar in einem Knoten genannten Personen können weitere an der Textherstellung oder -veröffentlichung beteiligt sein (vgl. Jakobs, 2003, S. 239). Die Webauftritte von Institutionen haben z. B. einen Urheber (die Organisation selbst), der von mehreren Interessengruppen vertreten wird (z. B. vom Vorstand oder der Abteilung für Öffentlichkeitsarbeit), die als Auftraggeber fungieren, um die Website von einer Gruppe von Technikern, Autoren und Webdesignern oder einer externen Agentur entwickeln zu lassen.

eine spezifische Person und führt Publikationen und Forschungsprojekte auf; als Nennung des Autors wird jedoch ein *anderer* Name angegeben, da ein Techniker oder eine Hilfskraft hiermit beauftragt wurde (vgl. Abschnitt 10.5.11). Eben diese Nennung kann jedoch auch fehlen. Der Rezipient eines Dokuments hat dann die Möglichkeit, den HTML-Quelltext des Dokuments zu untersuchen, in dem unter Umständen der Autorenname innerhalb eines `meta`-Elements enthalten ist (vgl. Abschnitt A.4.4). Einige HTML-Editoren fügen den Namen des Produzenten, der daher nicht zwangsläufig mit dem Besitzer der persönlichen Homepage identisch sein muss, automatisch ein. Sind derartige Angaben nicht verfügbar, kann der Rezipient aufgrund der URL des Dokuments Rückschlüsse auf die Identität des Autors ziehen (allgemein geht es dabei um Fragen der Aktualität, Verlässlichkeit, Korrektheit, Glaubwürdigkeit etc., vgl. Fogg et al., 2001, 2003, und Harrison, 2002). Falls der Produzentenname nicht angegeben ist, das Dokument sich jedoch im Adressraum eines universitären Instituts oder eines Arbeitsbereiches befindet, dürfte der Autor ein Angehöriger der entsprechenden Organisationseinheit sein, d. h. die Institution selbst fungiert als veröffentlichende Stelle der angebotenen Texte und Informationen, die hierdurch einen offiziellen Status erhalten und die individuelle Autorschaft in den Hintergrund drängen: Sobald dem Rezipienten deutlich ist, dass es sich bei einem Hypertext z. B. um die offiziellen Informationen eines Instituts handelt, spielt es eine nur untergeordnete Rolle, ob die Dokumente vom Geschäftsführer, einem wissenschaftlichen Mitarbeiter oder einem Techniker angelegt und gepflegt werden. Dies bezieht sich jedoch nur auf generelle Angaben – ein von einer Hilfskraft verfasstes Grußwort könnte durchaus einen Kohärenzbruch verursachen. Falls sich das Dokument in einem persönlichen Bereich des Webservers befindet (in den meisten Fällen erkennbar an der Tilde, ~, innerhalb der URL) ist es in einigen Fällen möglich, über die Benennung des „persönlichen“ Adressraums den Namen des Autors zu identifizieren.[93]

Hammwöhner diskutiert die Akzeptabilität von Hypertexten auf der Basis der Kommunikationsmaxime (Grice, 1975).[94] Hammwöhner (1997, S. 54 f.) argumentiert, dass das Kooperationsprinzip in der Formulierung von Grice nicht auf Hypertexte anwendbar ist und modifiziert es entsprechend: „Gestalte den Hypertext so, daß ein Leser einschätzen kann, ob ein Hypertextsegment für ihn relevant ist und zu welchem Zweck es mit dem aktuellen Knoten verknüpft wurde.“ Auf dieser Basis werden vier hypertextspezifische Ausprägungen der Konversationsmaximen dargestellt (nach Hammwöhner, 1997, S. 55):

[93] Beispielsweise werden an der Justus-Liebig-Universität Gießen persönliche Homepages nicht über den Namen, sondern über die UNIX-Kennung im Hochschulrechenzentrum adressiert (z. B. `g91063`). Mittels eines Verzeichnisdienstes wie X.500, der im WWW recherchierbar ist, kann der Rezipient mit etwas Mühe die kryptische Kennung einem Personennamen zuordnen. Bei den (meist durch Werbung oder geringe monatliche Beträge finanzierten) privaten Webspace-Dienstleistern ist es in fast allen Fällen – rechtliche Schritte ausgenommen – unmöglich, den Namen des Anbietes eines persönlichen Webangebots zu ermitteln.

[94] Grice (1975) leitet die vier nachfolgend nach Bußmann (2002, S. 379) dargestellten Kommunikationsmaximen aus dem Kooperationsprinzip ab, das besagt, dass Kommunikationspartner ihre Beiträge der jeweiligen Kommunikationssituation und dem akzeptierten Kommunikationsziel anpassen sollten. (i) Maxime der Quantität: Mache deinen Beitrag so informativ wie erforderlich. (ii) Maxime der Qualität: Versuche deinen Beitrag so zu machen, dass er wahr ist. (iii) Maxime der Relation: Mache deinen Beitrag relevant. (iv) Maxime der Modalität: Sei klar und deutlich. Haas und Grams (1998a, S. 485) erläutern implizit die Anwendung dieser Prinzipien auf die Produktion und Rezeption von Hypertexten im *World Wide Web*.

- *Maxime der Quantität:* Gib alle Informationen, die zur Einschätzung der aktuellen Navigationsmöglichkeiten erforderlich sind, aber nicht mehr!

  Hiermit sind Metainformationen über Verknüpfungen gemeint, z. B. das Thema des Zielknotens. In HTML-Dokumenten kann dies beispielsweise durch mit visuellen Mitteln typisierte Hyperlinks erfolgen. Auch die Struktur des Hypertextes selbst ist von Bedeutung, um sie mit einer Sitemap transparenter zu gestalten.
- *Maxime der Qualität:* Biete nur verlässliche Verknüpfungen an oder gib zumindest den Grad der Verlässlichkeit zu erkennen!

  Diese Maxime gilt laut Hammwöhner für maschinell generierte Linknetze, bei denen ein Hyperlink durch ein Verlässlichkeitsmaß ergänzt werden könnte. Sie kann auch auf manuell erstellte HTML-Dokumente übertragen werden, indem z. B. die Verlinkung „stabiler" Dokumente präferiert wird (vgl. Nielsen, 1999, S. 249 ff.).[95]
- *Maxime der Relation:* Biete nur Verknüpfungen an, die für den Leser relevant sind!

  Hammwöhner (1997, S. 55) geht in seiner Bewertung dieser Maxime implizit von unsequenzierten Hypertexten aus: „Offensichtlich kann der Autor dies nicht entscheiden, da er kein verläßliches Bild des Lesers oder [der] Lesesituation hat. [...] Dem Autor ist aber abzuverlangen, die Grundlage für die notwendigen Auswahlprozesse zu schaffen." In den Geltungsbereich dieser Maxime fallen Hyperlinks ohne jeglichen Bezug zum Thema des aktuellen Knotens oder Verweise auf Knoten, die eine andere Lesart des Konzepts diskutieren, das der Hyperlinkanzeiger enthält.
- *Maxime der Modalität:* Gib zu erkennen, welche diskursiven Ziele mit einer Verknüpfung oder einem Knoten erreicht werden sollen, um dem Leser (oder dem Hypertextsystem) die Möglichkeit zur Bewertung von Navigationsalternativen zu geben!

  Diese Maxime wird von Hammwöhner nicht kommentiert. Im WWW können häufig Hyperlinks beobachtet werden, die keine explizite Kennzeichnung des vom Autor verfolgten diskursiven Ziels besitzen. In vielen Fällen ist dieser Umstand durch die Absicht zu erklären, den z. B. von einer Grafik neugierig gemachten Leser, durch ein unerwartetes Linkziel zu überraschen. Die Erwartbarkeit einer Äußerung ist nach de Beaugrande und Dressler (1981) ein wesentlicher Aspekt der Informativität.

Die modifizierten Kommunikationsmaximen stellen Prinzipien dar, nach denen Hypertexte und Verknüpfungen gestalten werden sollten, so dass sie für den Rezipienten eine bestmögliche Akzeptabilität besitzen. Hammwöhner diskutiert sie unter der Überschrift „Akzeptabilität"; ebenso sind sie der Intentionalität zugehörig. Ein wichtiges Instrument zur Bestimmung der Akzeptabilität (und weiterer Textualitätskriterien) von Hypertexten stellen Usability-Untersuchungen dar (vgl. Nielsen, 1995b, 1999).

[95] Im WWW kann ein Knoten als „flüchtig" oder „instabil" bezeichnet werden, wenn z. B. explizit gekennzeichnet ist, dass er nach einem gewissen Zeitraum wieder entfernt wird, wenn es sich um einen Knoten handelt, dessen Webserver häufig nicht erreichbar ist oder wenn der Autor weiß, dass der umgebende Hypertext häufig reorganisiert wird. Bei einem „stabilen" Dokument kann der Autor aus mehreren Gründen der Ansicht sein, dass der Knoten langfristig erreichbar sein wird. Ein Beispiel: Die Einstiegsseite des Webauftritts eines großen Software-Herstellers kann als stabiler eingestuft werden als ein Hypertext (ein Webserver mit einem Namen, der nicht www lautet, kann ein weiteres Indiz für Instabilität sein), der z. B. dynamisch aus einer Datenbank generiert wird und über Fehler einer spezifischen Software informiert (vgl. McMillan, 2001).

Informativität

Die Informativität betrifft das „Ausmaß der Erwartetheit bzw. Unerwartetheit oder Bekanntheit bzw. Unbekanntheit/Ungewißheit der dargebotenen Textelemente." (de Beaugrande und Dressler, 1981, S. 10 f.).[96] Die Informativität kann sich auch auf die Textstruktur beziehen. Nach Hammwöhner (1997, S. 56) steht sie daher in einem „Widerstreit" zu Kohäsion und Kohärenz, „die eine Textstruktur konstituieren, damit die Vorhersagbarkeit erhöhen und die Informativität verringern", weil die Aufmerksamkeit des Lesers auf die konventionelle Textstruktur gelenkt wird, „die für die Einführung von Information vorgesehen ist." (ebd.). Hammwöhner übertragt diesen Aspekt auf Hypertexte:

> Entsprechend werden sich auch in Hypertexten Muster einer adäquaten Informationsverteilung ausprägen müssen. Derartige Muster sind weitgehend konventionell. Sie können sich daher für ein vergleichsweise junges Medium wie Hypertext noch nicht ausgeprägt haben. [...] Informationsmuster können sich [...] innerhalb eines Knotens ausprägen und auf einer globalen, knotenübergreifenden Ebene. (Hammwöhner, 1997, S. 56)

Obwohl er nur vage von „konventionellen Informationsmustern" spricht, bezieht sich Hammwöhner hier auf ein Konzept, das mit Textstrukturmustern oder Textsorten vergleichbar ist. Und sich da das Medium Hypertext erst noch in der Entwicklungsphase befindet, im Gegensatz zum Medium Papier also noch sehr jung[97] ist, konnten sich noch keine Konventionen bilden. Dennoch befindet sich im Anschluss an die oben reproduzierte Passage ein Indikator für den Umstand, dass sich erste Muster entwickeln. Hammwöhner stellt fest,

> daß innerhalb eines Knotens der informativste Teil meist im Zentrum der Präsentation liegt, während konventionelle oder aufmerksamkeitsleitende Textteile eher an den oberen oder unteren Rand rücken. Daß diese Aufteilung bereits verinnerlicht ist, zeigt sich daran, daß Seiten mit ungewöhnlichem Aufbau Aufmerksamkeit auf sich ziehen. Hier liegt die Informativität nicht im Inhalt, sondern in der unerwarteten Aufbereitung. (Hammwöhner, 1997, S. 56 f.)

Mit diesem sehr generellen Beispiel entfernt sich Hammwöhner deutlich vom Konzept der Textsorte, denn es bezieht sich auf prototypische Hypertexte und einen prototypischen Knotenaufbau, der zu hinterfragen ist. So kann die Beobachtung, der informativste Teil liege meist im Zentrum der Präsentation, als Widerspruch zur Informativität gesehen werden, das ein rezipientenabhängiges Kriterium ist. Nicht jeder Rezipient wird eine identische Auffassung darüber haben, welcher Teil eines Knotens der „informativste Teil" ist. Im „Zentrum der Präsentation" wird sich derjenige Teil befinden, den der Produzent als wichtigsten Teil erachtet – eine derartige „Konvention" ist nahe liegend. Die produzentenseitige Einschätzung der Wichtigkeit eines Knotenteils ist aber nicht mit seiner rezipientenabhängigen Informativität gleichzusetzen. So vage das von Hammwöhner vorgebrachte Beispiel auch sein mag,

[96] Storrer (2004b) zufolge ist Informativität als Indiz für Textualität problematisch, weil *jede* Äußerung als informativ anzusehen ist – dieser Aspekt wird auch von de Beaugrande und Dressler (1981, S. 149) thematisiert: „Informativität [...] ist immer [...] vorhanden [...]." Storrer (2004b, S. 22) zieht den Schluss, dass ein tautologisches Kriterium „streng genommen überflüssig" ist.

[97] Freisler (1994, S. 39) bezeichnet Hypertext als ein „Medium im Inkunabelstatus". Als Inkunabeln (auch: Wiegen- oder Frühdrucke) werden Druckwerke aus der Frühzeit des Buchdrucks (bis etwa 1500) bezeichnet.

es ist zweifelsohne ein Hinweis dafür, dass auch in Hypertexten textsortenähnliche Konzepte existieren können. Weitere Indikatoren werden in Abschnitt 3.6 diskutiert.

Hammwöhner bespricht auch die Neuigkeit als Aspekt der Informativität: Es kann das Problem auftreten, dass ein Benutzer einen bereits rezipierten Knoten erneut anwählt. Im WWW werden in einem bestimmten Zeitraum besuchte Hyperlinks von fast allen Browsern farblich von nicht besuchten Hyperlinks abgesetzt.[98] Eine weitere Eigenschaft ist von Bedeutung: Da viele Websites einer ständigen[99] Modifikation unterliegen, ist „ein effizienter Zugang zu den neuen und deshalb informativeren Teilen" eines Hypertextes wesentlich (Hammwöhner, 1997, S. 57). HTTP bietet lediglich die Möglichkeit, das Datum der letzten Änderung einer HTML-Datei zu übermitteln; es kann über eine Browserfunktion wie „Page Info" abgerufen werden. Sobald ein Dokument dynamisch erzeugt wird, entfällt diese Angabe jedoch. Die zweite Option setzt voraus, dass der Produzent das Datum der letzten Änderung in das Dokument aufnimmt (vgl. Storrer, 2001c, S. 189, S. 194). Dieser Vorgang kann, je nach Webserver, auch automatisch erfolgen; bei einem CMS wird das Modifikationsdatum des Knotens meist automatisch integriert. Im Falle umfangreicher Hypertexte ist dem Anwender jedoch nicht mit der globalen Angabe der letzten Änderung des gesamten Hypertextes geholfen, weshalb häufig Hyperlinks oder Textfragmente visuell betont werden (z. B. durch animierte Grafiken mit der Beschriftung „neu!"), damit die Aufmerksamkeit unmittelbar auf die aktualisierten oder neuen Knoten gelenkt wird.

Situationalität

Die Situationalität bezieht sich auf „Faktoren, welche einen Text für eine aktuelle oder rekonstruierbare Kommunikationssituation relevant machen" (de Beaugrande und Dressler, 1981, S. 169). Sie wird von de Beaugrande und Dressler an Dialogen exemplifiziert, wobei Faktoren wie Situationskontrolle und -lenkung diskutiert werden. Nur in einem Gespräch ist es möglich, eine unmittelbare Anpassung des Kommunikats auf exakt diese Situation vorzunehmen (vgl. Hammwöhner, 1997, S. 57). Storrer (2004b, S. 21) zufolge ist unklar, „wie über das Vorliegen oder Nicht-Vorliegen des Kriteriums entschieden werden kann; diese Unschärfe mag dazu geführt haben, dass es in der Nachfolge sehr unterschiedlich gedeutet worden ist."

Hammwöhner (1997, S. 57) hebt den Nachteil geschriebener Texte hervor: Diese sind meist an eine allgemeine Situation angepasst, es liegt Situationsentbundenheit vor. Hypertext wiederum ist zwar schriftlich fixiert, aufgrund seiner größeren Flexibilität besteht jedoch die Möglichkeit, Knoteninhalte situationsabhängig anpassen zu können. Adaptive Hypertexte bzw. Hypertextsysteme sind in der Lage, kontextabhängige und „situationsadäquate" (ebd.) Entscheidungen über den Inhalt eines zu präsentierenden Knotens vornehmen zu können. Dieses Thema wird insbesondere im Zusammenhang mit adaptiven E-Learning-Systemen diskutiert (vgl. z. B. Leutner, 1997, und Lobin et al., 2003), die sich – basierend auf einem

[98] Diese Informationen beziehen Browser aus der History-Liste, in der Zeitstempel und Adressen der geladenen Knoten protokolliert werden. Bei den meisten Browsern ist die zeitliche Begrenzung konfigurierbar, meist umfasst sie zwischen sieben und 14 Tage. Nicht besuchte Hyperlinks werden üblicherweise blau hervorgehoben, wohingegen besuchte Links in einem dunkleren Lila visuell abgetönt werden. Mit CSS-Angaben können diese Einstellungen geändert werden. Auf einigen Websites werden für besuchte und noch nicht besuchte Hyperlinks identische Farben verwendet, was die Navigation empfindlich stört (vgl. auch Storrer, 2001c, S. 198).

[99] Dies gilt für die im Korpus enthaltenen Dokumente nicht (vgl. Abschnitt A.3.6, S. 728 ff.).

expliziten Wissensvortest oder impliziten Rückmeldungen, die in einem Benutzermodell resultieren (vgl. Lobin, 1999a) – dem Wissensstand des Lerners anpassen und ein spezifisches Gebiet z. B. auf einer rudimentären Ebene oder auf Expertenniveau vermitteln. Adaptive Hypertexte in dieser sehr spezifischen und komplexen Ausprägung sind im WWW kaum zu finden. Manuelle und automatische Anpassungen an die Kommunikationssituation können aber dennoch festgestellt werden. Storrer (2004b, S. 21) geht auf unterschiedliche „Sichten" ein, die mit Hilfe texttechnologischer Methoden aus einer Dokumentbasis erzeugt werden können. Im WWW werden derartige *Single-Source-Publishing*-Lösungen unter anderem auf Websites von Online-Zeitungen eingesetzt, um eine Artikelansicht mit Navigationselementen und Werbung und eine Druckansicht anzubieten (vgl. Fußnote 25, S. 74, und Storrer, 2001c, S. 180). Auch die Bereitstellung von Dokumenten im WML-Format für WAP-fähige Endgeräte basiert auf derartigen Verfahren. HTTP ist zwar ein zustandsloses Protokoll, Cookies[100] gestatten es dem Webserver jedoch, im Browser einen benannten Zustand mitsamt Zeitstempel zu hinterlegen, der bei nachfolgenden HTTP-Requests ausgewertet werden kann, um z. B. den Inhalt eines HTML-Dokuments situationsbezogen zu modifizieren. Auf diese Weise kann das Datum des letzten Aufrufs gespeichert werden, um dem Benutzer bei einem erneuten Zugriff die seit diesem Datum durchgeführten Änderungen zu präsentieren.

Intertextualität

Die Intertextualität ist für die Untersuchung von Hypertextsorten von zentraler Bedeutung (vgl. Schütte, 2004a, S. 110–115). Sie bezieht sich auf die „Abhängigkeiten zwischen Produktion bzw. Rezeption eines gegebenen Textes und dem Wissen der Kommunikationsteilnehmer über andere Texte" (de Beaugrande und Dressler, 1981, S. 188). Dieses Wissen betrifft neben spezifischen Texten auch Textsorten. Ein weiterer Aspekt hängt mit Beziehungen zwischen Texten zusammen (vgl. Hess-Lüttich, 1997): Ein einzelner Text wird als Intertext aufgefasst, der nur einen Bestandteil eines größeren Beziehungsgeflechts zahlreicher Texte darstellt, die durch intertextuelle Verknüpfungen verbunden sind. Eine spezifischere Auffassung von Intertextualität bezeichnet den Bezug auf konkrete Textexemplare (Prätexte). Explizite intertextuelle Verweise sind obligatorische Konstituenten in Textsorten wie *Rezension* oder *Leserbrief*, implizite Verknüpfungen werden in Parodien verwendet.

Intertextuelle Verknüpfungen der ersten Form besitzen eine Affinität zu Hypertext, denn „Hypertextsysteme unterstützen den einfachen Wechsel zwischen intertextuell verlinkten Teilen von Texten, die in der Printwelt in verschiedenen Büchern publiziert waren", was insbesondere „für digitale Editionen [...] aufeinander bezogener Textsammlungen sowie die Verlinkung mehrerer Typen von Nachschlagewerken" interessant ist (Storrer, 2004b, S. 19; vgl. auch Bush, 1945a). Wenn von der idealtypischen Situation ausgegangen wird, dass jeder Knoten kohäsive und kohärente Geschlossenheit aufweist, erscheint es plausibel, einzelne Knoten als Texte aufzufassen.[101] Ausgehend von dieser Perspektive differenziert Huber

[100] Cookies, genauer gesagt, der *HTTP State Management Mechanism* (RFC 2109, RFC 2964, RFC 2965) sind nicht Bestandteil der Spezifikation von HTTP (RFC 2616). Aktuelle Browser besitzen Werkzeuge zur Verwaltung von Cookies (z. B. um ihre Inhalte zu inspizieren oder Einträge zu löschen), und es ist ebenfalls möglich, das automatische Akzeptieren von Cookies während der Abarbeitung eines HTTP-Requests zu deaktivieren.

[101] In einem ähnlichen Zusammenhang kommt Ziegler (2004, S. 170) zu dem Schluss, dass Hypertext nicht mehr ist als „Text mit anderen Mitteln" (d. h. Hyperlinks).

(2002, S. 71) Intertextualität in Bezug auf Hypertexte: Sie manifestiert sich in Hyperlinks. Huber favorisiert die Option, Knoten als Teiltexte aufzufassen, weil eine übergeordnete Textfunktion des Hypertextes sowie „in der Regel auch designtechnische bzw. paratextuelle Mittel“ vorliegen, wodurch „ein gemeinsames Erscheinungsbild der Knoten eines Hypertextes erreicht wird.“ (ebd.). Folglich nimmt er an, dass nur dann von Intertextualität gesprochen werden kann, „wenn mittels externer Links auf Knoten anderer Hypertexte verwiesen wird.“ (ebd.). Derartige Hyperlinks bezeichnet Huber als „direkte und explizite, hypertextspezifische Intertextualismen“ (ebd.). Als „implizite Intertexualismen“ (ebd.) werden Verweise auf dem Rezipienten aller Wahrscheinlichkeit nach bekannte Konzepte bezeichnet, die jedoch *nicht* als Hyperlink realisiert sind.[102] Auch Hammwöhner (1997) bezieht das Kriterium der Intertextualität auf das Medium Hypertext, nähert sich der Problematik jedoch von einer texttypologischen Perspektive (vgl. Abschnitt 3.6.2).[103]

3.5.4 Einsatz von Metaphern

In sehr vielen Arbeiten über Hypertext- und Multimediatechnologien befinden sich metaphorische Vergleiche, die die Möglichkeiten, aber auch die Gefahren des Mediums erläutern sollen.[104] Hess-Lüttich (1997, S. 139) beschäftigt sich etwa mit dem Interface-Design moderner grafischer Oberflächen (vgl. auch Panko und Panko, 1998) und ermittelt viele „semio-

[102] Huber (2002, S. 71) führt den Satz „Erste theoretische Überlegungen führten zum System Memex.“ als Beispiel an. Ist „Memex“ mit einer digitalen Version von „As we may think“ verknüpft, liegt ein „expliziter Intertextualismus“ vor: „Dies bedeutet nicht, daß Hypertexte völlig frei von impliziten Intertextualismen wären.“ Das Wort könnte „in einem Fachartikel [...] auch ohne expliziten Link auf den Ursprungstext erscheinen, da der Autor von einem Fachpublikum erwartet, diesen Terminus (und den zugehörigen Aufsatz) zu kennen. Da jedoch gerade im WWW das Instrument der extra-hypertextuellen Links existiert, wird dies in der Regel auch zum Explizieren dieser Zusammenhänge benutzt.“ (ebd.). Das Beispiel ist zu hinterfragen, da es sich vom Konzept Hypertext entfernt; gemeint ist ein als *Hypertext* realisierter Fachartikel. Fachartikel zeichnen sich durch die Verwendung von Literaturverweisen aus und es dürften nur wenige Beiträge zum Thema Hypertext existieren, die bei der (ersten) Verwendung des Wortes „Memex“ *keinen* Verweis auf Bush (1945a) enthalten. Dieses Beispiel eines „expliziten Intertextualismus“ ist nicht etwa hypertext-, sondern vielmehr textsortenspezifisch.

[103] Sager (1997, S. 119) beschäftigt sich mit einer allgemeineren Ausprägung von Intertextualität: „Der Hypertext [...] externalisiert [...] die gedankliche Verknüpfung, die Heuristik des Lesers wird nun in [...] das programmierte Netz der Verknüpfungen [...] hinausverlagert. [...] In der hypermedialen Kultur wird nun auch die in der kognitiven Heuristik gedanklich aktualisierte Intertextualität in die materiell physikalische Netzstruktur des [...] Speichermediums hinausverlagert und bereits heute in on-line Netzen bereit gestellt. In ihnen liegen [...] gleichsam vorgefertigte Heuristiken vor, durch die der Benutzer virtuell wandern kann.“ Dieser den „naturalistic associationism“ (Dillon, 1996, S. 28; vgl. Abschnitt 3.4.1) reflektierende Standpunkt ist kritisch zu hinterfragen: Wenn Sager von „on-line Netzen“ spricht, so ist zweifelsohne das WWW gemeint. Die „gleichsam vorgefertigte[n] Heuristiken“ entsprechen Hyperlinks (z. B. in Hotlists). Diese können von einem Rezipienten in der Regel nicht modifiziert werden, so dass „die gedanklich aktualisierte Intertextualität“ gerade *nicht* „in die materiell physikalische Netzstruktur des Speichermediums hinausverlagert“ wird, denn der Rezipient folgt ja lediglich bereits existierenden Verknüpfungen, die ihrerseits auf der subjektiven Intertextualität des Produzenten basieren und darüber hinaus keinesfalls einen idealtypischen Status besitzen.

[104] Bruce (1999) zeigt anhand von 37 Interviews mit australischen Wissenschaftlern, dass die in den Massenmedien verwendeten Metaphern zur Umschreibung der Funktionsweise des Internet (z. B. „information infrastructure“, „infobahn“, „communication superhighway“ etc.) nur einen geringen Einfluss auf die von den Probanden aufgeführten Metaphern besitzen. Fast die Hälfte der befragten Wissenschaftler vergleicht das Internet mit einem „information store“ oder einer Bibliothek. Sieben Metaphern konnten nicht zugeordnet werden (z. B. „fruit shop“, „a sort of infinite depth of bubbles“ und „short wave radio“ Bruce, 1999, S. 192).

tische Lösungen", die metaphorisch und kulturabhängig sind (z. B. „file", „folder", „desktop" und „trash"): „Die damit erstrebte »Benutzerfreundlichkeit« soll die kognitive Belastung des *modus operandi* reduzieren und die Konzentration auf die Inhaltsverarbeitung erleichtern." Lutz (1995, S. 157) konzentriert sich auf die Hypertextliteratur und beobachtet ebenfalls viele Metaphern an „zentralen Stellen" (ebd.), die aus den Gebieten des Reisens, des Bewegens und des Erkundens stammen (ausführlich hierzu Cölfen et al., 1997, S. 257 ff.):

> Von *Reisen* durch das *Hyperland* ist da die Rede, *to be lost in hyperspace* wird als schwerwiegendes Problem bezeichnet. Bei diesen Gefahren ist es kein Wunder, daß der *traveller* eine *guided tour* unternehmen muß, um sich nicht zu verlaufen (obwohl er doch seine *bookmarks* gesetzt und seine *footprints* hinterlassen hat). Auch die Typologie der Navigationsstile beim Lesen von Hypertexten paßt ins Bild [...]: *Scanning – browsing – searching – exploring – wandering*. Ich vermute, daß diese auffällig metaphernreiche Sprache [...] mit der Begeisterung über die Möglichkeiten des neuen Textmediums zusammenhängt. (Lutz, 1995, S. 157)

Von der Verwendung metaphorischer Ausdrücke zur Charakterisierung[105] der Eigenschaften von Hypertext ist der Einsatz kohärenzstiftender Metaphern zu unterscheiden (vgl. Abschnitt 3.5.3), die den Benutzern anhand einer vertrauten Rahmensymbolik einen Überblick über die Struktur des Hypertextes ermöglichen (vgl. Schmid-Isler, 2000).[106] Die übergreifende Metapher für diese Struktur ist das Netz, das aus Knoten und Verknüpfungen besteht (vgl. Storrer, 1999a, S. 50). Die Rezeption von Hypertexten ist eher von der Raummetapher geprägt (vgl. Fleming, 1998, S. 2, und Hammwöhner, 1997, S. 66).

Vora und Helander (1997, S. 885) diskutieren Techniken, um den Verstehensprozess des Lesers zu unterstützen: Metaphern werden an erster Stelle genannt, weil ihr Erfolg empirisch nachweisbar ist.[107] Vora und Helander schreiben ihnen zwei Vorteile zu: (i) Metaphern ermöglichen es dem Autor, die Struktur des Hypertextes mit einem intuitiven Wiedererkennungswert zu veranschaulichen, (ii) vertraute Konzepte werden auf nicht vertraute Knoten und Kanten abgebildet; die Metapher ermöglicht es Lesern, bekanntes Wissen und vorhandene Fähigkeiten auf eine unbekannte Domäne zu transferieren (vgl. Rouet und Levonen, 1996, S. 12 f.). Als häufige Beispiele zählen Vora und Helander Karteikarten, Bücher, Enzyklopädien, Bibliotheken, Reisen, Städte und Karten auf, wobei die Analogien aber als häufig

[105] Nach Storrer (1999a, S. 50) ermöglichen Metaphern „überhaupt erst, sich über Bedienung und Strukturierung von Hypertexten zu verständigen und entsprechende Metatexte abzufassen." Rouet und Levonen (1996, S. 12) sind der Auffassung, dass viele aktuelle Arbeiten zum Thema Hypertext keine theoretische Fundierung aufweisen, was sie als Ursache für den Metapherngebrauch deuten: „There is neither a general theory of hypertext, nor a model of the cognitive processes involved in reading hypertext. There is a large gap between theories of knowledge or discourse and the actual hypertext systems. [...] Given this lack of foundation, researchers have tried to characterize hypertext by pointing out its similarities and differences with linear (printed) text. Analogies and metaphors play a central role in these efforts to understand the specifity of hypertext."

[106] McKnight et al. (1991, S. 81 ff.) betrachten Metaphern in Hypertexten aus kognitionspsychologischer Sicht. Nach Sager (1997, S. 120) entsprechen Metaphern in Hypertexten „Techniken der antiken Mnemonik".

[107] Für Sager (2000) besitzen Metaphern einen derart hohen Stellenwert, dass er sie als eines von fünf Mitteln zur Erzeugung von Textualität in Hypermedia-Anwendungen ansieht: „(i) ein einheitliches Screendesign [...], (ii) eine durchgehende Metapher, mittels derer die verschiedenen medialen Elemente präsentiert werden (etwa als zu durchwandernde Räume), (iii) das [...] einheitliche Navigationssystem, (iv) [...] Überblicksscreens, (v) die dadurch vorgenommenen Einschränkungen der universellen Montage der Screens untereinander, (vi) ein einheitliches Cursordesign und Cursorverhalten." (Sager, 2000, S. 596, vgl. Fußnote 37, S. 78).

nicht offensichtlich („tenuous“, ebd.) bezeichnet werden. Außerdem kann eine einzelne Metapher zu restriktiv sein, weshalb der Einsatz mehrerer Metaphern vorgeschlagen wird.[108] Vora und Helander weisen auf die sorgfältige Auswahl von Metaphern hin, denn schlechte, d. h. nicht intuitive Metaphern können – ebenso wie überbeanspruchte Metaphern (Fleming, 1998, S. 67 f., Nielsen, 1999, S. 180) – eine kontraproduktive Wirkung besitzen (vgl. Pfammatter, 1998, S. 67). Daher sei ein Interface mit „well-designed features“ (Vora und Helander, 1997, S. 885) dem Einsatz von Metaphern häufig vorzuziehen.

Rosenfeld und Morville (1998, S. 150 f.) unterscheiden drei Typen von Metaphern, die im Webdesign[109] eingesetzt werden können: (i) Organisationsmetaphern lehnen sich an Institutionen an. Auf der Website eines Kraftfahrzeughändlers könnten z. B. Rubriken wie „Neuwagen“, „Gebrauchtwagen“, „Reparatur und Service“ und „Ersatzteile“ existieren. (ii) Funktionale Metaphern fokussieren Prozesse und Aufgaben, die Besucher in einer bekannten Umgebung durchführen. Als Beispiel wird das Durchstöbern der Regale einer Bibliothek genannt. (iii) Visuelle Metaphern verwenden bekannte grafische Elemente. Ein Online-Verzeichnis von Adressen könnte z. B. einen gelben Hintergrund und stilisierte Telefon-Icons verwenden, um eine Verbindung zu den Gelben Seiten herzustellen.

Storrer (1999a, S. 51, 2003, S. 287, 2004a) zählt ebenfalls häufig eingesetzte Metaphern auf (weitere Beispiele liefern Horn, 1989, Nielsen, 1999, S. 180 ff., Siegel, 1999b, S. 21, Bucher, 1999, S. 27 f., Thalheim und Düsterhöft, 2000, und Crijns, 2001): In der Hypertextliteratur sowie in Lernsystemen werden häufig Metaphern aus dem Bereich der Printmedien verwendet: „Elektronische Bücher und Karteikästen werden in virtuelle Bibliotheken eingestellt, in elektronischen Katalogen kann geblättert werden, elektronische Notizblöcke und Terminkalender übernehmen ähnliche Funktionen wie ihre papierenen Vorgänger.“ (ebd.). Die Buchmetapher eignet sich nach Ansicht von Storrer eher für komplexe Hypertexte: „Sie dient als funktionales Modell für den Umgang mit Texten, knüpft an vertraute Arbeits- und Rezeptionsformen an und verfügt dabei über zusätzliche Leistungsmerkmale und Charakteristika.“ (ebd.). Typisch ist für komplexe Hypertexte, dass peritextuelle Merkmale des Buches übernommen werden (Inhaltsverzeichnis, Register, Glossar), ihr Gesamterscheinungsbild und ihre Funktionalität wird aber deutlich erweitert bzw. modifiziert (vgl. Fußnote 109).

3.5.5 Typologisierungen, Funktionen und Positionen von Hyperlinks

Hypertexte unterscheiden sich von traditionellen Texten insbesondere durch Hyperlinks. Die an der Verknüpfung von Knoten beteiligten Ebenen werden nachfolgend diskutiert.

[108] Storrer (2004a) merkt an, dass eine Vermischung metaphorischer Ebenen im WWW zum Alltag gehört: Der Terminus „Homepage“ umfasst Storrer zufolge die Ebenen „Reise“ und „Buch“, und auch Raum- und Interaktionsmetaphern sind eng miteinander verknüpft (vgl. Storrer, 1999b, S. 2, sowie Dürscheid, 2000, S. 70).

[109] In einer Studie professionell gestalteter Einstiegsseiten machen Nielsen und Tahir (2002, S. 2 f.) nicht weniger als acht „of the more common metaphors for homepages“ aus: „Magazine cover“, „Your face to the world“, „Artwork“, „Building lobby“, „Company receptionist/concierge“, „Book table of contents“, „Newspaper front page“ und „Brochure“. Das Fazit von Nielsen und Tahir (2002, S. 3) lautet: „All these metaphors have some truth to them, but each has ways in which it differs from the true nature of homepages. [... I]t is hard to design a homepage because it must have aspects of all the metaphors.“ Einige der aufgeführten Beispiele sind nicht als Metaphern, sondern vielmehr als Anlehnungen an Textsorten aufzufassen (vgl. Abschnitt 3.6).

Typologisierungen und Funktionen von Hyperlinks

Hyperlinks tragen zur Bedeutung eines Hypertextes bei (Ricardo, 1998, S. 146) und können auf der Ebene ihrer Typen unterschieden werden (vgl. Huber, 2002, S. 108–175).[110] Storrer (2001b, S. 99 f.) differenziert z. B. zwischen navigatorischen und thematischen Hyperlinks: Mit Hilfe navigatorischer Links gelangen Anwender von Inhaltsknoten zu strukturbezogenen Knoten. Thematische Links verbinden Knoten aufgrund inhaltlicher Aspekte. Nielsen (1999, S. 53) unterscheidet „structural navigation links", die die Hypertextstruktur verdeutlichen, „associative links", die weitere Informationen zu einem Thema liefern und „see also lists of additional references", die dem Benutzer eine Auswahl alternativer Themen anbieten. Abbildung 3.3 zeigt die von Kuhlen (1991) vorgeschlagene Typologie.

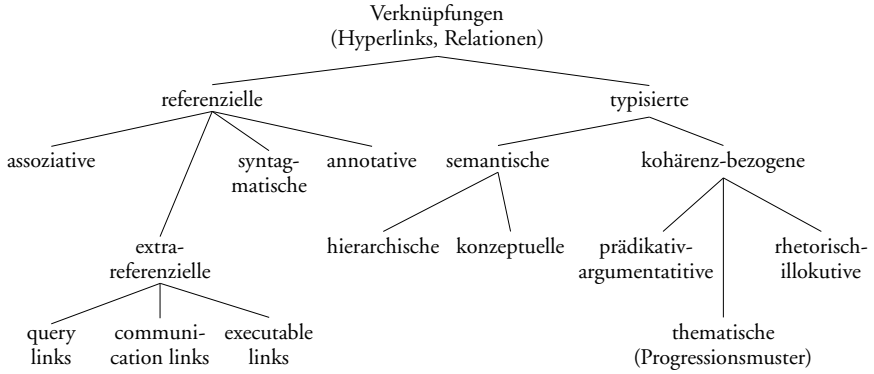


Abbildung 3.3: Typologisierung von Hyperlinks (nach Kuhlen, 1991, S. 106)

Referenzielle Verknüpfungen sind nach Kuhlen (1991, S. 104 f.) weder explizit spezifiziert noch typisiert oder strukturiert. Sie verbinden Knoten, um das assoziative Navigieren zu unterstützen, besitzen eine semantische oder argumentative Funktion und können daher Kohärenz stiften. Eine Unterscheidung referenzieller Typen kann nur aufgrund der „Qualitäten der [verketteten] Einheiten" (ebd., S. 113) erfolgen (vgl. Abbildung 3.3, und Kuhlen, 1991, S. 258 ff.). Die von Kuhlen nicht in die Typologie aufgenommenen Metaverknüpfungen „verbinden die hypertextspezifischen Übersichts- bzw. Metainformationsmittel, z. B. dynamische Inhaltsverzeichnisse [...], mit den informationellen Teilen" (ebd., S. 114). Anmerkungsverknüpfungen entstehen bei der Hypertextierung von Texten, die Fußnoten enthalten.

Typisierte Verknüpfungen besitzen eine explizite Struktur und werden vornehmlich zur hierarchischen Organisation von Knoten eingesetzt.[111] Hierarchische und konzeptuelle Hyperlinks fasst Kuhlen (1991, S. 118) als semantische Verknüpfungen auf, die zur „Relatio-

[110] Kuhlen (1991, S. 107) weist darauf hin, „daß es keinen besten Weg" zur Erstellung einer Typologie von Hyperlinks gibt, denn „zu unterschiedlich sind die möglichen Sichten [...]." Die Aufgabe unterliegt ähnlichen Freiheiten wie die Konstruktion einer Texttypologie (vgl. Abschnitt 2.3).

[111] Kuhlen (1991, S. 106) merkt (mit Bezug auf Conklin, 1987, S. 35) an, dass Knotenhierarchien „dem Grundgedanken der Hypertextidee zuwiderlaufen, die doch gerade die vernetzte Informationsstruktur zum Ziel hat [...]. Untersuchungen an experimentellen Hypertextanwendungen haben jedoch ergeben, daß faktisch die Mehrheit der darin erstellten Verknüpfungen hierarchische und damit taxonomisch-systematisierende Funktionen erfüllten". Conklin zufolge kommen solche Verknüpfungen den Lesestrategien für linear organisierte Texte entgegen, die Hierarchisierung ist eine „nicht aufzugebende Ordnungshilfe" (Kuhlen, 1991, S. 106).

nierung von Konzepten verwendet“ werden, wobei „die Strukturierung von Metainformationen, wie Registern, Inhaltsverzeichnissen, globalen und lokalen Übersichten“ im Vordergrund steht (ebd., S. 106 f.). Argumentative Verknüpfungen erläutert Kuhlen anhand eines Hypertextsystems, das die Modellierung von Streitgesprächen ermöglicht. Vertretene Positionen können durch Hyperlinks verbunden werden, die explizite Typeninformationen wie „responds-to“, „supports“, „objects-to“ und „generalize“ mit sich führen: „Auf diese Weise können baumartige Strukturen von Streitpunkten, Unterstreitpunkten, Antworten, Unterantworten, Argumenten und Unterargumenten »top-down« [...], aufgebaut werden.“ (Kuhlen, 1991, S. 121). Diese Graphen erinnern an RST-Strukturbäume (Mann und Thompson, 1988); die RST ist jedoch nicht für dialogische Texte geeignet (vgl. Abschnitt 2.2.5 sowie Storrer, 1997, Lobin, 1999a, Storrer, 2001c, S. 201, und Huber, 2002, S. 145 ff.). Getypte Links weisen eine konzeptionelle Nähe zu wissensbasierten Systemen, insbesondere semantischen Netzen auf, da ebenfalls Konzepte durch getypte Relationen miteinander verbunden werden (vgl. Abschnitt 13.2). Es wird ein weiterer Zusammenhang deutlich: Beim Lesen eines linear organisierten Textes baut der Rezipient sukzessive ein Modell der Textwelt auf, das insbesondere von der Kohärenz eines Textes determiniert wird. De Beaugrande und Dressler (1981, Kapitel 5) empfehlen daher einen an semantische Netze angelehnten Formalismus, um die Konzepte eines Textes und die zwischen ihnen herrschenden Relationen zu modellieren. Nach Hammwöhner (1997, S. 48) besteht bezüglich der prädikativen Kohärenz kein prinzipieller Unterschied zwischen Hypertexten und linearen Texten, weil „die semantischen Beziehungen in Texten generell netzwerkartig“ sind.

Storrer (2001c, S. 200 f.) führt eine Unterscheidungsebene ein, die sich auf die relative Position des Ziels bezieht. Externe Hyperlinks führen aus einem Hypertext heraus, wohingegen interne Hyperlinks die Knoten eines Hypertextes verknüpfen. Interne Hyperlinks können unterschieden werden in intratextuelle Links, die Teile eines Knotens verknüpfen, und intertextuelle Links, die mehrere Knoten verbinden. Letztgenannte führen zu „Paratexten und Hilfstexten bzw. zu thematisch verwandten Hypertexten in dem von den Autoren kontrollierten Informationsraum“ (Storrer, 1999a, S. 39). Weiterhin unterscheidet Storrer (2001c, S. 200 f.) zwischen *node-to-node-*, *node-to-point-*, *point-to-node-* und *point-to-point-*Hyperlinks (vgl. Storrer, 2001b, S. 99, und Pfammatter, 1998, S. 57).[112] Darüber hinaus werden statische von dynamischen Links unterschieden: Statische Links besitzen ein vorgegebenes Ziel. Bei dynamischen Links wird das Ziel zur Laufzeit berechnet.

Die Ratgeberliteratur betont, die Grundlage einer guten Website sei „a well-designed hierarchy. In this hypertextual world of nets and webs, such a statement may seem blasphemous, but it's true.“, wobei Rosenfeld und Morville (1998, S. 37) der Argumentation von Conklin folgen (vgl. Horn, 1989, Nürnberg et al., 1997, Nordbotten und Nordbotten, 1999, und Dalal et al., 2000). Rosenfeld und Morville (1998, S. 40) stellen fest: „[H]ypertext is rarely a good candidate for the primary organization structure. Rather, hypertext can be used to complement structures based upon the hierarchical or database models.“ Benutzerstudien und Analysen von Protokolldateien bestätigen, dass die hierarchische Navigation präferiert wird.

[112] *Node-to-node-* und *node-to-point*-Links können auch im WWW existieren: Während *point-to-node-* und *point-to-point*-Links von einer spezifischen Position, d. h. einem Linkanzeiger ausgehen, scheinen auf der Knotenebene angesiedelte Links keine klar definierte Absprungsstelle zu besitzen. Durch das HTML-Element `link` wird ein Knoten innerhalb eines Hypertextes verortet (vgl. S. 79). Besitzt eine Webseite derartige Elemente, blendet der Browser eine zusätzliche Menüzeile ein (in *Mozilla* wird sie *Site Navigation Bar* genannt), die unter anderem die *node-to-node*-Links „Top“, „Up“, „First“, „Previous“, „Next“ und „Last“ bereitstellt.

Nach Storrer (2001c, S. 196 f.) existieren drei Ebenen[113] der Hypertexterstellung, die beim Umgang mit Hyperlinks beachtet werden müssen: (i) Auf der Inhaltsebene kann ein explizit oder implizit[114] typisierter Hyperlink spezifizieren, in welcher Relation der Ausgangspunkt zum Endpunkt steht (vgl. Bucher, 1999, S. 22 ff.). (ii) Die Interaktionsebene betrifft die Frage, wie ein Hyperlink aktiviert werden kann und wie das Linkziel sichtbar gemacht wird.[115] (iii) Auf der Präsentationsebene wird bestimmt, wie Hyperlinks kenntlich gemacht werden und auf welche Weise der Typ des Ziels bzw. Hyperlinks veranschaulicht wird. Der nachfolgende Abschnitt geht auf die dritte Ebene genauer ein.

Positionen und Objekttypen von Hyperlinks

Hyperlinks können aufgrund verschiedener Merkmale unterschieden werden. Hierzu zählen die Objekttypen, die als Anzeiger in Erscheinung treten, sowie ihre Positionen innerhalb eines Knotens. Als Hyperlinkanzeiger können Textsegmente, Grafiken und HTML-Formulare fungieren. Storrer (2001b, S. 100) unterscheidet vier Typen, die entweder in den Fließtext integriert oder separat positioniert werden können: (i) Der erste Objekttyp betrifft Zeichenketten, vornehmlich Wörter, Phrasen und Sätze beliebiger Länge. (a) Sind textuelle Hyperlinks separat positioniert, so sind sie häufig Teil von Navigationshilfen, die navigatorische Hyperlinks umfassen (vgl. Storrer, 2001b, S. 100).[116] Navigationshilfen werden häufig separat am oberen oder linken Rand positioniert (vgl. Rosenfeld und Morville, 1998, S. 59). (b) Textlinks können weiterhin in den Fließtext eines Dokuments eingebunden sein, wobei Storrer (2001b) zwischen textintegrierten und metakommunikativen Verfahren der Einbettung unterscheidet.[117] Bei integrierten Verfahren ist der Link Teil des Fließtextes, wobei

[113] Storrer (2001b, S. 97) führt drei alternative Ebenen ein: (i) Durch die Link-Kennzeichnung erkennt der Rezipient, welche Objekte Hyperlinks sind. (ii) Die Link-Explikation betrifft die Tatsache, dass der Autor deutlich machen sollte, was bei der Aktivierung eines Links passiert. (iii) Im Rahmen der Link-Positionierung muss sich der Autor für die bestmögliche Stelle zur Anbringung des Linkanzeigers entscheiden.

[114] Eine implizite Typisierung liegt vor, wenn z. B. externe Hyperlinks in einer anderen Farbe dargestellt oder von einem Icon flankiert werden (vgl. Storrer, 2001c, S. 198, und Huber, 2002, S. 167–175). Nach Storrer (1999a, S. 60) ist die fehlende Unterstützung typisierter Hyperlinks einer der Gründe dafür, „dass sich die Struktur der im WWW publizierten Hypertexte immer noch stark am Printtext orientiert."

[115] Storrer (2001c, S. 197) unterscheidet nach Kuhlen (1991) die eingebettete Anzeige (das Ziel wird entweder anstelle des Linkanzeigers in den Knoten eingefügt oder in einem Pop-up-Fenster dargestellt), die parallele Anzeige (das Ziel wird parallel zum aktuellen Knoten dargestellt) und die ersetzende Anzeige (das Ziel ersetzt den Knoten), die im WWW überwiegt. Parallele Anzeigen werden häufig bei externen Links verwendet. Eine andere Form der parallelen Anzeige findet in Webseiten mit Framesets statt: Die Aktivierung eines Hyperlinks im Navigationsframe bewirkt eine parallele Darstellung des Linkziels im Inhaltsframe. Die eingebettete Anzeige ist z. B. bei Bildergalerien üblich: In einem Knoten mit Vorschaubildern (*thumbnails*) wird nach einem Klick das korrespondierende Bild in einem Pop-up-Fenster in voller Größe dargestellt.

[116] Navigationshilfen fungieren als Orientierungs- und Suchhilfen (Weingarten, 1997b, S. 233 ff.). Sie präsentieren Metainformationen über die Struktur, heben den aktuellen Standort hervor, tragen zur Stiftung von Kohärenz bei und sind „gegenüber dem eigentlichen Text auf einer Metaebene situiert [...]. Dieser Ausbau der metatextuellen Ebene reagiert [...] auf den Strukturverlust auf der lokalen sprachlichen Ebene." (ebd., S. 234).

[117] Da die Inhalte eines HTML-Dokuments den Kategorien „Fließtext" und „Navigationsleiste" nicht eindeutig zugeordnet werden können, liegen Zwischenstufen vor (z. B. die Auflistung von Informationen in Form eines Literaturverzeichnisses; vgl. S. 90). Ist der Titel eines Artikels verlinkt, kann von einer Integration gesprochen werden. Befindet sich jedoch ein Link wie „PDF" am Ende des Literatureintrags, handelt es sich weder um eine Integration noch um eine separate Positionierung (vgl. Abschnitt 8.5, sowie Amitay, 1998, 2000a, 2000b).

der Kontext zusätzliche Informationen über das Ziel und eventuell auch den Linktyp liefert (Storrer, 2001b, S. 103). Von textintegrierten Verfahren unterscheidet Storrer metakommunikative Einbettungsverfahren. Diese zeichnen sich durch die Thematisierung der *Bedienung* des Hypertextes aus, was häufig durch den deiktischen Ausdruck „hier" geschieht, wie z. B. in „Klicken Sie hier, um . . ." (vgl. Storrer, 2001b, S. 100 f.). Die Explikation solcher Links übernimmt ausschließlich der Kontext. Stilratgeber lehnen dieses Verfahren ab, da es vom Inhalt ablenkt. Storrer kommt zu dem Schluss, dass das metakommunikative Verfahren in einigen Fällen „das angemessenere, weil explizitere" ist (ebd.), seine „pauschale Ächtung" solle „nicht unkritisch übernommen werden" (ebd.), denn für unerfahrene Nutzer sei es „eine naheliegende Strategie, um den Rezipienten sicher an die Absprungstelle zu führen." (ebd.).[118] (ii) Der zweite Objekttyp sind Grafikdateien, die in unterschiedlichen Formaten eingebettet werden und als Hyperlinkanzeiger fungieren können. Nach Storrer (2001b) werden sie insbesondere für navigatorische Links verwendet, für die bereits einige Konventionen existieren. So verweist eine Lupe meist auf ein Suchformular und ein stilisiertes Haus führt zurück zur Einstiegsseite (vgl. Bucher, 2001, S. 164 f.). Icons stellen jedoch nur einen Teil der als Anzeiger benutzten Grafiken dar. Häufige Verwendung finden Logos, Diagramme, Abbildungen, *thumbnails* und in Grafikprogrammen entworfene Schriftzüge. (iii) Animierte Grafiken werden Storrer zufolge häufig für E-Mail-Links eingesetzt (z. B. stilisierte Briefe, die zyklisch in einen Briefkasten eingeworfen werden). Auch Werbebanner werden oft als animierte Grafiken realisiert. Storrer (2001b, S. 102) sieht hier eine Parallele zur Bandenwerbung bei Sportveranstaltungen, bei der ebenfalls in regelmäßigen Abständen andere Produkte zu sehen sind; bei Werbebannern kann jedoch immer nur *ein* Produkt (mit wechselnden Motiven) beworben werden, weil nur ein einziger Hyperlink zur Verfügung steht. (iv) Sensitive Grafiken sind entweder statische oder animierte Grafiken, in denen mehreren Bereichen Hyperlinks zugeordnet werden können. Funktional sind *image maps* „vor allem dann, wenn die Analogie zur Landkarte noch greift, d. h. für geografische Übersichten, im übertragenen Sinne für kognitive Landkarten über thematische »Gebiete«, für die Darstellung von Ganzheiten, zu deren Teilen man per Mausklick detailliertere Information abrufen kann." (ebd.).

3.5.6 Textdesign und Webdesign

Die Ebene des Text- und Webdesigns betrifft eine Vielzahl von Aspekten. Die nachfolgenden Ausführungen konzentrieren sich auf die wesentlichsten Unterschiede zwischen Printtexten und Hypertexten im *World Wide Web*.[119] Es wurden bereits diverse Merkmale angerissen, die mit dem Webdesign zusammenhängen, das als Teil des Peritextes eines Knotens aufzufassen ist. Das Webdesign besitzt einen erheblichen Anteil an der Bildung lokaler und globaler Kohärenz, da eine konsistente Gestaltung die Ganzheit des Hypertextes vermittelt, was unter

[118] Diese Ansicht ist aus drei Gründen zu hinterfragen: Die Wörter „klicken" und „hier" tragen keine Bedeutung und sind als redundant einzustufen; Usability-Handbücher schlagen vor, überflüssige Wörter zu vermeiden, da sie das Überfliegen einer Seite und das problemlose Identifizieren von Hyperlinks erschweren. Weiterhin gilt die Aktion des Klickens nicht für alle Anwendungssituationen; neben der Maus kann ein Browser auch mit einem Touchpad oder der Tastatur bedient werden. Abschließend ist es nicht notwendig, Hyperlinks an die Bedürfnisse vermeintlich unerfahrener Benutzer anzupassen (vgl. Abschnitt 3.6.7).

[119] Die gestalterisch-technischen Aspekte werden von der Ratgeberliteratur detailliert erläutert (vgl. z. B. Fleming, 1998, Goodman, 1998, Rosenfeld und Morville, 1998, Siegel, 1999b, Nielsen, 1999 und Reiss, 2000).

anderem durch Navigations- und Orientierungsmittel erfolgt (vgl. Storrer, 2001c, S. 192).[120] Fleming hebt die Bedeutung des Interface-Designs hervor:

> The interface is the intermediary between users and content, an interpreter and guide to the complexities of a site. In the graphical environment of the Web, interface design has to do with constructing visual meaning. The happy marriage of architecture and interface – of logical structure and visual meaning – creates a cohesive user experience. This marriage is crucial to helping users around on the Web. (Fleming, 1998, S. 63)

Bucher (1996, S. 33) stellt im Zusammenhang mit neuen, modular gestalteten Formen der Text- und Informationsanordnung in den Printmedien fest: „Verständlichkeitsurteile sind nicht mehr nur textbezogen, sondern präsentationsbezogen: Neue Kriterien sind Übersichtlichkeit, Anschaulichkeit, Informationsportionierung, Nutzungskomfort und Erschließbarkeit [...] für eine selektive Lektüre."[121] Bei einer Betrachtung des Textdesigns von Online-Zeitungen betont Bucher (1999, S. 12), dass die „Modularisierung zu Textformen [führt], deren Strukturen durch grafische Mittel visualisiert werden. Solche Texte mit optischen Strukturierungshilfen kann man als *visuelle Texte* bezeichnen." Storrer (2001b, S. 89) zieht einen ähnlichen Schluss: „Webgestaltung heißt, [...] aus den [multimedialen] Elementen ein bildschirmgerechtes Ensemble zu flechten." Schmitz (2001, S. 214) beschäftigt sich ebenfalls mit Online-Zeitungen und stellt fest: „Der Bildschirm wird als Textschirm benutzt. Möglichst viel Text wird in möglichst kleinen Portionen feilgeboten und mit grafischen Elementen und Bildern zu einem Plakat komponiert." (vgl. auch Abschnitt 4.6.4).

Modularisierung von Text- und Informationseinheiten

Die spezifischen Eigenschaften der Textualität von Hypertexten (vgl. Abschnitt 3.5.3) sind mitverantwortlich dafür, dass in Knoten oftmals eine modularisierte Anordnung textueller und informationeller Einheiten praktiziert wird (vgl. Siegel, 1999a, S. 249 f.).[122] Weingarten (1997b, S. 216) bezeichnet Hypertext als „eine neuartige Textstruktur, die aus einem Cluster von Textsegmenten besteht." Ein Cluster bildet Weingarten (1997b, S. 217) zufolge „eine minimale Ordnung der Schriftsegmente, Cluster sind frei auf der Schriftfläche verteilt." Weiterhin werden die Begriffe der Aggregation und der Integration eingeführt: „Aggregation bedeutet, daß die Beziehung zwischen sprachlichen Elementen nicht markiert ist – es handelt sich um einen niedrigen sprachlichen Ordnungsgrad. Integration bedeutet, daß durch formale

[120] Schmitz (1997, S. 33) vertritt eine Extremposition: „Ein fortlaufend geschriebener Text bezieht seine Kohärenz aus dem Sinn, der Inhalt einer Bildschirmseite dagegen zunächst aus der Optik."

[121] Die Beiträge auf einer Zeitungsseite können simultan wahrgenommen werden, während „der Bildschirm [...] nur einen begrenzten Einblick in das Gesamtangebot erlaubt", Bucher (2000, S. 683) von „informationeller Kurzsichtigkeit" spricht. Die Metapher des „informationellen Tunnelblicks" scheint adäquater zu sein.

[122] Es ist zu hinterfragen, ob der Begriff „Modul" auf den Gesamtinhalt eines Knotens zu beziehen oder feinkörniger aufzufassen ist, so dass sich mehrere Module in einem Knotens befinden (dies betrifft die Granularität eines Knotens, vgl. Pfammatter, 1998, S. 54). Storrer (2000b) geht von der ersten Lesart aus (vgl. Fußnote 52, S. 84), wohingegen sich sowohl Bucher (1999, 2000) als auch Weingarten (1997b) mit seinem Cluster-Begriff (vgl. auch Bucher, 2000, S. 678) an der zweiten Lesart orientieren. Im Folgenden wird ein geringfügig abstrakterer, an die Informationsmodellierung und Typografie angelehnter Ansatz vertreten, der Module auf – im weitesten Sinne – makrostrukturelle Komponenten bezieht, die Positionen in Knoten besetzen (vgl. Kapitel 5).

Kennzeichen, z. B. durch Flexion, grammatische Funktionswörter, orthographische oder kohäsive Mittel, die Beziehungen zwischen sprachlichen Elementen gekennzeichnet werden." (ebd.). Diese Kennzeichen fallen in Hypertexten zu großen Teilen weg, d. h. die sprachliche Integration nimmt ab, weshalb Hypertexte auf der globalen und auf der Knotenebene Textcluster aufweisen, die aggregativ zusammengeführt werden (ebd., S. 227). Derartige Anordnungsformen sind nicht auf dieses Medium begrenzt. Bucher (1996, S. 35) geht auf typografische und gestalterische Tendenzen in den Printmedien ein und stellt fest: „Aus dem Langtext wird ein Cluster aus verschiedenen visuellen und textlichen Darstellungsformen. Die lineare Präsentation wird durch eine modulare Informationsaufbereitung abgelöst."

Merkmale und Aufgaben des Textdesigns

Die Modularisierung von Knoten, d. h. die Verwendung kurzer, prägnant formulierter und funktional auf der Bildschirmfläche verteilter (Weingarten, 1997b, S. 231) Text- und Informationseinheiten ist eng mit dem von Bucher (1996) eingeführten Begriff Textdesign verknüpft: „Die Gestaltung modularer, mehrmedialer und hypertextueller Formen der Informations- und Wissensvermittlung wird als »Textdesign« bezeichnet [...]." (Bucher, 1999, S. 13). Gerade Tageszeitungen und wöchentlich erscheinenden Magazine tendieren dazu, erläuternde Teiltexte typografisch vom Haupttext abzusetzen, Informationsgrafiken aufzunehmen und Farbräume als Leitsysteme für einzelne Ressorts zu verwenden. Storrer (2001c, S. 182) hebt den Stellenwert dieser Aspekte für Hypertext hervor: „Die im Ansatz des Textdesign [...] angelegte modulare Informationspräsentation *wird im Hypertext zum dominanten Gestaltungsprinzip*: Kleine Info-Module [...] sind durch Links verknüpft, die der Nutzer nach Bedarf und Interesse aktivieren kann." (Hervorhebung hinzugefügt, G. R.).[123] Generell geht es beim Textdesign nach Bucher (1996, S. 32) „immer um die Frage nach dem Verhältnis von Form, Inhalt und Funktion, der Verzahnung von Layout, Textgestaltung und Kommunikationszielen."[124] Textdesign ist also nicht auf den grafischen Gestaltungsaspekt beschränkt.[125] Vielmehr spielen die Informationsziele eine wesentliche Rolle, d. h. „Textdesign ist eine Strategie, um die Lücke zwischen Layout und Text, zwischen Seitengestaltung und Beitragsgestaltung, zwischen Inhalt und Form zu schließen." (Bucher, 1996, S. 33).

Wesentliche Elemente des Textdesigns sind Layout und Typografie (vgl. Sager, 2000), hiermit verbunden ist die modulare Anordnung von Texten und Teiltexten, deren Stil und Formulierungscharakteristika von den übergeordneten Informationszielen abhängig sind. Storrer (2001c, S. 183) weist darauf hin, dass eine Modularisierung von Artikeln erreicht werden kann, indem Gebrauch von der Abschnittsgliederung gemacht wird: „Nach dem Prinzip

[123] Es ist zu beachten, dass *jedes* HTML-Dokument ein gewisses Webdesign besitzt, ebenso wie z. B. jede gedruckte Textseite typografische Eigenschaften aufweist (vgl. Fleming, 1998, S. 64).

[124] Bucher (1996) spricht die Faktoren Inhalt, Form und Funktion an. Diese stellen die Bezugsebene der Arbeiten zum Themenkomplex *Digital Genres* dar, die in Kapitel 4 thematisiert werden (vgl. Abschnitt 2.3.7).

[125] Bucher (1999, S. 31) betont das Ineinandergreifen von Inhalt, Form und Funktion und stellt dar, weshalb Textdesign nicht auf die Form reduziert werden kann, denn „Design-Aufgaben [sind] mehr als Verpackungsaufgaben: selbstexplikative Linksymbole, verständliche Linkbeschriftungen, plausibel strukturierte Sitemaps, visuelle Texte, mit optischer Strukturierung und auf Anforderungen der Internet-Kommunikation abgestimmt, lassen sich nicht mehr auf den Formaspekt [...] beschränken." (vgl. auch Rosenfeld und Morville, 1998, S. 70). Das (in Deutschland erstmals durch das Bauhaus umgesetzte) gestalterische Paradigma „form follows function" gilt im *World Wide Web* daher nur in eingeschränkter Weise (vgl. auch Schmid-Isler, 2000).

»One-idea-per-paragraph« [...] sollte jeder Abschnitt [...] nur genau einen Aspekt [...] behandeln". Morkes und Nielsen (1998) raten dazu, den scannenden Leser (siehe Fußnote 28, S. 76) zu unterstützen, d. h. längere Texte zusammenzufassen, Zwischenüberschriften und Listen zu verwenden, wichtige Textteile visuell hervorzuheben und Inhaltsverzeichnisse zu verwenden, da die Benutzbarkeit und Akzeptanz eines Hypertextes durch den gezielten Einsatz dieser Elemente verbessert werden kann. Bucher (1996, 1999, 2000, 2001) nennt weitere Gestaltungsmittel, die die modulare Textgestaltung positiv beeinflussen: Überschriften, Zwischentitel, Vorspann, Farbleitsysteme, Logos, Piktogramme, Icons, Inhaltsleisten, Frames, Flyouts, Marginalien, Seitenkennzeichnungen, Fettauszeichnungen, optische Gliederungshilfen, Linien und Spalten. Neben der grafischen Komponente visualisieren diese Gestaltungsmittel die Strukturierung des Informationsangebots. Bucher (2000) unterscheidet daher die Informations- von der operationalen Erschließungsebene. Im Kontext eines Hypertextsystems entsprechen diese Ebenen Bucher (2000, S. 676) zufolge „der Unterscheidung zwischen der Informationsbasis einerseits und dem Management- oder Navigationssystem andererseits." Zusätzlich differenziert Bucher (2000, S. 676 f.) „Reichweiten des Textdesigns": (i) beitragsinterne Formen beziehen sich auf Überschriften und Zwischentitel; (ii) modulinterne Formen beziehen sich auf das Layout, Kastenlinien und Pop-up-Fenster; (iii) beitragsübergreifende Formen des Textdesigns sind Signets, Leitfarben und Themenlogos, die Beiträge zu Themen bündeln; (iv) makrostrukturelle Formen sichern die Kohärenz des Gesamtangebots, insbesondere das Seitenlayout, Gliederungsmittel, Inhaltsverzeichnisse und Sitemaps (vgl. Boardman, 2005, S. 31). Im Hinblick auf die modulinterne Form des Textdesigns weist Fleming (1998, S. 64) darauf hin, dass die Visualisierung von Informationshierarchien ein wichtiger Faktor für die Gestaltung eines guten Webdesigns ist: „Visual hierarchies show relationships between elements on a page." Die Einflussgrößen sind hierbei unter anderem (i) die Größen der einzelnen Komponenten, (ii) ihre Position und (iii) farbliche Gestaltung. Die relative Größe eines grafischen oder textuellen Elements kommuniziert Angaben zu seiner Wichtigkeit: Größere Elemente sind wichtiger als kleinere. Überschriften werden z. B. in einer größeren Schrift dargestellt als Fließtext. Auch die Position eines Elements trifft Aussagen über seine Wichtigkeit, denn ein oben links befindliches Element dürfte wichtiger sein, als ein Objekt, das unten rechts positioniert wird. Die farbliche Gestaltung spielt ebenfalls eine wichtige Rolle: Ein leuchtendes Rot vor einem weißen Hintergrund erlangt eher Aufmerksamkeit und signalisiert gleichermaßen Wichtigkeit als z. B. ein heller Grauton.

Nielsen weist im Rahmen einer Plenumsdiskussion auf einen wichtigen übergreifenden Aspekt hin: „Also, your pages are only a minute fraction of the pages seen by any user. In fact, the user experience is dominated by a feeling of "being on the Web" or interacting with the Web as a whole, so your design must form part of this larger whole." (Shneiderman et al., 1998, S. 92).[126] In gewisser Hinsicht plädiert Nielsen für eine konsistente Gestaltung einer Website, es solle eine Anpassung an bekannte Strukturen, d. h. die Gesamtheit anderer, thematisch verwandter Websites stattfinden. Das Design einer Website kann nur dann „part of

[126] Auch die Ratgeberliteratur rät einstimmig dazu, den Benutzer in das Zentrum zu stellen, vgl. z. B. Fleming (1998, S. 2): „To begin solving these pressing problems – user frustration, disappointment, and confusion – we need to begin thinking about this medium in new ways. Instead of obsessing over new gimmicks and digital doodads, we should focus on removing long-standing obstacles between users and goals. Instead of focussing on clicks or hits, we should focus on the entire *user experience*."

this larger whole“ sein, wenn sich die Strukturen und die Gestaltungsprinzipien auf einer funktional-thematischen Ebene ähneln, wenn eine Website also, wie die nachfolgenden Ausführungen zeigen werden, auf einer spezifischen Hypertextsorte basiert (vgl. Abschnitt 3.6.7).

Im Hypertext herrschen für das Textdesign neue Anforderungen: „Zeitungsdesign ist fixiert, Screen-Design variabel und muss deshalb hinsichtlich seiner möglichen Erscheinungsformen durchdacht werden, damit der Leser [...] einheitliche Orte und darüber sinnvolle Ganzheiten entdecken oder aufbauen kann.“ (Schmitz, 2001, S. 213). Im WWW bedeutet gutes Textdesign, die Website konform zu den Standards HTML und CSS zu halten und mit den verbreiteten Browsern zu testen, da sich Unterschiede in der Darstellung ergeben können. Die Variabilität bezieht sich auch auf Faktoren wie die Bildschirmauflösung, die Farbtiefe und die Breite und Höhe des Browserfensters. Das Textdesign sollte nicht von diesen Faktoren abhängig sein, so dass in unterschiedlichen Nutzungssituationen eine identische Darstellung erzielt wird. Die zweite von Schmitz genannte Anforderung betrifft die Bildung globaler Kohärenz. Ein wenig „durchdachtes“ Textdesign kann Kohärenzbrüche verursachen, wobei, wie Ipsen (2001, S. 76) anmerkt, die Adäquatheit des Textdesigns eine wichtige Rolle spielt: „Wird das Design dem Kontext nicht gerecht, so mag ein Hypertext abgelehnt werden, obwohl der Inhalt an sich informativ ist.“[127] In diesem Zusammenhang ist Schmitz (2001, S. 211) der Auffassung, dass in den bisher vorgelegten Textdefinitionen ein „triviales wie entscheidendes Merkmal übersehen wurde“, womit „die Einheit des Ortes, an dem Zeichen stehen“ gemeint ist. Im Kontext eines Hypertextes bedeutet dies, dass der Rezipient im Falle der Wahrnehmung eines „einheitlichen Ortes“ erwarten muss, dass alle „an diesem Ort anzutreffenden Zeichen zusammengehören“ (ebd.). Die gemeinsame Anordnung der Zeichen „stiftet die lektüreleitende Hypothese, dass die hier versammelten Zeichen einen sinnvollen Zusammenhang ergeben, also Text bilden.“ (ebd.).

Funktionen von Bildern in Hypertexten

Besondere Bedeutung für das Text- und Webdesign haben alle Arten von Bildern, z. B. Illustrationen, Logos, Icons, Schriftzüge, zugleich dekorative und funktionale Navigationselemente.[128] Freisler (1994, S. 31) ist der Ansicht, dass sich die klare Grenze zwischen Schrift und Bild auflöst und spricht vom „Synästhetisierungsprozess“, den er als „inhärenten Teil des Hypertextbegriffs“ ansieht.[129] Schmitz (2003, S. 263) fügt hinzu, dass „[d]ie starre Barriere zwischen Textlektüre und Bildbetrachtung fällt“, was darin begründet ist, dass nur einzelne

[127] Bucher (1999, S. 31) fordert, dass zukünftige Kohärenztheorien dem Textdesign mehr Platz einräumen sollten, denn der Begriff „ist [...] als integrative Etikette gemeint: Prinzipien für das Verstehen medialer Kommunikation müssen nicht nur erklären, wie Beitragseinheiten zusammenhängen, sondern auch, wie diese Zusammenhänge mit Hilfe des Text- und Webdesigns zu finden sind.“

[128] Fortanet et al. (1998, 1999) und Crijns (2001) erläutern die Funktion von Bildern in der Online-Werbung.

[129] Bittner (2003, S. 110) greift diesen Begriff auf: „Text, Grafik, Ton, Bewegung, Layout und das Zusammenwirken dieser Elemente haben [...] zu einer neuen Form von »Text« geführt, die nurmehr synästhetisch und räumlich zu erfassen ist. [...] In dieser Form von Text spielt die Schrift eine andere Rolle als in traditionellen analogen Texten, indem sie etwa Strukturierungsfunktionen an Grafik bzw. grafische Elemente oder Verweisfunktionen an Links abgibt.“ Eine genauere Spezifizierung der „neuen Form von Text“ erfolgt dabei nicht. Die genannten Beispiele sind jedoch weder Hypertext-, noch Hypermedia-, noch *World Wide Web*-spezifisch: Die Abgabe von „Strukturierungsfunktionen an Grafik“ ist ein typisches Mittel des Textdesigns, und Verweisfunktionen werden im WWW nicht exklusiv an Hyperlinks „abgegeben“, sondern lediglich durch sie unterstützt.

Teile „innerhalb dieses vieldimensionalen Zeichengeflechts [...] als Texte in jenem klassischen Sinne gelten“ können (Schmitz, 2001, S. 219). Storrer (1999b, S. 4) weist darauf hin, dass „Bild und Text im WWW ein Ensemble bilden, das auf die Rezeption am Bildschirm und deshalb auf eine ganzheitliche Wahrnehmung als Bild ausgelegt ist.“ Die visuellen Komponenten dienen häufig zur Orientierung, sie sind behilflich, die in Bezug auf lineare Texterwartungen deformierten Texte und Teiltexte zu „reparieren, die sich aus der Diskrepanz der zu kleinen Bildschirmfläche und dem multimedialen Durcheinander einer Überfülle an Zeichen ergeben.“ (Schmitz, 2001, S. 219). Ein gezielter Einsatz von Bildern und eine konsistente Gestaltung kann sich positiv auf die Kohärenzbildung auswirken, sie ist aber auch dafür verantwortlich, dass Texte ihrerseits visuellen Charakter annehmen und im Extremfall eine Transformation zu Bestandteilen von Bildern erleben können (Schmitz, 2003, S. 261).

3.5.7 Textlinguistische Analyse von Hypertexten

Huber (2002) schlägt ein Modell zur Analyse von Hypertexten vor. Mehrere Ansätze werden „zu einem konsistenten textlinguistischen Analysemodell für Hypertexte verschmolzen“ (Huber, 2002, S. 100), das „deskriptive Analysekriterien“ bereitstellt (ebd., S. 102), um ein „Gespür für die vorhandenen Ebenen, Beschreibungsaspekte und Analysekategorien [zu] vermitteln, die bei der Untersuchung von Hypertexten eine Rolle spielen.“ (ebd., S. 230).[130] Die Textdefinition von Brinker (2001, S. 17) bildet – mit Bezug auf Storrer (2000b), die die Entwicklung eines neuartigen Textbegriffs für die Charakterisierung von Hypertexten für unnötig erachtet – die theoretische Basis von Hubers (2002, S. 45) Hypertextdefinition: „Hypertexte sind im elektronischen Medium realisierte, tendenziell nicht-lineare und potentiell multimedial ausgerichtete *Texte*.“ Die Tabellen 3.1 und 3.2 stellen das Modell im Überblick dar (Huber, 2002, S. 103 f., vgl. auch Brinker, 2004, S. 149). Es wird deutlich, dass eine *top-down*-Analyse angestrebt wird, die auf der Ebene des Hypertextes ansetzt, der als funktionales Ganzes (Huber, 2002, S. 100) aufgefasst wird und kontextuelle, kommunikativ-funktionale, konventionelle, strukturelle bzw. paratextuelle und intertextuelle Aspekte umfasst. Auf der darunter befindlichen Ebene können Knoten, Abschnitte und Sätze analysiert werden, die identische Inventare von Beschreibungsaspekten aufweisen. Huber (2002) zufolge stellt das Modell ein erweiterbares Grundgerüst dar.[131] Eine derartige Erweiterung nimmt der Verfasser mit einem „Modul“ (Huber, 2002, S. 101) zur Charakterisierung von Hyperlinks selbst vor (vgl. Tabelle 3.2). Das *eigentliche* Modell dient also der Beschreibung der Ebenen „Hypertext“, „Knoten“, „Absatz bzw. Abschnitt“ und „Satz“ (Huber, 2002, S. 103).

Das von Huber (2002) vorgeschlagene Analysemodell ist ein gelungener Versuch zur Etablierung eines kanonischen Inventars von Beschreibungsebenen, der jedoch in verschiedenen

[130] Huber (2002) exemplifiziert das Modell anhand mehrerer Hypertexte und fokussiert visuell typisierte Hyperlinks. Untersucht werden zwei Versionen der HTML-Anleitung „SelfHTML“ (`http://de.selfhtml.org`), das grammatische Informationssystem „GRAMMIS“ (`http://hypermedia.ids-mannheim.de`), das Sportportal „Sport1“ (`http://www.sport1.de`) und die Website des W3C (`http://www.w3.org`). Angaben über die Anzahl der analysierten Hypertexte oder Webseiten werden nicht gemacht.

[131] Huber (2002, S. 231) grenzt das Analysegerüst wiederholt von Modellen zur Erklärung des menschlichen Textverstehens (Kintsch und van Dijk, 1978, van Dijk und Kintsch, 1983) ab. Vielmehr sieht Huber sein Modell als den Beitrag einer Einzeldisziplin für eine nur interdisziplinär zu bewältigende Aufgabe an, denn „von seiten der Textlinguistik“ seien in Bezug auf Hypertext „noch längst nicht alle Verhältnisse geklärt“ (ebd.).

Ebene	Beschreibungsaspekt	Analysekategorie (und evtl. Unterkategorien)	
Hypertext	Kontextuelle Aspekte	Kommunikationssituation	• Kommunikationsform • Handlungsbereich • Autor (Person und Intention) • Zielgruppe(n) • Nötiges Welt- und Fachwissen
	Kommunikativ-funktionale Aspekte	Gesamt-Textfunktion	
	Konventionelle Aspekte	• Textsorte • (Schemataaktivierung durch Textsorte)	• (Textsortenwissen)
	Strukturelle bzw. paratextuelle Aspekte	• Knoten- und Link-Anzahl • Knoten- und Link-Typen (soweit vom Autor definiert)	
		Anwenderspezifischer Peritext	Browser-Software
		Verlegerischer Peritext	Screendesign und Metaphorik
		Allgemeiner Peritext	• Gliederung • Hypertext-interne Navigation • Definierte Lesepfade • Glossare • Suchmaschinen
	Intertextuelle Aspekte	• Implizite Einbettung in einen größeren textuellen Komplex aufgrund der Kommunikationssituation • Explizite Intertexualismen durch extra-hypertextuelle Verweise	
Knoten	Kommunikativ-funktionale Aspekte	Knoten-Textfunktion	Verhältnis zur Gesamt-Textfunktion
	Thematisch-strukturelle Aspekte	• Thema • Themenentfaltung	• Makroproposition • Referentielle Bewegung • Isotopieebenen
	Grammatisch-strukturelle Aspekte	• Kohäsion als Grundlage für Kohärenz • Mögliche kohäsive Lücken (füllbar durch Inferenzen?)	
Absatz	Kommunikativ-funktionale Aspekte	Absatz-Textfunktion	Verhältnis zur Knoten-Textfunktion
	Thematisch-strukturelle Aspekte	• Thema • Themenentfaltung	• Makroproposition • Referentielle Bewegung • Isotopieebenen
	Grammatisch-strukturelle Aspekte	• Kohäsion als Grundlage für Kohärenz • Mögliche kohäsive Lücken (füllbar durch Inferenzen?)	
Satz	Kommunikativ-funktionale Aspekte	Satzfunktion	Verhältnis zur Absatz-Textfunktion
	Thematisch-strukturelle Aspekte	• Thema • Themenentfaltung	• Makroproposition • Referentielle Bewegung • Isotopieebenen
	Grammatisch-strukturelle Aspekte	• Kohäsion als Grundlage für Kohärenz • Mögliche kohäsive Lücken (füllbar durch Inferenzen?)	

Tabelle 3.1: Ein textlinguistisches Analysemodell für Hypertexte (nach Huber, 2002, S. 103)

Ebene	Beschreibungsaspekt	Analysekategorie (und evtl. Unterkategorien)	
Link	Aspekte des Bezugsbereichs	• Aspekte des Ausgangs-Knotens, die mittels Link selektiert werden • Aspekte des Ziel-Knotens, die mittels Link verknüpft werden • Deixis	
	Kommunikativ-funktionale Aspekte	Link-Funktion	• Relation, die der Link zwischen Funktion der Ausgangsressource und der Funktion der Zielressource etabliert • Illokutionäre Beziehung, die über die Knotengrenze hinweg etabliert wird
	Thematisch-strukturelle Aspekte	Themenentfaltung	• Thematische Einbettung des Link-Textes in die Ausgangsressource • Entfaltung des Themas, die mittels des Links über die Ressourcengrenze hinaus geht • Propositionale Beziehung, die über die Knotengrenze hinweg etabliert wird
	Grammatisch-strukturelle Aspekte	Kohäsion	• Kohäsive Einbettung des Link-Textes (bzw. Link-Ankers) in die Ausgangs-Ressource • Grammatische Strukturen, die durch den Hyperlink über die Ressourcengrenze hinaus wirken • Mögliche kohäsive Lücken beim Übergang zwischen den Ressourcen (füllbar durch intendierte und elaborative Inferenzen?)
	Paratextuelle und metasprachliche Aspekte	Signalisierung der Existenz	• Semiotische Hilfsmittel jeder Art, mit denen der Link in irgendeiner Art kenntlich gemacht oder näher spezifiziert wird • Relation zu Beschreibungsaspekten, auf die sich die Spezifizierung bezieht
	Browser-spezifische Aspekte	Traversal-Verhalten	• Verhaltensweise der Browser-Software • Bezugnahme auf kommunikativ-funktionale Aspekte • Bezugnahme auf metasprachliche und paratextuelle Aspekte

Tabelle 3.2: Ein textlinguistisches Analysemodell für Hypertexte (nach Huber, 2002, S. 104)

Aspekten zu kurz greift. Dies betrifft zunächst die offene Architektur: Das Modul zur Analyse von Hyperlinks wird implizit als optional bezeichnet, was überrascht, schließlich sind Verknüpfungen *die* zentrale Eigenschaft von Hypertexten. Die Annahme unterschiedlicher Ebenen ist plausibel: Ein Hypertext besteht aus Knoten, die den Status von Teiltexten besitzen, die wiederum aus Absätzen oder Abschnitten bestehen, die Sätze enthalten. Trotz der Fundierung auf einer pragmatisch-funktionalen Textdefinition erscheint diese Vorgehensweise zu stark im strukturalistischen Textbegriff verankert zu sein; sie könnte jedoch durch den Versuch Hubers erklärt werden, das Modell mit etablierten textlinguistischen und kognitionspsychologischen Methoden kompatibel zu halten (z. B. der RST oder dem Ansatz von Kintsch und van Dijk, 1978). Das nachfolgende Fazit geht genauer auf die Problematik ein, dass die traditionelle Auffassung vom Text als Sequenz von Sätzen für Hypertexte nur noch eine eingeschränkte Gültigkeit besitzt.[132] Unberücksichtigt bleiben auch unterschiedliche Positionierungen der textuellen Bestandteile von Knoten, multimediale Objekte werden nur am

[132] Als Vorgriff ist hier im Detail zu kritisieren, dass Huber (2002) nicht auf unterschiedliche Typen von Abschnitten oder Sätzen eingeht (oder eine weitere Ebene annimmt), z. B. Überschriften, Adressangaben, isolierte Hyperlinks oder Navigationshilfen. Diese Elemente sind als „Satz“ zu konzeptualisieren.

Rande thematisiert. Bezüglich der globalen Ebene eines Hypertextes fallen Unklarheiten auf, so wird nicht deutlich, weshalb sich der Beschreibungsaspekt des „verlegerischen Peritextes“ (vgl. Abschnitt 3.5.2) mit seinen Unterkategorien „Screendesign“ und „Metaphorik“ gerade *nicht* auf die Knotenebene bezieht, schließlich können in unterschiedlichen Knoten durchaus unterschiedliche Metaphern verwendet werden – gleiches gilt für das Webdesign. Auch die Wechselwirkungen zwischen den Ebenen werden nicht unmittelbar deutlich, z. B. bezüglich des Einflusses einer spezifischen „Knoten-Textfunktion“ auf die „Gesamt-Textfunktion“. Huber (2002, S. 101) stellt einen vermeintlichen Vorteil seines Modells heraus, das sich seiner Ansicht nach „sowohl zur Analyse der Hypertext-Struktur als auch zu der einzelner Lesepfade eignet – die Einteilung in Ebenen, Aspekte und Kategorien ist für beide Perspektiven relevant.“ Diese These wird nicht exemplifiziert, vielmehr ist davon auszugehen, dass sich das Modell gerade nicht für diese Zwecke eignet, da kein Mechanismus vorgesehen ist, um die Strukturierung der einzelnen Knoten einer Hypertextbasis zu erfassen. Zwar existiert innerhalb des Aspekts „Allgemeiner Peritext“ eine Unterkategorie „Gliederung“, diese wird jedoch nicht spezifiziert: „Begreift man den gesamten Hypertext als *einen* Text, so spiegelt die Knoten-Struktur die vom Autor gewünschte inhaltliche Gliederung dieses Textes zum Teil wider. Die Knotenüberschriften können somit mit Teilüberschriften traditioneller Texte verglichen werden.“ (Huber, 2002, S. 84). In einer „exemplarischen Makrostrukturanalyse“ eines (fiktiven) Hypertextes geht Huber (2002, S. 88–91) ebenfalls nur am Rande auf dieses Thema ein. Besonders deutlich wird dieses Problem in Bezug auf die Behauptung, das Modell könne „einzelne Lesepfade“ erfassen – dies kann nur mit einem graphentheoretischen Ansatz erfolgen. Außer Acht gelassen wurde bislang die Analysekateogrie „Textsorte“, die lediglich für die Ebene des Hypertextes vorgesehen ist. Dass diese Auffassung ebenfalls zu kurz greift, wird Abschnitt 3.6.9 verdeutlichen. Für Huber (2002, S. 80) „nimmt das Problem der Textsorten jedoch allgemein eine eher untergeordnete Rolle ein.“ Zugleich beschäftigt sich Huber mit der Frage, „ob Hypertext als eigene Textsorte zu sehen ist.“ (ebd.).[133]

3.5.8 Fazit

Bevor im Folgenden das Thema Hypertextsorten fokussiert wird, sind die bisherigen Beobachtungen zu den Spezifika von Hypertexten in Relation zum Textbegriff zu setzen (vgl. Kapitel 2). Jucker (2000, S. 24) beobachtet eine „Abwendung vom Langtext und eine Bevorzugung von Schnipseltexten [...]. Damit wird das Konzept eines Textes immer schwerer fassbar [...]. Die traditionelle Vorstellung, dass ein Text von einem einzelnen Autor geschrieben wird, verliert damit zunehmend an Gültigkeit.“ Mit Bezug auf das gesamte WWW gelangt Jucker (2000, S. 27) zu der Ansicht: „Die Texte auf den einzelnen Rechnern sind nicht mehr selbständige Texte, sondern Teiltexte eines unüberschaubaren und sich ständig weiter entwickelnden Globaltextes.“ Schmitz (2001, S. 218 f.) ist der Meinung, dass der „klassische Textbegriff“ an „multimedialen [...] und hypermedialen Zeichenmengen“ gleichsam „zer-

[133] Als weiterer Vorgriff ist diese Frage entschieden zu verneinen, wie die nachfolgende Charakterisierung von Huber (2002, S. 81) verdeutlicht: „Im WWW sind mit großer Häufigkeit Informations- (Nachrichten, Berichte, Dokumentationen, Sachbücher etc.) und Appellationstexte (Werbung) zu finden, gelegentlich auch Texte mit Obligationsfunktion (Verträge). Während Kontakttexte wie Kondolenzen oder Danksagungen den Weg in das [WWW] seltener finden, sind Deklarationstexte (Testament, Ernennungsurkunde) nahezu gar nicht vertreten.“

schellt", weshalb er „für einen (mono-, poly- oder gegebenenfalls multimedialen) Textbegriff" plädiert. Entgegengesetzter Auffassung ist Storrer (2000b, S. 223), die die Ansicht vertritt „dass kein neuer Textbegriff benötigt wird, dass begriffliche Differenzierungen ausreichen, um Hypertexte als textuelle Gebilde mit ganz spezifischen Eigenschaften an einen pragmatisch und funktional fundierten Textbegriff anzubinden."[134] Ich schließe mich der Argumentation von Storrer an, weil die von Jucker geäußerten Vorbehalte und die von Schmitz geforderten Aspekte nicht im Widerspruch zu modernen, an der Pragmatik orientierten Textdefinitionen stehen (vgl. Huber, 2002): Jucker umschreibt zunächst nichts anderes als eine zunehmende Modularisierung von Texten, die häufig als inhärenter Bestandteil von Hypertext angesehen wird. Der Umstand, dass Jucker die Selbstständigkeit einzelner Texte gefährdet sieht, hängt vermutlich mit dem typischen Nutzerverhalten im WWW zusammen, bei dem die vollständige Rezeption einer Hypertextbasis eher die Ausnahme als die Regel sein dürfte. Stattdessen ist es vom fortwährenden Springen zwischen einzelnen Knoten, dem regelmäßigen Besuch favorisierter Websites (häufig aktualisierte, d. h. „offene Hypertexte" im Sinne von Storrer, 2000b), der Auswahl von URLs aus der Bookmark-Liste und dem Einsatz von Suchmaschinen geprägt. Dieses Rezeptionsverhalten ändert jedoch nichts an der Eigenständigkeit eines Hypertextes, der von einem oder mehreren Autoren auf der Grundlage eines bestimmten Themas mit einer spezifischen Zielvorstellung verfasst und im WWW publiziert wird.[135]

Die vorangegangenen Ausführungen haben gezeigt, dass die Rezeption eines Hypertextes veränderten Kohärenzbedingungen unterliegt, die in großen Teilen auf der visuellen Präsentation einzelner Knoten im Kontext eines Hypertextes basieren und sich weniger explizit im Text selbst manifestieren. Die „Einladung zu flexibler Sinnsuche" (Schmitz, 2001) gilt somit durchaus eher für prototypische[136] Hypertexte als für linear organisierte Texte, jedoch ist dies noch kein hinreichender Grund für die Einforderung eines neuen Textbegriffs. Vielmehr sind, wie Storrer (2000b, S. 244) anmerkt, Modifikationen an tradierten Konzepten vorzunehmen: „Die Beschäftigung mit Hypertext gibt [...] Anlass, sich von zwei Vorstellungen zu verabschieden, die v. a. den strukturalistischen Textbegriff geprägt haben". Gemeint ist zunächst die Vorstellung von der Verkettung von Sätzen zu Abschnitten und von Abschnitten zu einem Text (vgl. Abschnitt 2.2.1). Storrer (2000b, S. 244) zufolge sollte sich „die Perspektive erweitern hin zu den verschiedenen Dimensionen der Textverflechtung, zur Beschreibung von Textmustern und -architekturen." Zum zweiten sollte die „Vorstellung vom abgeschlos-

[134] Siehe auch die vier unterschiedlichen Lesarten des „hyper" in „Hypertext", die Storrer (2000b, S. 231 ff.) als „Mehr-als-Text", „Noch-nicht-Text", „Text-in-Bewegung" und „Interaktiver Text" beschreibt.

[135] Ich beziehe mich hier nicht auf prototypische, unsequenzierte Hypertexte, sondern auf *beliebige* Webseiten (kommunikationsorientierte Hypertextsorten wie das Diskussionsforum oder das Gästebuch werden jedoch explizit ausgenommen, vgl. Abschnitt 4.6.8).

[136] Freisler (1994, S. 37) stellt „die Unterschiede zwischen linearen Texten und Hypertexten" in tabellarischer Form dar (ebd., S. 38 f.), wobei die Globalstruktur („eher netzwerkartig"), das Symbolsystem (der Synästhetisierungsgrad ist „eher hoch"), die Themenentfaltung („Typisch ist die Existenz verschiedener voneinander unabhängiger Strukturen der Themenentfaltung"), Kohäsion („explizite Verweise", „keine feste Leseabfolge") und Kohärenz (von einer „Gesamtkohärenz kann nur noch in eingeschränktem Maße gesprochen werden") diskutiert werden. So plausibel diese direkte Gegenüberstellung auch sein mag, ihr Geltungsbereich bezieht sich ausschließlich auf prototypische Hypertexte. Bezüglich des Aspekts der sozialen Akzeptanz vertritt Freisler (1994, S. 39) die Ansicht, dass für „spezielle Bereiche der Informationsübermittlung" von einer „steigenden Akzeptanz bzw. sogar von einer *Erwartungshaltung von Seiten des Lesers* auszugehen" ist (Hervorhebung hinzugefügt, G. R.). Das nachfolgende Kapitel geht genauer auf dieses Thema ein.

senen Text und von statisch fixierten Textgrenzen“ (ebd.), wie Storrer zu Recht fordert, von einer holistischen Perspektive abgelöst werden, in der Texte als funktionale Ganzheiten angesehen werden, die in übergreifende Handlungszusammenhänge integriert sind. Die für eine pragmatische Sichtweise wesentlichen Kategorien Zweck und Institution sind geeignet, „die im WWW als »Sites« bezeichneten Ganzheiten und die darin eingebundenen Hypertexte in Struktur und Funktion zu beschreiben und [die] Offenheit und Dynamik von Hypertexten in ihrer kommunikativen Leistung adäquat zu erfassen.“ (Storrer, 2000b, S. 245).

Hypertextsorten wurden bislang nur am Rande angesprochen. Sowohl Schmitz (2001) als auch Storrer (2000b) liefern in diesem Zusammenhang implizite Hinweise, die andeuten, dass im WWW Veränderungen festzustellen sind. Schmitz (2001, S. 218) vermutet, dass die „neuen technischen Bedingungen [...] Textsorten [...] nach sich [ziehen]“. Und Storrer vertritt im Rahmen einer Diskussion der Unterschiede von Text und Hypertext den Standpunkt,

> dass es in einer Zeit medialer Übergänge und raschen Textsortenwandels weniger darauf ankommt, nach trennscharfen Kriterien und harten Definitionen zur Ausgrenzung von nicht zur Textlinguistik gehörigen Gegenständen zu suchen. Es gilt vielmehr, *die neuen Textarten und Kommunikationsformen im Internet* in Beziehung zu setzen mit Textarten und Kommunikationsformen, die bereits im Rahmen der Textlinguistik, der interdisziplinären Textverstehens- und Textproduktionsforschung oder der Gesprächsanalyse untersucht worden sind. (Storrer, 2000b, S. 235; Hervorhebung hinzugefügt, G. R.)

Storrer spricht in ihrem Plädoyer für textlinguistische Untersuchungen digitaler Medien von „neuen Textarten im Internet“ – derartige Hinweise auf die Existenz von Hypertextsorten sind insbesondere in der neueren deutschsprachigen Literatur zu finden, in der das Medium Hypertext bzw. der Internet-Dienst *World Wide Web* aus linguistischer Sicht betrachtet wird, die Hinweise bleiben dabei jedoch meist relativ vage. Der nachfolgende Abschnitt geht in detaillierter Form auf Indikatoren ein, die die Existenz von Hypertextsorten nahe legen.

3.6 Indikatoren für die Existenz von Hypertextsorten

In textlinguistischen Arbeiten wird Hypertext vielfach als Textsorte bezeichnet (vgl. Fußnote 1, S. 65). Im Folgenden werden insgesamt 54 Indikatoren diskutiert, die auf eine Existenz von *Hypertexttypen* und *Hypertextsorten* hindeuten. Diese Termini werden zunächst als im Medium Hypertext, genauer gesagt, im Hypertextsystem *World Wide Web* realisierte Textklassen (vgl. Abschnitt 2.3.2) verstanden, die einen sehr großen Geltungsbereich und nur wenige differenzierende Merkmale (Hypertexttypen) sowie einen geringeren Geltungsbereich und entsprechend mehr distinktive Merkmale besitzen (Hypertextsorten). Aus Illustrationsgründen kann eine Gegenhypothese formuliert werden: Hypertexte im WWW können *nicht* aufgrund distinktiver Merkmale voneinander unterschieden werden, sie sind in konzeptioneller, struktureller und sprachlicher Hinsicht nicht differenzierbar (also identisch), weshalb das Medium Hypertext selbst als Textsorte eingestuft werden muss.

Die Indikatoren, die diese bereits intuitiv äußerst unplausibel erscheinende Gegenhypothese widerlegen, stammen aus unterschiedlichen Themengebieten und verwenden nur teilweise linguistische Termini. Dennoch können zahlreiche Konzepte herausgearbeitet werden, die Begriffe wie Hypertexttyp und Hypertextsorte reflektieren. Dieser Abschnitt fungiert als

konzeptuelle Überleitung zwischen der in Kapitel 2 eingeführten textlinguistischen Ebene, den hypertexttheoretischen (Abschnitt 3.3) sowie linguistisch ausgerichteten Begrifflichkeiten (Abschnitt 3.5) und den nachfolgenden Kapiteln. Abschnitt 3.6.1 geht auf die Navigation in Texten und Hypertexten ein. Abschnitt 3.6.2 diskutiert den Einfluss von Textsorten auf die Konversion, woraufhin Abschnitt 3.6.3 Indikatoren für die Entstehung von Konventionen im WWW präsentiert. In Abschnitt 3.6.4 schließen sich Hinweise an, die insbesondere mediale und linguistische Aspekte von Hypertextsorten berühren. Die Abschnitte 3.6.5 und 3.6.6 nennen Indikatoren, die aus den Bereichen Informationsarchitektur, Webdesign und Metadatenschemata stammen. Abschnitt 3.6.7 erläutert den Zusammenhang zwischen Hypertextsorten und Erwartungshaltungen. Aus der Sammlung von Indikatoren werden abschließend zwei initiale Typologien abgeleitet (Abschnitt 3.8).

3.6.1 Zur Navigation in gedruckten Dokumenten und Hypertexten

McKnight et al. (1991, S. 67 ff.) gehen aus kognitionspsychologischer Perspektive auf Faktoren ein, die die Navigation und Orientierung in Dokumenten betreffen. Hierbei ist der Begriff des Schemas von Bedeutung (vgl. Abschnitt 2.2.6): Menschen eignen sich Schemata über Konzepte an, die als Orientierungsrahmen für das zukünftige Handeln und Navigieren dienen. McKnight et al. exemplifizieren dies an Schemata von Städten, in denen sich Straßen, Gebäude, Bürgersteige, Industrie- und Wohngebiete, Einwohner, Kirchen und Restaurants befinden. Die Menschen, die in einer Stadt leben, besitzen ausgeprägtes Wissen und können die kürzeste Route zwischen zwei Punkten beschreiben. Solche kognitiven Karten entstehen in mehreren Entwicklungsstufen, die die initiale Orientierung anhand salienter Landmarken und das Erlernen rudimentären sowie detaillierten Überblickswissens umfassen. Derartiges Schemawissen eignen sich Menschen auch für Papierdokumente an (ebd., S. 70 ff.): Anhand eines Buchumschlags kann der Inhalt eingeschätzt werden, auf den ersten Seiten können ein Inhaltsverzeichnis und Angaben über die Autoren erwartet werden, Seitennummern erlauben die Ansteuerung einzelner Seiten und in Fachbüchern befindet sich am Ende ein Literatur- und Schlagwortverzeichnis. Die generische Strukturierung von Textsorten eignen sich Menschen aufgrund des individuellen Umgangs mit Textexemplaren an – van Dijk und Kintsch (1983) sprechen von Superstrukturen. Deren Existenz wurde anhand empirischer Versuche mit verschiedenen Textsorten gezeigt, wobei die Platzierung von Textabschnitten und die Benennung einzelner Kapitel und Unterkapitel einen signifikanten Einfluss auf das Verstehen eines Textes haben, sofern sie der kanonischen Superstruktur entsprechen.

Wenn das räumliche Navigieren auf den Umgang mit einem neuen Buch übertragen wird, können Analogien gefunden werden, die den Aufbau kognitiver Karten des Informationsraums betreffen (McKnight et al., 1991, S. 72 ff.). Generische Elemente wie Kapitelüberschriften, Zusammenfassungen, Seitenzahlen und Inhaltsverzeichnisse fungieren als Landmarken, die Hinweise auf die aktuelle Position im Text liefern. Es wurde gezeigt, dass derartige saliente, kontextuelle Textbestandteile zur Orientierung verwendet werden: Leser können sich daran erinnern, dass sich wichtige Detailinformationen an spezifischen Textpositionen befinden. Die als Landmarken operierenden Textteile dienen der Orientierung und zur Benennung der Position. Die Analogien zum Aufbau kognitiver Karten – insbesondere hinsichtlich des in Texten unnötigen Routenwissens – sind jedoch begrenzt (ebd., S. 73).

McKnight et al. (1991, S. 75 ff.) beziehen die bislang nur für Papierdokumente geltenden Beobachtungen auf Hypertexte. Diese Überlegungen entstanden im Jahr 1991 und obwohl zur gleichen Zeit noch an den Basistechnologien des WWW gearbeitet wurde (Berners-Lee et al., 1992), können verschiedene Indikatoren für die Existenz von Hypertextsorten identifiziert werden.[137] Zunächst charakterisieren McKnight et al. (1991)[138] die Rahmenbedingungen zur Bildung von Schemata in digitalen Informationsräumen:

> 1. The concept of a schema for an electronic information space is less clear-cut than those for physical environments or paper documents. Electronic documents have a far shorter history than paper and the level of awareness of technology among the general public is relatively simple [...]. Exposure to information technology will almost certainly improve this state of affairs, but even among the contemporary computer-literate it is unlikely that the type of generic schematic structures that exist for paper documents have electronic equivalents of sufficient generality. (McKnight et al., 1991, S. 75)

McKnight et al. vertreten den Standpunkt, dass das Schemakonzept in digitalen Dokumenten nicht klar umrissen ist, weil sie, im Vergleich zu Papierdokumenten, noch relativ jung sind (sich also noch keine Schemata etablieren konnten) und ein Großteil der Menschen nicht mit ihnen in Kontakt kommt. Der zunehmende Umgang mit der Informationstechnologie könne jedoch, so McKnight et al., die Situation schlagartig ändern, wobei sich jedoch ihrer Ansicht nach keine so generellen Schemata wie bei traditionellen Dokumenten entwickeln dürften. Diese im Jahr 1991 durchaus gültigen Aussagen müssen unter Bezugnahme auf den Erfolg des WWW revidiert werden. Dieser deutet an, dass sich mittlerweile auf einer zunächst nicht näher bestimmbaren Granularitätsstufe Schemata etabliert haben könnten.

In Bezug auf das Medium Hypertext erläutern McKnight et al. (1991, S. 75 f.) verschiedene Probleme: Der Anwender kann den Umfang eines Hypertextes (im Gegensatz zu einem Buch) nicht unmittelbar einschätzen, dies gilt auch für seinen Inhalt, sein Alter und die Anzahl der Zugriffe. Auch der grundlegende Modus operandi ist in diesem Medium unterschiedlich: Während das Umblättern einer Seite in jedem gedruckten Text zur nachfolgenden Seite führt, muss dies in einem Hypertext aufgrund seiner multilinearen Organisation nicht notwendigerweise der Fall sein, d. h. eine möglicherweise vorliegende Erwartungshaltung kann aufgrund der medieninhärenten Varianz in puncto Organisation und Darstellung von Knoten verletzt werden.[139] Dennoch räumen die Verfasser ein, dass gerade häufige Benutzer dieses Mediums Schemata interiorisiert haben könnten:

[137] Es wird (vereinfachend) davon ausgegangen, dass die Konzepte „Schema“ und „Textsorte“ in diesem Kontext, in dem es lediglich um die Existenz konventionalisierter, generischer Abstraktionen von Dokumentstrukturen geht, synonym verwendet werden können. Dillon und Vaughan (1997) hinterfragen in einer Folgearbeit die Navigationsmetapher und schlagen vor, dass unter anderem das Genre (im Sinne der *North American Genre Theory*, vgl. Abschnitt 2.3.7) Einzug in Evaluationen von Hypertexten finden sollte, um empirisch bei Benutzerbeobachtungen erhobene Daten über die Navigation in Informationsräumen mit primär soziokulturell geprägten Superkonzepten zu komplementieren.

[138] Die in diesem Abschnitt als Indikatoren aufgeführten Zitate werden typografisch abgesetzt und fortlaufend nummeriert, um Querverweise zu spezifischen Hinweisen zu ermöglichen.

[139] Diese Ausführungen gelten zwar allgemein für Hypertext- und Hypermediaanwendungen, in dem skizzierten Ausmaß jedoch nicht für HTML, denn „the original Web/HTML model has one big advantage: All Web pages look and work more or less the same. You see something black, you read it. You see something gray, that's the background. You see something blue or underlined, you click on it. When you use a set of *traditional* Web sites, you don't have to learn anything new.“ (Greenspun, 1999, S. 132; Hervorhebung hinzugefügt, G. R.).

> 2. At the current stage of development, it is likely that users or readers familiar with hypertext will have a schema that includes such attributes as linked nodes of information, non-serial structures, and perhaps even potential navigation difficulties! The manipulation facilities and access mechanisms available in hypertext will probably occupy a more prominent rôle [sic] in their schemata for hypertext documents than they will for readers' schemata of paper texts. As yet, empirical evidence for such schemata are lacking. (McKnight et al., 1991, S. 76)

Wird der Ansicht gefolgt, dass die Zugriffsmöglichkeiten eine wichtige Rolle für die Etablierung von Schemata spielen, so kann dies als weiteres Indiz für die Existenz von Hypertextsorten im WWW gelten, denn die verbreiteten Browser besitzen sehr ähnliche Navigationsfunktionen und stellen Dokumente in den meisten Fällen (nahezu) identisch dar.[140] Dennoch sind die Gemeinsamkeiten mehrerer Hypertexte nach McKnight et al. aufgrund der gesteigerten Varianz dieses Mediums auf einer konzeptuell höheren Ebene anzusiedeln:

> 3. [U]sers' schemata of hypertext environments are likely to be 'informationally leaner' than those for paper documents. This is attributable to the recent emergence of electronic documents and comparative lack of experience interacting with them as opposed to paper texts for even the most dedicated users. The lack of standards in the electronic domain compared to the rather traditional structures of many paper documents is a further problem for schema development [...]. (McKnight et al., 1991, S. 77)

Alle genannten Einschränkungen können zum aktuellen Zeitpunkt als nicht mehr gültig bezeichnet werden: Mittlerweile gehen sehr viele Menschen – nicht nur die „most dedicated users" – im beruflichen und privaten Alltag mit elektronischen Dokumenten, E-Mails, Webseiten, Textverarbeitungsdateien etc., wie selbstverständlich um. Ein Mangel an Standards – eine grundlegende Prämisse für die allmähliche Entwicklung von Textsorten – existiert ebenfalls nicht mehr, wie die nachfolgenden Abschnitte zeigen werden.[141]

3.6.2 Der Einfluss von Textsorten auf die Konversion

Die Hypertextierung eines vorhandenen Textes wird als Konversion bezeichnet (vgl. Abschnitt 3.3.6). Es stellt sich die Frage, welchen Einfluss die Textsorte des vorliegenden Textes auf diesen Prozess besitzt. Die Literatur gibt auf diese Frage übereinstimmende Antworten, von denen einige nachfolgend exemplarisch aufgeführt werden. Textsorten, denen eine deutlichere Affinität für eine Umsetzung als Hypertext inhärent ist, dürften auch häufiger im WWW zu finden sein, wohingegen dies bei Exemplaren von Textsorten, bei denen die Konversion nur mit viel Aufwand oder ohne ersichtlichen Mehrwert durchgeführt werden kann, seltener der Fall sein dürfte. Es handelt sich bei diesen Indizien somit primär eben nicht um

[140] McKnight et al. (1991, S. 78 ff.) gehen auch auf den Aufbau kognitiver Karten bei der Rezeption von Hypertexten ein. Es werden Studien diskutiert, in denen Navigationsprobleme festgestellt wurden, weshalb Landmarken als Orientierungsmittel vorschlagen werden, um den Aufbau kognitiver Karten zu unterstützen (hierzu raten auch Rosenfeld und Morville, 1998, S. 50 f.). Unter textlinguistischer Perspektive sind derartige Landmarken als kohärenzstiftende Überblicks- und Navigationshilfen sowie Bestandteile des Textdesigns aufzufassen (vgl. die Abschnitte 3.5.3 und 3.5.6).

[141] Der „lack of standards" bei McKnight et al. (1991, S. 77) bezieht sich nicht auf Technologiestandards, sondern auf standardisierte Textstrukturen, eben die „rather traditional structures of many paper documents".

Indikatoren für die Existenz von Hypertextsorten, sondern vielmehr um Hinweise darauf, aus welchen Klassen von Textsorten die Entstehung von Hypertextsorten zu erwarten ist.

Rada (1991, S. 58 ff.) unterscheidet zwischen „clearly" und „implicitly structured text" (ähnlich in Rada, 1995, S. 25). Zur ersten Kategorie gehören insbesondere Verzeichnisse:

> 4. A prime example of clearly structured text is a *directory*. Technical manuals, dictionaries, encyclopedias, [...], and bibliographies are like directories. (Rada, 1991, S. 58)

Sofern ein eindeutig strukturierter Text vorliegt, kann eine Konversion leicht durchgeführt werden. Die zweite Kategorie, implizit strukturierter Text, bezieht sich auf

> 5. text whose explicit logical structure is minimal. The extreme case is an *essay* which has no subdivisions or other logical decomposition. A *novel* likewise may often be an extended stream of consciousness for which the logical structure is not suggested in the layout of the document and is not indicated by the markup language which may have been used on the computer. To translate implicitly structured text into hypertext requires substantial human effort to be invested in characterizing the relations among components of the document. (Rada, 1991, S. 58 f.)

Kürzere, in sich kohärente Konglomerate von Texten sind also eher für eine Umsetzung als Hypertext geeignet. Freisler (1994) führt ein Beispiel an, das sich wie roter Faden durch die Hypertextliteratur zieht (vgl. hierzu auch S. 76):

> 6. Hypertexte scheinen nicht neu zu sein, denn es existieren strukturelle Ähnlichkeiten zwischen Hypertexten und bekannten Textsorten (z. B. Lexika), die so bestechend sind, daß es einen auch nicht wundert, daß diese Textsorte zu den ersten gehörte, bei denen eine Konversion in ein Hypertextformat ernsthaft versucht wurde. (Freisler, 1994, S. 19)

Lutz (1995, S. 159, S. 162) nennt weitere für eine Umsetzung als Hypertext prädestinierte Textsorten und geht auf generelle Eigenschaften ein, die für eine Konversion vorgesehene Texte idealerweise besitzen sollten:

> 7. Sämtliche Texte, bei denen schneller Zugriff auf definierte Teiltexte interessant ist, sind potentielle Kandidaten. Die Voraussetzung dazu sind abgrenzbare Einheiten (sowohl formal als auch inhaltlich), die mit anderen Einheiten in Zusammenhang stehen. Damit reicht das Spektrum vom Idealfall Lexikon über Lehrtexte bis hin zu allen deskriptiven Texten (sämtliche technische Texte, wissenschaftliche Texte, aber auch Gesetzestexte in ihrer paragraphenorientierten, wirklichkeitsschaffenden Diktion). (Lutz, 1995, S. 162)

Es sind demnach die Textsorten geeignet, die aus vielen kohäsiv geschlossenen und in sich kohärenten Subtexten bestehen. Querverweise können unmittelbar auf Hyperlinks übertragen werden, um die häufig als Idealform aufgefasste unsequenzierte Strukturierung eines Hypertextes zu erzielen (vgl. Abschnitt 3.5.1).[142] Die weiteren von Lutz genannten Textsorten weichen bereits deutlich von diesen Voraussetzungen ab, so sind z. B. wissenschaftliche Texte üblicherweise linear organisiert (können aber aufgrund ihres konventionalisierten Aufbaus durchaus auch in alternativen Reihenfolgen rezipiert werden, z. B. erst die Zusammenfassung, dann die Literaturliste und anschließend die Schlussfolgerungen). Lutz (1995, S. 159) geht auf weitere Charakteristika ein, die sich auf die Ebene des Texttyps beziehen:

[142] Vora und Helander (1997, S. 905) nennen drei Kriterien, die der vorhandene Text im Falle einer Hypertextierung aufweisen sollte: „(i) A large body of information organized into numerous fragments, (ii) The fragments relate to each other, (iii) The user needs only a small fraction at any time."

8. *Deskriptive Texte* dürften sich am besten für Hypertextierung eignen. Bei *argumentativen Texten* entsteht das Problem, daß die Argumentationslinie nicht durch die Trennung in mehrere Knoten durchbrochen werden darf [...]. *Narrative Texte* schließlich dürften aufgrund der sequentiellen Anordnung der Inhalte am wenigsten für die Hypertextierung geeignet sein, wobei es allerdings bereits eine Tradition der Hyperfiction gibt (der Leser *erliest* sich seinen eigenen Roman). Voraussetzung dazu sind jedoch klar komponierte Einsprungstellen bei sonstigem Verlust inhaltlicher Stimmigkeit. (Lutz, 1995, S. 159)

Hypertextsorten dürften sich primär in Bezug auf deskriptive Texte bilden. Narrative Texte erfahren nur dann einen Mehrwert, wenn der Autor über das schlichte Abbilden der Kapitel einer Erzählung in monosequenzierte Knoten hinausgeht und Modifikationen des Textes vornimmt, um zahlreiche „klar komponierte Einsprungstellen" und somit Unsequenziertheit zu erzielen (Smoliar und Baker, 1997). Kuhlen (1991, S. 174 ff.) geht auf Textsorten ein, die für eine Konversion geeignet und weniger geeignet sind, weist jedoch darauf hin, dass eine solche Zuordnung „problematisch" ist (ebd.). Die Begründung, weshalb diese Textsorte geeigneter erscheint als jene, führt letzten Endes zu der Frage, welche Merkmale einen „guten" Hypertext auszeichnen (vgl. Fußnote 69, S. 91).[143] Kuhlen (1991, S. 176) vermeidet eine Beantwortung und nimmt an, „daß es zunächst einmal ein Gewinn sein kann, wenn Texte der [geeigneten] Art konvertiert werden." Vielmehr geht es Kuhlen um eine Aufstellung von „Konversionsargumenten" (ebd.), die z. B. als Kriterien herangezogen werden können, wenn beabsichtigt wird, eine aufwändige und kostenintensive Konversion vorzunehmen. Tabelle 3.3 zeigt die von Kuhlen (1991, S. 175 f.) aufgestellten Listen.[144]

Hammwöhner (1997, S. 59 ff.) beschäftigt sich mit dem Einfluss von Textsorten auf das Medium Hypertext und geht zunächst von deskriptiven, narrativen und argumentativen Texten aus. Es existieren Hypertextsysteme, die „gezielt einzelne dieser Textsorten unterstützen" (ebd., S. 60). Hammwöhner nimmt an, dass für eine linguistische Charakterisierung des Mediums Hypertext kein neuer[145] Textbegriff benötigt wird, denn

9. allein die systematische Beziehung zwischen Texttypologie und diskursiver Funktion garantiert [...] die Anwendbarkeit der Texttypen auf Hypertexte, denn die Ziele des Erklärens, Erzählens oder Argumentierens werden nicht [...] durch einen Wechsel der medialen Aufbereitung verändert. (Hammwöhner, 1997, S. 60)

[143] Diese Frage kann für einen generischen Hypertext bzw. das Medium Hypertext ebenso wenig pauschal beantwortet werden wie für einen generischen (gedruckten) Text bzw. das Medium Papier.

[144] Interessanterweise bezeichnet Kuhlen (1991) Gesetzestexte als Textsorte, die für eine Umsetzung als Hypertext ungeeignet sind, wohingegen Lutz (1995, S. 162) eine entgegengesetzte Meinung vertritt (vgl. Indikator 7).

[145] Mit Bezug auf Landow (1992, S. 101) geht Hammwöhner (1997) auf Veränderungen ein, die narrative Texte aufgrund der nichtlinearen Organisation erfahren. Jedoch ist diese „nicht-lineare Erzählweise [...] aber nicht dem Hypertext vorbehalten, sondern für moderne Literatur [...] nachgerade als prägend anzusehen." (ebd., S. 60). Dies betrifft die mediale und konzeptionelle Linearität (vgl. Storrer, 2000b, sowie Abschnitt 3.5.1), was durch einen Vergleich mit dem Medium Film besonders deutlich wird (vgl. Mancini und Shum, 2004, und Essid, 2004). Dieses wird dominiert von einer unüberwindbaren medialen Linearität (vgl. den von Nelson, 1965, S. 96, eingeführten Terminus „hyperfilm"), d. h. nichtlineare Erzählstrukturen entsprechen konzeptioneller Nichtlinearität: *Citizen Kane* (Orson Welles, 1941) und *Pulp Fiction* (Quentin Tarantino, 1994) beziehen einen Großteil ihrer Faszination aus dem Spiel mit unterschiedlichen chronologischen Ebenen. Besonders ausgeprägt ist die konzeptionelle Nichtlinearität in *Memento* (Christopher Nolan, 2000). *Alle* Szenen dieses Films sind entgegengesetzt zur Chronologie der dargestellten Ereignisse montiert. Die DVD-Ausgabe bietet eine versteckte Funktion, um den Film in der chronologischen Reihenfolge abzuspielen.

Geeignete Textsorten	Ungeeignete Textsorten
• Texte, deren Inhalte sich in distinkte Blöcke zergliedern lassen, z. B. Lexika oder technische Handbücher • Texte, die nach einheitlichen Kategorienschemata strukturiert sind • Texte, die benennbare und formalisierbare Beziehungen zwischen Informationseinheiten aufweisen, z. B. „siehe auch"-Relationen in Handbüchern oder Enzyklopädien • Texte, die strukturelle Metainformationen, wie Register, Glossare, Abkürzungsauflösungen, enhalten • Texte mit statischen, weitgehend abgeschlossenen beziehungsweise abgesicherten Wissensstrukturen • Lerntexte, die aus didaktischen Gründen argumentativ klar aufgebaut und aus sich heraus verständlich sind, sich also wenig auf externe Referenzen abstützen	• Textsequenzen, bei denen durch die Entlinearisierung ein heillos verwirrendes Netz aus atomisierten Einheiten und bedeutungslosen Verknüpfungen entsteht • Textsorten mit Wissensstrukturen, die sich ständig ändern, daher strukturell sowie inhaltlich aktualisiert werden müssen, z. B. Gesetzestexte, Vorschriften • Textsorten, die auf eine bestimmte Präsentationsform angewiesen sind, z. B. Argumentationen, Gedankenentwicklungen, essayistische Texte, Kriminalromane • Textsorten, die eine organische inhaltliche Strukturierung aufweisen, die ohne Beeinträchtigung der Gesamtaussage nicht aufgebrochen werden kann, z. B. Kunstprosa • Ausführliche Beurteilungen, Wertungen und Interpretationen von Faktenmaterial bzw. Beweisführungen und Herleitungen von Begründungszusammenhängen

Tabelle 3.3: Für eine Konversion in das Medium Hypertext geeignete und ungeeignete Textsorten (nach Kuhlen, 1991, S. 175 f.)

Da deskriptive, narrative und argumentative Texte nur selten in Reinform existieren, geht Hammwöhner (1997, S. 61) auf eine Typologie ein, die auf Textemen als „geschlossene Textsegmente mit homogener Funktion" basiert. Bezüglich der Rezeptionsreihenfolge wird zwischen freiem, narrativem, restringiertem und koordiniertem Text unterschieden, wobei Hammwöhner eine „Entsprechung von Hypertextknoten und Textemen als funktional abgeschlossene Textteile postuliert" (ebd.). Hierdurch können die Beschränkungen von Textsorten hinsichtlich der positionellen Vorkommen von Textemen bzw. Teiltexten im Medium Hypertext durch eine Typisierung der Knoten explizit gemacht werden.

> 10. Thematisch zusammenhängende Teile des Hypertexts werden zu Analoga von Texten zusammengefaßt, indem sie in Strukturknoten oder ähnliche aggregierende Hypertextobjekte eingefügt werden. Die Lesereihenfolge [...] kann dabei einerseits durch eine explizite Ordnung innerhalb des Strukturknotens im Sinne eines Pfades vorgegeben sein oder durch Verknüpfungen [...]. (Hammwöhner, 1997, S. 61 f.)

Strukturknoten fassen thematisch ähnliche Knoten zusammen, ihre Verknüpfung kann Restriktionen unterliegen. Hammwöhner (1997, S. 62) differenziert zwischen freiem (keine Restriktionen), fixiertem (nur die vorgegebene Verknüpfung ist möglich), restringiertem (es können alternative Lesestrategien vorliegen) und koordiniertem Text (Vertauschung von Lesesequenzen ist möglich; vgl. Abbildung 3.4).[146] Zwei Aspekte werden deutlich: Zwischen dem Sequenziertheitsgrad eines Hypertextes und seiner Hypertextsorte existiert eine Verbindung (vgl. Indikator 27, S. 133) und es ist möglich, dass in einen Hypertext mehrere Hypertextsorten eingebettet sein können (vgl. die Abschnitte 3.6.5 und 4.5).

[146] Abbildung 3.4 zeigt fünf „Textsegmente" (Hammwöhner, 1997, S. 63), d. h. Gruppen eigenständig lesbarer Hypertextknoten (A–E). Die Strukturierungsformen sind schematisch und entsprechen drei „Texttypen" (ebd.), die Nummerierung gibt die Lesereihenfolge an. Um einzelne Knoten können sich Metatexte gruppieren, siehe Strukturknoten C, zu dem Hammwöhner weiter ausführt: „Knoten C ist fixiert, er erlaubt keine Abweichung im Leseverhalten." Hier liegt ein Fehler vor, gemeint ist E. Die Knoten A und C sind restringierte Texte, die interne Abweichungen gestatten, während B und D freies Navigieren erlauben.

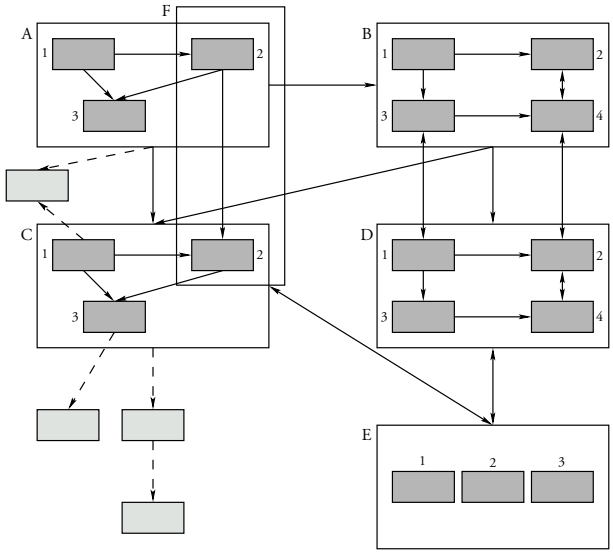


Abbildung 3.4: Idealisierte Hypertextstruktur (nach Hammwöhner, 1997, S. 63)

3.6.3 Zur Entstehung von Konventionen im World Wide Web

Abstrakt betrachtet umfassen Textsorten konventionalisierte Textstrukturmuster und kommunikative Funktionen. Schmitz (2001, S. 207) diskutiert Online-Zeitungen und bezeichnet sie als „noch ungewohnte Medien“:

11. Wie es bei jedem neuen Medium (Schrift, Buch, Telefon, Radio, Fernsehen u. a.) anfangs der Fall war, bilden sich neue Formen der Darstellung und Anordnung von Informationen derzeit erst noch heraus. (Schmitz, 2001, S. 207)

Wird dieses Argument unter Beachtung des noch jungen Alters sowie der Verbreitung des WWW konsequent weitergeführt, können die „neuen Formen der Darstellung und Anordnung von Informationen“ letzten Endes in Hypertextsorten resultieren. Auch Hammwöhner (1997, S. 51) geht auf diesen Aspekt ein, bezieht sich dabei aber nicht speziell auf das WWW:

12. Konventionen über den Aufbau von Hypertexten [...] sind zur Zeit allenfalls im Entstehen begriffen [...]. (Hammwöhner, 1997, S. 51)

Storrer (2001c, S. 174) stellt – ebenfalls im Kontext von Online-Zeitungen – ein zentrales Verbindungsglied zwischen alten und neuen Medien heraus:

13. Die Mediengeschichte hat gezeigt, dass der Umgang mit neuen Medien anfänglich immer von Darstellungsformen alter Medien geprägt ist, dass sich erst mit der Zeit mediengerechte journalistische Formen herausbilden [...]. Dieser Prozess ist im WWW noch im Gange. (Storrer, 2001c, S. 174)

Wie Hammwöhner (1997) bezieht sich auch Storrer (2001c) nicht unmittelbar auf die Dimension von Textsorten, sondern auf „mediengerechte journalistische Formen“, sie fokussiert also einen Texttyp. Wenn Storrer feststellt, dass der Prozess ihrer Herausbildung noch

im Gange ist, ist dies zweifelsohne so zu interpretieren, dass er bereits angestoßen wurde, dass derartige Formen in der Entstehung befindlich sind.[147] Nielsen und Tahir (2002, S. 3) beschäftigen sich mit Homepages und bestätigen diese Ansicht:

14. Websites need to represent many more services on the homepage because the genre is less established than newspapers. (Nielsen und Tahir, 2002, S. 3)

Wenn die Textsorte also „less established than newspapers" ist, sollten Konventionen zumindest in ersten Ansätzen festzustellen sein. Freisler (1994, S. 20) gibt einen Hinweis auf mögliche Vorbilder:

15. Während die ersten Drucke zunächst versucht haben, Handschriften nachzuahmen, versuchen noch viele Hypertexte, gedruckte Bücher nachzuahmen, weil sie so (zumindest kurzfristig) mit der Zustimmung der Leser rechnen können. (Freisler, 1994, S. 20)

Dieser Artikel ist zu einer Zeit erschienen, in der das WWW außerhalb informatischer Institute und Rechenzentren kaum bekannt war, weshalb sich Freisler, wie Hammwöhner, in genereller Weise auf das Medium Hypertext bezieht. Da die Leser eines Hypertextes mit Büchern bzw. auf Papier gedruckten Texten vertraut sind, können Begriffe wie „Seite", „Inhaltsverzeichnis" und „Glossar" metaphorisch verwendet werden, um die Akzeptanz zu erhöhen (vgl. Abschnitt 3.5.4). Storrer (2001c, S. 192) betont jedoch – mit Bezug auf Gestaltungskonventionen in Tageszeitungen – sieben Jahre nach der Veröffentlichung des Beitrags von Freisler (vgl. Indikator 13):

16. Im WWW haben sich solche konventionalisierten Gestaltungsformen noch nicht herausgebildet. (Storrer, 2001c, S. 192)

Es stellt sich nun die Frage, ob denn überhaupt Regularitäten in WWW-basierten Hypertexten zu identifizieren sind, oder ob es sich lediglich um gleichsam amorphe Gebilde ohne gemeinsame Binnenstrukturen handelt. Einen vagen Indikator liefern van Berkel und de Jong:

17. Usually, the important nodes, like introductory text, image maps and table of contents are at the very top of the hierarchy. The user can use it as a reference point that facilitates the processing which enables her/him to navigate purposefully through all information nodes. Therefore it is necessary that all pages have an anchor to the top page. (van Berkel und de Jong, 1999, S. 34)

Van Berkel und de Jong bestimmen „important nodes", womit sowohl inhaltliche („introductory text") als auch navigatorische Knoten („table of contents") gemeint sind. Ein zentrales Merkmal von Hypertext ist seine potenzielle Unsequenziertheit. Van Berkel und de Jong stellen jedoch die *hierarchische* Anordnung heraus, die vom Anwender benutzt werden kann, um sich in einem Hypertext zu bewegen. Hierarchische Strukturen sind Baumstrukturen und besitzen eine Wurzel. Die „top page" fungiert in einer Hypertextbasis als Einstiegsseite, als

[147] Siehe auch die Einschätzung des Webdesigners Clement Mok: „Right now, in the absence of web conventions, we're relying on crutches. [...] The house icon is about home, and so on. This situation is partly the result of technical limitations in giving the physical sense of space in three dimensions. The minute you have that, I think some of the more lame conventions of underlining and 'home' will probably go away. At some point, it will become very implicit, but there is no set convention at this point in time." (Fleming, 1998, S. 22).

Startpunkt, den alle Knoten referenzieren sollten, so dass der Anwender einen abgebrochenen Rezeptionsvorgang in der Einstiegsseite erneut aufnehmen kann.

Die prinzipiell arbiträren Sequenziertheitsgrade von Hypertexten (vgl. Abschnitt 3.5.1) werden von van Berkel und de Jong ignoriert, sie beziehen sich explizit auf hierarchische Strukturen, die eine Einstiegsseite sowie wichtige und unwichtige Knoten enthalten. Dieser Indikator ist sehr vage und kann mit der Kombination aus Titelseite bzw. Umschlag und Inhaltsverzeichnis in gedruckten Texten verglichen werden (vgl. Zhang et al., 2000, S. 165). Beide Komponenten können in sehr vielen unterschiedlichen Textsorten existieren, d. h. diese – ausschließlich auf den Peritext bezogenen – Textmuster sind vielfach inhärente und obligatorische Bestandteile *mehrerer* Textsorten (vgl. Indikator 31). Einen möglichen Grund für das „Fehlen etablierter Textmuster und Textsorten" liefert Storrer (1999a, S. 45):

18. Ein zentrales Problem bei der Gestaltung von Hypertexten ist das Fehlen etablierter Textmuster und Textsorten. Gerade im WWW haben sich noch keine einheitlichen Gestaltungsmuster herausgebildet, was u. a. auch daran liegt, dass sich die technischen Rahmenbedingungen stets verändern und erweitern. Die Rezipienten müssen sich bislang deshalb auf jeder Site, bei uneinheitlicher Site-Gestaltung sogar bei jedem Hypertext, neu orientieren. (Storrer, 1999a, S. 45)

Storrer (1999a) zufolge existieren im WWW *keine* „etablierten Textmuster und Textsorten". Als Grund wird die in der fortwährenden Weiterentwicklung befindliche Technologie angegeben.[148] Der Beitrag schließt mit einem scheinbaren Widerspruch:

19. Dazu kommt, dass sich funktionell sinnvolle Muster der Benutzerführung und Informationsstrukturierung im WWW erst langsam herausbilden. Hier ergeben sich interessante neue Aufgaben für den Bereich der Textlinguistik, der sich mit der Beschreibung und Erklärung von Textmustern befasst. (Storrer, 1999a, S. 62)

Storrer (1999a) ist also nicht der Ansicht, dass im WWW *keinerlei* Textsorten beobachtet werden können, die Betonung liegt vielmehr auf dem Merkmal der fehlenden Etabliertheit und Einheitlichkeit vorhandener Prototextsorten. Foltz (1996), der das Verstehen von Texten und Hypertexten untersucht, beschäftigt sich mit einem ähnlichen Aspekt wie Storrer (vgl. Indikator 19)und geht auf einen zusätzlichen Faktor ein:

20. In linear texts, there are a variety of common narrative schemata employed. [...] Although readers of hypertexts may not currently be able to rely on a familiar narrative schema, this may change in the future. As hypertexts become more accepted and widespread, writers of hypertext may develop standard rhetorical styles. Readers who are then familiar with those rhetorical styles can use that knowledge to help in their structuring of the information in an effective manner. (Foltz, 1996, S. 118)

Foltz (1996) verwendet den Begriff des „narrative schema". Je bekannter und vertrauter ein Vertextungsschema ist, desto besser kann es die Prozesse der Produktion und Rezeption von

[148] Auch Bucher (2001, S. 165) spricht diesen Aspekt an: „Die Entwicklungsdynamik [...] ist so hoch, dass sich Standardisierungen in der Verwendung der Gestaltungsmittel nur sehr schwer einstellen." (vgl. auch Fußnote 21, S. 73). Schütte (2004a, S. 122) schließt sich der Auffassung Storrers an, denn „[m]edienspezifische Textformen können sich erst im Laufe der Zeit entwickeln." Dennoch ermittelt Schütte (2004a, S. 135) in ihrer Arbeit Charakteristika der „mediumsspezifischen Textsorte Homepage" (vgl. Abschnitt 4.6.2).

Texten unterstützen. Foltz argumentiert, dass sich „standard rhetorical styles“ (ebd.) herausbilden werden, sobald Hypertexte eine gewisse Verbreitung und Akzeptanz gefunden haben. Eben diese Faktoren gelten ohne Zweifel für das *World Wide Web*, so dass nun unter anderem zu untersuchen ist, welche „standard rhetorical styles“ in diesem Medium existieren.

3.6.4 Mediale und linguistische Aspekte von Hypertextsorten

Der Begriff Hypertextsorte wird in der linguistischen Hypertextliteratur nur selten explizit genannt. Schlobinski (2000b) untersucht Hypertexte im WWW in Bezug auf ihre Hyperlinkstrukturen und führt ein graphentheoretisch ausgerichtetes Analyseinstrumentarium ein:

21. Netzliteraturen lassen sich hinsichtlich ihrer Gerüste analysieren, differenzieren und typologisieren! Hier wären in einem ersten Schritt Hyperfictiondokumente [...] zu analysieren. Aus der empirischen Analyse ließen sich prototypische Strukturen für unterschiedliche HyperTextsorten [sic] etc. gewinnen. (Schlobinski, 2000b, S. 820)

Schlobinski (2000b) verwendet den Begriff der Hypertextsorte, er wird jedoch nicht präzisiert. Es ist erstaunlich, dass sich Schlobinski auf „Hyperfictiondokumente“ (ebd.) bezieht, von denen er annimmt, dass sie auf Hypertextsorten basieren. Hypertexte aus dem Bereich der Hyperfiction – Horn (1989, S. 32) spricht von „branching stories“ – beruhen auf dem künstlerischen, ästhetisch wirkenden Umgang mit Hypertext, weshalb eine Standardisierung oder Konventionalisierung von Verknüpfungsstrukturen in diesem Bereich, in dem Standards jeglicher Art in Frage gestellt und Normen ignoriert oder ad absurdum geführt werden (vgl. de Almeida Santos, 1999), nur in Ansätzen erwartet werden kann (vgl. Abschnitt 3.6.2).[149] Abschließend bezieht sich Schlobinski auf einen umfassenderen Geltungsbereich:

22. Sites, Dokumente, sog. Hyperfiction sollten einer konkreten Strukturanalyse und ihre Resultate einer Typologisierung unterzogen werden. Es wird sich meiner Meinung nach erweisen, dass zum einen sehr heterogene Strukturmuster vorliegen, die mit wenigen Metaphern wie Linie, Baum und Rhizom nicht ausreichend beschrieben werden können, zum anderen, dass bestimmte HyperTextsorten mit bestimmten Strukturmustern verbunden sind. (Schlobinski, 2000b, S. 824 f.)

Die „Strukturmuster“ beziehen sich vornehmlich auf Sequenzierungstypen. Schlobinski weist darauf hin, dass parallel eine Analyse „semantisch-textueller Strukturen“ (ebd.) erfolgen sollte.[150] Schlobinski äußert zusätzlich die Ansicht, dass eine enge Verbindung zwischen Sequenzierungstypen und spezifischen Hypertextsorten besteht.

23. Die Positionierung von Links ist stark kontextabhängig. Neben der [...] Interdependenz vom Anzeigemodus hängt die Entscheidung ab von der Zielsetzung, dem anvisierten Publikum, vom Genre und vom Umfang der Präsentation. (Storrer, 2001b, S. 109)

[149] Boardman (2005, S. 34) bestätigt: „Whether there will ever develop a novel form unique to the Web, exploiting the Web's conventions (which are *not* quite the same as other forms of hypertext), remains to be seen.“

[150] Hierbei stellt sich die Frage, ob „heterogene Strukturmuster“ (Schlobinski, 2000b, S. 824), deren Komplexität über die in der Graphentheorie etablierten metaphorischen Bezeichnungen hinausgeht, überhaupt noch als Muster im engeren Sinne aufgefasst werden können (vgl. Bernstein, 1998).

Storrer (2001b) thematisiert die Positionierung von Linkanzeigern und stellt Einflussfaktoren dar, zu denen auch das „Genre“ gehört. Der Begriff wird weder konkretisiert oder exemplifiziert, kann aber als Synonym für Text- oder Hypertextsorten interpretiert werden. Der nachfolgende Hinweis betrifft die technische Ebene der Realisierung von Webseiten:

24. XML als Ergänzung zu HTML wird das WWW noch einmal gewaltig verändern, da XML den Gestaltungsspielraum [...] erweitert [...]. Aus textlinguistischer Sicht ist vor allem die mit XML gegebene Möglichkeit interessant, Links und Module zu typisieren und textsortenspezifische Hypertextstrukturen zu spezifizieren. (Storrer, 2001b, S. 94)

Storrer vertritt den Standpunkt, dass XML und die flankierenden Standards wie die XML *Linking Language* (DeRose et al., 2001) „textsortenspezifische Hypertextstrukturen“ ermöglichen wird.[151] Wenn die Existenz von Hypertextsorten angenommen wird, ergibt sich bezüglich dieser These die Frage, wie derzeit, d. h. mit den Mitteln von HTML, derartige textsortenspezifische Hypertextstrukturen erstellt werden – die dem Medium Hypertext inhärenten Hyperlinks nehmen diesem Indikator zufolge eine wesentliche Rolle ein. Schmitz (2003, S. 259) bezieht sich auf eine weitere inhärente Komponente: die simultane Verwendung von Text und Bild. Schmitz schreibt Bildern einen derart hohen Stellenwert zu, dass er von „Hyper-Text-Bild-Sorten“ spricht.[152] Dieser Terminus wird anhand der Einstiegsseite des Webauftritts eines Kraftfahrzeugherstellers erläutert, die

25. als typische Homepage eines konsumorientierten Unternehmens gelten [kann]. Herkömmliche Funktionen von Werbung, Public Relations und Information werden in einer beweglichen Mischung von Plakat, Pinnwand, Inhaltsverzeichnis und Wegweiser miteinander verknüpft. (Schmitz, 2003, S. 259)

Die Webpräsenzen „konsumorientierter Unternehmen“ können als Hypertexttypen charakterisiert werden, da ihre Funktionen identisch sind (vgl. Nielsen und Tahir, 2002), jedoch entwickeln sich, abhängig von der Branche, spezifischere Hypertextsortenklassen und Hypertextsorten – z. B. ähneln die Webauftritte von Online-Zeitungen und vielen Informationsportalen einander in Bezug auf ihre Funktionalität und die Anordnung der Informationen (vgl. Indikator 38 sowie Abschnitt 4.6.4). Storrer (2000b, S. 238) gibt ein weiteres Beispiel:

26. Um Dopplungen bei der Themenbehandlung zu vermeiden, haben sich eigene Hypertextsorten herausgebildet, z. B. die sog. FAQs [...]. (Storrer, 2000b, S. 238)

FAQ-Dokumente (*Frequently Asked Questions*) bündeln von Benutzern häufig gestellte Fragen zu einem bestimmten Thema und liefern gleichzeitig die Antworten. FAQ-Dokumente

[151] Storrers Ansicht, dass „XML [...] das WWW noch einmal gewaltig verändern“ wird, teile ich aus verschiedenen Gründen nicht in diesem Ausmaß: XML-basierte Technologien sind sehr komplex und kostenintensiv und durch die fortwährende Verabschiedung zusätzlicher Standards verkompliziert sich die Situation zusätzlich (vgl. Abschnitt 13.2). Weiterhin sind die traditionellen Technologien (HTML, CSS und *JavaScript*) zur Realisierung der meisten Anwendungen ausreichend. Im Übrigen werden XML-Dokumente zur Darstellung derzeit noch nach HTML konvertiert, d. h. der „Gestaltungsspielraum“ erweitert sich nicht durch XML, sondern erst durch die möglichst vollständige Implementierung der Standards, die XML begleiten, in den großen Browsern.

[152] Schmitz (2003, S. 259) umschreibt den Begriff wie folgt: „Textsorten, in denen Bilder eine entscheidende Rolle spielen und die folglich als Text-Bild-Sorten klassifiziert werden müssten, [sind] in der Sprachwissenschaft kaum gewürdigt worden. [...] Allein die Untergruppe der Hyper-Text-Bild-Sorten eröffnet ein riesiges Forschungsfeld, das bisher erst von wenigen Linguistinnen und Linguisten betreten wurde.“

existieren zwar mittlerweile auch im WWW, es ist jedoch nicht zutreffend, dass sie sich als „eigene Hypertextsorte herausgebildet“ haben. Sie sind vielmehr als adaptierte oder transformierte Hypertextsorten aufzufassen, denn ihr Ursprung befindet sich in den Newsgroups des 1979 ins Leben gerufenen Usenet (*UNIX User Network*).[153] Dennoch sind FAQ-Dokumente auch im WWW ein interessantes Phänomen.

27. Die Sequenzierung von Textkonstituenten ist dabei nur eine von verschiedenen Strukturierungsoptionen, die [...] in verschiedenen Textsorten eine mehr oder weniger bedeutende Rolle spielt. (Storrer, 2000b, S. 244)

Storrer (2000b) bezieht sich auf traditionelle Textsorten und fordert, dass sich die Textlinguistik von der Vorstellung loslöst, Texte als Satz- und Abschnittssequenzen zu konzeptualisieren. Dies gilt nicht speziell für das Medium Hypertext, sondern zeigt Anknüpfungspunkte zwischen hypertexttheoretischen Konzepten und der Textlinguistik auf (vgl. Abschnitt 3.6.2).

28. Beim aktuellen Stand des WWW lässt sich [...] gut beobachten, wie Gestaltungsformen [...] aus herkömmlichen Medien [...] im neuen Medium wieder verwendet werden und wie sich daraus ein medientypischer Stil erst herausbildet. Da dabei viel experimentiert, gemischt und gespielt wird, ist das WWW eine Fundgrube für Beispiele von Textmustermischungen und den damit erzeugten Stilwirkungen. (Storrer, 2001b, S. 91 f.)

Die angesprochene Verwendung bekannter Gestaltungsformen im WWW (vgl. Indikator 15) kann einerseits bewusst erfolgen (wie z. B. beim Einsatz der von Freisler, 1994, S. 20, erwähnten Buchmetapher), andererseits auch auf unterbewusst ablaufenden Prozessen basieren. Der zweite Aspekt betrifft die Mischung von Textmustern, die im nachfolgenden Kapitel näher thematisiert wird (vgl. Indikator 25). Des Weiteren geht Storrer (2001b, S. 94) auf die Relevanz der Produktionsbedingungen ein:

29. Viele Entwicklungsumgebungen liefern [...] Schablonen zur Homepage-Gestaltung mit, die nur noch aufgefüllt werden müssen. Diese tragen, da sie ohne großen Zeit- und Geldaufwand zur eigenen Homepage führen, [...] zur Herausbildung und Verfestigung von Strukturierungs- und Gestaltungsmustern im WWW bei. (Storrer, 2001b, S. 94)

Viele HTML-Editoren unterstützen Anfänger durch Vorlagen (vgl. Abschnitt 3.3.6). Da das Webdesign, die Entscheidung zur Präsentation spezifischer Informationseinheiten und ihre Positionierung – gerade diese Dokumenteigenschaften werden von Vorlagen vorgegeben – für die Charakterisierung von Hypertextsorten relevant sind, könnte eine Analyse der `meta`-Elemente von Webseiten in Verbindung mit einer maschinellen Identifizierung von Hypertextsorten zur Überprüfung der von Storrer aufgestellten These eingesetzt werden.[154]

[153] Im Usenet existieren tausende von Newsgroups, in denen Interessierte zu spezifischen Themengebieten miteinander asynchron kommunizieren, wobei gerade von Anfängern viele grundlegende Fragen gestellt werden. Der Zweck eines FAQ-Dokuments besteht darin, diese rudimentären Fragen vorab zu beantworten. Es gehört zum guten Stil, erst zu recherchieren, ob eine Gruppe ein FAQ-Dokument pflegt, dieses gegebenenfalls zu konsultieren, und erst wenn die gewünschte Antwort dort nicht gefunden wird, sollte die Frage in der Newsgroup selbst gestellt werden, so dass sich die aktiven Teilnehmer auf die interessanten Fragen und Diskussionen konzentrieren können (vgl. Pfaffenberger, 1995, S. 199, Bins und Piwinger, 1997, S. 237, und Abschnitt 4.5.3).

[154] Wie bereits in Abschnitt 3.3.6 dargestellt wurde, hinterlassen die meisten HTML-Editoren einen identifizierenden Fingerabdruck in einem `meta`-Element des editierten Dokuments (vgl. Abschnitt A.4.7). Das skizzierte Szenario zur Überprüfung der These von Storrer (2001b, S. 94) kann nur dann realisiert werden, wenn die Hypertextsortenidentifizierung eine hohe Präzision aufweist.

3.6.5 Informationsarchitektur und Webdesign

Web-Publishing-Ratgeber enthalten eine Vielzahl von Indikatoren, die die Existenz von Hypertextsorten implizieren (vgl. Storrer, 2001b, S. 91). Im Zentrum dieses Bereichs steht die grafische Gestaltung von Webpräsenzen zur Realisierung visuell ansprechender und bestmöglich benutzbarer Dokumente („page design" nach Nielsen, 1999) sowie die generelle Gliederung von Websites, die Aufteilung in unterschiedliche Bestandteile und die Verknüpfung dieser Teile untereinander („site design" nach Nielsen, 1999). Häufig wird der letztgenannte Aspekt unter dem Begriff *Information Architecture* zusammengefasst.[155] Eine Aufgabe der Informationsarchitektur betrifft die Strukturierung einer Website. Nach Rosenfeld und Morville zeichnen sich deren Inhalte vor allem durch ihre Heterogenität aus:

> 30. Most web sites [...] are highly heterogenous in two respects. First, web sites often provide access to documents and their components at varying levels of *granularity*. A web site might present articles and journals and journal databases side by side. Links might lead to pages, sections of pages, or to other web sites. Second, web sites typically provide access to documents in *multiple formats*. You might find financial news, product descriptions, employee home pages, image archives, and software files. (Rosenfeld und Morville, 1998, S. 24 f.)

Die Heterogenität bezieht sich unter anderem auf „document formats" (ebd.): Wirtschaftsnachrichten, Produktbeschreibungen, persönlichen Homepages und Bildarchiven sind unterschiedliche Eigenschaften inhärent, was wiederum im Kontext von Hypertextsorten impliziert, dass eine Website *mehrere* Hypertextsorten umfassen kann. Möglicherweise können hierbei Regularitäten festgestellt werden, was den Status der eingebetteten Hypertextsorten innerhalb der übergreifenden Ganzheit betrifft (vgl. Abschnitt 3.6.2). Rosenfeld und Morville beschäftigen sich mit verschiedenen Navigationshilfen:

> 31. The table of contents and the index are the state of the art in print navigation. Given that the design of these familiar systems is the result of testing and refinement over the centuries, we should not overlook their value for web sites. (Rosenfeld und Morville, 1998, S. 65)

Die hierarchische Organisierung einer Website sei der netzartigen Strukturierung vorzuziehen (vgl. Fußnote 111, S. 108), weshalb ein Inhaltsverzeichnis als – dem Anwender aus gedruckten Texten wohlbekannte – Überblicks- und Navigationshilfe dienen kann (vgl. die Erläuterungen zu den Indikatoren 15 und 17). Es sollte, wie in gedruckten Texten, nur die obersten Hierarchiestufen umfassen, klar gegliedert sein und einen unmittelbaren Zugang mittels Hyperlinks bieten. Außerdem sollten Inhaltsverzeichnisse primär funktional konzipiert werden; ihre visuelle Gestaltung sollte in den Hintergrund treten, damit der Leser nicht mit unnötigen Ladezeiten konfrontiert wird, die z. B. durch Grafiken entstehen. Der Index ist im Gegensatz zum Inhaltsverzeichnis weniger strukturiert und enthält eine Liste alphabetisch sortierter Schlagwörter, weshalb er für Websites geeignet ist, die keinen hierarchischen

[155] Rosenfeld und Morville (1998, S. 11) definieren den Informationsarchitekten als diejenige Person, die „mission and vision" einer Website spezifiziert sowie „content and functionality" festlegt: „The information architects focus on the design of organization, indexing, labeling, and navigation systems to support browsing and searching throughout the site." (ebd., S. 20; vgl. auch Reiss, 2000).

Aufbau besitzen. Der Produzent des Index hat – der gedruckten Variante vergleichbar – die Aufgabe, diejenigen Wörter und Phrasen zu antizipieren, die die Rezipienten interessieren könnten. Im WWW kommt dem Index dabei eine spezifische Funktion zu:

> 32. In selecting items for the index, keep in mind that an index should point only to destination pages, not navigation pages. [...] They are often heavy on links and light on text. In contrast, destination pages contain the content that users are trying to find. The purpose of the index is to enable users to bypass the navigation pages and jump directly to these content-bearing destination pages. (Rosenfeld und Morville, 1998, S. 67)

Es existieren demnach mindestens drei funktionale Dokumenttypen: (i) Der Zugriffstyp betrifft das Inhaltsverzeichnis und den Index, die dem Benutzer die Möglichkeit geben, sich einen Überblick über den gesamten Hypertext zu machen und Knoten gezielt anzusteuern.[156] (ii) Zum Navigationstyp können „navigation pages" (ebd.) gerechnet werden, die ebenfalls einen Zugriffscharakter aufweisen, dieser gilt jedoch nur für einen spezifischen Teil eines Hypertextes. (iii) Der Inhaltstyp umfasst Dokumente, die Inhalte präsentieren. Rosenfeld und Morville (1998, S. 67 ff.) gehen mit der Sitemap, der Guided Tour und der Gateway-Page auf drei weitere Ausprägungen ein (vgl. Reiss, 2000, S. 130–140). Eine Sitemap ist eine grafische Repräsentation der Hypertextstruktur und soll ebenfalls den Zugriff unterstützen:

> 33. Unlike tables of contents and indexes, maps have not traditionally been used to facilitate navigation through bodies of text. Maps are typically used for navigating physical rather than intellectual space. This is significant for a few reasons. First, users are not familiar with the use of site maps. Second, designers are not familiar with the design of site maps. Third, most bodies of text (including most web sites) do not lend themselves to graphical representation. (Rosenfeld und Morville, 1998, S. 67)

Sitemaps stellen mit ihrer metaphorischen Anbindung an Karten eine Herausforderung dar, da die beteiligten Kommunikationspartner mit ihnen nicht vertraut sind – es handelt sich um eine Hypertextsorte ohne unmittelbares Pendant im Printbereich.[157] Die Guided Tour dient weniger der Unterstützung des Zugriffs als der Anleitung von Benutzern, die die Inhalte, den Umfang oder die Navigationselemente einer Website noch nicht kennen:

> 34. A guided tour should feature linear navigation [...], but a hypertextual navigation bar may be used to provide additional flexibility. The tour should combine screenshots of major pages with narrative text that explains what can be found in each area of the web site. [...] Remember that a guided tour is intended as an introduction for new users and as a marketing opportunity [...]. Many people may never use it, and few people will use it more than once. For that reason, you might consider linking to the tour from the gateway page rather than the main page. (Rosenfeld und Morville, 1998, S. 69)

[156] Fleming (1998, S. 58) bezeichnet Knoten dieses Typs als „Shortcuts": „One important thing to understand about shortcuts is that they don't work very well by themselves, generally. They make excellent added tools, but often lack the flexibility to serve as standalone navigation systems that work for a variety of users."

[157] Obwohl Sitemaps intuitiv als plausible Navigations- und Überblickshilfen betrachtet werden können, geben Nielsen und Tahir (2002, S. 44) zu bedenken: „Because it's not currently clear whether sitemaps really help users navigate, we recommend including a site map only if substantial resources are allocated for its design and if it's been extensively tested with real users performing real navigation tasks." Pohl und Purghathofer (2004) zeigen, dass visuelle Überblickshilfen den Autoren bei der Erstellung unsequenzierter Hypertexte behilflich sein können. Yip (2004) weist nach, dass permanent sichtbare Sitemaps die Benutzer in der Navigation insbesondere komplexer Websites besser unterstützen als Sitemaps, die explizit angesteuert werden müssen.

Es existieren mehrere charakteristische Merkmale: Die Navigation in einer Guided Tour findet primär monosequenziert statt, sie enthält Reproduktionen der wesentlichsten Inhaltsseiten eines Hypertextes und sie wird nur von wenigen Benutzern in einem begrenzten Umfang rezipiert. Zusätzlich erläutern Rosenfeld und Morville die Gateway-Page:

35. Web sites sometimes have a gateway page that first-time users encounter before reaching the main page. This gateway might serve as a splash page with fancy graphics and animation, as an audience-selection page that sends users to the appropriate area of a site, or as a preview page that shows users what they will get if they subscribe to that particular web site. (Rosenfeld und Morville, 1998, S. 69)

Eine Gateway-Page oder auch Splash-Seite kann noch vor der eigentlichen Einstiegsseite als grafisch opulent gestalteter Blickfang dienen: „Eine Eingangstür – auch *Splash Screen* genannt – läßt sich schnell herunterladen und erzählt den Leuten, was drinnen los ist. Einer guten Eingangstür sollte man schwer den Rücken kehren können. Präsentieren Sie ein Bild, das Ihr Publikum einfängt und es hineinzieht." (Siegel, 1999b, S. 16).[158] Häufig enthalten derartige Seiten nur einen einzigen Hyperlink, der zur Einstiegsseite führt (vgl. Siegel, 1999b, S. 148, für ein Beispiel). Gelegentlich kann auch die Auswahl der Sprache, in der ein Hypertext präsentiert wird, ausgehend von einer Splash-Seite durchgeführt werden (Nielsen und Tahir, 2002, S. 44, sprechen in diesem Fall von einer „Routing Page"). Siegel stellt mit dem Eingangstunnel eine Mischform aus Splash-Seite und Guided Tour dar:

36. Wenn Besucher Ihre Site betreten, geben Sie ihnen lieber die Möglichkeit zu einer kurzen Tour, als sie direkt in die Site zu lassen. Ich nenne diese Tour *Eingangstunnel*. Sie bauen eine Erwartungshaltung beim Besucher auf, während er sich auf das Herz der Site zubewegt. (Siegel, 1999b, S. 18)

Der Eingangstunnel führt zur „Kernseite", die die Aufgabe besitzen, Leser mit Inhalten „zu ködern und zu verführen." (ebd.). Bucher beschäftigt sich eher mit der grafischen Gestaltung von Webauftritten und vergleicht das Textdesign traditioneller und digitaler Zeitungen:

37. Im Falle der gedruckten Zeitungen hat sich zwischen Blattmachern und Lesern ein gemeinsames Strukturwissen über den Aufbau von Titelseiten etabliert, wie es sich bei Online-Zeitungen noch nicht entwickeln konnte. (Bucher, 1999, S. 15)

Obschon diese Aussage für das Jahr 1999 noch eine gewisse Gültigkeit besitzen mag, kann sie mittlerweile nicht mehr uneingeschränkt vertreten werden, da sich in den vergangenen Jahren Konventionen etabliert haben. Behme weist auf die „heilige Dreispaltigkeit" hin:

38. Insbesondere größere Sites mit Tendenz zum Portal [...] haben in den letzten Jahren das Design ihres Webauftritts vereinheitlicht, so dass man fast schon von webweiter Langeweile sprechen kann. Die ›heilige Dreispaltigkeit‹ der Seiten gibt dem Webmaster nicht

[158] Usability-Experten vertreten bezüglich der Verwendung von Splash-Seiten einen eindeutigen Standpunkt: „Take users to your "real" homepage when they type your main URL or click a link to your site. Splash screens must die." (Nielsen und Tahir, 2002, S. 28). Da derartige Seiten (Runkehl et al., 1998, S. 173, sprechen von der „Pre-Homepage") mittlerweile nicht mehr sehr verbreitet sind, sollten sie aus einem weiteren Grund nicht verwendet werden: „Not only are splash pages an annoyance, but given how rarely they occur on the Web these days, they also violate users' expectations." (Nielsen und Tahir, 2002, S. 44). Schütte (2004a) geht ausführlich auf Splash-Seiten ein, die sie – in Anlehnung an Runkehl et al. (1998, S. 173) – als „Pre-Page" bezeichnet.

> mehr allzu viele Möglichkeiten, weil er an festgesetzten Stellen bestimmte Informationen vorsehen muss: Werbung zu allererst, Site-Navigation links und rechts, eigentlicher Inhalt darf in die Mitte. Das ist allerdings nicht nur ein Trend zum designerisch Langweiligen, sondern auch eine Erleichterung für viele Anwender, die oft unabhängig von der besuchten Seite fast schon wissen, wo sie was finden. (Behme, 2000a, S. 55 f.)

Bei vielen Anbietern hat sich ein Webdesign etabliert, das auf drei funktional differenzierten Spalten basiert (vgl. Abbildung 3.5). Behme ist der Ansicht, dass Webgestalter kaum noch eine andere Wahl haben, als diesen Layout-Typ einzusetzen, der einer Konvention unterliegt, was in ähnlicher Weise auch für diverse Druckerzeugnisse gilt (Tageszeitungen, Monografien, Sammelbände etc.). Nielsen geht mit dem Logo auf einen spezifischen Bestandteil des Layouts ein, für den er eine Plazierung in der linken oberen Ecke empfiehlt:

> 39. On all interior pages, the logo should be [...] linked to the home page. Unfortunately, not all users understand the use of the logo as a link to the home page, and it will take a while until this convention is fully established. So for the next few years, it will also be necessary to have an explicit link named "home" on every page. (Nielsen, 1999, S. 178)

Die sich noch in der Entstehung befindliche Konvention spart einen Teil der Bildschirmfläche ein, der sonst zur expliziten Kennzeichnung eines textuell als „Homepage" oder „Home" bzw. als Icon markierten Links eingesetzt werden müsste. Greenspun (1999) unterscheidet

> 40. two broad categories of Web sites: The first is *Web Publishing*. These are sites that are vaguely magazinelike and include, as a degenerate case, the typical corporate product catalog site. The second broad category is *Web-based Services*. These are sites that do a job for a user, such as a site that keeps a dog's medical record and sends out email reminders when the dog needs immunizations. The first category is older and more familiar [...]. (Greenspun, 1999, S. 5)

Der „ältere und bekanntere" Typ betrifft „zeitschriftenähnliche Websites", die – meist auf der Blaupause von Druckerzeugnissen – lediglich Informationen und Texte präsentieren (vgl. Fußnote 159, S. 139). „Web-based Services" stellen einen neueren Typ dar. Diese Sites besitzen spezifischere Funktionen, ihnen ist eine umfassendere Interaktivität inhärent. Greenspun vertritt den Standpunkt, dass Webangebote nur dann erfolgreich sind, wenn ihre Produzenten die potenziellen Fragen der Benutzer antizipieren und durch die Website bestmöglich beantworten können (Rosenfeld und Morville, 1998, S. 2, sprechen von „user-centered awareness"; vgl. Indikator 26). Greenspun exemplifiziert dies am Beispiel der persönlichen Homepage: Es können diverse Gründe existieren, weshalb eine persönliche Homepage aufgerufen wird, so könnte sich z. B. ein Besucher mit dem Homepage-Besitzer in Verbindung setzen wollen, die beiden sind verabredet, der Gast hat aber keine Wegbeschreibung, oder ein Journalist benötigt biografische Informationen des Anbieters:

> 41. So you obviously need to include your phone number, mailing address, and possibly fax number. [...] So you should have a map to your house [...]. So you'd better hope that there is a portrait of you [...]. You'll be getting a wake-up call unless you remembered to put a copy of your résumé online. (Greenspun, 1999, S. 6 f.)

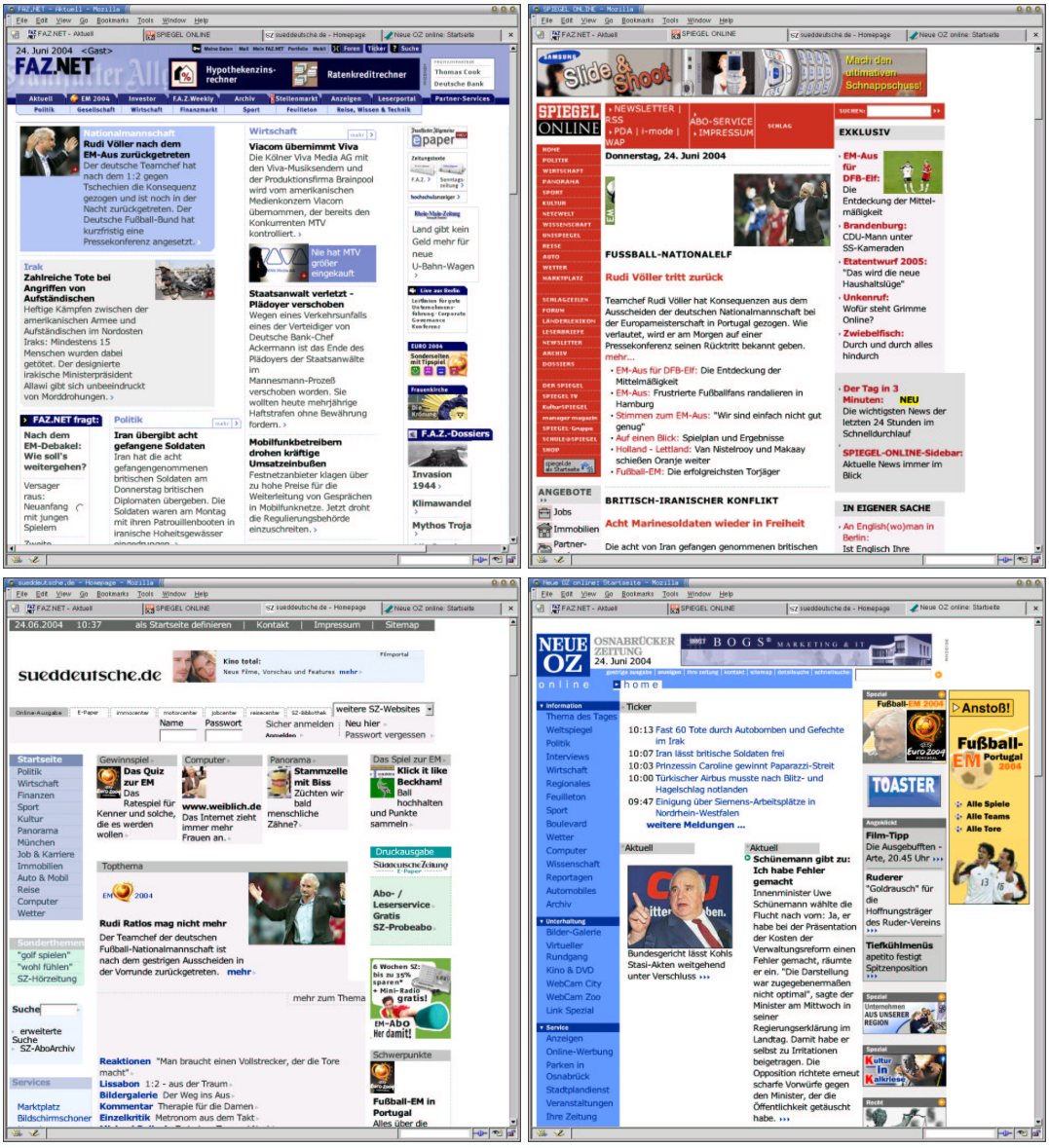


Abbildung 3.5: Beispiele für das etablierte dreispaltige Layout vieler Online-Zeitungen

Die nachfolgenden Kapitel werden zeigen, dass die genannten Informationen tatsächlich hochfrequente Bestandteile persönlicher Homepages sind. Dass gerade in den Anfangsjahren des WWW die Fokussierung auf den Benutzer *nicht* stattgefunden hat, zeigt Greenspun mit einem Vergleich dieses Mediums mit dem Fernsehen, denn das WWW ist

> 42. like early television. People didn't understand the new medium, so they stuck a camera at the back of a live theater, recorded the movements and speech of the actors, and broadcast the result. On the Internet, companies have produced Web sites by sticking a camera in front of their marketing and sales brochures. (Greenspun, 1999, S. 8)

Viele der ersten Websites zeichnen sich durch eine nicht mediengerechte Adaption gedruckter Broschüren aus (vgl. die Indikatoren 13, 15, 25 und 28).[159] Die Ratgeberliteratur unterscheidet zwar mehrere Typen von Websites, es handelt sich jedoch nicht um übergreifende, konventionalisierte Etiketten: Fleming (1998) differenziert zwischen „Shopping", „Community", „Entertainment", „Identity", „Learning" und „Information Sites". Siegel (1999b) exemplifiziert seine Webdesign-Philosophie an den Typen „private Site", „Schaufenster" (kommerzielle Website) und „Fotogalerie". In einem anderen Band geht Siegel (1999a, S. 155 ff.) auf unterschiedliche Typen von Websites aus der Sicht einer Webagentur ein: „Pro-Bono-Sites" werden z. B. für wohltätige Zwecke kostenlos entwickelt und dienen als Referenzprojekte. „Business-to-Business-Sites" dienen der Unternehmenskommunikation, einige „sehen aus wie schlampig erstellte Werbebroschüren, andere weisen interaktive, voll geschäftstaugliche Anwendungen vor." (ebd., S. 156). Produkte werden mit Hilfe von „Retail/Commerce-Sites" präsentiert. „Consumer-Sites" werden von Firmen oder staatlichen Einrichtungen in Auftrag gegeben und sollen Kundenbindung und Markentreue erzeugen. „Vanity-Sites" dienen der Werbung, z. B. einer Person, eines Unternehmens, eines Vereins oder eines Filmes; sie sind primär informativer Natur und meist nur temporär online. „Periodicals" sind Websites für beliebige Typen von Zeitschriften. „Entertainment-Sites" werden über Anzeigenkunden finanziert. Siegel (1999a, S. 157) geht auch auf verschiedene Subtypen ein: „Es gibt viele Arten von Entertainment-Sites, von Magazinen über Sites, die Filmproduktionen begleiten, bis hin zu Rätseln, Spielen, Comics, Soaps und vieles mehr." „Intranets" schließlich sind firmeninterne Netzwerke, zu denen auch Websites mit Funktionen gehören, die sich nur an die Mitarbeiter richten. Bei Intranets wird mehr Wert auf umfassende Funktionalität und Ressourcenreichtum als auf das Webdesign gelegt. Die Beispiele zeigen, dass nicht länger pauschal von „Homepages" oder „Websites" gesprochen wird. Vielmehr besitzen die unterschiedlichen Typen von Webangeboten Bezeichnungen und charakteristische Eigenschaften:

> 43. Info-Sites präsentieren endlos [sic] Textseiten und Listen mit Blickfangpunkten und habe vorne die zu erwartende Homepage (*News* | *About Us* | *Catalog* | *Faq* | *Help*). (Siegel, 1999b, S. 26)

[159] Für Websites, die den Charakter einer Werbebroschüre besitzen, hat sich der Ausdruck *brochure-ware* eingebürgert: „The original approach to web marketing was to treat the website as the company's brochure with the homepage serving as the lead-in to the brochure." (Nielsen und Tahir, 2002, S. 3). Rosenfeld und Morville (1998, S. 165) geben zu bedenken: „One page from a print brochure does not necessarily map onto one page on the Web." (vgl. Jones und Nye, 1995, S. 80, Crowston und Williams, 1997, S. 30, Palmer und Griffith, 1998, S. 50, Haas und Grams, 1998b, S. 103, Middleton et al., 1999, S. 220, Shepherd und Watters, 1999, S. 5 sowie Crowston und Williams, 2000, S. 201). Dürscheid (2000) vergleicht die Printbroschüre des Instituts für Deutsche Sprache (Mannheim) mit seinem Webauftritt.

Der Typ „Info-Site“ präsentiert Informationen eines Unternehmens (vgl. Abschnitt 4.6.2).[160] Siegel zufolge hat sich eine kanonische Benennung der Bestandteile etabliert: „News“ umfasst Neuigkeiten, „About Us“ stellt den Anbieter vor, „Catalog“ verweist auf den Produktkatalog, das „FAQ“ beantwortet die wichtigsten Fragen der Leser (vgl. Indikator 26) und unter „Help“ befindet sich z. B. eine Anleitung zur Betrachtung der digital verfügbaren Handbücher der Produkte.[161] Nach Ansicht von Siegel sollten Info-Sites

> 44. Möglichkeiten sowohl zum Herumschauen als auch zur konkreten Suche bieten. [...] Immer häufiger bieten diese Seiten ein individuelles Profil, daß der Anwender so ausfüllt, das er den von ihm gewünschten Inhalt enthält. (Siegel, 1999b, S. 28)

Für den zuletzt genannten Aspekt hat sich der Begriff Personalisierung eingebürgert. Viele Websites erlauben es dem Besucher, bestimmte Teile (z. B. einer modular aufgebauten Einstiegsseite) auszublenden und diese individuelle Ansicht auch bei späteren Besuchen beizubehalten. Derartige Websites stellen die Inhalte aus Datenbanken zusammen. Siegel (1999b, S. 30) führt weitere Merkmale dynamischer Websites an:

> 45. Dynamische Sites entwickeln sich zur Norm im Informationsbereich. [...] Die Site wird [für den Benutzer] arbeiten, ihn per E-Mail über interessante Neuigkeiten informieren, eine maßgeschneiderte Seite für ihn anfertigen, die ihn fortan begrüßt, und stets die gleichen Interessen des Benutzers mitberücksichtigen, wenn er die Site durchblättert. Eine gute dynamische Site präsentiert Möglichkeiten, neue Dinge zu erfahren und neue Angebote zu sehen, während sie gleichzeitig versucht, 90% der Bedürfnisse von Stammkunden auf den ersten beiden Seiten abzudecken. (Siegel, 1999b, S. 30)

Das Paradebeispiel derartiger, auf impliziten Daten, d. h. nicht auf vom Benutzer explizit hinterlegten Profilen, operierender Sites sind die Websites der Firmengruppe *Amazon*, die registrierte Kunden auf der Einstiegsseite namentlich begrüßen (per Cookie-Technologie), kürzlich angesehene Artikel anzeigen und Produkte auflisten, die den Kunden aufgrund ihrer Ähnlichkeit zu bereits getätigten Käufen interessieren könnten.

Eine Website (und auch eine Webseite) kann aus Modulen zusammengesetzt sein, die auf unterschiedlichen Hypertextsorten basieren (vgl. Indikator 30). Interessant sind in diesem

[160] Nielsen und Tahir (2002, S. 12) bestätigen die Konvention teilweise: „Include a homepage link to an “About Us” section that gives users an overview about the company and links to any relevant details about your products, services, company values, business proposition, management team, and so forth. [...] Include a “Contact Us” link [...] that goes to a page with all contact information for your company.“ Bei der Diskussion einer spezifischen Homepage bemängeln sie: „Conventionally, “About Us” is the last option in a navigation bar, not the middle.“ (ebd., S. 215). Boardman (2005, S. 18) beschäftigt sich mit den vertikalen Navigationshilfen von Online-Zeitungen („sidebars“): „An aspect of this categorisation is the ‘About us’ and ‘Contact us’ sidebar, separated from the hyperlinks to the other main sections of the newspaper. This has become a standard feature of many institutional websites, repeated as a *motif* on most of the pages.“ Darüber hinaus beobachtet Boardman auch den „*search dialogue*“ als „repeated motif“, der durch eine Platzierung am oberen Seitenrand hervorgehoben wird (ebd., S. 19). Crijns (2001, S.280) bezeichnet „*online-order*-Formulare“ und „*feedback-e*-Formulare“ als „auf die Aktivierung ausgerichtete, stark formalisierte Textsorten“. Auf die verbleibenden von Siegel (1999b, S. 26) genannten Kategorien gehen Nielsen und Tahir in anderen Ratschlägen ebenfalls ein.

[161] An die Kategorie „Help“ können unterschiedliche Erwartungshaltungen bestehen: Ein Benutzer, der Schwierigkeiten mit der Navigation hat, könnte eine Anleitung erwarten. Ein Kunde, der ein Problem mit einem bereits zugestellten Produkt hat, erwartet eine Beantwortung seiner Frage. Ein anderer Kunde könnte die angesprochene digitale Version eines Benutzerhandbuches erwarten (vgl. Nielsen und Tahir, 2002, S. 48).

Zusammenhang die Erläuterungen von Siegel (1999a, S. 206 ff.) zu einem in Agenturen häufig eingesetzten Verfahren, die Entwicklung eines Webauftritts mit Hilfe einer „Project-Site“ zu organisieren. Siegel (1999a, S. 209) nennt wichtige Bestandteile einer Projektsite: „Kalender“ (Projektübersicht mit Phasen und Meilensteinen), „Terminplan“ (bündelt Arbeitspakete), „Chronologie“ (dokumentiert den Projektfortschritt und umfasst Notizen und technische Dokumente), „Kontaktinformationen“ (beinhalten die Namen, E-Mail-Adressen und Telefonnummern der beteiligten Mitarbeiter), „Ressourcen“ (Verträge, Protokolle, Rechnungen) und „Hilfe“ (erläutert z. B. die Navigationsmetapher). Diese Beispiele zeigen, dass auch Exemplare vieler traditioneller Textsorten in Websites Verwendung finden.

3.6.6 Metadatenschemata

Metadatenschemata spezifizieren standardisierte Vokabulare, die zur Integration von Metadaten in digitale Dokumente verwendet werden können (vgl. Schmidt, 2004). Das im Kontext des WWW prominenteste Metadatenschema, *Dublin Core*, wurde von der *Dublin Core Metadata Initiative* entwickelt (DCES, 2003, DCMT, 2004, vgl. Abschnitt A.4.4). Ihre Empfehlungen werden – ähnlich wie die vom W3C veröffentlichten Standards – als *Recommendations* publiziert. Für den Bereich der Hypertextsorten ist das DCMI *Type Vocabulary* interessant: „The DCMI Type Vocabulary provides a general, cross-domain list of approved terms that may be used as values for the Resource Type element to identify the genre of a resource.“ (DCTV, 2004). Die Version vom 14.06.2004 spezifiziert 12 Werte: „Collection“, „Dataset“, „Event“, „Image“, „InteractiveResource“, „MovingImage“, „PhysicalObject“, „Service“, „Software“, „Sound“, „StillImage“ und „Text“. Der hohe Abstraktionsgrad dieser Begriffe erklärt sich durch den Geltungsbereich des *Dublin Core*. Er soll sich nicht nur auf das WWW beziehen, sondern als fach- und medienübergreifender Standard fungieren und z. B. auch in Bibliothekssystematiken Verwendung finden können. Der Wert „Text“ wird wie folgt definiert: „A text is a resource whose content is primarily words for reading. For example – books, letters, dissertations, poems, newspapers, articles, archives of mailing lists.“

Aussagekräftiger ist ein älteres Metadatenschema (VW 96, 1996), das auf einer Vorversion des *Dublin Core* basiert[162] und 21 Werte für das innerhalb des HTML-Elements `meta` anzugebende Element `VW96.ObjectType` spezifiziert (das korrespondierende DC-Element wurde früher `ObjectType` genannt, vgl. Caplan, 1995).[163] Tabelle 3.4 stellt die im VW 96-Schema enthaltenen Werte dar. Die 21 Begriffe sind sehr heterogener Natur und können fünf Klassen zugeordnet werden: (i) Zunächst fallen die Bezeichnungen traditioneller Internet-Texttypen auf, die schon vor der Entwicklung des WWW verwendet wurden (FAQ, RFC, HOWTO) und vermutlich aus eben diesem Grund in das Schema aufgenommen wurden. (ii) Die zweite Gruppe betrifft die Technologien PGP und VRML („World“, „RealWorld“, „Keybank“). VRML war ein früher Versuch, eine Markupsprache für dreidimensionale Objekte einzuführen (vgl. Gobbetti und Turner, 1997). Aufgrund mangelnder Unterstützung durch die verbreiteten Browser und des aufwändigen Prozesses zur Anfertigung von VRML-Welten konnte

[162] Fragmente dieser älteren Version finden sich z. B. in dem 1999 veröffentlichten RFC 2413 unter dem Eintrag „Resource Type“ wieder. Die aktuelle Bezeichnung für dieses *Dublin Core*-Element lautet `DC.Type`.

[163] Dieses unter `http://vancouver-webpages.com/META/VW96-schema.html` erhältliche Metadatenschema wurde für einen Crawler entwickelt (vgl. Abschnitt 7.2.1). Zusätzlich lehnt es sich an ein Metadatenschema an, das an der Queen's University of Belfast entwickelt wurde (vgl. `http://quis.qub.ac.uk/overview/metatags.html`).

Wert	Beschreibung	Wert	Beschreibung
Homepage	The Web homepage for a company, organisation, or individual	Search Engine	An online search engine
FAQ	A Usenet/Web FAQ – list, with answers	Hypercatalog	An online, moderated, categorized list of online resources
RFC	An Internet RFC (Request for Comments)	Keybank	A repository for PGP or other encryption keys
Document	General document	Manual	An online manual for a program, piece of equipment, etc.
World	General VRML world	Book	An online book
RealWorld	VRML representation of some real-world object.	Database	An online database, searchable but not necessarily browsable
Index	A browsable index of related documents	Journal	An online, refereed journal
Magazine	An online magazine, not necessarily accepting public submissions	Catalog	A catalog of items in stock or for sale, not necessarily online
Mall	An online shopping mall where products may be purchased	Linecard	A list of product names carried by a distributor/manufacturer
Dictionary	An online dictionary	HOWTO	An Internet/Usernet HOWTO document
Archive	An archive of software or other files		

Tabelle 3.4: Die möglichen Werte zur Charakterisierung von Informationsobjekten auf der Grundlage des Metadatenschemas VW 96 (1996)

sich das Format nicht durchsetzen. PGP ist ein Verfahren zur Verschlüsselung und Signierung von Dokumenten. Anders als VRML wird PGP auch heute noch eingesetzt (z. B. zur Verschlüsselung sensitiver E-Mails). (iii) Die dritte Kategorie bezieht ihr Vokabular aus traditionellen Textklassen („Magazine“, „Dictionary“, „Manual“, „Journal“, „Catalog“ sowie die beiden abstrakteren Klassen „Book“ und „Document“), deren häufige Verwendung im WWW offenbar Anlass genug war, sie in das Schema zu integrieren. (iv) Die Begriffe „Archive“ und „Database“ sind isoliert zu betrachten. Datenbanken stellen zwar eine etablierte Technologie dar, sie können aber keinesfalls als Textklasse (wie bei der dritten Gruppe) gelten, weil Benutzer nur über eine Webschnittstelle auf entfernte Datenbanken zugreifen können; hier dürfte der technologische Aspekt wesentlich dazu beigetragen haben, sie als separaten Wert zu berücksichtigen. Gleiches gilt für Software-Archive, die bereits vor dem WWW eingesetzt wurden.[164] (v) Die verbleibenden sechs Werte beziehen sich primär auf das WWW: Der Begriff „Homepage“ ist, wie auch die Erklärung in Tabelle 3.4 zeigt, mehrdeutig. Er bezieht sich auf *beliebige* Einstiegsseiten (vgl. Abschnitt 4.6.1). Als „Index“ wird eine navigierbare Gruppe von verwandten Dokumenten bezeichnet – gemeint sind themenspezifische Listen von Hyperlinks. Der Begriff „Search Engine“ bezieht sich auf einen weiteren Typ, den auch unerfahrene Nutzer aufgrund ihres zentralen Stellenwerts zur Durchführung von Recherchen schon nach kurzer Zeit als eigenständig klassifizieren dürften. Das Gegenstück hierzu ist der „Hypercatalog“, der sich auf hierarchisch organisierte Kataloge bezieht. Der Begriff „Mall“ bündelt digitale Kaufhäuser, einige Jahre nach der Erstellung des VW 96-Schemas hat sich hierfür der Begriff E-Commerce eingebürgert (vgl. Runkehl et al., 1998, S. 181 ff.); „Linecard“ bezieht sich auf die Produktpalette eines Unternehmens. Es fällt auf, dass dieses Metadatenschema die schon von Gelegenheitsbenutzern intuitiv differenzierbaren Typen von Knoten bzw. Hypertexten umfasst: Dokumente, die beliebige Inhalte enthalten, werden von Hyperlinklisten („Document“ vs. „Index“) und Suchmaschinen von Katalogen abgegrenzt („Search Engine“ vs. „Hypercatalog“), Einstiegsseiten („Homepage“) scheinen, ebenso wie E-Commerce-Angebote („Mall“), konzeptuell eine eigenständige Kategorie darzustellen. Hinzu kommen die unter (iii) genannten traditionellen Textklassen sowie die im Internet seit den achtziger Jahren etablierten Klassen der ersten Gruppe.

[164] Beispielsweise auf öffentlich zugänglichen FTP-Servern. Im VW 96-Schema wird nicht deutlich, ob es sich beim „Archive“ um einzelne Archivdateien oder Sammlungen von Archivdateien handelt.

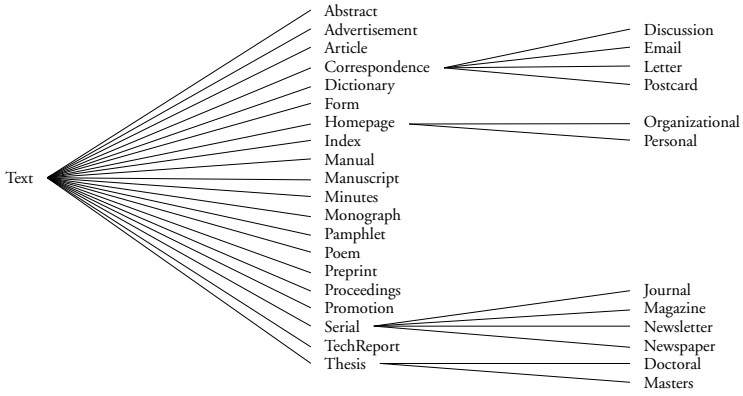


Abbildung 3.6: Ein „standard set of genre types“ als Spezifizierungsvorschlag des *Dublin Core*-Elements „Resource Type“ (nach Tennant, 1997)

Einen Anknüpfungspunkt an dieses Schema stellt der initiale Vorschlag zur Spezifizierung des DC-Elements *Resource Type* mit einem „standard set of genre types“ dar (Tennant, 1997). Das Arbeitspapier stellt zwei Optionen vor: Der „Minimalist Approach“ (charakterisiert als „When nothing but the least will do“) spezifiziert die sechs Werte „Image“, „Sound“, „Software“, „Data“, „Interactive“ und „Text“. Der „Structuralist Approach“ („Why leave anything to chance?“) schlägt – neben diesen sechs Werten – Subklassifizierungen auf bis zu zwei Hierarchieebenen vor (vgl. Abbildung 3.6).[165] Im Vergleich zu dem VW 96-Schema fällt auf, dass der DC-Vorschlag weder Begriffe wie RFC oder FAQ noch WWW-spezifische Textklassen wie Listen von Hyperlinks oder Angebotstypen wie „Mall“ oder „Search Engine“ enthält.[166] Die Ausnahmen betreffen den Begriff „Homepage“, der laut Tennant (1997) entweder auf eine Organisation oder eine Person bezogen werden kann, sowie die Gruppe „Correspondence“: Zwei Subklassen („Letter“, „Postcard“) stammen aus dem traditionellen Bereich, „Discussion“ und „Email“ beziehen sich auf digitale Kommunikationsmedien. Alle weiteren „genre types“ sind traditionellen Textklassen zuzuordnen („Article“, „Minutes“, „Poem“ etc.), wobei auffällt, dass einige Klassen aus dem wissenschaftlichen („Preprint“, „Proceedings“, „Thesis“) sowie dem öffentlichen Bereich stammen („Advertisement“, „Promotion“).[167]

Es kann festgehalten werden, dass sich das VW 96-Schema fast ausschließlich auf HTML-Dokumente bezieht, während sich der DC-Vorschlag an traditionellen Textklassen orientiert.[168] Des Weiteren existieren spezifische Hypertextsortenklassen: Der Begriff „Homepage“ wird in beiden Schemata benutzt, wobei Tennant (1997) eine Subklassifizierung in Home-

[165] Solche hierarchisch gestuften Begriffe werden bei einer Anwendung, d. h. bei der Auszeichnung eines spezifischen Dokuments, durch Punkte abgetrennt, z. B. `Text.Thesis.Doctoral` (vgl. Abschnitt A.4.4). Abbildung 3.6 enthält lediglich die „Text“-Hierarchie. Ebenfalls relevant ist die Unterteilung des Werts „Interactive“, der die Subklassen „Chat“, „Games“, „Multimedia“ und „VR (Virtual Reality)“ enthält.

[166] Der von Tennant (1997) benutzte Begriff „Index“ bezieht sich – im Gegensatz zum VW 96-Schema – nicht auf Listen von Hyperlinks, sondern auf „a summary list of other items“.

[167] „Advertisement“ bezieht sich nicht auf Werbeanzeigen, sondern auf „for example, a job posting“ (Tennant, 1997). „Promotion“ wiederum ist „material that promotes a product, service, or organization“ (ebd.).

[168] Der „Structuralist Draft“ wurde, wie schon die eingangs skizzierte aktuelle Version zeigt (DCTV, 2004), von der DCMI nicht verfolgt, da der „Minimalist Approach“ als flexibler eingestuft wurde und es kaum möglich ist,

pages von Organisationen und Personen vornimmt. Textsorten wie RFC, FAQ und HOWTO, die bereits vor dem Erfolg des WWW im Internet etabliert waren, besitzen für den Autor des VW 96-Schemas einen so hohen Stellenwert, dass sie als eigenständige Begriffe aufgenommen wurden (vgl. Indikator 26). In beiden Schemata sind die traditionellen Textklassennamen auffällig. Hervorzuheben sind auch die Werte, die sich auf bestimmte Funktionen bzw. Inhalte beziehen, z. B. „Archive“, „Search Engine“, „Hypercatalog“ und „Keybank“.

3.6.7 Erwartungshaltungen an Hypertexte

Rezipienten haben Erwartungen gegenüber Exemplaren spezifischer Textsorten. Storrer geht auf diesen Aspekt im Rahmen einer medienunabhängigen Diskussion der Kohärenz ein:

> 46. Die Kohärenzbildung wird [...] geleitet durch Erwartungen an die Art und Weise, wie die Segmente eines Textes in einem Medium angeordnet werden. Diese „Textmuster“, „Textsorten“ oder „Superstrukturen“ genannten Anordnungsformen lassen sich zwar meist funktional-pragmatisch erklären, haben dennoch häufig konventionellen Charakter und sind zum Teil sogar institutionell standardisiert. (Storrer, 1999a, S. 42)

Ein elementarer Faktor von Textsorten betrifft somit die Erwartbarkeit spezifischer Bestandteile eines Textes an etablierten Positionen. In Hypertexten werden häufig Metaphern eingesetzt, um Erwartungshaltungen, die an traditionelle Textsorten bestehen, zu transferieren:

> 47. Textsortenmetaphern tragen nicht unerheblich dazu bei, das Fehlen physisch greifbarer Textgrenzen zu kompensieren und eine Menge von Modulen funktional und thematisch auf eine Ganzheit (z. B. ein digitales Wörterbuch oder ein digitales Vorlesungsverzeichnis) zu beziehen. Das durch die Metapher aktivierte Vorwissen steuert die Vorerwartung an bestimmte Handlungsabläufe, sodass sich der Nutzer lediglich die veränderten und erweiterten Funktionen merken muss, durch die sich das metaphorisch konstituierte Objekt von seinem Pendant im „real life“ unterscheidet. (Storrer, 2003, S. 287)

Storrer geht davon aus, dass sich ein „digitales Wörterbuch“ oder ein „digitales Vorlesungsverzeichnis“ nur metaphorisch an das jeweilige Pendant im Bereich der gedruckten Texte anlehnt, um den Rezipienten zu unterstützen. Der Terminus „Textsortenmetapher“ ist kritisch einzuschätzen, da – um bei Storrers Beispielen zu bleiben – lediglich Exemplare traditioneller Textsorten in ein anderes Medium transferiert werden. Die Texte werden vollständig, mit erweitertem Funktionsumfang und effizienteren Zugriffsmöglichkeiten in einer anderen medialen Aufbereitungsform publiziert, wodurch der Status einer metaphorischen Anbindung deutlich überschritten wird, da identische kommunikative Funktionen vorliegen. Dem eigentlichen Argument ist natürlich zuzustimmen: Durch die Kenntniss der Textsorten *Vorlesungsverzeichnis* und *Wörterbuch* kann der Leser Inhalte und Gliederungen antizipieren – auch im WWW. Ein wesentlicher Faktor betrifft das Vorwissen: Personen, die das WWW

eine vollständige Liste aller potenziell mit Hilfe des *Dublin Core* auszuzeichnenden Textklassen zu erstellen, was Greenspun schon 1995 festgestellt hat: „What if we locked a bunch of librarians and a handful of programmers in a room together and made them think up every possible slot that any Web document could ever want to fill. They'd come out with a list of thousands of fields, each one appropriate to at least a small class of documents. This wouldn't work because the committee could never think of all the useful fields. Five years from now, people are going to want to do new, different, and unenvisioned things with the Web and Web clients.“

nur selten nutzen, dürften lediglich eine einfache Übertragung (im Sinne von Kuhlen, 1991, vgl. Abschnitt 3.3.6) antizipieren, wohingegen erfahrene Anwender z. B. von einem digitalen Vorlesungsverzeichnis erwarten, dass einzelne Veranstaltungen mit einem Kommentartext, der persönlichen Homepage des Dozenten und einer E-Learning-Plattform verknüpft sind, die digitale Handapparate bereitstellt (vgl. Heines et al., 2003, sowie Abschnitt 12.7.1).

Webdesign-Ratgeber betonen, dass sich ein Informationsangebot inhaltlich und gestalterisch von anderen Angeboten absetzen sollte, um Eigenständigkeit und Individualität zu demonstrieren. Usability-Ratgeber helfen Produzenten, intuitiv zu benutzende Angebote zu erstellen, z. B. durch die Anlehnung an Metaphern, so dass auch ungeübte Benutzer in die Lage versetzt werden, sich schnell zurechtzufinden. Mittlerweile gehen jedoch Usability-Experten wie Nielsen und Tahir davon aus, dass beim Entwurf einer neuen Website gerade die Gruppe der unerfahrenen Erstbenutzer praktisch keine Rolle mehr spielt, denn „users spend most of their time on *other* sites than your site“ (Nielsen und Tahir, 2002, S. 37). Es ist weiterhin mehr als unwahrscheinlich, dass eine in der Entwicklung befindliche Website tatsächlich die *zuerst* besuchte Website eines bestimmten Benutzers ist:

> 48. In general, though, the fact remains that users will have seen a very large number of homepages by the time they arrive at your site for the first time. And by this time, users have accumulated a generic mental model of the way homepages are supposed to work, based on their experiences on these other sites. (Nielsen und Tahir, 2002, S. 37)

WWW-Benutzer rezipieren neue Websites auf der Grundlage bereits besuchter Angebote und besitzen somit ein „generic mental model“, das aber nicht nur „the way homepages are supposed to work“ (ebd.), sondern auch die Parameter der Gestaltung, der Positionierung von Bestandteilen und der erwartbaren Inhalte betrifft.[169] Nielsen und Tahir (2002, S. 38) charakterisieren die Erwartungshaltung des Durchschnittsbenutzers:

> 49. The average user expectation upon encountering a site for the first time is that the site is probably going to be a disappointment. Users invest very little time (often on the order of 10 seconds) looking over a new site in the hope it will be one of the rare good ones. But if the site seems too strange or too difficult, or if it's not apparent how the site applies to their immediate concerns, they'll be out of there as fast as they can click their mouse. (Nielsen und Tahir, 2002, S. 38)

Wenn eine Website also eine Erwartungshaltung nicht erfüllen kann, weil sie eine unübliche Gestaltung aufweist oder schwierig zu bedienen ist, wird sie in der Regel unmittelbar ignoriert. Zu dieser Aussage gelangen Nielsen und Tahir aufgrund einer empirischen Analyse von 50 kommerziellen Homepages und den umfassenden Erfahrungen, die sie bei der Beobachtung von WWW-Benutzern[170] gemacht haben und raten:

[169] Haas und Grams (1998a, S. 486) weisen darauf hin, dass sowohl in Bezug auf die medienspezifischen Konventionen, als auch hinsichtlich spezifischer Websites mentale Modelle aufgebaut werden: „Experience with the Web in general helps build expectations, such as that an icon labeled “home” will lead to a top-level page within the site [...]. A reader may also gain experience within a site or a page, for example, learning that every link whose anchor is the name of a town leads to a map of the town.“

[170] Bucher (2001, S. 165) merkt an, dass unterschiedliche Grade der „Webkompetenz“ existieren: „So zeichnen sich prototypische Nutzergruppen, wie Erstbenutzer, Gewohnheitsnutzer, Laien und Profis, gerade dadurch aus, dass sie über jeweils unterschiedliche Wissensvoraussetzungen verfügen.“ (Bucher, 2001, S. 149).

> 50. [The] homepage has to follow standard user interface design conventions, because users won't have time to learn anything new. [...] If you divert their attention – to even the smallest degree – from finding content to having to learn something new, you lose. [... The] average first-time visitor to your site won't be a novice user in the true sense of the term. The user will typically have a good deal of experience with other homepages and will be acquainted with the way most other pages work. To the extent your homepage works similarly, users will feel welcome and will understand the familiar design conventions. (Nielsen und Tahir, 2002, S. 38)

Die Benutzerfreundlichkeit der Einstiegsseite und der Website hängt somit von dem Grad der Befolgung der Konventionen ab, die sich für den jeweiligen Typ etabliert haben. Dieser Aspekt weist eindeutig auf die Existenz von Hypertextsorten hin und bildet die Kernaussage des gesamten Bandes *Homepage Usability* (Nielsen und Tahir, 2002), die durch die bereits erwähnte empirische Analyse gestützt wird, deren Ergebnisse in unterschiedlich gewichteten Richtlinien umgesetzt werden; Abschnitt 4.6.2 stellt diese Analyse vor. Neben der Einstiegsseite existieren weitere Typen, wobei, wie Bucher ebenfalls anhand empirischer Beobachtungen zeigt, die Benutzer spezifische Verknüpfungserwartungen haben:

> 51. Die Sequenzmuster, die Online-Nutzer unterstellen, betreffen die regelhafte Abfolge der [...] charakteristischen Seitentypen wie Einstiegsseite, Themenseite, Inhaltsseite, Archivseite, Selbstdarstellungsseite oder Orientierungsseite. So kann die Inhaltsseite im Sequenzverlauf nicht vor der Einstiegs- oder der Themenseite stehen, und wird umgekehrt aber von diesen beiden als Fortsetzungsseite gewissermaßen regelhaft erzwungen. (Bucher, 2001, S. 157)

Benutzer sind also sowohl in der Lage, verschiedene Typen von HTML-Dokumenten zu unterscheiden, als auch gewisse Abfolgen dieser Typen zu erwarten. Nur in den seltensten Fällen ist z. B. schon die Einstiegsseite zugleich eine „Inhaltsseite", vielmehr dient sie im Regelfall dazu, einen Überblick über das Gesamtangebot zu bieten und thematisch sortierte Sprungpunkte zu Inhaltsseiten bereitzustellen – die Einstiegsseite dient gleichzeitig als „Orientierungsseite" (vgl. Indikator 32). Folgt auf eine Einstiegsseite keine Inhaltsseite, wird, wenn auch nur kurzzeitig, die Erwartungshaltung nicht erfüllt.

3.6.8 Zusammenfassung

Bezüglich der Annäherung an das Konzept der Hypertextsorte sind verschiedene Aspekte deutlich geworden: Mittlerweile greifen sehr viele Anwender regelmäßig auf das WWW zu, so dass sich spezifische Vorstellungen sowohl über den Aufbau und die Gestaltung als auch über die Inhalte und Funktionen unterschiedlicher Typen von Websites und HTML-Dokumenten etabliert haben. Websites können einen oder mehrere Hypertexte umfassen, die Einstiegsseiten besitzen. Die Homepage ist mit einer Kombination aus Inhaltsverzeichnis, Buchumschlag und Einführung bzw. Zusammenfassung vergleichbar, sie dient dem Rezipienten zur Orientierung über den Inhalt und die Strukturierung der Hypertextbasis. Entgegen der prototypischen Realisierung eines Hypertextes als Menge unsequenziert verknüpfter Knoten werden im WWW Hypertexte vornehmlich hierarchisch organisiert, weil sie hierdurch für den Produzenten leichter zu erstellen und zu pflegen und für den Rezipienten einfacher zu erschließen sind, da sich die hierarchische Strukturierung an Textsorten aus dem Printbereich

anlehnt; diese werden darüber hinaus auch häufig im Rahmen einer metaphorischen Anbindung verwendet. Ein derartiger Einsatz von Metaphern erfolgt oft aufgrund des Einsatzes spezieller HTML-Editoren, die dem Produzenten vorgefertigte Schablonen zur Verfügung stellen, die vom Autor lediglich mit Inhalt gefüllt werden müssen. Neben der Einstiegsseite und Webseiten, die einen separaten Index oder ein Inhaltsverzeichnis umfassen, haben sich die Hypertextsorten Sitemap und Guided Tour herausgebildet, die eindeutig definierte Funktionen und Gestaltungsmerkmale und keine Entsprechungen in anderen Medien besitzen. Dies gilt auch für die Gateway- bzw. Splash-Seite, die sich bezüglich ihrer Funktion als um Aufmerksamkeit werbender Blickfang mit einem optisch aufwändig gestalteten Buchumschlag vergleichen lässt. Während sich die genannten Hypertextsorten im WWW entwickelt haben, können die Hypertextsorten FAQ, RFC und HOWTO als transformierte oder auch adaptierte Textsorten gelten, da sie bereits vor der Einführung des WWW verwendet wurden. Bezüglich der Charakterisierung von Hypertextsorten ist von Bedeutung, dass eine übergeordnete Hypertextsorte, z. B. die persönliche Homepage, Textexemplare weiterer Hypertextsorten beinhalten kann, z. B. eine Sitemap, einen Lebenslauf und ein FAQ-Dokument. Die Betrachtung des Einflusses der Textsorte eines vorhandenen Textes auf seine Konversion verdeutlicht, dass sich für diesen Zweck insbesondere deskriptive Texte eignen, die aus in sich abgeschlossenen Teiltexten bestehen. Die Guided Tour zeigt, dass Zusammenhänge zwischen Hypertextsorten und Sequenzierungsformen bestehen.

3.6.9 Fazit

Es haben sich auf mehreren Ebenen Benennungen unterschiedlicher Typen von Websites und Dokumenten etabliert. Problematisch sind jedoch die Ursprünge dieser Etiketten, die von Linguisten (Abschnitte 3.6.2 bis 3.6.4), professionellen Webdesignern und Usability-Experten (Abschnitt 3.6.5) sowie Programmierern und Bibliothekaren (Abschnitt 3.6.6) stammen und sich eben nicht auf die realen Gegebenheiten, d. h. den Sprachgebrauch der WWW-Benutzer und ihren kommunikativen Alltag stützen (vgl. Abschnitt 2.3.2). Darüber hinaus haben einige der Indikatoren lediglich den Status idealtypischer Empfehlungen: Sie stammen aus der Ratgeberliteratur zur Gestaltung intuitiv benutzbarer Websites, so dass ebenfalls kein Bezug zu den realen Gegebenheiten vorliegt. Weiterhin ist deutlich geworden, dass die Benennung unterschiedlicher Typen von Websites noch klaren Grenzen unterworfen ist. Nielsen und Tahir gehen auf einen Aspekt ein, der in ihrem Band *nicht* thematisiert wird:

> 52. We don't address special design considerations for vertical industry segments, such as homepages for software companies, conferences, or dentists. For every industry or type of company, there will be many detailed guidelines that address the ways customers of such companies expect to interact with websites and the best ways to serve those users' needs. We cannot provide a set of generic vertical guidelines. [...] The only way to generate vertical design guidelines is to study each industry's users and their tasks. (Nielsen und Tahir, 2002, S. 8)

Obwohl sich die Autoren primär auf kommerziell ausgerichtete Websites beziehen, machen sie zwei entscheidende Aspekte klar: Erstens existiert bezüglich der Typisierung von Websites auch eine vertikale Dimension: Die Webauftritte von z. B. Software-Firmen, Konferenzen und Ärzten unterscheiden sich in vielerlei Hinsicht, weshalb sich spezifische Konventio-

nen gebildet haben, sie unterliegen somit zwangsläufig einer Binnendifferenzierung (vgl. Ho, 1997).[171] Zweitens können Usability-Richtlinien für diese Segmente nur durch Benutzerstudien erstellt werden, die die konkreten Probleme und Bedarfe der Anwender ermitteln, um sie in die Gestaltung segmentspezifischer Ratschläge zu integrieren. Die Untersuchungsdomäne der vorliegenden Arbeit – Webangebote von Universitäten – kann in diesem Sinne als ein Segment der vertikalen Ebene aufgefasst werden. Es ist nun aber nicht das Ziel, einen Katalog von Ratschlägen zur Gestaltung und Strukturierung universitärer Websites zu erstellen, sondern diese von einer textlinguistischen Perspektive aus auf Vorkommen von Hypertextsorten zu analysieren.[172] Greenspun hat schon 1995 erkannt, dass im WWW unterschiedliche „document types" existieren und geht auf ein weiteres Ziel der vorliegenden Arbeit ein:

> 53. The META tag in HTML [...] can be exploited to implement a document typing system. We need to develop a hierarchy of document types to facilitate implementation of programs that automatically process Web documents. (Greenspun, 1995)

Die automatische Verarbeitung von HTML-Dokumenten unter Berücksichtigung ihrer Hypertextsorten kann z. B. dadurch ermöglicht werden, dass diese Information im Dokument selbst hinterlegt wird. Greenspun schlägt das Element `meta` vor, d. h. die Produzenten müssten diese Information selbst explizit angeben.[173] Hierfür müsste ihnen jedoch eine möglichst präzise Hierarchie dieser Dokumenttypen zur Verfügung stehen, die – ähnlich den in Abschnitt 3.6.6 vorgestellten Metadatenschemata – zur Annotation verwendet werden kann:

> 54. Regardless of how the hierarchy is maintained, developing the initial core taxonomy is a daunting task. The taxonomies developed by librarians are only a partial solution because they do not generally concern themselves with the sorts of ephemera that constitute the bulk of Internet traffic. If we don't get the core taxonomy right, we won't reap the benefits of useful standard software. (Greenspun, 1996)

Etablierte Klassifikationen aus der Bibliothekswissenschaft sind Greenspun zufolge nicht ausreichend, weil sie einen Großteil der im WWW existenten Dokumente nicht beschreiben können. Die Entwicklung einer Hierarchie von „document types" wird als „gewaltige Aufgabe" bezeichnet. Im Zuge der Betrachtung der Metadatenschemata in Abschnitt 3.6.6 wurde deutlich, dass in der Vergangenheit durchaus komplexe Vorschläge vorgelegt wurden (z. B. Tennant, 1997), derzeit wird jedoch ein sehr abstraktes Schema verfolgt. Die aktuelle *Dublin Core*-Version enthält lediglich den Wert „Text", der durch benutzerspezifische (d. h. nichtstandardisierte) Verfeinerungen ergänzt werden kann.[174]

[171] Abschnitt 4.6.2 stellt eine empirische Analyse der Homepages kommerzieller Webangebote vor (Nielsen und Tahir, 2002). In einer vergleichbaren Studie untersuchen Zhang et al. (2000) die Homepages der Websites von etwa 200 Firmen, die aus zehn vertikalen Kategorien stammen: „general merchandiser", „computer software", „electronics and electrical equipment", „telecommunications", „commercial banks", „food and drug", „wholesalers", „motor verhicles and parts", „specialist retailers" und „forest products" (ebd., S. 167).

[172] Diesbezüglich könnte eine Benutzerstudie durchgeführt werden, um die Benennungen unterschiedlicher Typen von HTML-Dokumenten zu ermitteln. Kapitel 4 wird zeigen, dass derartige Studien kein geeignetes Instrument sind, um eine möglichst detaillierte Liste von Hypertextsorten zu erstellen.

[173] Die Analyse der Vorkommen von `meta` im Korpus zeigt, dass Autoren nur selten Metadaten explizit angeben (vgl. Abschnitt A.4.4). Vielmehr werden sie von einigen HTML-Editoren automatisch eingetragen und geben nicht explizit Auskunft über den Inhalt oder Typ des erstellten Dokuments.

[174] Crijns (2001, S. 288) gibt einen weiteren Grund zur Systematisierung „aller mediumspezifischen Textsorten" an: Eine solche „Inventur" könne die Analyse der Wirksamkeit digitaler Werbung unterstützen.

3.7 Zusammenfassung

Dieses Kapitel geht ausführlich auf das Thema Hypertext ein und fokussiert die bislang von der Textlinguistik zu diesem Thema vorgelegten Grundlagenarbeiten, so dass – in den wesentlichen Grundzügen – der aktuelle Kenntnisstand der linguistisch orientierten Hypertextforschung sichtbar wird.[175] Es wird deutlich, dass nahezu ausschließlich mit abstrakten Konzepten hantiert wird, die nicht auf empirischen Befunden basieren. Die abstrakten Konzepte beziehen sich in aller Regel auf die prototypische Ausprägung eines Hypertextes als unsequenzierte Menge von Knoten. Dieses Kapitel deutet an, dass Hypertexte dieser prototypischen Form im WWW eher zu den Randerscheinungen zählen (vgl. Rehm, 2002a, S. 303 f.). Es existiert folglich eine Diskrepanz zwischen den grundlegenden Prämissen textlinguistischer Arbeiten und den realen Gegebenheiten, wie sie sich in dem mit Abstand größten Hypertextsystem präsentieren. Den im WWW herrschenden Mangel an denjenigen Funktionen (typisierte Hyperlinks, Annotationen etc.) und Konzepten (allen voran die multilineare, d. h. unsequenzierte Organisationsform), die von der Hypertexttheorie als inhärente Charakteristika von Hypertexten dargestellt werden, nimmt Pang (1998) zum Anlass, die These aufzustellen, dass Hypertext – in seiner prototypischen Ausprägung – eine nicht existente Technologie darstellt und dass dieser Umstand von den Befürwortern von Hypertext nicht erkannt wird (vgl. Fußnote 20, S. 73). Pang (1998) diskutiert die Arbeiten von Bolter (1991) und Landow (1992), die von einem literaturwissenschaftlichen Standpunkt ausgehen: Der Rezipient eines (prototypischen) Hypertextes erlese sich seinen eigenen Text auf einem individuell gewählten Pfad, der sich als gleichsam individuell gestalteter Text manifestiere, was oftmals als Transformation des Lesers zum Autor oder als Konvergenz dieser Rollen beschrieben wird (z. B. spricht Huber, 2002, S. 47, von einer „Ko-Autorenschaft“, die dem Leser eines Hypertextes zugestanden werden kann; vgl. Simanowski, 2004, sowie Abschnitt 4.6.5, S. 241 ff.). Dieses Konzept bezieht sich in der Argumentation der Hypertexttheoretiker primär auf narrative Texte, d. h. auf Hyperfiction. Die Anfertigung eines Hyperfiction-Hypertextes ist wiederum alles andere als trivial und nur von sehr erfahrenen Autoren realisierbar, die mit den technischen und konzeptionellen Grundlagen von Hypertext vertraut sein müssen. Daher verwundert es auch nicht, dass (narrative) Hypertexte in dieser prototypischen Ausprägung im *World Wide Web* kaum zu finden sind.[176] Selbstverständlich sind auch alle anderen Websites in gewisser Hinsicht nicht- bzw. multilinear strukturiert, da es z. B. dem Leser einer privaten Homepage freigestellt ist, zuerst die Hobbys einer Person, ihren Lebenslauf oder erst das Gästebuch zu inspizieren – der wesentliche Aspekt ist jedoch, dass die für narrative Hypertexte so ausschlaggebende individuelle Knotenauswahl und insbesondere die hiermit verbundene *Reihenfolge* der Rezeption bei derartigen Gebrauchs- bzw. Informationshypertextsorten nur eine untergeordnete Rolle spielt (vgl. Storrer, 2004b, S. 35 f.), weil durch die enthaltenen Knoten keine thematische Progression ausgedrückt wird. Antos und Tietz (1997b, S. viii) stellen fest, dass „ein Herzstück der Textlinguistik“ – gemeint ist die Textsortenlinguistik – „auf der Stelle zu treten“ scheint. Dieser Befund gilt in noch dras-

[175] Ausgenommen davon sind die (wenigen) Arbeiten, die in den nachfolgenden Kapiteln diskutiert werden, z. B. Eckkrammer (2001), Bittner (2003), Jakobs (2003) und Schütte (2004a).

[176] Siehe auch Boardman (2005, S. 33 f.): „It was thought that the Web would popularise the hypertext novel, to the extent that the web form would rival the printed novel, but this has not happened to date.“

tischerer Weise für den textlinguistischen Untersuchungsgegenstand Hypertext, sowohl in Bezug auf das Hypertextsystem *World Wide Web* als auch hinsichtlich einer Verbindung der beiden Konzepte, d. h. der Beschäftigung mit Hypertextsorten.

Abschließend werden die wesentlichen Eigenschaften von Hypertexten im *World Wide Web* zusammengefasst: Die Mehrzahl aller Websites basiert in konzeptioneller Hinsicht nicht auf unsequenzierten, sondern auf streng hierarchischen Organisationsformen. Die visuellen und thematischen Abgrenzungen zwischen einzelnen Websites verschwimmen nicht etwa, sie sind vielmehr sehr deutlich ausgeprägt. Die wenigen unsequenzierten Hypertexte müssen derzeit noch als Randerscheinungen aufgefasst werden. Viele Websites basieren zu großen Teilen auf mehreren traditionellen Textsorten aus dem Printbereich (vgl. Huber, 2002, S. 81), d. h. webbasierte Hypertexte beinhalten eher E-Texte (im Sinne von Storrer, 2003) als prototypische Hypertexte. Die „modulare Informationsaufbereitung" (Bucher, 1996) und das Vorhandensein von „Schnipseltexten" (Jucker, 2000) beziehen sich primär auf Einstiegsseiten von Hypertexten, bei untergeordneten Knoten sind diese Konzepte insbesondere im Bereich des Peritextes zu beobachten, z. B. bei Navigationsleisten, Logos, Suchfunktionen und Werbebannern. Eine Modularisierung der eigentlichen Inhalte in multilinear rezipierbare Komponenten findet eher selten statt, d. h. Hypertextknoten sollten zunächst als eigenständige, in sich kohärente Texte aufgefasst werden, die der von Kuhlen (1991) geforderten kohäsiven Geschlossenheit unterliegen und über Hyperlinks zu einem Hypertext mit einem übergeordneten Thema sowie einer globalen Textfunktion zusammengefügt werden. Mit anderen Worten: Die in einem Hypertext enthaltenen Knoten können als lokal kohärente und kohäsiv geschlossene Teiltexte mit spezifischen Inhalten und Funktionen konzeptualisiert werden.

3.8 Fazit – Initiale Hypertextsortentypologien

Die in Abschnitt 3.6 diskutierten Indikatoren wurden zur Konstruktion einer initialen Typologie von Hypertextsorten eingesetzt, die in Abbildung 3.7 dargestellt ist. Diese Typologie bezieht sich ausschließlich auf unterschiedliche Typen von Hypertexten (d. h. Websites) im WWW und wurde primär aus den Indikatoren der Bereiche Informationsarchitektur und Webdesign abgeleitet (Abschnitt 3.6.5). Es ist zu beachten, dass Hypertexte weitere Hypertexte einbetten können, in der Realität existieren also Überschneidungen. Zusätzlich stellt Abbildung 3.8 eine ebenso initiale Typologie von Hypertextknotentypen dar. Diese zweite Typologie bezieht sich ausschließlich auf die Knotenebene, also auf einzelne HTML-Dokumente und wurde ebenfalls aus den in Abschnitt 3.6 genannten Indikatoren abgeleitet. Die beiden Typologien streben die initiale Systematisierung der unterschiedlichen in Abschnitt 3.6 genannten Kategorien an. Aus diesem Grund sind sie nicht als vollständig zu verstehen.

Die Typologie von Hypertexttypen (Abbildung 3.7) unterscheidet auf der ersten Ebene zwischen kommerziellen und nichtkommerziellen Websites. Die Klasse der kommerziellen Websites umfasst drei Kategorien, die Unterschiede bezüglich der jeweiligen Zielgruppe aufweisen. An Firmen gerichtete Websites werden als *Business-to-Business-Sites* bezeichnet, im Intranet einer Firma können zusätzlich z. B. *E-Learning-Sites* oder auch *Project-Sites* (die nicht nur zur Entwicklung von Webpräsenzen, sondern zur Abwicklung verschiedener Projekte eingesetzt werden können) existieren. An den Endkunden richten sich *Consumer-* und

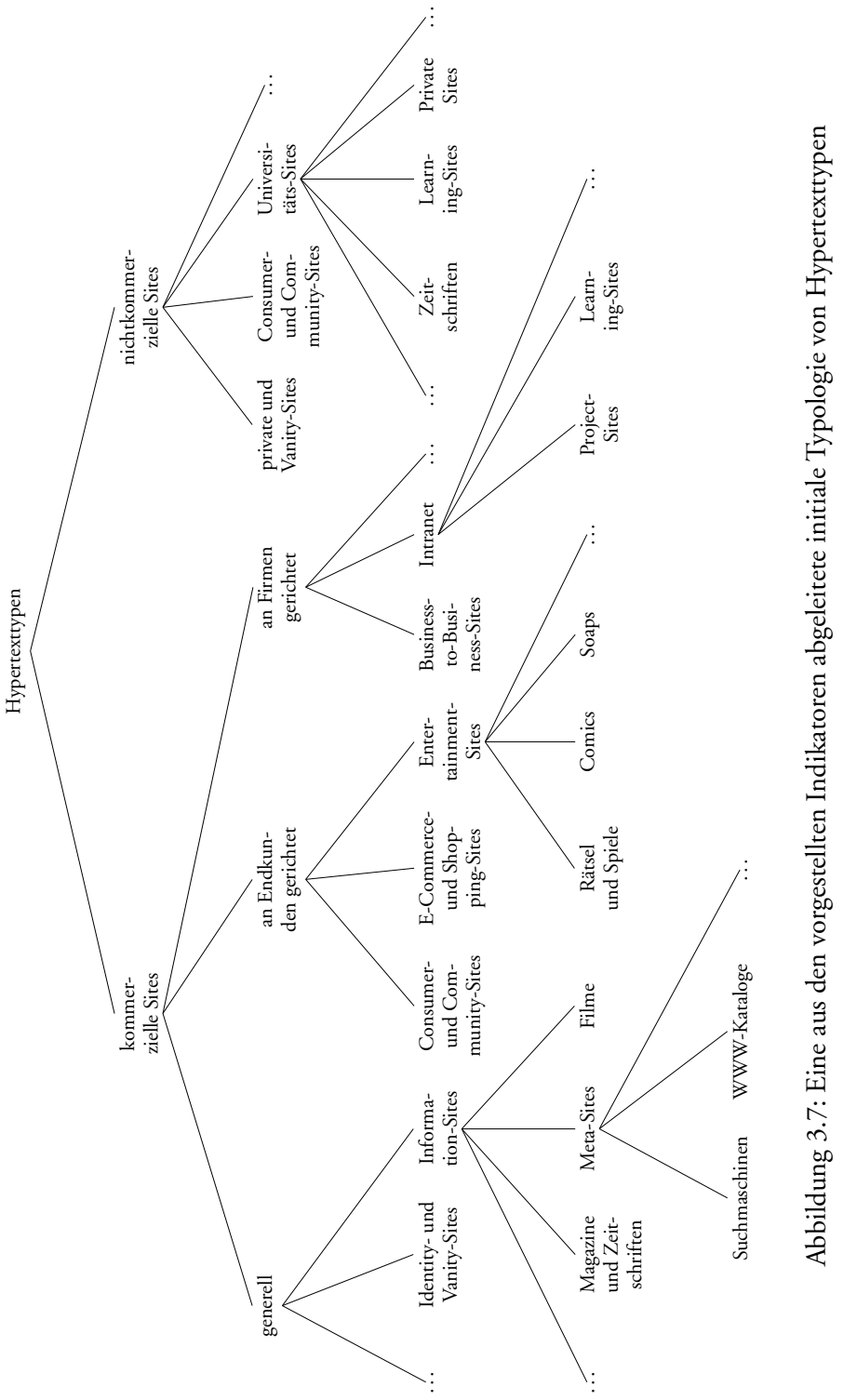


Abbildung 3.7: Eine aus den vorgestellten Indikatoren abgeleitete initiale Typologie von Hypertexttypen

Community-Sites (präsentieren vornehmlich Informationen und bieten Diskussionsfunktionen an, z. B. Foren), *Entertainment-Sites* (dienen der Unterhaltung und dem Zeitvertreib mit der Intention der Imagepflege und der Stärkung einer bestimmten Marke) und *E-Commerce-* und *Shopping-Sites* (für den Online-Einkauf). Keine Einschränkung hinsichtlich einer spezifischen Zielgruppe liegt bei *Identity-* und *Vanity-Sites* vor, die primär der Werbung dienen und eine Firma, ein Produkt oder eine Person des öffentlichen Lebens vorstellen. *Information-Sites* werden ebenfalls zu Werbezwecken eingesetzt und bieten darüber hinaus einen spezifischen Mehrwert: Zu einem beworbenen Film können z. B. Biografien von Regisseur und Schauspielern eingesehen und Trailer abgerufen werden. *Zeitschriften und Magazine* (umfasst auch *Zeitungen*) stellen einen Teil des redaktionellen Angebots online zur Verfügung. Eine Sonderstellung nehmen *Suchmaschinen* und *WWW-Kataloge* ein. Diese präsentieren zwar auch eigene Inhalte, ihre wichtigste Funktion ist jedoch das Recherchieren nach weiteren Angeboten, die eben nicht Teil des eigenen Angebots sind – diese Typen von Websites werden daher als *Meta-Sites* bezeichnet. Die nichtkommerziellen Sites umfassen Angebote, die nicht aus primär wirtschaftlichen Gründen existieren. Hierzu gehören *private Sites*, die den jeweiligen Autor, seine Familie oder auch ein Hobby vorstellen, von Vereinen für Verbraucher aufgebaute Informationsportale (*Consumer-Sites*) und auch die *Sites von Universitäten* (und sonstigen Forschungseinrichtungen). Diese können wiederum Websites von *Zeitschriften*, *E-Learning-Sites* und auch *persönliche Homepages von Angehörigen* umfassen. Es wird deutlich, dass diese Typologie einen sehr rudimentären Charakter besitzt.[177] Es existieren z. B. auch Suchmaschinen und Kataloge, die aus Forschungsprojekten oder Initiativen von Freiwilligen hervorgegangen sind, und zu Filmen oder auch Musikgruppen gibt es eine Vielzahl von Websites, die von Fans betrieben werden. Trotz aller Fragmentiertheit zeigt die Typologie, dass mittlerweile zahlreiche Kategorien von Websites unterschieden werden können.

Die in Abbildung 3.8 dargestellte Typologie von Hypertextknotentypen bezieht sich primär auf die Ebene des einzelnen HTML-Dokuments, wobei zwischen Navigationsknoten, Strukturknoten und Inhaltsknoten unterschieden wird.[178] Die erste Klasse umfasst Knotentypen, die dem Zugriff und der Navigation dienen, wozu die *Einstiegsseite*, die *Sitemap* und der *Index* gehören.[179] Die Klasse der Strukturknoten (Terminus nach Hammwöhner, 1997, vgl. Abschnitt 3.6.2) umfasst Hypertextknotenklassen, die sich konzeptionell zwischen den Navigations- und den Inhaltsknoten befinden, da sie Binnensequenzierungen beinhalten (*Eingangstunnel*, *Guided Tour*) bzw. in eine übergeordnete Sequenzierung eingebettet sind (*Splash-Seite*). Der *Eingangstunnel* besteht aus *mehreren* Knoten, die linear sequenziert sind und den Leser sukzessive in die Website führen (diesbezüglich ist gelegentlich auch von einem *forced march* die Rede). Einer Monosequenzierung unterliegen auch die einzelnen HTML-Dokumente einer *Guided Tour*, deren Rezeption jedoch optional ist. Die *Splash-Seite* führt

[177] In diese Typologie wurden fast alle in Abschnitt 3.6 genannten Bezeichnungen von Websites integriert. Die Ratgeberliteratur zu den Themen *Information Architecture* und *Web-Publishing* bezieht sich dabei fast ausschließlich auf kommerzielle Websites, was die Dominanz dieser Kategorie in der Typologie erklärt.

[178] Die in kursiver Schrift dargestellten Kategorien stellen instanziierbare Hypertextknotensorten dar, die einen niedrigeren Abstraktionsgrad als Hypertextknotentypen besitzen (vgl. Abschnitt 2.3.2). Die in der Grundschrift dargestellten Kategorien beschreiben strukturierende Klassen.

[179] Es könnte eine weitere Subklassifizierung vorgenommen werden, die sich auf den Skopus eines Navigationsknotens bezieht: Der *Index*, die *Sitemap* und die *Einstiegsseite* beziehen sich in der Regel auf die gesamte Hypertextbasis, wohingegen die *Hyperlinkliste* meist nur einen spezifischen Teil des Hypertextes betrifft.

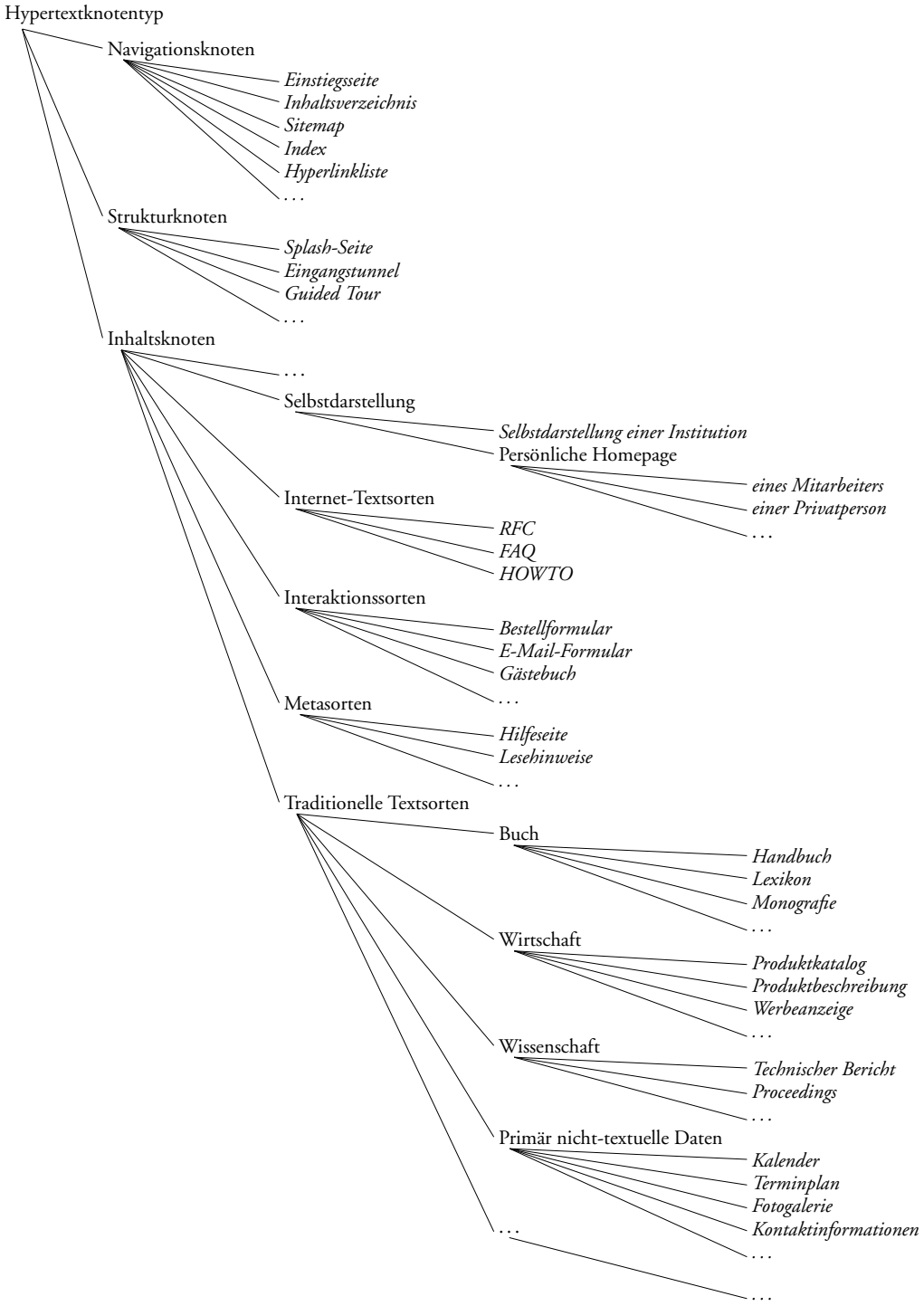


Abbildung 3.8: Eine aus den vorgestellten Indikatoren abgeleitete initiale Typologie von Hypertextknotentypen

unmittelbar zur *Einstiegsseite*. Weiterhin sind der *Eingangstunnel* und die *Guided Tour* (gegebenenfalls auch die *Splash-Seite*) inhaltlich markiert, da sie Informationen über die anbietende Website vermitteln. Die dritte Klasse (Inhaltsknoten) umfasst Hypertextknotensorten, die Inhalte und Informationen enthalten. Die Kategorie Interaktionssorten bündelt diejenigen Sorten, die eine aktive Interaktion des Benutzers erfordern, insbesondere durch das Ausfüllen eines HTML-Formulars, z. B. zur Bestellung von Waren oder zur Hinterlassung eines Gästebucheintrags. Die Metasorten umfassen Hypertextknotensorten, die z. B. technische (*Hilfeseite*) oder konzeptionelle Hinweise (*Lesehinweise*) zu einem spezifischen Hypertext geben. Die Klasse Selbstdarstellung umfasst *Selbstdarstellungen von Institutionen* sowie *persönliche Homepages von Mitarbeitern einer Institution* und *private Homepages*.[180] Die traditionellen Textsorten wurden bereits in Abschnitt 3.6.6 diskutiert, hinzugekommen ist die Kategorie der primär nichttextuellen Daten, die Textsorten wie *Kalender*, *Fotogalerie* und *Kontaktinformationen* umfasst. Abschließend sei erneut darauf hingewiesen, dass diese initiale Typologie – ebenso wie Abbildung 3.7 – keinen Anspruch auf Vollständigkeit besitzt und lediglich die generelle Komplexität aufzeigen und einen ersten Strukturierungsvorschlag darstellen soll.[181] Sie umfasst beinahe alle Hypertextknotensorten, die in Abschnitt 3.6 angesprochen wurden.[182] Es ist weiterhin ersichtlich, dass eine gemeinsame Einordnungsinstanz nicht immer berücksichtigt werden konnte (insbesondere im Bereich der Inhaltsknoten), weil prinzipiell arbiträre Strukturierungen möglich sind – die dargestellte Strukturierung orientiert sich an den Binnenstrukturen der in den einzelnen Abschnitten vorgestellten Indikatoren.[183] Die Typologie bezieht sich auf die Ebene des einzelnen HTML-Dokuments. Es ist jedoch zu betonen, dass Mischungen der aufgeführten Hypertextknotensorten in ein- und demselben Knoten sehr häufig beobachtet werden können, so kann z. B. eine persönliche Homepage *Kontaktinformationen*, eine *Hyperlinkliste* und ein *Inhaltsverzeichnis* enthalten. Die Grenze zwischen Hypertexttypen und Hypertextknotentypen ist also keinesfalls so klar definiert, wie die beiden getrennten Typologien dies andeuten.[184] Die nachfolgenden Kapitel gehen in detaillierter Form auf diese Problematik ein.

[180] Exemplare der Sorte *Selbstdarstellung einer Institution* können z. B. durch den in Indikator 43 dargestellten Link „About Us“ erreicht werden. Derartige Texte können auch in gedruckten Firmenbroschüren beobachtet werden und könnten somit in der Kategorie der traditionellen Textsorten vertreten sein. Da sie aber als ein institutionsbezogenes Gegenstück zur persönlichen Homepage aufgefasst werden können, wurden diese beiden Sorten in der Kategorie Selbstdarstellung zusammengefasst.

[181] Schütte (2004a, S. 152) reproduziert eine vier Kategorien umfassende Typologie von Homepages, die einem Webdesign-Ratgeber entnommen wurde, und merkt an, dass diese die „Schwierigkeiten und Unschärfen [dokumentiert], die mit Typologisierungsversuchen von WWW-Dokumenten verbunden sind, insbesondere, wenn diese Versuche sich dem Eindruck der Beliebigkeit nicht ganz entziehen können und überdies allgemeine, netzweite Gültigkeit beanspruchen.“

[182] Einige traditionelle Textsorten wurden aus Platzgründen nicht aufgenommen.

[183] Die traditionellen Textsorten nehmen einen Großteil der Typologie ein. In der *Web-Publishing-* und Webdesign-Literatur werden vornehmlich Hypertextknotensorten wie die *Sitemap* oder die *Splash-Page* diskutiert, gerade weil die Möglichkeiten zur Realisierung von Inhaltsknoten so flexibel und vielseitig sind.

[184] Auch wenn hierdurch der eigentliche Mehrwert des Mediums neutralisiert wird, spricht in technischer Hinsicht letzten Endes nichts dagegen, eine mehrere Knoten umfassende Hypertextbasis in nur einem einzigen Knoten zu aggregieren und lediglich innerhalb dieses Knotens zu verknüpfen.

4

Hypertextsorten und Digital Genres

4.1 Einleitung

Dieses Kapitel geht auf die theoretischen Aspekte von Hypertextsorten ein, die in den bislang zu diesem Themenkomplex veröffentlichten Arbeiten diskutiert werden und stellt an verschiedenen Anknüpfungspunkten Erweiterungen vor. Dies betrifft z. B. verschiedene Studien, in denen zufällig zusammengestellte Stichproben von HTML-Dokumenten mit dem Zweck der Sammlung unterschiedlicher Web-Genres analysiert werden, aber auch übergreifende Aspekte wie die an der Entwicklung von Hypertextsorten beteiligten Faktoren.[1] Darüber hinaus werden die Charakteristika der bislang untersuchten Hypertextsorten vorgestellt und diskutiert. Dieses Kapitel beschränkt sich auf deskriptive Arbeiten, die oftmals das Potenzial der automatischen Kategorisierung von Hypertextsorten betonen; die implementierten und naturgemäß auf sehr viel gröberen Konzeptionen beruhenden maschinellen Verfahren werden in Kapitel 14 thematisiert. Fast alle der hier diskutierten Arbeiten beziehen sich auf die *North American Genre Theory* (vgl. Abschnitt 2.3.7). Die seit 1968 jährlich stattfindende *Hawai'i International Conference on System Sciences* (HICSS) – speziell der erstmals 1997 ausgerichtete Minitrack „Genres in Digital Documents" innerhalb der Sektion „Digital Documents" – hat sich als internationales Podium zur Präsentation neuer Ansätze etabliert.[2]

[1] In diesem sowie den sich anschließenden Kapiteln werden verschiedene englischsprachige Arbeiten thematisiert, die die Begriffe „Genre" und „Web-Genre" verwenden; im Folgenden werden diese äußerst generischen Termini (vgl. Abschnitt 2.3.7) bei der Diskussion der entsprechenden Arbeiten beibehalten. Der Begriff „Digital Genres" bezieht sich auf Textsorten digitaler Dokumente und subsumiert somit „Web-Genre". Die Begriffe „Hypertexttyp", „Hypertextsorte", „Hypertextknotentyp" und „Hypertextknotensorte" werden entsprechend der in Abschnitt 3.8 verwendeten Lesarten benutzt. Kapitel 5 führt präzise Definitionen ein.

[2] Von 1997 bis 2000 wurde jeweils der Minitrack „Genre[s] in Digital Documents" veranstaltet. In den Jahren 2001 bis 2003 wurde dieser nicht aufgelegt, einige relevante Arbeiten wurden jedoch in den thematisch verwandten Minitracks „Digital Documents: Understanding and Communication" (2001/2002) und „Digital Libraries" (2003) vorgestellt. Seit der HICSS-37 (2004) existiert der eingangs genannte Minitrack wieder. Der Begriff des „Digital Documents" wird dabei sehr unspezifisch aufgefasst und berührt Aspekte der Textverarbeitung, der *Computer-Mediated Communication* und des WWW.

In der Eröffnungsrede des Minitracks hob Nunberg hervor, wie umfangreich der potenzielle Einflussbereich digitaler Medien auf das heterogene Konzept „Genre“ ist:

> [Genres] can be characterized by their formal properties (that is, properties of format, language, presentation), by their schemes of content organization, by their social functions, and by the communities who use and interpret them. Electronic media have the potential to alter any and all of these characteristics, changing the way documents are read, organized, distributed, and used. (Nunberg, 1997, S. 2)

Weil es sich bei Genres um sehr heterogene Einheiten handelt, betrachten die vorgelegten Arbeiten sehr viele unterschiedliche Eigenschaften dieses Konzepts. Eine homogene Forschungslinie hat sich bislang – mit Ausnahme des Bezugs auf die *North American Genre Theory* – nicht etabliert. Da es sich um ein noch junges Forschungsfeld handelt, wird die verfügbare, teilweise schwer zugängliche und in deutschsprachigen Arbeiten bislang nicht zur Kenntnis genommene Literatur ausführlich dargestellt. Abschnitt 4.2 geht auf Anwendungen der Genre-Theorie auf digitale Medien ein, woraufhin Abschnitt 4.3 die Herausbildung und Etikettierung von Hypertextsorten diskutiert. Abschnitt 4.4 thematisiert Analysen zur Sammlung von Web-Genres auf der Grundlage von Stichprobenuntersuchungen, woraufhin Abschnitt 4.5 ihre Kerneigenschaften diskutiert. Abschnitt 4.6 geht ausführlich auf diejenigen Hypertexttypen bzw. -sorten ein, zu denen spezifische Darstellungen vorliegen.

4.2 Untersuchungen von Digital Genres

Die innerhalb der *North American Genre Theory* vertretene Auffassung des Konzepts Genre (vgl. Abschnitt 2.3.7) kann zwar im Vergleich zu aktuellen Text- und Textsortentheorien als relativ abstrakt bezeichnet werden, sie erweist sich jedoch bei der Untersuchung digitaler Medien als außerordentlich flexibel, wie die nachfolgend dargestellten Ansätze zeigen (vgl. auch Firth und Lawrence, 2003). Es werden grundlegende Parameter digitaler Genres eingeführt, die für die nachfolgende Diskussion relevant sind.[3]

4.2.1 Digital Genres und computervermittelte Kommunikation

Erickson (1997) untersucht den Web-basierten Chatraum „Café Utne“ (`http://cafeutne.org`) und wendet die Genre-Theorie zur Charakterisierung dieser Dialogform an. In diesem Zusammenhang weist Erickson den vornehmlich in der populärwissenschaftlichen Literatur präsenten Begriff der „virtual community“ zurück, weil eine Analyse unter Genre-Gesichtspunkten automatisch mit einer Fokussierung des Zwecks einer kommunikativen Handlung,

[3] Verschiedene Arbeiten beschäftigen sich mit weiteren digitalen Genres: Procter und Goldenberg (1998) untersuchen den Einsatz von Genres wie z. B. „reference interaction“ und „peer assistance“ in einer Benutzerschnittstelle für digitale Bibliotheken. Von Westarp et al. (1999) gehen auf digitale Jahresberichte von Unternehmen ein. Fox et al. (1999) diskutieren Abschlussarbeiten und Dissertationen, die in einem digitalen Archiv hinterlegt werden. Fortanet et al. (1998, 1999) untersuchen Werbung im WWW (vgl. auch Runkehl et al., 1998, S. 185–204). Marlow (2004) behandelt Trouble-Tickets, die von den in einem *Network Operations Center* beschäftigten Technikern zur Protokollierung von Zwischenfällen angefertigt werden und zeigt Gemeinsamkeiten und Unterschiede zu CMC-basierten Medien auf. Sæbø und Päivärinta (2005) analysieren ein politisches Diskussionsforum im Hinblick auf vier „e-Democracy“ Modelle.

ihrem Inhalt und ihrer Form verbunden sei. In vielen Bereichen der Online-Kommunikation steht Erickson zufolge nicht der bloße Smalltalk, sondern der Austausch von Informationen und Meinungen im Mittelpunkt, so dass das Kommunikat selbst in den Vordergrund tritt. Erickson kommt zu dem Schluss, dass „Café Utne" ein „conversation-as-document model" verwendet. Die Unterhaltung wird den Rezipienten, die als potenzielle Produzenten fungieren, als einzelne Webseite präsentiert, die wie ein gedrucktes Dokument schnell überflogen werden kann. Derartige asynchrone Unterhaltungen sind – im Gegensatz zu fast allen synchronen Kommunikationsmedien – persistenter Natur, sie sind dauerhaft zugreifbar.

Erickson (1999) analysiert einen weiteren Diskussionsstrang, in dem 69 Teilnehmer abwechselnd die einzelnen Zeilen mehrerer Limericks geschrieben haben; die Analyse basiert auf 2 109 Beiträgen, die in einem Zeitraum von einem Jahr verfasst wurden.[4] Eine sich schnell etablierte Konvention betrifft den Beginn eines neuen Limericks: Der erste Beitrag, der das verteilte Schreiben von Limericks initiiert hat, forderte die Teilnehmer auf, jeweils nur eine Zeile beizusteuern. Nach etwa einer Woche hat ein Autor einen Limerick mit der fünften Zeile beendet, einen Trennstrich in den Beitrag aufgenommen und daraufhin die erste Zeile eines neuen Limericks begonnen. Da die anderen Teilnehmer keine Einwände vorbrachten, konnte sich diese Konvention etablieren. Nach vier Monaten wurden fast alle Limericks basierend auf dieser „Last-First rule" geschrieben. Erickson geht mit der typografischen Auszeichnung von Randgesprächen, die zur Thematisierung der sich langsam entwickelnden Regeln dieses Genres initiiert wurden, auf eine weitere Konvention ein.[5] In einer weiteren Arbeit untersucht Erickson (2000) das synchrone und asynchrone Kommunikation unterstützende CMC-System „Babble", insbesondere werden mehrere thematisch und funktional eingeschränkte Diskussionen und schriftlich geführte Gespräche in einer Gruppe von 19 Benutzern als einzelne Genres konzeptualisiert.[6] Aufbauend auf den Begriffen des Genre-Systems (Bazerman, 1994) und des innerhalb einer Diskursgemeinschaft etablierten Genre-Repertoires (Orlikowski und Yates, 1994), führt Erickson den Terminus der „genre ecology" ein, der insbesondere auf die „conversational genres" von CMC-Systemen ausgerichtet ist und die Beziehungen und Abhängigkeiten zwischen ihnen bezeichnet.

Ein zweiter Forschungsstrang beschäftigt sich mit der Analyse von E-Mail-Systemen. Bergquist und Ljungberg (1999) untersuchen 66 E-Mails, die ein Mitarbeiter eines schwedischen Informationstechnologie-Unternehmens innerhalb einer Woche gesammelt hat, um die elek-

[4] Erickson (1999) untersucht die Beiträge der ersten sechs und letzten zwei Monate, um Konventionen zu ermitteln. In diesem Zeitraum wurden 327 Limericks erfolgreich produziert; fünf Limericks wurden abgebrochen.

[5] In der Tat spricht Erickson (1999) vom „On-Line Participatory Limerick Genre". Da die *North American Genre Theory* die innerhalb einer Diskursgemeinschaft etablierten Konventionen in den Mittelpunkt stellt, kann die von Erickson untersuchte „Dialogform" durchaus als Genre bezeichnet werden, da es irrelevant ist, wie viele Teilnehmer die Diskursgemeinschaft umfasst. Siehe hierzu auch Rosso (2005, S. 35): „Theoretically, two or three people can constitute a [discourse community]. If they communicate distinctively via web pages in recurring situations amongst themselves, then strictly speaking, this is a web genre."

[6] Diekmannshenke (2004) analysiert Chats, in denen sich Politiker den Fragen der Teilnehmer stellen, wobei in der Regel ein Moderator die Beiträge selektiert. Der Chat stellt eine eigene Kommunikationsform dar, die spezifische Typen herausbildet. Diekmannshenke (2004, S. 134 f.) kommt zu dem Schluss, dass es sich bei dem nach Abschluss eines derartigen Chats im WWW publizierten „Transskript eines Politik-Chats" erstens um einen „Text im Sinne der Textlinguistik" und zweitens um eine „Textsorte" handelt, für die „die spezifische und durch den Moderator konstruierte Abfolge der Beiträge der am Chat beteiligten Gruppen Moderator, Chatter und Politiker" charakteristisch ist (ebd.).

tronischen Arbeitspraktiken eines Angestellten dieser Firma bezüglich der Genres zu analysieren, die er verwendet oder mit denen er konfrontiert wird. Im Gegensatz zu Yates und Orlikowski (1992) und Erickson (2000) gehen Bergquist und Ljungberg davon aus, dass per E-Mail geführte Dialoge und informelle sowie verwaltungsbezogene Unterhaltungen nicht als selbstständige Genres anzusehen sind, sondern vielmehr mit dem Zweck eingesetzt werden, sich über ein zu verwendendes Genre auszutauschen. Zusätzlich werden zahlreiche Beispiele instanziierter Genres in den 66 untersuchten Mails beobachtet: In den externen Nachrichten des Angestellten finden Bergquist und Ljungberg die Genres „Newsletter", „Sales promotion letter", „Call for participation", und „Digest". Die internen Nachrichten werden in die Genres „Internal Job Proposal", „Call for participation", „Organizational announcement", „Project report", „Project report summary", „Oral presentation of project report" und „Claim" eingeteilt. Boudourides et al. (2002) analysieren in ähnlicher Weise 1 104 Nachrichten der Mailing-Liste eines europäischen Verbundprojekts auf der Grundlage eines Schemas von 12 Form- und 24 Funktionsmerkmalen und finden acht Genres, die sich in Teilen deutlich von der von Bergquist und Ljungberg (1999) aufgestellten Liste unterscheiden: „Dialogue", „Team Announcement", „Socializing", „Distribution of Completed Work", „Reminder", „Group Decision", „Distribution of Project Work" und „Criticisms". Gruber (2000) hinterfragt, ob in wissenschaftlichen E-Mail-Diskussionslisten (vgl. Schütte, 2004b) veröffentlichte Beiträge ein „scholarly email posting"-Genre darstellen und untersucht mehrere Themenstränge der „Linguist-List" und der „Ethno-List" mit qualitativen und quantitativen Verfahren. In quantitativer Hinsicht ähneln sich die beiden Diskussionslisten sehr, sie vereinigen sowohl Merkmale für konzeptionelle Mündlichkeit (vgl. Koch und Oesterreicher, 1994, sowie die Abschnitte 2.2.7 und 8.3) als auch schriftsprachliche Elemente. Dennoch kann Gruber zufolge nicht eindeutig von einem einzigen neuen Genre gesprochen werden, da Unterschiede hinsichtlich der Bezugnahme auf vorhergehende Beiträge bestehen, weshalb die jeweiligen Beiträge als Instanzen von Subgenres betrachtet werden.

4.2.2 Digital Genres und Computer-Supported Collaborative Work

Die Bereiche *Computer-Mediated Communication* (CMC) und *Computer-Supported Collaborative Work* (CSCW) sind eng miteinander verbunden. CSCW kann als ein Anwendungsszenario von CMC-Systemen aufgefasst werden, in dem primär Prozesse des institutionalisierten gemeinschaftlichen Arbeitens und Lernens im Vordergrund stehen. CSCW-Systeme werden häufig unter dem Stichwort *Groupware* zusammengefasst.

Yates et al. (1997) untersuchen den Einsatz von *Team Room* (eine Software zur Unterstützung der Gruppenkommunikation innerhalb von *Lotus Notes*) in einem US-amerikanischen Technologieunternehmen. Das Werkzeug wird in drei Gruppen eingesetzt, wobei Yates et al. insgesamt 492 Nachrichten für Analysen zur Verfügung stehen. Die Teams verwenden drei Genre-Systeme, wobei 43% aller Nachrichten Instanzen mindestens eines Systems sind: Das Genre-System „Meeting Documentation" besteht aus den Genres „meeting logistics", „meeting agenda" und „meeting minutes". Während die Agenda und das Protokoll etablierte Genres darstellen, bezieht sich „meeting logistics" auf die Planungsprozesse, die vor einem Treffen erfolgen (Ankündigung oder Verschiebung eines Termins etc.). Yates et al. (1997, S. 53) sind der Ansicht, dass die drei Genres einen identischen Abstraktionsgrad besitzen und sich auf

das allgemeinere „meeting genre“ beziehen. Das zweite Genre-System wird als „Collaborative Repository“ bezeichnet und umfasst die Genres „placeholder“ und „response“. Ausgehend von einer „placeholder“-Instanz als Diskussionsgrundlage werden die anderen Mitglieder gebeten, Anmerkungen und Verbesserungsvorschläge beizusteuern. Das dritte Genre-System, „Collaborative Authoring“, besteht aus den Genres „circulated draft“ und „reaction to draft“. Yates et al. zeigen, dass ehemals auf traditionellen Medien (persönliches Gespräch, Memo etc.) basierende Prozesse der Gruppenkommunikation durch ein hierfür entwickeltes Werkzeug unterstützt werden können. Während vergleichbare Werkzeuge die unterschiedlichen Kommunikationsmöglichkeiten meist explizit vorgeben (und somit einschränken), erfolgte die Strukturierung der drei Genre-Systeme durch die Gruppenmitglieder selbst.

Ausgehend von einer Untersuchung der 1 018 Nachrichten der *Lotus Notes* Datenbanken, die von zwei Projektgruppen einer Versicherungsgesellschaft eingesetzt werden, entwickeln Schultze und Boland (1997) eine Differenzierung von „hard“ und „soft information genres“, worunter unterschiedliche Grade der Ausprägung korrespondierender Regelsysteme verstanden werden: Je mehr Mitglieder einer Diskursgemeinschaft an der Definition von Regeln und Normen eines Genres beteiligt sind, desto eher wird das Genre verhärtet. Diese Differenzierung ist als Kontinuum aufzufassen. Schultze und Boland gehen von den vier Dimensionen „Substance“ (Inhalt), „Medium“, „Language“ und „Structure“ aus, hinsichtlich derer härtere und weichere Genres charakterisiert werden können. Härtere Genres umfassen z. B. Kommunikationsereignisse mit einer etablierten Bedeutung oder die Darstellung vergangener Ereignisse zur Unterstützung einer Argumentation. Weichere Genres enthalten eher subjektive Meinungsäußerungen oder noch nicht abgeschlossene kommunikative Ereignisse. In härteren Genres wird präzise und objektiv argumentiert, während die Produzenten in weicheren Genres häufig mehrdeutige, subjektiv gefärbte und kontroverse Aussagen tätigen, die sprachlich individuell gestaltet und vom „Ich“-Erzähler dominiert werden.

Yates und Sumner (1997) hinterfragen, ob die rapide Entwicklung digitaler Technologien einen destabilisierenden Effekt hinsichtlich der Verwendung von Genres mit sich führt. Es ist schließlich möglich, Dokumente zu modifizieren, ohne sichtbare Spuren zu hinterlassen, weshalb nicht gewährleistet ist, dass sie über längere Zeiträume unverändert bleiben. Der hierdurch entstehenden Bürde, eine gewisse Festigkeit („fixity“) der Kommunikation zu erzeugen (vgl. Jucker, 2004), wird Yates und Sumner zufolge mit einem stärkeren Rückgriff auf Genres begegnet. Zwar haben die digitalen Technologien Prozesse der Destabilisierung initiiert, da jedoch, so Yates und Sumner, die Unterschiede zwischen Produzenten und Rezipienten zunehmend verschwimmen, ist das Resultat dieser Prozesse nicht etwa der Zusammenbruch von Genres, sondern die Demokratisierung ihrer Entstehung: Mehr und mehr Benutzer bringen sich in die Erstellung von Texten ein, so dass sich innerhalb von Diskursgemeinschaften immer spezifischere Genres herausbilden, die ihre Kommunikationsbedürfnisse zunehmend besser unterstützen. Yates und Sumner exemplifizieren diese These anhand zweier Studien: Innerhalb des CMC-Konferenzsystems einer Fernuniversität haben sich in vier sozialen Kontexten fünf Genres etabliert: „Focused interactive interpersonal discussion“ und „Open interactive interpersonal communication“ (Kontext: „Chat“), „Academic's discussion“ („Academic“), „Student/Tutor discussions“ („Teaching“) und „Non-Academic discussion“ („Administrative“). Eine zweite Studie untersucht die Evolution des Genres „Design Representation“ (für telefonbasierte Dialogsysteme), das sich von einer textbasierten Form

unter zunehmender Verfügbarkeit neuer Technologien (z. B. Datenbanken und Werkzeuge zur Erstellung von Flussdiagrammen) zu einem komplexen Genre entwickelt hat, in dessen Zentrum ein Ablaufdiagramm steht, das von Tabellen und Textabschnitten flankiert wird.

4.2.3 Digital Genres und Document Management

Karjalainen et al. (2000) untersuchen *Enterprise Document Management*-Systeme und schlagen eine Analyse der verwendeten Genres auf der Grundlage von Metadaten vor (z. B. Zielgruppe, Einsatzbereich etc.): Nur ein explizit repräsentiertes Genre-Repertoire mitsamt assoziierter Metadaten sei eine solide Basis für EDM-Systeme (vgl. Tyrväinen und Päivärinta, 1999). Als Fallstudie dient eine finnische Organisation, die zunächst hinsichtlich ihrer Organisationseinheiten analysiert wird, woraufhin in Kooperation mit Angestellten einzelne Genres identifiziert werden. Es ist zu beachten, dass etablierte, „harte" Genres spezifischere Metadaten umfassen als „weiche" Genres (vgl. Schultze und Boland, 1997, und Karjalainen und Salminen, 2000). Es wird ein Genre angenommen, sobald mindestens zwei Mitarbeiter im Arbeitsalltag eine gemeinsame Benennung verwenden. Anschließend werden von den Angestellten auf der Basis von Formularen die mit Genres assoziierten Metadaten erhoben. Durch Zuhilfenahme einer Tabellenkalkulation, in der die Genres sowie ihre Metadaten abgetragen werden, wird entschieden, ob eventuell essenzielle, jedoch noch weiche Genres durch normierende Regeln verhärtet werden sollten. Diese Methode ignoriert zwar Genre-Systeme, doch kann die gesamte Tabelle als Formalisierung des Genre-Repertoires der Organisation betrachtet werden. Sehr zur Überraschung von Karjalainen et al. werden 850 Genres mit jeweils 19 Metadatenwerten identifiziert, deren Verwendungshäufigkeit, Etabliertheitsgrad bezüglich ihrer Benennung und Wichtigkeit für die Organisation zur Aufstellung eines Leistungskatalogs des EDM-Systems verwendet wird. Honkaranta und Lyytikäinen (2003) benutzen eine modifizierte Version dieser Methodologie, in der XML-DTDs zur Generalisierung der Genres einer Kirchengemeinde verwendet werden. Ermittelt werden 54 unterschiedliche Genres, in deren Zentrum der „monthly service calendar" steht; die Genres „working plan", „year calender", „vacation list", „newspaper announcement" und „outdoor announcement" sind Honkaranta und Lyytikäinen zufolge eng miteinander verbunden. Diese Genres werden gemeinsam mit den als Domänenexperten fungierenden Mitarbeitern der Gemeinde analysiert, wobei Metadaten erhoben und einzelne Bestandteile expliziert werden: Der „monthly service calendar" enthält z. B. ein „topic" namens „service" mit „subtopics" wie „date", „name" und „target audience". Die Ergebnisse werden in Form von DTDs formalisiert und können ebenfalls als Verhärtungen von Genres aufgefasst werden (vgl. auch Honkaranta, 2003).

4.2.4 Fazit – Zum Verhältnis von Medium und Textsorte

Die in diesem Abschnitt vorgestellten Studien sind auch für die Charakterisierung von Hypertextsorten relevant: Die Differenzierung zwischen harten und weichen Genres (Schultze und Boland, 1997) bezieht sich auf die Verortung digitaler Genres auf einem Kontinuum, dessen Pole eine Art maximal ausgeprägte Standardisierung und den Status von Prototextsorten repräsentieren. Die Anwendung von Genres durch die Mitglieder einer Diskursgemeinschaft resultiert in ihrer zunehmenden Verhärtung. Werden alle Benutzer des WWW als Diskursgemeinschaft (bzw. eine Menge von Diskursgemeinschaften) aufgefasst, kann postuliert

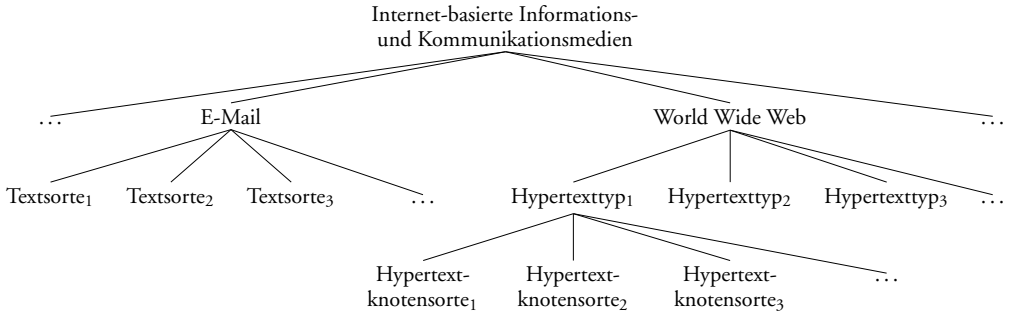


Abbildung 4.1: Text- bzw. Hypertextsorten in unterschiedlichen Internet-Diensten

werden, dass die von sehr vielen Benutzern frequentierten Angebotstypen (z. B. Suchmaschinen, Kataloge, Buchhändler und Auktionshäuser) standardisierteren Genre-Regeln unterliegen als Typen, die nur von wenigen Benutzern besucht werden; letztere basieren auf weichen Genres, deren Anzahl überwiegt. Darüber hinaus wird deutlich, dass das Medium einen Einfluss auf die Entwicklung digitaler Genres besitzt (vgl. Yates und Orlikowski, 1992): Ein Wechsel der eingesetzten Software kann mit der Formulierung neuer Genre-Regeln verbunden sein. Bezüglich der Analysen von Genres, die in der Organisationskommunikation verwendet werden, nehmen Honkaranta und Lyytikäinen (2003) eine Operationalisierung durch ausführliche Interviews und Gruppengespräche vor, in denen Domänenexperten Auskunft über die Benennungen und Merkmale der von ihnen tagtäglich verwendeten Genres geben. Eine vergleichbare Methodologie zur Ermittlung einer detaillierten Typologie von Hypertextsorten mit einem hohen Abdeckungsgrad ist nicht durchführbar (vgl. Abschnitt 4.4).

In Kapitel 3 wurde die Frage gestellt, ob Hypertext als Textsorte aufzufassen ist, da diese Meinung wiederholt vertreten wird (vgl. Fußnote 1, S. 65). Sie dürfte auf den größeren Einfluss der technologischen Komponente zurückzuführen sein: Wenn in der Textlinguistik von Texten oder Textsorten die Rede ist, sind fast ausnahmslos geschriebene oder gedruckte Texte gemeint. Die Tatsache, dass Papier dabei als Medium fungiert, wird nur in den seltensten Fällen explizit gemacht. Ebenso jedoch, wie Papier als *einzelnes* Trägermedium mit *mehreren* Ausprägungen (vom Schmierzettel über den Kassenbon zu einer Schreibmaschinenseite hin zum Werbeplakat an einer Bushaltestelle) aufzufassen ist, existieren aufgrund der durch die Informationstechnologie gegebenen Flexibilität im Internet *zahlreiche* digitale Informations- und Kommunikationsmedien, die in der Informatik als Internet-Dienste bezeichnet werden und in gleicher Weise lediglich als Transportmittel von Texten, Daten und Informationen fungieren (vgl. Abbildung 4.1).[7] Bergquist und Ljungberg (1999) finden z. B. in 66 E-Mails Instanzen der unterschiedlichsten Textsorten. Neben E-Mail existieren weitere Medien wie z. B. das Usenet (vgl. Fußnote 153, S. 133), das WWW, IRC und Gopher, was möglicher-

[7] Die Abbildung visualisiert diesen Umstand anhand der Internet-Dienste *World Wide Web* und E-Mail (vgl. auch Orlikowski und Yates, 1994, Ziegler, 2002, und Fußnote 68, S. 59). Zusätzlich erläutert die Abbildung die in Abschnitt 3.8 eingeführte Unterscheidung von Hypertexttyp und Hypertextknotensorte. Die Konzepte Hypertexttyp und Hypertextsorte beziehen sich auf eine Website als funktionales Ganzes, die wiederum unterschiedliche HTML-Dokumente umfasst, die Hypertextknotensorten zugeordnet werden können, die ihrerseits von abstrakteren Hypertextknotentypen subsumiert werden (vgl. Kapitel 5).

weise den Umstand zu erklären hilft, dass gerade in den Anfangsjahren der Forschungen im Bereich *Computer-Mediated Communication* und in aktuellen textlinguistischen Arbeiten das jeweils betrachtete digitale Medium mit dem Konzept der Textsorte identifiziert wurde.[8] Bereits Yates und Orlikowski haben auf diese Problematik hingewiesen:

> [T]he concept of medium has often been confused with that of genre. Confusion arises when researchers compare genres of communication (e. g., memos or bulletins) with communication media (e. g., electronic mail or fax). Genres, however, may be physically created, transmitted, and stored in various media. Thus, comparing memos with electronic mail, for example, confounds the concept of communication medium with that of communication genre. (Yates und Orlikowski, 1992, S. 310)

Einen wesentlichen Beitrag zu dieser Diskussion liefert Schütte (2004a, S. 142), die mit Bezug auf Brinker (2001, S. 138 f.) die „WWW-spezifische Struktureinheit Website" als Kommunikationsform auffasst, weil sie – gemäß der Definition Brinkers (vgl. auch Abschnitt 2.3.5) – nur situativ und medial festgelegt, aber weder inhaltlich noch kommunikativ-funktional markiert ist (ähnlich bei Schönefeld, 2001, S. 57). Websites besitzen Schütte zufolge eine monologische Kommunikationsrichtung (vgl. Fußnote 79, S. 94): Der Kontakt zwischen Produzent und Rezipient verläuft zeitlich versetzt, räumlich getrennt, wird vom Medium Schrift dominiert und kann zusätzlich multimediale Elemente enthalten.[9] Im nächsten Schritt erfolgt nach Schütte (2004a, S. 143) eine thematische Spezifizierung der Kommunikationsform Website (eine kommunikativ-funktionale Spezifizierung wird nicht erwähnt), die somit „[z]ur Textsorte wird", die „selbst wiederum verschiedene Textsorten [enthält], für die sie, ähnlich wie das Fernsehen oder die Zeitung, als Trägerstruktur fungiert." (vgl. Abschnitt 4.5.2). Als Beispiele für WWW-spezifische, aber von der Website unabhängige Textsorten nennt Schütte die Homepage, die Sitemap sowie das FAQ-Dokument. Obwohl das Thema eine wichtige Rolle spielt, muss die Argumentation, dass eine Kommunikationsform bereits durch eine „thematisch-inhaltliche Spezifizierung" zu einer Textsorte wird (Schütte,

[8] Häufig wird auch das Internet als „Medium" oder „Kommunikationsmedium" bezeichnet, z. B. von Dürscheid (2004, S. 142): „Heterogen ist das Medium [Internet, G. R.] auch in technischer Hinsicht: Erstmals können sowohl alte als auch neue Kommunikationsformen (Telefon, Fax, Brief, Videokonferenz, E-Mail, Chat, Newsgroups, Instant Messaging u. a.) in einem Medium kombiniert werden." Bei den von Dürscheid aufgezählten Kommunikationsformen handelt es sich nicht um Teile eines Mediums „Internet", sondern um Internet-Dienste, die auf standardisierten Spezifikationen basieren (z. B. RFCs). Die von Dürscheid angesprochene Kombinationsmöglichkeit der Dienste ist nicht per se gegeben: Alle Internet-Dienste werden zwar naturgemäß über eine Software bedient, die sich auf einem Rechner mit Internet-Anschluss befindet (Mobiltelefone, PDAs und sonstige Endgeräte eingeschlossen), die Kombination bezieht sich jedoch einzig und allein auf die Tatsache, dass mehrere dieser Applikationen gleichzeitig auf ein- und demselben Rechner aktiv sein können. Es ist auch zu erwähnen, dass moderne Browser wie z. B. *Mozilla* zwar den Zugriff auf mehrere Internet-Dienste ermöglichen (WWW, Usenet, E-Mail und Chat), es handelt sich aber trotz der einheitlich erscheinenden Oberfläche um getrennte Applikationen mit spezifischen Protokollen.

[9] Es ergibt sich unmittelbar die Frage, in welchem Verhältnis das abstrakte Konzept Hypertext, der Internet-Dienst bzw. das Hypertextsystem WWW und das Medium Schrift (Brinker, 2001, S. 138) zueinander stehen. Es wird davon ausgegangen, dass es sich bei einem Hypertext- bzw. Hypermediasystem um ein Medium handelt, das weitere Medien umfassen kann – der Terminus Multimedia schließt die Schrift ausdrücklich mit ein. Die Manifestation dieser Bündelung von Medien findet in der Kommunikationsform Website des Internet-Dienstes WWW statt. Die Begriffe „Website" und „Hypertext" (im Sinne eines konkreten Textexemplars, das eine funktionale Ganzheit darstellt) werden im Folgenden als Synonyme verstanden.

2004a, S. 143), kritisch hinterfragt werden. Nicht das Thema ist diesbezüglich der determinierende Faktor, sondern die kommunikative Funktion. Schließlich kann *ein* Thema durchaus in Exemplaren *mehrerer* Textsorten behandelt werden.

Neben dem Verhältnis von Medium (bzw. Kommunikationsform) und Textsorte ist der Faktor der alltagssprachlichen Etikettierung einzubeziehen: Eine E-Mail wird meist als „E-Mail", „Mail" oder „Nachricht" bezeichnet, d. h. die eigentliche Textsorte tritt, obgleich sie in vielen Fällen zweifelsohne vorhanden ist, in den Hintergrund – vermutlich aufgrund des Mangels an individuellen Gestaltungsmöglichkeiten, der diesem Medium inhärent ist. Salient wird die Textsorte nur bei besonders markanter Ausprägung, z. B. im Falle des Erhalts einer Rechnung oder einer Einladung. Nach Ansicht von Honkaranta und Lyytikäinen (2003, S. 109) werden die Extensionen der Konzepte „Text" und „Dokument" aufgeweicht: Von einigen Rezipienten wird ein per E-Mail eingegangenes Memo als „Memo" bezeichnet, andere bezeichnen es schlicht als „E-Mail".[10] Noch verworrener, zugleich jedoch auch transparenter wird das Problem der Benennung bei Erhalt einer E-Mail ohne eigentlichen Textinhalt, der jedoch ein Anhang beiliegt, der ein Dokument beinhaltet, das mit einer Textverarbeitung erstellt wurde und alle formalen Kennzeichen eines Briefes aufweist (z. B. typografisch abgesetzte Angaben über Absender und Adressat, eine förmliche Anrede und eine Verabschiedungsfloskel). In diesem Fall tritt das Medium E-Mail in den Hintergrund und die Textsorte – aufgrund ihrer formalen Kennzeichnung – in den Vordergrund, so dass nicht mehr von der „Mail", sondern tatsächlich von einem „Brief" oder einer „Mail", der ein „Brief" beiliegt, die Rede ist. Diese Beobachtungen können auf Hypertextsysteme übertragen werden, denn ebenso wie Textexemplare unterschiedlicher Textsorten im Medium E-Mail existieren, können im WWW Exemplare unterschiedlicher Hypertextsorten beobachtet werden.

4.3 Zur Entstehung und Etikettierung von Hypertextsorten

Die Entstehung von Hypertextsorten ist noch weitgehend unerforscht.[11] Shepherd und Watters (1998) haben eine Evolutionstaxonomie vorgelegt, die von vielen nachfolgenden Arbeiten als Referenz zu diesem Themenkomplex übernommen wurde. Bevor diese Taxonomie vorgestellt und durch zwei zusätzliche Modelle komplementiert wird, thematisiert der nachfolgende Abschnitt die alltagssprachliche Etikettierung von Hypertextsorten.

4.3.1 Aspekte der Etikettierung von Hypertextsorten

Der komplexe Entstehungsprozess von Textsorten, ihre allmähliche Formierung und Variation wird vornehmlich aus sprachgeschichtlicher Perspektive diskutiert. Heinemann und Viehweger (1991, S. 144) weisen darauf hin, dass mit „der Herausbildung und Spezifizierung von

[10] Im Bereich traditioneller Medien besitzt dieses Phänomen durchaus Parallelen: Wissenschaftliche Arbeiten im weitesten Sinne (Zeitschriftenartikel, Konferenzbeiträge, Vorträge etc.) werden häufig als „Papier" bezeichnet, doch wohl kaum jemand würde eine Steuererklärung, einen Bußgeldbescheid, ein Zeugnis, einen Sammelband, die Tageszeitung oder den Wetterbericht in dieser Art benennen.

[11] Es existieren jedoch einige Studien, in denen spezifische Webseiten über längere Zeiträume beobachtet werden, (Eriksen und Ihlström, 2000, Ryan et al., 2003). Diese Studien werden in Abschnitt 4.6 dargestellt.

Kommunikationsbereichen [...] weitere Klassifizierungen entstanden [sind], die das Textsortenpotential einer Gemeinschaft erweitern." Es kann nun argumentiert werden, dass durch die Etablierung des Kommunikationsbereiches *World Wide Web* und seine globale Akzeptanz neue „Klassifizierungen" entstanden bzw. noch in der Entwicklung befindlich sind. Der Begriff bezieht sich nach Heinemann und Viehweger auf „Alltagsklassifikationen, die innerhalb einer menschlichen Gemeinschaft erreicht und mit Lexikonzeichen belegt wurden, die das Wissen über eine bestimmte Textsorte »kondensieren«." (ebd.). Auch Sandig (1997, S. 26) geht davon aus, dass eine Bezeichnung vorliegen muss: „Der gesellschaftlichen Relevanz entsprechend gibt es in der Gesellschaft (mindestens) eine Benennung [für eine Textsorte]."

Abschnitt 2.3.2 hat bereits angedeutet, dass im Kontext des WWW nur wenige Benennungen existieren. „Homepage" oder die abstrakteren Termini „Webseite" bzw. „Seite" können sich prinzipiell auf sehr viele unterschiedliche Typen von Knoten beziehen und treten, die institutionelle Dimension des referenzierten Informationsangebots näher bestimmend, eher als Teil von Komposita auf, z. B. „Uni-Homepage", „Firmen-Homepage" oder „Vereinsseite". Die Begriffe „private Homepage" oder „persönliche Homepage" werden meiner Ansicht nach im Alltag (z. B. in Gesprächen oder E-Mails) deutlich seltener als die angesprochenen Komposita verwendet und die Extension des Begriffs „Portal", der seit etwa 1997 benutzt wird, ist nicht eindeutig bestimmbar, häufig wird er im Alltag synonym zu „Homepage" verwendet.[12] Eher wird direkt auf die Typen bzw. Bezeichnungen von Informationsangeboten oder die Namen der zugehörigen Firmen oder Organisationseinheiten verwiesen (z. B. „Google", „Suchmaschine", „Seite von Microsoft", „Instituts-Homepage" etc.). Dass sich die Einstiegsseiten der Webauftritte von Universitäten und Firmen in grundsätzlicher Weise voneinander unterscheiden, spielt für die Benutzer des Mediums WWW, also für die Rezipienten der jeweiligen Angebote, offenbar keine entscheidende Rolle, denn spezifischere lexikalische Etiketten wie z. B. – in Bezug auf die traditionellen Textsorten – *Geschäftsbrief*, *Kochbuch*, *Arztrezept*, *Telegramm* oder *Zeitschriftenartikel* haben sich im WWW bislang nicht herauskristallisiert. Die genannten Textsorten zeigen einen möglichen Grund für diesen Umstand: Die zugehörigen Textexemplare unterscheiden sich, obgleich ausnahmslos im Medium Papier realisiert, in materieller Hinsicht (z. B. Durchschnittsumfang, Funktion, Transportweg etc.) – das *look and feel* solcher Textexemplare ist vollkommen unterschiedlicher Natur.[13] Im WWW hingegen basiert jedes Dokument auf der Sprache HTML und unterliegt somit sehr einschränkenden Rahmenbedingungen, was in einem sehr konformen Umgang, einem nahezu identischen *look and feel* einer jeden Webseite resultiert (vgl. Fußnote 139, S. 123). Abgesehen von den haptischen Unterschieden kann über weitere Gründe spekuliert werden: Das häufig mit der Metapher des Surfens umschriebene Navigieren im Netz findet in den meisten Fällen allein statt, d. h. eine einzelne Person verwendet einen Rechner und rezipiert zu einem spezifischen Zweck oder aus purem Zeitvertreib verschiedene Webseiten. Es kommt

[12] Brandl (2002) hat Interviews mit zehn Experten durchgeführt, in denen Typen von Websites aufgeführt werden sollten. Alle Probanden nennen zwar das „Portal", eine eindeutige Definition kann aber nicht ermittelt werden; die Gemeinsamkeiten beziehen sich auf seine Funktion als Eingangstor ins WWW. Insgesamt werden 36 Bezeichnungen erhoben, die Brandl in 11 Typen zusammenfasst (vgl. Abschnitt 4.4.5).

[13] Nielsen (1995b, S. 5) benutzt den *look and feel*-Begriff zur Charakterisierung von Hypertextsystemen: „When asked whether I would view a certain system as hypertext, I would not rely so much on its specific features, command, or data structures, but more on its user interface "look and feel"."

hinzu, dass sich nur ein Bruchteil der Menschen, die das WWW verwenden, in diesem Medium auch als Textproduzenten betätigt.[14] Der ordnende, sortierende oder informierende Umgang mit Webseiten findet teilautomatisiert statt: Eigene Lieblingsdokumente – „Lesezeichen", „Bookmarks", „Favoriten" – werden auf Knopfdruck mit Titel und Adresse in eine Liste eingetragen, die bequem nach meist inhaltlichen Kriterien gepflegt werden kann. Wenn man einen Freund oder Arbeitskollegen über eine relevante oder interessante Webseite informieren möchte, kann ein materieller Austausch – vom Ausdruck der Seite abgesehen – nicht stattfinden. Da URLs sehr lang sind und einen komplizierten Aufbau besitzen, bleiben die Benutzer häufig bei den Kommunikationsdiensten des Internet und integrieren einen Link in eine auf der persönlichen Homepage befindliche Hyperlinkliste[15] oder verschicken eine Mail mit der oder den URLs (eingefügt per *Copy & Paste*), flankiert von einem knappen Hinweis wie z. B. „Schon bekannt?". In solchen Fällen entsteht kein Bedarf, die URL zusätzlich mit einem abstrakteren Etikett zu versehen. Diese Gründe sind nur einige Indikatoren dafür, dass für die Diskursgemeinschaft der WWW-Benutzer keine Notwendigkeit besteht, unterschiedliche Typen von Dokumenten mit spezifischeren Etiketten als z. B. „Homepage", „Firmen-Homepage" oder „Uni-Homepage" zu belegen, weil das alltägliche Rezipieren von Webseiten in zwischenmenschlicher Hinsicht einer kommunikativen Isolation unterliegt.

Der rezipierende Umgang mit dem WWW hat Auswirkungen, die sich aber nicht auf die Etikettierung beziehen: „Textsorten [...] stellen [...] ein Potential, ein bestimmtes Reservoir an Kenntnissen dar, auf die die Mitglieder einer menschlichen Gemeinschaft in ihrer sprachlichen Tätigkeit zurückgreifen. Struktur und Umfang dieses zur Lösung von Kommunikationsaufgaben bestimmten Potentials [...] werden ganz entscheidend durch die Kommunikationsbedürfnisse determiniert, die in einer menschlichen Gemeinschaft bestehen." (Heinemann und Viehweger, 1991, S. 144). Auch ohne explizite Etiketten entwickeln Benutzer Erwartungen an unterschiedliche Typen von Webseiten, so kann z. B. antizipiert werden, dass auf der persönlichen Homepage eines Wissenschaftlers eine Publikationsliste und eine Aufstellung der Forschungsschwerpunkte zu finden sind. Auf der Einstiegsseite des Webauftritts einer Hochschule erwartet der Leser, rasch die Verweise zu ihren Fakultäten, Instituten oder zum Angebot der Universitätsbibliothek finden zu können.[16]

Gleichzeitig lassen sich diejenigen WWW-Benutzer, die auch als Produzenten von Angeboten tätig sind, in der Strukturierung, Gestaltung und Verschriftlichung der von ihnen publizierten Inhalte von dem „Reservoir an Kenntnissen" leiten, das sie bei der Benutzung des WWW sukzessive aufgebaut haben und kontinuierlich weiter aufbauen. Die Usability-

[14] Die kleinere Gruppe der Produzenten (insbesondere professionelle Webdesigner) kennt und verwendet durchaus ein spezifischeres Vokabular von Bezeichnungen, wie z. B. anhand der Ratgeberliteratur gezeigt werden kann (vgl. Abschnitt 3.6.5). Dieses Vokabular wird jedoch nicht immer in den Textexemplaren selbst eingesetzt, d. h. die alltäglichen Benutzer haben üblicherweise keine Kenntnis von ihrer Existenz.

[15] Walker (2000, S. 107) ermittelt anhand einer empirischen Untersuchung diverse Intentionen, die Autoren privater Homepages mit ihrem Webangebot verfolgen. Einer der Befragten pflegt seine Liste von Lesezeichen auf der eigenen Homepage: „It's my public hotlist. When I tell people there's something useful on the web, instead of reciting URLs and sounding like a space alien, I can just tell them where to find it on my page."

[16] Shepherd und Watters (1999, S. 1 f.) geben ein weiteres Beispiel: „[U]sers expect to interact with a Web-based search engine by inputting a search string, viewing a dynamically composed set of responses [...], viewing selected Web pages, and revising the search string [...]. They expect the search engine to search its indices, to compose the set of responses, and to retrieve and display the selected Web pages. Once users have used one such search engine, they can easily transfer these expectations of functionality to other instances of this genre."

Literatur rät, in spezifischen Typen von Webseiten bzw. Websites rekurrent auftretende modulartige Bestandteile durch ein konventionalisiertes Vokabular zu kennzeichnen: „Frequent features should have familiar names. This makes it easier [...] to find these features" (Nielsen und Tahir, 2002, S. 45). Konventionalisierte Etiketten sind also primär auf der Ebene der Benennung einzelner Komponenten zu beobachten. Für die Herausbildung spezifischer Hypertextsortenbezeichnungen zur Differenzierung unterschiedlicher Typen von Websites in der Alltagskommunikation besteht hingegen für das Gros der WWW-Benutzer – diejenigen Personen, die nicht in professioneller Weise große Webauftritte erstellen oder betreuen – keine Notwendigkeit.[17] Dies erklärt, weshalb die Autoren der Studien, die in Abschnitt 4.4 vorgestellt werden, immer wieder Bezeichnungen von identifizierten Hypertextsorten kreieren (ersichtlich anhand von Aussagen wie „A related genre we called [...]", Crowston und Williams, 2000, S. 208). In einigen Fällen konnten sich Namen von Hypertextsorten aufgrund ihrer Verwendung in Textexemplaren etablieren (wie z. B. in „Willkommen auf der *persönlichen Homepage* von [...]" oder „Dies ist die *persönliche Homepage* von [...]"). Die in Teil III dargestellten Analysen gehen genauer auf den Aspekt der Etikettierung ein.

4.3.2 Zum Prozess der Entstehung von Hypertextsorten

Abbildung 4.2 stellt eine modifizierte Version der von Shepherd und Watters (1998) aufgestellten Taxonomie zur Entstehung von „cybergenres" dar.[18] Die Taxonomie ist in ihrer ursprünglichen Ausprägung nicht auf das Medium Hypertext beschränkt, sondern bezieht sich prinzipiell auf *alle* digitalen Medien.[19] Die Taxonomie wird als „fuzzy" (ebd., S. 97) bezeichnet: Die ersten drei Subklassen sind nicht als trennscharfe Kategorien, sondern eher als eine Art evolutionäres Kontinuum aufzufassen; die durchbrochenen Pfeile repräsentieren „evolutionary paths" (ebd.) (vgl. Holloway, 1987). Shepherd und Watters (1998, S. 98) unterscheiden zwischen traditionellen Genres und neuen Hypertextsorten, die als „not like any existing genre in any other medium" charakterisiert werden.[20] Die im WWW existen-

[17] Haas und Grams (1998b, S. 100) reißen diesen Aspekt nur an: „In conversation, we use terms such as *home page* or *zine article* to describe Web pages, and rarely need to clarify what we mean." Crowston und Williams (1997, S. 34) thematisieren den Aspekt in der Darstellung ihrer Stichprobenanalyse: „Other pages represented types of communication that are stereotyped, but not usually named [...]."

[18] Jakobs (2003, S. 246 und S. 249) geht auf die „Herausbildung" und „Entwicklung neuer Hypertextsorten" ein und bezieht sich auf die von Kuhlen (1991, vgl. Abschnitt 3.6.2) vorgeschlagenen Konversionstypen, ohne jedoch den eigentlichen Vorgang der Entstehung zu thematisieren. Unklar ist diesbezüglich die Relevanz zweier weiterer Aspekte: „Im ersten Jahrzehnt des 21. Jahrhunderts wird die Erstellung von Hypertexten leichter durch die Erweiterung von Textverarbeitungsprogrammen durch Komponenten, die [...] XML unterstützen – eine wichtige Voraussetzung für Hypertext. E-Mail kann seit neuestem auch als HTML-Dokument verschickt werden." (ebd., S. 249). Während der Einfluss von XML bereits in Fußnote 151 (S. 132) diskutiert und zurückgewiesen wurde, kann auch die *technische Funktion* des Verschickens elektronischer Post als HTML-Dokument als irrelevant für die *Entstehung* von Hypertextsorten eingestuft werden. Dass jedoch die *Kommunikationsform* E-Mail die individuelle Ausgestaltung von Hypertextexemplaren beeinflusst, wird Kapitel 9 zeigen.

[19] Erstaunlicherweise sprechen Shepherd und Watters (1998, S. 97) nur von einem einzelnen Medium (vgl. auch Abschnitt 4.2.4): „The combination of the computer and the Internet has been such a powerful trigger that it has resulted in the emergence of a new class of genre, which we call cybergenre, existing in this new medium."

[20] In der Taxonomie befinden sich – mit Ausnahme des als „cybergenre" etikettierten Wurzelknotens – keine spezifizierenden Nomina an den Knoten (Shepherd und Watters, 1998, S. 98). Im Hinblick auf die Entwicklung von Hypertextsorten ist jedoch eine Unterscheidung zwischen traditionellen Textsorten und Hypertextsorten

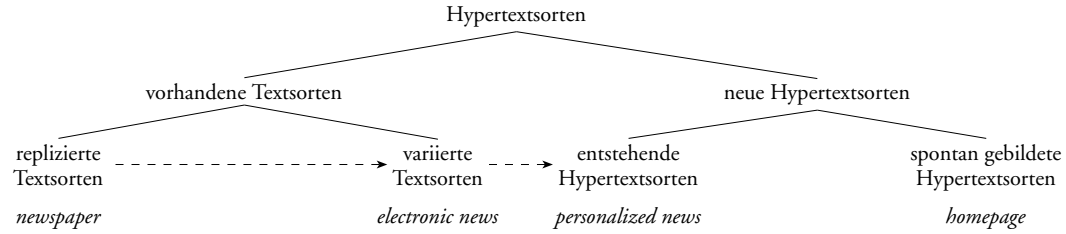


Abbildung 4.2: Evolution von Hypertextsorten (nach Shepherd und Watters, 1998, S.98)

ten Instanzen traditioneller Textsorten basieren auf replizierten oder variierten Textsorten aus anderen Medien. Im digitalen Medium wird eine traditionelle Textsorte meist vollständig nachgeahmt, ohne die technologischen Möglichkeiten auszuschöpfen. Die evolutionäre Entwicklung wird, so Shepherd und Watters (1998, S. 98), von der „new functionality afforded by the new medium“ vorangetrieben und kann in der dritten Phase in neuen Hypertextsorten münden, die in anderen Medien nicht existieren bzw. nicht existieren können.[21] Aus diesem Grund wird die Ansicht vertreten, dass „cybergenres“ nicht nur durch ihren Inhalt und ihre Form, sondern zusätzlich durch ihre Funktionalität charakterisiert werden müssen, die als „capabilities now available through the new medium“ aufgefasst werden (ebd., S. 99). Der langsame Ausbau primär funktionsorientierter Eigenschaften gewährleistet demnach eine gewisse Festigkeit (vgl. Yates und Sumner, 1997) hinsichtlich Inhalt und Form, so dass die Rezipienten weiterhin in der Lage sind, die variierte Textsorte oder entstehende Hypertextsorte als neue Ausprägung einer traditionellen Textsorte zu erkennen (Shepherd und Watters, 1998, S. 102, vgl. auch Watters und Shepherd, 1997a). Zusätzlich ist die Entstehung gänzlich neuer Hypertextsorten möglich, denn „the new medium supports the spontaneous creation of new genres that have never existed in other media.“ (Shepherd und Watters, 1998, S. 98).

Shepherd und Watters (1998, S. 99) führen nur vage Unterscheidungskriterien zwischen den einzelnen Kategorien an: Ein Cybergenre ist als repliziert aufzufassen, wenn das Exemplar dem traditionellen Genre bezüglich Inhalt und Form entspricht und durch das digitale Medium keine neue Funktionalität entsteht. Exemplare variierter Textsorten hingegen bedienen sich einiger Funktionen des digitalen Mediums, d. h. Inhalt und Form unterscheiden sich in substanzieller Weise vom Quellmaterial.[22] Als Beispielvariante eines „standard text document“ wird „a hyperlinked document with images or video components“ angegeben.[23] Neue

hilfreich, weshalb die Knotenbeschriftungen übersetzt und um diese Begriffe ergänzt wurden; auf eine Unterscheidung von Hypertexttypen und Hypertextknotensorten wird verzichtet. Shepherd und Watters verwenden die Begriffe „extant“, „novel“, „replicated“, „variant“, „emergent“ und „spontaneous“.

[21] Shepherd und Watters (1999, S. 2) präzisieren den Begriff „functionality“: „Although *purpose* is not an attribute of our characterization of cybergenre, functionality cannot be discussed without reference to the goal or purpose of the genre. [...] Therefore, purpose must be viewed only from the perspective of the author of the site and thus, the functionality incorporated into the site is driven by this purpose.“

[22] Ihlström und Åkesson (2004) analysieren 85 schwedische Online-Zeitungen und erklären, dass sie wegen der Ähnlichkeit zu ihren traditionellen Pendants keine neuen Hypertextsorten darstellen. Es werden Exemplare beobachtet, die auf replizierten und variierten Textsorten basieren. Zur Repräsentation einer dritten Gruppe wird ein als „progressed“ bezeichneter Knoten jenseits von „variant“ eingeführt (vgl. Abschnitt 4.6.4).

[23] Der entscheidende Aspekt, welche Textsorte bzw. welches Cybergenre das „standard text document“ besitzt, wird von Shepherd und Watters nicht thematisiert.

Hypertextsorten basieren vollständig auf der Funktionalität des neuen Mediums. Sie können entweder den in der Taxonomie angedeuteten evolutionären Prozess durchlaufen oder sich als eigenständige Hypertextsorten ohne Pendant in anderen Medien etablieren.[24] Für die Kategorie der spontan entstandenen Hypertextsorten geben Shepherd und Watters die Beispiele „home page", „hot list" und „FAQ" an, wobei jedoch nicht diskutiert wird, inwiefern ein FAQ-Dokument auf die Funktionalität eines digitalen Mediums angewiesen ist.

Das von Shepherd und Watters vorgeschlagene Modell verdeutlicht die Entwicklung von Hypertextsorten, doch machen vier Aspekte eine Erweiterung notwendig. Im Vordergrund steht nicht die hierarchische Anordnung von Klassen und Subklassen, sondern die Evolution von Hypertextsorten aus vorhandenen Textsorten, die von spontan gebildeten neuen Hypertextsorten flankiert werden. Die Darstellung in Form einer Taxonomie lenkt von der wesentlichen Aussage des Modells ab. Des Weiteren ist die Konzeptualisierung des Entwicklungsprozesses als replizierte, variierte und entstehende (Hyper)textsorten unvollständig, schließlich können entstehende Hypertextsorten weitere evolutionäre Schritte erfahren. Eine zukünftige Veränderung dieses finalen Stadiums des Kontinuums kann von der Taxonomie nicht reflektiert werden; bei der Verwendung und Entwicklung von Textsorten handelt es sich um einen zyklischen Prozess, der auch als solcher zu repräsentieren ist. Weiterhin ist der Skopus der Taxonomie unspezifisch, Shepherd und Watters beziehen sich potenziell auf *alle* digitalen Informations- und Kommunikationsmedien. Des Weiteren fehlen in der Taxonomie die wichtigsten Einflussfaktoren, die Veränderungen einer Textsorte bewirken.

Zur Entstehung von Hypertextsorten bei der Konvertierung von Dokumenten

Die im Folgenden eingeführten zyklischen Modelle zur Entstehung von Hypertextsorten beziehen sich auf HTML-Dokumente im WWW, wobei zwischen der automatischen und der manuellen Herstellung von Dokumenten unterschieden wird (vgl. Abschnitt 3.3.6).

Der erste Fall betrifft die maschinelle Konvertierung verfügbarer Dokumente, die im *World Wide Web* publiziert werden sollen, nachdem sie ursprünglich z. B. für den Ausdruck angefertigt wurden. Der Kernaspekt des in in Abbildung 4.3 dargestellten Modells betrifft die Tatsache, dass nicht die automatisch erstellten HTML-Dokumente, sondern sowohl das zugrunde liegende Quelldokument als auch die Konvertierungssoftware Änderungsprozessen unterliegen. Modifikationen des Quelldokuments können anhand eines Lebenszyklus konzeptualisiert werden (Lobin, 2000).[25] Im Quelldokument vorgenommene Modifikationen machen

[24] Shepherd und Watters (1998, S. 99) führen zusätzlich die beiden Merkmale „persistent" (Inhalte und Form sind statisch) und „virtual" (Inhalte und Form werden dynamisch generiert) ein, die für neue Hypertextsorten gelten. Die Evolutionsschritte von replizierten, über variierte hin zu entstehenden Hypertextsorten werden anhand von Online-Zeitungen dargestellt (Shepherd und Watters, 1998, S. 100 f., vgl. auch Abbildung 4.2). Als Beispiel einer „virtual instantiation of news content" werden Diskussionsforen und die personalisierte Filterung von Inhalten auf der Grundlage von Agenten genannt.

[25] Als Quelldokumente kommen Texte in beliebigen Formaten wie z. B. LaTeX, troff, *Microsoft Word* und *Open Office* in Frage. Der von Lobin (2000) beschriebene *Document Lifecycle* bezieht sich auf SGML- und XML-Instanzen: In der Phase der Datenerfassung wird eine DTD erstellt (Strukturierung), an die sich die zyklisch ablaufenden Prozesse der Bearbeitung, Betrachtung und Revisionierung sowie gegebenenfalls eine Modifizierung der Strukturierung anschließen. Die Betrachtung kann mit der Konvertierung in ein Ausgabeformat (HTML, PDF etc.) abgeschlossen werden. Diese vier Phasen gelten nicht für alle genannten Dokumentformate, so kann z. B. beim Einsatz von *Word* keine Strukturierungsstufe stattfinden, da es nicht XML-basiert ist.

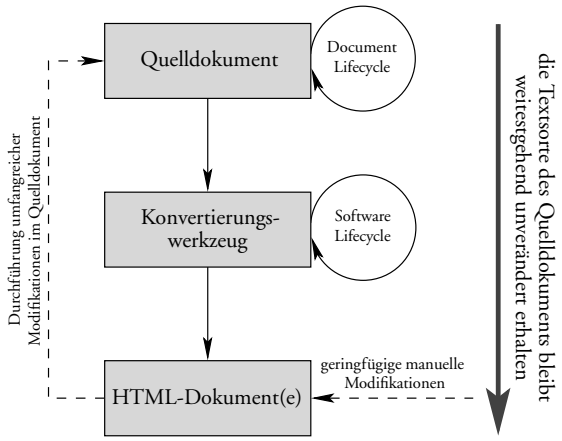


Abbildung 4.3: Der Einfluss der Textsorte bei der automatischen Konvertierung von Dokumenten in die *Hypertext Markup Language*

eine erneute Konvertierung notwendig, um eine aktuelle HTML-Version zu erstellen und auf dem Webserver zu publizieren. Das HTML-Dokument selbst wird also nicht bzw. nur selten direkt manipuliert. Etwaige Modifikationen beziehen sich primär auf peritextuelle Aspekte, z. B. das Entfernen maschinell eingefügter Navigationselemente. Das Zieldokument selbst ist nicht Teil des *Document Lifecycle*, sondern lediglich eines seiner Teilresultate.

Auch die Konvertierungssoftware selbst unterliegt einem Entwicklungsprozess, der als *Software Lifecycle* bezeichnet wird: Der Hersteller veröffentlicht in unregelmäßigen Abständen aktualisierte Versionen mit neuen Konvertierungsfunktionen, die sich unmittelbar auf die maschinell generierten HTML-Dokumente auswirken können.[26] Bei konvertierten Webseiten sind keine oder nur marginale Änderungsprozesse bezüglich der jeweiligen Textsorte auszumachen. Eine Auswirkung betrifft lediglich funktionale, peritextuelle Aspekte (z. B. die automatische Generierung und Einbindung von Navigationshilfen). Die Entstehung eigenständiger Hypertextsorten durch automatische Konvertierungsprozesse ist nicht zu erwarten.[27] Viele Hypertextsorten korrespondieren also unmittelbar mit traditionellen Textsorten wie z. B. *wissenschaftlicher Artikel*, *Publikationsliste* oder *Protokoll*. Trotz der Fundierung auf der Textsorte des Ursprungsdokuments werden diese in der vorliegenden Arbeit als Hypertextsorten bezeichnet, um die Bedeutung des Mediums WWW für die HTML-Dokumente zu reflektieren, die das Resultat von Konvertierungsprozessen sind (vgl. Fußnote 7, S. 266).

Die bisher thematisierten Aspekte beziehen sich auf die Konvertierung eines Dokuments, das typischerweise als einzelne Datei vorliegt. Umfangreiche Datenbestände werden oftmals

[26] Dieser Aspekt wird bei frei verfügbaren Open-Source-Konvertierungswerkzeugen wie z. B. *LaTeX2HTML* (vgl. Goosens und Rahtz, 1999) oder den XSLT- bzw. DSSSL-Stylesheets für *DocBook*-Dokumentinstanzen (vgl. Walsh, 1999) besonders deutlich, deren Funktionsumfang permanent erweitert wird.

[27] Diese Feststellung bezieht sich auf den Großteil der entsprechenden Dokumente im WWW, d. h. auf prototypische Konvertierungsprozesse (vgl. die im Korpus befindliche Menge von mehr als 400 000 HTML-Dokumenten, die mit Hilfe der Textverarbeitung *Word* angefertigt wurden, vgl. Abschnitt A.4.7). Nur in Einzelfällen münden sehr komplexe Transformationsprozesse und hochgradig strukturierte Quelldaten in Textstrukturen, die keinerlei Korrespondenzen zum Quelldokument aufweisen.

von mehreren Personen in einer Datenbank gepflegt und daraufhin maschinell in ein Druckformat überführt (z. B. das tabellarische Vorlesungsverzeichnis einer Hochschule). Derartige Datenbestände werden mit vergleichbaren Verfahren für eine Publikation im WWW aufbereitet, es liegt also ebenfalls eine automatische Konvertierung vor. Das in Abbildung 4.3 dargestellte Modell bezieht sich auch auf derartige Prozesse, jedoch ist zu beachten, dass sich der *Document Lifecycle* auf Modifikationen der Datenbankinhalte und möglicherweise Änderungen ihrer Tabellenstrukturen bezieht. Der Konverter kann statisch oder dynamisch arbeiten: Bei einer statischen Transformation werden die Inhalte der Datenbank verwendet, um HTML-Dokumente zu erzeugen, die anschließend auf dem Webserver publiziert werden. Bei der dynamischen Transformation rezipiert der Benutzer keine statischen HTML-Dokumente, sondern z. B. *PHP*-Skripte, die zur Laufzeit Datenbankinhalte in dynamisch generierte HTML-Dokumente integrieren. In beiden Fällen müssen Entscheidungen getroffen und Verfahren implementiert werden, die die Präsentation der Daten betreffen. Diesbezüglich greifen Entwickler häufig auf diejenigen charakteristischen Muster und typografischen Merkmale zurück, die in gedruckten Vertretern der jeweiligen Textsorte beobachtet werden können (vgl. Indikator 47, S. 144). Auch bei der maschinellen Konvertierung von Datenbankinhalten bleibt die Textsorte des „Quelldokuments“ somit weitestgehend unverändert erhalten. Aufgrund der meist gegebenen Notwendigkeit, einen Konverter zu implementieren, kann jedoch eine größere Flexibilität beobachtet werden (vgl. Abschnitt 3.3.6).

Eine ähnliche Situation liegt bei der maschinellen Erstellung eines HTML-Archivs von Mailing-Listen vor, jedoch existiert in diesem Fall kein *Document Lifecycle*: Wenn eine E-Mail an eine Mailing-Liste geschickt wird, kann auf einem Webserver automatisch ein Konvertierungsprozess initiiert werden, der diese E-Mail maschinell nach HTML überführt und in das vorhandene Archiv integriert. Für diesen Zweck sind zahlreiche Werkzeuge frei verfügbar. Somit wird dem ursprünglichen Autor die Möglichkeit entzogen, die E-Mail zu überarbeiten, da er in der Regel keinen Schreibzugriff auf das Archiv besitzt (vgl. Abschnitt 12.6.6).

Zur Entstehung von Hypertextsorten bei der manuellen Erstellung von Dokumenten

Abbildung 4.4 stellt die Phasen und Einflussfaktoren der Entwicklung von Hypertextsorten bei der manuellen Anfertigung von Webseiten in schematischer Form dar (vgl. auch Yates und Orlikowski, 1992, Yoshioka et al., 2001, sowie Abbildung 2.9, S. 59). Es wird nicht zwischen Hypertextsorte und Hypertextknotensorte unterschieden, weil die jeweiligen Prozesse identisch sind, jedoch unterschiedliche Abstraktionsstufen betreffen. Der hohe Abstraktionsgrad von Abbildung 4.4 tritt im Vergleich zu Modellen der Entwicklung traditioneller Textsorten, wie sie in sprachhistorischen Arbeiten vorgeschlagen werden (z. B. von Gaberell, 2000, S. 169 f.), besonders deutlich zu Tage. Das Modell ist als initialer Vorschlag der Modellierung des zyklischen Entwicklungsprozesses von Hypertextsorten zu verstehen. Verschiedene Studien, in denen Webseiten über einen längeren Zeitraum beobachtet werden, bestätigen das Modell indirekt (vgl. z. B. Ryan et al., 2003, und Emigh und Herring, 2005, S. 9). Eriksen (1997) vergleicht drei skandinavische Online-Zeitungen, deren im Frühjahr 1996 untersuchte Webpräsenzen sich in einer sehr frühen Entwicklungsstufe des hier vorgestellten Modells befinden, Eriksen spricht dabei von einer „Experimentierphase“ (vgl. Abschnitt 4.6.4).

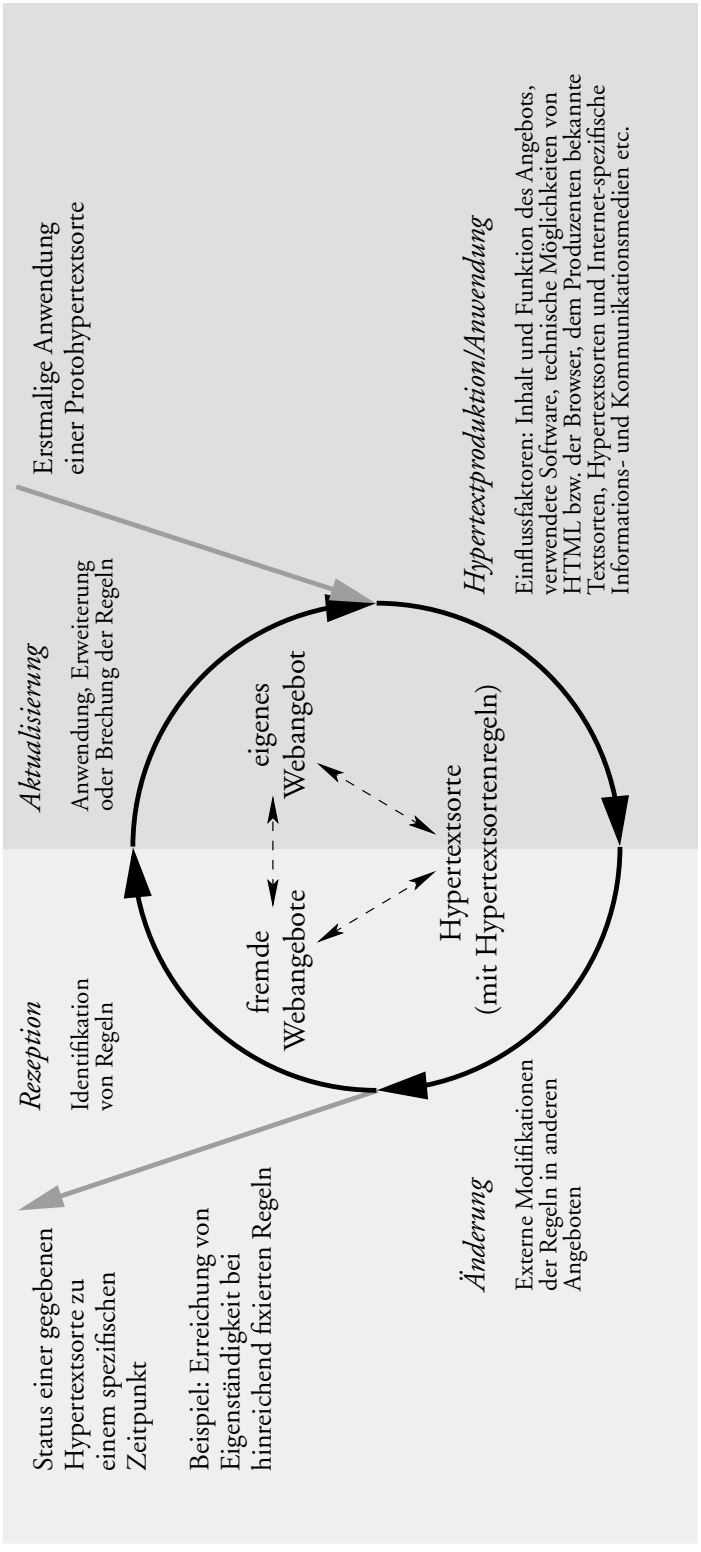


Abbildung 4.4: Zyklisches Modell der Phasen und Einflussfaktoren der Entwicklung von Hypertextsorten bei der manuellen Erstellung von Dokumenten

Die Entwicklung unterliegt dem zyklischen Prozess der *Rezeption* von Hypertexten, die auf Hypertextsorten mit zugehörigen Konventionen basieren (in der Abbildung als Regeln bezeichnet), sowie der *Produktion* von Hypertexten (vgl. Abschnitt 2.3.7).[28] Die im Modell rechts dargestellten Phasen der Anwendung und Aktualisierung gehen von einem generischen Produzenten aus, der einen Hypertext erstellt und in der Folgezeit pflegt und aktualisiert. Der Autor wird von verschiedenen Faktoren beeinflusst. Hierzu zählen unter anderem die beiden wesentlichen Aspekte des darzustellenden Inhalts und seiner Funktion sowie die Software, mit der die Dokumente erstellt werden, die technischen Möglichkeiten von HTML, kulturelle Spezifika und die Kommunikationssituation innerhalb einer Diskursgemeinschaft (vgl. Furuta und Marshall, 1996).[29] Zu den Einflussfaktoren gehören auch die dem Produzenten explizit oder implizit bekannten Text- und Hypertextsorten sowie generelle Merkmale Internet-spezifischer Informations- und Kommunikationsmedien (vgl. Teil III).[30] Auch Eckkrammer (2001, S. 62) geht auf diesen Faktor ein, sie bezieht sich jedoch auf das Stöbern in einem Webangebot (z. B. von Kontaktanzeigen) und das anschließende Ausfüllen eines HTML-Formulars, denn „während des Surfens durch die Hypertextelemente einer Website [erfolgt] eine unbewußte Auseinandersetzung mit den Merkmalen, Vertextungsstrategien und Konventionen der Textsorten [...]. Der spätere Sender ist vorerst noch Rezipient von WWW-Texten, wobei ›instinktiv‹ eine Aneignung der Online-Vertextungsmuster erfolgt.“ Zusätzlich weist Eckkrammer darauf hin, dass Produzenten von Webseiten „nach wie vor eine sehr deutliche Prägung durch die traditionelle Schriftlichkeit und ihre Produkte beibehalten [...]. Somit wird das traditionelle Textwissen auch in den neuen medialen Umgebungen perpetuiert“ (ebd., S. 51, vgl. Amitay, 2000a, und Kallmeyer, 2000a, S. 297 f.).

Die beiden links dargestellten Phasen beziehen sich auf Webangebote, die von Dritten gepflegt werden, auf die ihnen zugrunde liegenden Hypertextsorten, deren Konventionen ebenfalls Veränderungsprozessen unterliegen sowie auf den Umstand, dass von generischen Produzenten sowohl thematisch verwandte bzw. relevante als auch andersartige Hypertexte rezipiert werden.[31] Produzenten identifizieren – bewusst oder unbewusst – Konventionen und integrieren diese, falls sie als wichtig erachtet werden und der eigene Hypertext in Bezug auf diese Konvention als defizitär eingeschätzt wird, in der Phase der Aktualisierung in das eigene Angebot, wodurch sich fremde Angebote und die eigene Webpräsenz einander

[28] Die Darstellung des Zyklus lehnt sich an den *Document Lifecycle* an (Lobin, 2000, vgl. Fußnote 25).

[29] Der Einfluss der verwendeten Software bezieht sich z. B. auf die Unterstützung von Gestaltungsvorlagen. Die zur Verfügung stehende Hardware wird im Modell nicht genannt, sie wirkt sich jedoch ebenfalls auf Hypertextsorten aus: Galt vor einigen Jahren noch eine Bildschirmauflösung von 640×480 Punkten als Standard, sind es mittlerweile $1\,024 \times 768$ oder $1\,280 \times 1\,024$ Pixel, wodurch ein breiteres und informationsdichteres Webdesign ermöglicht wird, sofern das Browserfenster auf die volle Bildschirmgröße aufgezogen wird. Auch die Bildschärfe nimmt immer mehr zu, was in der zunehmenden Verwendung kleinerer Schriften resultiert.

[30] Das Modell wurde zur Beschreibung der Entwicklung von Gebrauchshypertextsorten aus dem kommerziellen und nichtkommerziellen Bereich konzipiert; Abbildung 4.4 beschränkt sich bezüglich der Einflussfaktoren jedoch auf nichtkommerzielle Hypertexttypen. Die Faktoren, die den Aufbau kommerzieller Websites beeinflussen, werden von Kowtha und Choon (2001) diskutiert, die die statistische Analyse einer Befragung der Vorstände von 135 E-Commerce-Unternehmen aus Singapur präsentieren. Die Entwicklung der Website hängt insbesondere von der strategischen Ausrichtung der Firma ab, die ihrerseits von den vorhandenen Kompetenzen, der Unternehmensgröße und der Intensität des Marktes beeinflusst werden.

[31] Diese Phase subsumiert auch die Lektüre von Webdesign- oder Usability-Ratgeberliteratur, die einen Produzenten bei der Erstellung oder Aktualisierung ebenfalls beeinflussen kann (vgl. Schütte, 2004a, S. 332 f.).

annähern. Der Produzent geht davon aus, durch die Übernahme der Konvention das kommunikative Ziel des Hypertextes erfolgreicher realisieren zu können.

Die Aktualisierung von Webangeboten betrifft verschiedene Faktoren: Einerseits sind Autoren bestrebt, die Website bzw. spezifische Knoten der Website auf einem aktuellen Stand zu halten, andererseits erwarten Rezipienten tagesaktuelle Angebote. Falls ein Hypertext veraltet ist, macht sich dieser Umstand eventuell in einem Rückgang der Besucherzahlen bemerkbar. Für den Produzenten entsteht eine Art Aktualisierungszwang, der unter anderem auch das Anbieten eines zeitgemäßen Webdesigns und die Verfügbarkeit verschiedenster Komponenten betreffen kann. Dieser Zwang besteht insbesondere für Webangebote, die kommerzielle Interessen verfolgen, sowie für private und mit großem persönlichen Engagement erstellte Websites. Er kann jedoch auch durch eine öffentliche Sanktionierung initiiert werden, etwa durch ein Ranking der Webangebote eines spezifischen Bereiches (z. B. der Webauftritte aller Universitäten eines Landes, vgl. Abschnitt 6.3.7): Die bestplatzierten Angebote setzen die inhaltlichen und funktionalen Standards, die in der Folgezeit von den weniger gut platzierten Angeboten imitiert und erweitert werden, um beim nächsten Ranking eine bessere Platzierung zu erhalten, was wiederum der Verhärtung von Hypertextsortenregeln entspricht. Bucher (2004, S. 150) geht bezüglich des individuellen und gemeinsamen Wissens der Kommunikationsteilnehmer auf die Rezipientenseite ein und stellt fest, dass sich „[...] die Online-Kommunikation [...] als systematischer Wissensaufbau beschreiben [lässt], sowohl hinsichtlich des Strukturwissens als auch hinsichtlich des *content*-Wissens." Da Produzenten von Webangeboten auch Rezipienten von Webangeboten sind, besitzen auch Produzenten das von Bucher beschriebene individuell aufgebaute Wissen über spezifische Merkmale fremder Webangebote. Darüber hinaus wird betont, dass der Wissensaufbau, ebenso wie das hier vorgestellte Modell, „eine zyklische Struktur" besitzt (ebd., S. 152).

Jede Anwendung einer Hypertextsorte oder einer Konvention bewirkt ihre Stärkung innerhalb der Diskursgemeinschaft und erhöht die Wahrscheinlichkeit, dass sie von anderen Rezipienten, die ebenfalls als Produzenten tätig sind oder zu einem späteren Zeitpunkt sein werden, bewusst oder unbewusst wahrgenommen und in vergleichbaren Kommunikationssituationen bei der Produktion oder Aktualisierung eines Webangebots übernommen wird.[32] Basierend auf diesem zyklischen Prozess findet im Laufe eines längeren Zeitraums eine Verhärtung von Konventionen statt (vgl. Yates und Sumner, 1997), die schließlich in der Etablierung einer Hypertextsorte mit spezifischen Konventionen und möglicherweise einem kennzeichnenden Etikett mündet.[33] Insbesondere der Umstand, dass eine Website effizient aktualisiert werden kann, ist der entscheidende Grund dafür, dass sich bereits nach wenigen Jahren der Verfügbarkeit des WWW eigenständige Hypertextsorten bilden konnten.

Das Modell erklärt Beobachtungen von Jakobs (2003, S. 246), die in einer „vorsichtigen Bestandsaufnahme" bezüglich der Bandbreite von Hypertextsorten konstatiert, dass ein „un-

[32] Crowston und Williams (2000, S. 203) vertreten folgende Auffassung: „[T]he set of genres in use [...] is both a product and a shaper of the communicative practices of a community." Eriksen und Ihlström (1999, S. 289) fokussieren den Einfluss von Web-Genres: „When establishing a new site that serves a purpose similar to existing sites, the genre characteristics are copied and refined to reflect resemblance to an existing genre."

[33] Siehe hierzu auch Bucher (2004, S. 153) aus Sicht der Rezipienten: „Die Rezeption von Online-Angeboten besteht [...] darin, dass eine bestimmte Form oder Struktur auf Angebotsseite erkannt wird [...]. [...] Vom Standpunkt der *Usability* aus betrachtet sind demzufolge solche Online-Angebote nutzerfreundlicher, die [...] bekannte Nutzungsschemata anbieten und dadurch die Erschließung erleichtern und ökonomisieren".

überschaubares Konglomerat sehr heterogener Formen vor[liegt], die in unterschiedlichem Maße konventionalisiert sind." Jakobs ermittelt ein „breites Spektrum, das sich zwischen den Polen der Imitation und Innovation aufspannt". Diesen Aspekt diskutieren auch Androutsopoulos und Kraft (2003, S. 4): Die Autoren persönlicher Homepages „orientieren sich inhaltlich und formal aneinander bis hin zum direkten »Abkupfern« von vorbildhaft erscheinenden Sites. Gleichzeitig entwickeln sich rezipientenseitige Annahmen und Erwartungen über die typische persönliche Homepage. Neben der allmählichen Konventionalisierung von Gestaltungslösungen in einer Domäne ist jedoch mit einer beträchtlichen Variationsbreite zu rechnen, da kein starker Normierungsdruck herrscht [...]." Im Kontext universitärer Webangebote gelangen Androutsopoulos und Kraft (2003, S. 10) zu der Schlussfolgerung: „Aus der Vielzahl konkurrierender Vorschläge werden bestimmte Lösungen nachgeahmt, was letztlich zu einer Abnahme von Variation und der Verfestigung von Konventionen führt."

Der zyklische Aufbau des Modells ermöglicht es, den Entwicklungsstand einer Hypertextsorte zu beschreiben, so kann sie zu einem frühen Zeitpunkt als repliziert und zu einem späteren Zeitpunkt als variiert eingestuft werden (im Sinne von Shepherd und Watters, 1998). Der Moment, an dem von einer neuen und eigenständigen Hypertextsorte gesprochen werden kann, ist nicht exakt bestimmbar, da sie auf der Akzeptanz innerhalb einer Diskursgemeinschaft beruht (vgl. Crowston und Williams, 1997, S. 32). Autoren besitzen jedoch einen sehr breiten Spielraum: Gerade in Angeboten, die keinen offiziellen Charakter besitzen, wird mit Gestaltungsmerkmalen experimentiert und bereits etablierte Konventionen werden bewusst ignoriert. Offizielle Angebote hingegen unterliegen einem regelrechten Zwang, bestimmte Komponenten anbieten zu müssen (vgl. Nielsen und Tahir, 2002), z. B. eine interne Suchfunktion im Falle einer umfangreichen Website. In gleicher Weise wird von dem Webangebot eines universitären Fachbereichs erwartet, dass das aktuelle kommentierte Vorlesungsverzeichnis in digitaler Form verfügbar ist. Derartige (unterschiedlich granulare) Erwartungshaltungen sind unmittelbar auf etablierte Hypertextsorten zurückzuführen.

4.4 Studien zur Sammlung von Hypertextsorten

Die Arbeiten zu Hypertextsorten lassen sich in drei Kategorien einteilen: Die meisten Studien beschäftigen sich mit spezifischen Hypertexttypen (Abschnitt 4.6). Die Beiträge der zweiten Kategorie gehen auf ausgewählte Spezifika von Hypertextsorten ein (Abschnitt 4.5). Im Folgenden wird die dritte Gruppe diskutiert, die die Studien von Crowston und Williams (1997, 2000), Shepherd und Watters (1999) sowie Haas und Grams (1998a,b, 2000) umfasst, in denen zufällig ausgewählte Stichproben von Webseiten auf ihre Hypertextsorten untersucht werden. Die Analyse von Roussinov et al. (2001) wird von Interviews flankiert, Brandl (2002) führt Interviews mit Experten durch, Dewe et al. (1998) beziehen sich auf eine Fragebogenuntersuchung und Rosso (2005) kombiniert mehrere dieser Methoden.[34]

[34] Schönefeld (2001) untersucht mehrere spanische Websites auf Kohärenzaspekte und geht auch auf den Einfluss der jeweiligen Textsorte ein. Es werden verschiedene „neue Textsorten" postuliert: „Online-Zeitung", „Online-Werbung", „Online-Lehrbuch", „Online-Flugblatt", „Online-Zeitschrift (wissenschaftlich)" und „Online-(Selbst)präsentation" (ebd., S. 95 f.). Als Kriterium zur Binnendifferenzierung der zuletzt genannten Textsorte sieht Schönefeld den Emittenten an (z. B. Institutionen, Vereine, Firmen oder Einzelpersonen).

4.4.1 Die Studien von Crowston und Williams (1997, 2000)

Crowston und Williams (1997) wollen mit ihrer Analyse die Bandbreite digitaler Genres im WWW aufzeigen. Von der ausgeprägten Akzeptanz dieses Mediums wird vermutet, dass sie in etablierten Genres Veränderungsprozesse initiieren und neue Genres hervorbringen kann. Zusätzliche Impulse gehen von der rapiden technologischen Entwicklung und der unkomplizierten und effektiven Verfügbarkeit von Webangeboten aus. Als Stichprobe dienen 100 englischsprachige Webseiten, die im Januar 1996 mit Hilfe einer (mittlerweile nicht mehr verfügbaren) Funktion zur zufälligen Auswahl eines Dokuments aus dem Index der Suchmaschine *AltaVista* gesammelt wurden. Weiterführende Restriktionen wurden bei der Datensammlung nicht realisiert. Die Stichprobe enthält Dokumente, die aus 12 Ländern stammen und einen Umfang zwischen neun und 111 586 Wörtern besitzen. Crowston und Williams (1997, S. 33) geben an, dass die Untersuchung der Stichprobe hinsichtlich der vorhandenen Genres auf ihren eigenen Erfahrungen mit dem Internet basiert. Es werden 80 Vorkommen von „familiar genres“ (z. B. „book“, „FAQ“ und „script“), 11 Vorkommen von „new, but accepted genres“ („hotlist“, „home page“, „Web server statistics“)[35] und neun Textexemplare mit „unknown genres“ ermittelt.[36] Einige Dokumente enthalten zwar „stereotype“ (ebd., S. 34) Informationen, die Verfasser sind jedoch nicht in der Lage, die korrespondierenden Genres zu benennen. Es wird die Vermutung geäußert, dass sich diese Genres derzeit noch in der Entstehung befinden oder den Verfassern unbekannten Diskursgemeinschaften zugehörig sind. Abschließend äußern Crowston und Williams die Ansicht, dass sich das Konzept des Genres zwar gut auf eine Analyse von WWW-Dokumenten anwenden lasse, jedoch seien naturgemäß Probleme in Bezug auf Genres zu verzeichnen, die insbesondere hinsichtlich ihrer Form unterschieden werden (als Beispiele werden „brochure“, „booklet“ und „flyer“ genannt). Aus diesem Grund plädieren Crowston und Williams dafür, primär den Zweck („purpose“) der Veröffentlichung eines HTML-Dokuments zu fokussieren.

In einer späteren Arbeit (gleichen Titels) untersuchen Crowston und Williams (2000) eine Liste von 1 000 Webseiten aus 40 Ländern, die im Februar 1996 dem Index von *AltaVista* entnommen und im Mai 1996 gesammelt wurden; aufgrund der Verzögerung der Datensammlung und dadurch resultierender nicht mehr verfügbarer Dokumente enthält die analysierte Stichprobe 837 Webseiten. Restriktionen hinsichtlich der Sprache, in der ein Dokument verfasst ist, wurden nicht angewendet. Crowston und Williams identifizieren in der Stichprobe mindestens 64 Genres, die teilweise in eine Hierarchie eingeordnet werden, die auf dem *Art and Architecture Thesaurus* basiert.[37] Insgesamt finden Crowston und Williams

[35] Crowston und Williams (1997, S. 36) definieren die „hotlist“ als eine Liste von Hyperlinks zu Angeboten Dritter. Die „home page“ wird als „especially easily identified and commonly accepted genre“ bezeichnet (ebd.) und ist nach Crowston und Williams eine „web page presenting personal or organizational information or the page at the hierarchical top of a web site presenting such information“.

[36] In einer online erhältlichen Entwurfsversion des Artikels listen Crowston und Williams die 100 in der Stichprobe enthaltenen Dokumente mit knappen Kommentaren auf und geben die Bezeichnungen von 48 ermittelten Genres an (vgl. Tabelle 4.1, S. 176), die von kurzen Definitionen flankiert werden.

[37] Die vollständige Anzahl ermittelter Genres wird von Crowston und Williams (2000) nicht angegeben. Bei 47 Dokumenten (5,6%) konnte keine Zuweisung stattfinden, „because they were not in English and had ambiguous forms (24 pages), because we did not know the name for the genre (2 pages), or because the pages did not have a recognizable genre (21 pages). The latter included binary or other nontext documents.“ (ebd., S. 205). Crowston und Williams weisen wiederholt darauf hin, dass sich das Zuweisen von Genres bei be-

Book, Report, Newsletter, Essay, Pamphlet, Article, News wire article, Column, Memorial, Concert review, Product reviews, Ratings, Submission instructions, Table of contents, Index, Discography, Filmography, Regulation or rule, Product information, Government program description, Testimonial, University course listing, Problem set, Faculty information, Vitae, Publications list, List of research projects, Directory, Library acquisitions list, Order form, Meeting minutes, Box score, Chronicle, Script, Political party platform, Genealogy, Demographic data, Guide, Archive item, FAQ, Users' manual, Computer documentation, Source code, File directory listing, E-mail directory listing, Hot list, Home page, Server statistics	Crowston und Williams (1997)
books, photograph albums, pamphlets, business cards, catalogs, comment forms, order forms, URL submission forms, bookmark lists, course lists, discographies, e-mail lists, FAQs, file lists, filmographies, hotlists, project lists, publication lists, source codes, abstracts, indexes, table of contents, essays, chronicles, genealogies, scripts, classified advertisements, announcements, custom 404s, press releases, under construction, Web site moved, home pages, political party platforms, testimonials, eulogies, guidebooks, instructions, manuals, computer manuals, problem sets, prospectuses, ratings, accession records, accessions registers, minutes, resumes, rules (instructions), specifications, demographic data, course descriptions, box scores, Web server statistics, directories, faculty directories, topical home pages, reports, concert reviews, film reviews, product reviews, articles, newspaper columns, newswire articles, newsletters	Crowston und Williams (2000)
• *Topics*: Home page (business, celebrities), geographical location, special topics • *Publications*: Articles, publications, news bulletins • *Products*: Product information, product lists, advertisements, ratings, product reviews, order forms • *Educational material*: Glossary, course lists, instructional materials (guidebooks, instructions, manuals, problem sets, syllabus, tutorial page) • *FAQ*: FAQ	Roussinov et al. (2001)

Tabelle 4.1: Listen von Genres im WWW – Ergebnisse der Studien von Crowston und Williams (1997, 2000) sowie Roussinov et al. (2001)

(2000, S. 207) 507 Instanzen von „familiar genres" (60,6%), 239 Textexemplare von „new but accepted genres" (28,6%) und 44 Instanzen von „apparently new genres" (5,3%). Neben diesen drei Kategorien werden „reproduced genres" (z. B. „article", „FAQ", „meeting minutes" und „course description") und „adapted genres" eingeführt.

4.4.2 Die Studie von Roussinov et al. (2001)

Roussinov et al. (2001) streben die automatische Identifizierung von Web-Genres an, um Suchmaschinenanwendern eine zusätzliche Möglichkeit zur Filterung der Treffermenge zur Verfügung stellen zu können (vgl. auch Kwasnik et al., 2001). Als Voruntersuchung wurde eine Studie zur Beantwortung der Fragen durchgeführt, zu welchem Zweck Suchmaschinen eingesetzt werden und ob eine Korrelation zwischen dem Ziel ihrer Verwendung und den im WWW existenten Genres existiert. Im März 2000 wurden 184 Angehörige einer

stimmten Textexemplaren als sehr problematisch erwiesen hat, z. B. wenn in einem Dokument zwar ein Genre vorliegt, der Name jedoch nicht bekannt ist oder wenn die kommunikative Funktion eines Dokuments nicht bestimmt werden kann. Die 64 angegebenen Genres (ebd., S. 212–215) werden in Tabelle 4.1 dargestellt. Bei Crowston und Williams (ebd.) werden sie von kurzen Definitionen ergänzt. Nach eigenen Angaben basieren 30 der verwendeten Genre-Bezeichnungen auf der Intuition der Verfasser. Viele Arbeiten zum Komplex „Digital Genres" enthalten Listen von Genre-Bezeichnungen, in denen ausschließlich Pluralformen verwendet werden. Lediglich Rosso (2005, S. 82) geht auf diesen Aspekt ein.

Universität bei der Benutzung von Suchmaschinen beobachtet, parallel wurden Interviews durchgeführt und insgesamt 1 234 Webseiten gesammelt, die von den Probanden rezipiert wurden. Als Gründe für den Einsatz von Suchmaschinen werden unter anderem „Scholarly Research“ (22,95%), „Shopping“ (12,57%), „Cultural Arts Activities“ (7,10%), „Health“ (7,10%), „News“ (7,10%), „Computing“ (6,56%) und „Travel Planning“ (6,56%) genannt (Roussinov et al., 2001, S. 4). Basierend auf dem von Crowston und Williams (2000) vorgeschlagenen Inventar von Web-Genres wurden 1 076 Dokumenten sowohl von den Probanden als auch von den Versuchsleitern insgesamt 116 unterschiedliche Genres zugewiesen; 158 Dokumente konnten nicht zugeordnet werden.[38] Die Übereinstimmung beträgt lediglich 49,6%, was als Indiz dafür gewertet wird, dass das Genre-Konzept im Medium WWW „a little fuzzy“ ist. Daher wird argumentiert, dass bereits die grobe Zuordnung eines Dokuments zu einer „genre group“ für den Anwender hilfreich sein könnte, denn eine maschinelle Identifizierung von mehr als 100 Web-Genres sei mit aktuellen Technologien nicht durchführbar.[39] Die gefundenen Genres werden in die fünf Gruppen „Topics“, „Publications“, „Products“, „Educational material“ und „FAQ“ sortiert, wobei die Gruppen die ermittelten Gründe für die Verwendung von Suchmaschinen bestmöglich reflektieren sollen.[40] Für jede Gruppe werden zusätzlich exemplarische Genres sowie „recognition indicators“ angegeben, die innerhalb einer Implementierung zur Erkennung verwendet werden könnten.[41]

[38] Aus der Studie von Crowston und Williams (2000) stammen 72 der 116 Genres. Neu eingeführt wurden 44 Genres, zu denen insbesondere die Gruppe „Miscellaneous“ zählt, die z. B. „calendar“, „contact“, „discussion-board/forum“ und „weather page“ umfasst. Eine vollständige Liste aller Genres wird nicht angegeben.

[39] Diese Auffassung wird von Roussinov et al. (2001, S. 5 f.) ohne Angabe spezifischer Gründe als „plausible“ bezeichnet: „[I]t seems plausible that a large number of search tasks could be satisfied by documents of only a few groups of related genres, in which case distinguishing among documents in these groups would be almost as valuable as perfect recognition. In short, even a good guess as to the purpose of a page may help those overloaded with information pick out more useful pages.“

[40] Tabelle 4.1 zeigt die Beispiele, die von Roussinov et al. (2001, S. 8) für die fünf Gruppen genannt werden. Zahlreiche Aspekte fallen auf: Zunächst wirkt die Zusammenstellung sehr heterogen; es wird nicht diskutiert, weshalb eine Genre-Gruppe namens „FAQ“ angenommen wird, die aus dem Genre „FAQ“ besteht. Weiterhin ist auffällig, dass die persönliche Homepage vollständig fehlt; dieses Genre wurde von den Probanden zwar als irrelevant für Suchaufgaben eingestuft, ein solch salientes Web-Genre sollte sich jedoch in jedem Fall in einer derartigen Liste niederschlagen. Abschließend kann die Relevanz einiger der als Beispiele aufgeführten Genres für spezifische Suchaufgaben in Frage gestellt werden, so wird z. B. nicht motiviert, in welchem Kontext ein Anwender nach „order forms“ (oder gar „advertisements“) suchen könnte. Realistischer erschiene in diesem Fall zunächst die Suche nach E-Commerce-Anbietern einer spezifischen Branche (vgl. Ho, 1997) oder die Suche nach Informationen zu bestimmten Produkten.

[41] Roussinov et al. (2001, S. 6) erläutern diese Vorgehensweise wie folgt: „To develop the set of document genres, we first listed possible indicators of genre that could be automatically recognized with an accuracy of 60–100%. Second, we associated those indicators with each genre related to the major purposes in the sample [...]. At this point, these accuracy estimates are based solely on our intuition [...].“ Es muss jedoch bezweifelt werden, dass die angegebenen Indikatoren eine derartig hohe Trennschärfe aufweisen. Für die Gruppe „Topics“ werden z. B. die Indikatoren „url consists only of host name“, „short in length“ und „plenty of graphics and ads“ aufgeführt (Roussinov et al., 2001, S. 8). Genres der Gruppe „Educational material“ sollen durch vage Indikatoren wie „.edu domain“, „not many graphics or ads“ erkannt werden können. Zusätzlich schlagen Roussinov et al. eine (nicht implementierte) Benutzerschnittstelle zur Interaktion mit einer Suchmaschine vor, die Dokumente aufgrund ihrer Genres filtern kann. Obwohl explizit ausgesagt wird, dass eine Erkennung der fünf Genre-Gruppen ausreichend sei, enthält das abgebildete Interface eine Genre-Hierarchie mit drei der fünf Gruppen als Knoten der ersten und spezifischen Genres als Knoten der zweiten Ebene.

4.4.3 Die Studie von Shepherd und Watters (1999)

Shepherd und Watters (1999) analysieren 96 Webseiten, die mit Hilfe der (nicht mehr verfügbaren) Website `http://random.yahoo.com` zusammengestellt wurden, um das „functionality attribute“ genauer zu untersuchen. Die Dokumente der Stichprobe werden in die Genres „Home page“ (40% der Dokumente), „Brochure“ (17%), „Resource“ (35%), „Catalogue“ (5%) und „Game“ (3%) klassifiziert, zusätzlich wird das Genre „Search engine“ angenommen, für das aber keine Belege gefunden werden. Die von Shepherd und Watters postulierte Liste von Genres unterscheidet sich in puncto Granularität von den Genre-Listen von Crowston und Williams (1997, 2000) (vgl. Shepherd et al., 2004, S. 241), was durch die unspezifische Extension des Genre-Konzepts der *North American Genre Theory* erklärt werden kann (vgl. Abschnitt 2.3.7). Shepherd und Watters (1999, S. 2) gehen explizit auf diesen Aspekt ein: „These cybergenres are defined at a high level of abstraction and we recognize that there are more specific categories [...] within each of the six categories, for example, the personal home page and the corporate home page.“ Obwohl der Unterschied zwischen der „personal home page“ und der „corporate home page“ immer wieder betont wird (z. B. von Crowston und Williams, 2000), gehen auch Shepherd und Watters (ebenso wie z. B. Boardman, 2005, S. 106) davon aus, dass sie *einem* Genre namens „home page“ zugehörig sind.[42]

Tabelle 4.2 stellt die Ergebnisse der Analyse dar und verdeutlich den hohen Abstraktionsgrad der Untersuchung, der unter anderem durch sehr vage Merkmale wie „browsing“, „browse“, „hierarchical“, „shallow hierarchy“, „images“ oder „high-level of interactivity“ hervorgerufen wird. Shepherd und Watters (1999, S. 9) bilden die 48 von Crowston und Williams (1997) ermittelten Genres auf die fünf Cybergenres ab – die Ergebnisse dieser Zuordnung werden in Tabelle 4.2 als zweite Prozentangabe dargestellt. Die durchaus vorhandenen „significant differences in the proportions of each cybergenre from the previous work of Crowston and Williams to this research sample“ gehen Shepherd und Watters zufolge insbesondere auf „a terrific change on the Web over the past 2 years“ (ebd., S. 9) zurück sowie möglicherweise auf fehlerhafte Abbildungen seitens der Verfasser oder zu kleine Stichproben. Der Grund für die beobachteten Unterschiede liegt in keiner dieser Thesen, sondern vielmehr in der angewendeten Methodologie: Shepherd und Watters versuchen mit diesem Vergleich in gewisser Weise, detailliert spezifizierte Hypertextsorten auf eine sehr heterogene und kleine Menge von Hypertexttypen abzubilden. Das Cybergenre „Resource“ kann dabei aufgrund der umfassenden Extension seiner Definition als „information resource“ (ebd., S. 6) in fast allen Fällen als Zielkategorie verwendet werden, was ihren hohen prozentualen Anteil erklärt. Zusammenfassend äußern Shepherd und Watters (1999, S. 9) die Ansicht, dass nur relativ wenige Klassen von Cybergenres im WWW existieren, es wird aber weder eine Abgrenzung der Konzepte Genre (im Sinne von Crowston und Williams, 1997) und Cybergenre noch eine präzise Definition des letztgenannten Konzepts vorgenommen.

[42] Über die Gründe kann nur spekuliert werden, denn es handelt sich eindeutig um zwei unterschiedliche Hypertextknotentypen. Die Tatsache, dass Shepherd und Watters von *einer* Kategorie ausgehen, ist umso erstaunlicher, wenn die Tabelle mit den Frequenzen der Genres (die *separate* Einträge für „Personal Home Page“ und „Corporate Home Page“ enthält) und die Anmerkungen zur Datensammlung hinzugezogen werden: „For the purposes of this research, we viewed each such page as the root of a Web site, even if it was not the actual home page of a Web site.“ (ebd., S. 2), d. h. *jedes* untersuchte Dokument wird als Einstiegsseite aufgefasst, nur einige von ihnen besitzen aber den Verfassern zufolge das Genre „home page“.

Cybergenre	Content	Form	Functionality
Home Page (40%/10%)	information about person/company	introduction; hierarchical; images; animated images	browsing; e-mail
Brochure (17%/6%)	products and services	shallow hierarchy; high-impact visual	browsing; e-mail
Resource (35%/82%)	subject-specific information	hierarchical; images; video; audio	browsing; e-mail; search; discussion; interaction
Catalogue (5%/2%)	products and services	hierarchical; images	browse; e-mail ordering & inquiry; search; on-line ordering; on-line enquire
Game (3%/0%)	challenge to user; scenarios; rules	animation; audio; video; scenes	high-level of interactivity; collaborative computing
Search Engine	categories of sites; URLs	query box; lists of sites; virtual document	browse; search

Tabelle 4.2: Charakterisierungen von „Cybergenres“ – Ergebnis der Stichprobenanalyse von Shepherd und Watters (1999, S. 2, S. 8)

4.4.4 Die Studien von Haas und Grams (1998a, 1998b, 2000)

Abseits der HICSS-Konferenzreihe und mit nur eingeschränktem Bezug auf die *North American Genre Theory* haben Haas und Grams drei Artikel (1998a, 1998b, 2000) vorgelegt, die sich mit Dokumenttypen im WWW sowie der Typologisierung von Hyperlinks auseinander setzen und auf maschinelle Verfahren abzielen. Die Stichprobe besteht aus 75 zufällig aus dem Index von *AltaVista* ausgewählten, englischsprachigen Webseiten, den ca. 1 500 enthaltenen Hyperlinks und ihren Zielknoten.[43] Bei der Datensammlung wurden mehrere Restriktionen angewendet: Ein Dokument musste in englischer Sprache verfasst sein, keine pornografischen Inhalte enthalten und einen Dateityp aufweisen, der Hyperlinks ermöglicht. Die wesentlichen Faktoren für das Schema zur Klassifikation von Webseiten sind nach Haas und Grams (1998b, S. 100 f.) der Zweck bzw. die Funktion eines Dokuments, die Zielgruppe, der Inhalt, die enthaltenen Hyperlinktypen und Relationen zu verlinkten Seiten.

Abbildung 4.5 zeigt die von Haas und Grams (1998b, S. 102–104) ermittelte Typologie von „page types“, mit der kein Anspruch auf Vollständigkeit verbunden ist. Die Kategorien des Typs „organizational pages“ beziehen sich auf Dokumente, die Hypertextstrukturen visualisieren und Zugriffsmöglichkeiten anbieten.[44] Die Kategorie „documentation“ bün-

[43] Die Untersuchung erfolgte in drei Stufen (Haas und Grams, 1998b, S. 102), in denen die Verfasserinnen die Dokumente separat analysierten. Zunächst wurden 15 Dokumente und 346 abgehende Hyperlinks zur Erzeugung einer initialen Menge von Dokument- und Hyperlinktypen eingesetzt. In der zweiten Phase wurden 40 Dokumente und 941 Links mit „revised procedures and classes“ untersucht (ebd.). Basierend auf einem Vergleich der Ergebnisse der zweiten Phase wurden abschließend 20 Dokumente und 194 Links analysiert.

[44] Das „table of contents“ entspricht nach Haas und Grams (1998b, S. 103) einem traditionellen Inhaltsverzeichnis, das Inhalte in monosequenzierten Knoten auflistet. Ein „site content“-Knoten enthält Verweise zu weiterführenden Informationen. Er kann als Liste, Karte oder Ansammlung von Icons oder Buttons realisiert sein. Der „index“ hat die Aufgabe, beliebige Hyperlinks zur Verfügung zu stellen.

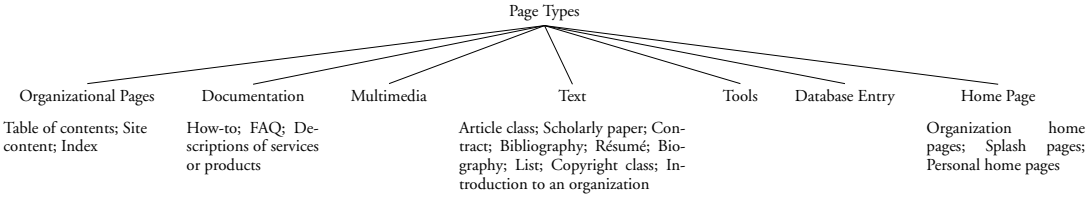


Abbildung 4.5: Typologie von „Page Types" – Ergebnis der Stichprobenanalyse von Haas und Grams (1998b, S. 102–104)

delt Dokumenttypen, die Referenzcharakter aufweisen. Die Gruppe „text" enthält Haas und Grams (1998b, S. 103) zufolge „most other types of writing", wobei angenommen wird, dass der Subtyp „article" weitere Typen subsumieren kann (z. B. „newspaper", „zine", „editorial", „commentary" und „press release"). Der Subtyp „introduction to an organization" kann Typen wie „mission statement" oder „description of the purpose of the organization" umfassen. Im Gegensatz zu Shepherd und Watters (1999), Crowston und Williams (1997) sowie Roussinov et al. (2001) gehen Haas und Grams (1998b) davon aus, dass der Typ „home page" separate Subkategorien für „organization home pages" und „personal home pages" beinhaltet. Während die als „multimedia" etikettierte Gruppe „sound, video, image, graphics, and other non-textual documents" umfasst (ebd., S. 104), gestatten es Webseiten der Kategorie „tools" dem Benutzer, Aufgaben durchzuführen (z. B. Suchwerkzeuge, E-Mail- und Bestellformulare). Der Typ „database entry" wird definiert als „a page containing highly structured information, such as might be found in a database." (ebd., S. 104), wobei als Beispiel die Website eines Online-Buchhändlers angegeben wird, auf der jedes Buch in Form einer separaten Webseite mitsamt Titel, bibliografischen Informationen und Preis vorgestellt wird.[45]

Abbildung 4.6 zeigt die von Haas und Grams (1998b, S. 104–106) vorgeschlagene Typologie von „link types", die auf der Positionierung eines Hyperlinks, seinem Ziel und dem Grund basiert, weshalb ihm ein Leser folgen könnte.[46] Die Typologie umfasst vier Superkategorien: Der Typ „navigation" enthält Links, die ausschließlich auf diejenige Website zeigen, auf der sich auch der Ausgangsknoten befindet, und somit einen Zugriff auf die Hypertextbasis erlauben. Der Subtyp „To site utility" subsumiert Links auf Angebote, die die Benutzung einer Website vereinfachen sollen (z. B. Suchwerkzeuge); „within a document" bündelt

[45] Es fallen verschiedene problematische Aspekte auf, die insbesondere die Heterogenität der Hierarchie betreffen und durch eine grafische Darstellung, auf die Haas und Grams verzichten, besonders deutlich werden. Der Typ „text" soll vermutlich ausschließlich traditionelle Textsorten umfassen, da für die im Internet etablierten Textsorten „how-to" und „FAQ" eine eigene Kategorie angelegt wurde; schließlich können auch einige der gefundenen traditionellen Dokumenttypen in spezifischen Situationen einen Referenzcharakter aufweisen. Weiterhin ist unklar, weshalb keine Binnenstrukturierung der Gruppen „multimedia" und „tools" vorgenommen wurde. Die Kategorie „database entry" ist kritisch zu hinterfragen, weil es dem Benutzer einer Suchmaschine, die eine automatische Identifizierung derartiger „page types" unterstützt, gleichgültig sein dürfte, ob ein Dokument „highly structured information" besitzt oder nicht. Im Falle des Beispiels eines Online-Buchhändlers hielte ich die Zuweisung der Kategorie „description of services or products" für adäquater.

[46] Nach Ansicht von Haas und Grams (1998b, S. 104) spielen Konventionen bei der Verknüpfung eine entscheidende Rolle: „Many types of links have become common, and their anchors do not need to provide a great deal of information, if one assumes that the reader has at least some familiarity with Web conventions." Ein als „Home" beschrifteter Link führt z. B. in praktisch allen Fällen zur Einstiegsseite eines Angebots.

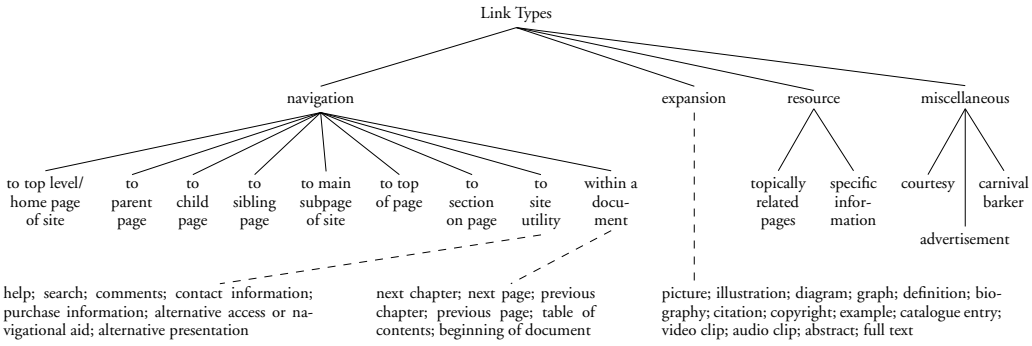


Abbildung 4.6: Typologie von „Link Types“ – Ergebnis der Stichprobenanalyse von Haas und Grams (1998b, S. 104–106)

Hyperlinks, die sich auf die dokumentinterne Navigation beziehen, wobei das Dokument durchaus in mehr als eine physikalische HTML-Datei aufgeteilt sein kann. Hyperlinks des Typs „expansion“ führen zu detallierteren Informationen bezüglich des Linkanzeigers. Der Typ „resource“ ist nach eigenen Angaben äußerst problematisch, denn „any target page that does not fit in one of the other classes is a potential resource for the reader, and has been recommended by the author by virtue of its inclusion on the page.“ (ebd., S. 105). Die Relation zwischen Ursprung und Ziel ist also als vage aufzufassen, dennoch differenzieren Haas und Grams die beiden Subtypen „topically related resource“ und „specific information“.[47] Der Typ „miscellaneous“ umfasst drei Subkategorien: „courtesy“-Verweise besitzen keine Relevanz für den eigentlichen Inhalt eines Dokuments und werden vom Autor als Zeichen der Anerkennung oder aus Höflichkeit integriert, z. B. gegenüber einem Sponsor oder aus vertraglichen Gründen. Links der Kategorie „carnival barker“ verknüpfen unter anderem Spiele und Wettbewerbe, um den Leser zum wiederholten Besuch einer Website zu bewegen. Links vom Typ „advertisement“ (die der ursprünglichen Taxonomie von Haas und Grams, 1998a, S. 492, hinzugefügt werden) beziehen sich auf Werbung. Zur Demonstration der Verbindung zwischen den Typologien von „page types“ und „link types“ zeigen Haas und Grams (1998b) verschiedene Korrelationen auf: Dokumente vom Typ „table of contents“ enthalten z. B. meist „within-document“ Hyperlinks, „site content“ Webseiten enthalten andere Hyperlinks der Kategorie „navigation“, wohingegen „index“-Webseiten primär „resource“-Verknüpfungen umfassen. Links des Typs „expansion“ zeigen meist auf Seiten der Typen „multimedia“ oder „text“. Korrelationen existieren auch bezüglich der Positionierung: Haas und Grams (1998b, S. 107) geben an, dass 84% der isolierten, d. h. nicht in Fließtext oder Listen, sondern z. B. in Navigationshilfen integrierten Links dem Typ „navigation“ zugehörig sind; 82% aller „navigation“-Hyperlinks wiederum sind isoliert angeordnet (vgl. Haas und Grams, 1998a, S. 493). Die von Haas und Grams vorgeschlagene Typologie von Hy-

[47] Der Typ „specific information“ führt einen Widerspruch zu „expansion“ ein: „This is a link [...] within the same site that provides information about a specific topic. The anchor and its context generally explicitly signal what can be found on the target page, with the assumption being that one would not follow the link unless one needed that information.“ (Haas und Grams, 1998b, S. 106). Es wird nicht deutlich, weshalb zwei Typen benötigt werden, schließlich kann „specific information“ auch als „expansion“ des Linkanzeigers aufgefasst werden (vgl. Haas und Grams, 1998a, S. 491).

perlinks (vgl. auch Kopak, 1999) bietet in Bezug auf den Typ „navigation“ einen hohen – wenngleich nicht bezüglich jedes Subtyps hinreichend motivierten – Grad der Differenzierung (vgl. auch die von Kuhlen, 1991, vorgeschlagene Typologie auf S. 108 sowie Thelwall, 2003). Da Korrelationen zwischen unterschiedlichen Links dieser Kategorie und spezifischen Hypertextknotensorten bestehen, bietet es sich an, die automatische Typisierung gegebener Hyperlinks als einen Parameter in die Identifizierung von Hypertextsorten zu integrieren.[48]

4.4.5 Die Studie von Brandl (2002)

Brandl (2002) untersucht Klassifikationen von Webangeboten aus einer kommunikationswissenschaftlichen Perspektive und erstellt auf der Basis von Interviews eine Liste von Typen und Subtypen.[49] Insbesondere geht es Brandl um die Frage, wie Rezipienten das WWW klassifizieren, welche charakteristischen Angebotstypen wahrgenommen werden und welche Merkmale dafür verantwortlich sind. In der ersten Phase wurden Interviews mit zehn Angehörigen der Universität München durchgeführt, die Experten in den Bereichen Webentwicklung, Webdesign oder Software-Entwicklung sind. Die Probanden wurden gebeten, charakteristische Typen von Websites, korrespondierende Merkmale und Beispiele zu benennen, wobei Brandl davon ausgeht, dass die besten Vertreter zuerst genannt werden und dass eine homogene Gruppe von Experten ein homogenes Inventar bereits interiorisierter Typen benennt. Die Interviews wurden von der Verfasserin strukturiert, sortiert und auf die Kernaussagen reduziert. Insgesamt ermittelt Brandl 11 übergreifende Typen, deren Bezeichnungen auf den 36 von den Experten benannten Kategorien basieren (vgl. Tabelle 4.3).[50]

Das „Portal“ ist der in den Interviews meistgenannte Typ und bezieht sich entweder auf inhaltlich spezialisierte Websites oder solche mit einem sehr breiten Themenspektrum. Ihre Gemeinsamkeiten „laufen allerdings in einem Punkt zusammen: in der ihnen zugrunde liegenden Metapher – dem ›Eingangstor‹ ins WWW.“ (ebd., S. 88). Sie bieten ein umfangreiches Angebot an primär informationsorientierten und nur sekundär unterhaltenden Inhalten. Portale stellen auch Dienste und Anwendungen zur Verfügung, insbesondere Suchmaschinen und Kataloge. Bezüglich des Typs „Firmenpräsenzen“ geben die Experten eine Reihe von Konventionen an (vgl. Abschnitt 3.6.5), z. B. Rubriken wie „›Wer sind wir‹, ›Produkte und Services‹, ›Referenzen‹, ›Presseinformationen‹ usw. Ebenso wichtig und typisch […] ist das Angebot von Kontaktinformationen sowie Erklärungstexte oder Hilfemenüs zu Produkten und Dienstleistungen.“ (ebd., S. 92). Den Befragten zufolge existieren bei Firmenpräsenzen

[48] In einer späteren Arbeit stellen Haas und Grams (2000, S. 190) die Hypothese auf, dass Hyperlinks als RST-Relationen analysiert werden könnten. Haas (persönliche Kommunikation) hat darauf hingewiesen, dass dieser Ansatz als nicht erfolgversprechend evaluiert wurde.

[49] Linguistische Termini werden von Brandl nicht verwendet, ebenso werden keine Arbeiten aus den Bereichen CMC oder *Digital Genres* referenziert. Brandl (2002, S. 54) äußert sich wie folgt zur Forschungslage: „Wahrgenommene Typen, Gattungen oder Genres von Webangeboten sind bislang ein weitgehend unerforschtes Feld.“ Es wird argumentiert, dass sich Interviews mit Experten anbieten, um eine erste Klassifikation aufzustellen. Der Begriff „Genre“ wird analog zu Typen von Fernsehangeboten verwendet. Der Terminus „Typ“ wird kognitionspsychologisch aufgefasst und bezieht sich auf die Prototypentheorie (vgl. Abschnitt 2.2.9).

[50] Die in der Tabelle in Klammern genannten Beispiele wurden von der Verfasserin hinzugefügt. Auffällig sind in den von Brandl abgeleiteten Haupt- und Untertypen von Websites verschiedene Korrespondenzen mit der aus den Indikatoren abgeleiteten rudimentären Typologie von Hypertexttypen (Abbildung 3.7, S. 151), z. B. bezüglich Unterhaltungs-, Informations- und E-Commerce-Angeboten.

Website-Typ	Beispiele
Portal	myYahoo, AOL, Netscape, Microsoft, Yahoo, Infoseek, Netguide
• Sportportal	Sport1
• Spieleportal	Zone.com
• Meinungsportal	Ciao, Dooyou
• Medienportal	Süddeutsche Online, Welt Online, Focus Online
• Portal aus Suchmaschinen-Tradition	Lycos, Altavista, Yahoo
• Portal als Einstiegsseite	TU-München, Kickz
Firmenseiten	Siemens
• Unternehmenspräsentation (seriös, Visitenkarte)	Media-Aktiv, Mercedes, Siemens
• Image-/PR-Seiten	Mercedes, CocaCola, Jägermeister
• Produktseiten	Apple, Jägermeister, Audi, VW
E-Commerce-Angebote	Amazon, BOL, Letsbuyit.com, neckermann, Galeria Kaufhof
• Shop-Angebote	Galeria Kaufhof, Zooplus, Amazon
• Auktions-Angebote	Ricardo, Buy & Sold, Ebay
• B2B-Plattformen	MySAP, Covisint
Informationsorientierte Angebote	Spiegel Online, Focus Online, Sport1
• Linklisten/Linksammlungen	Heise, (Kataloge der Suchmaschinen), Sportal
• Massenmedien	CNN, Spiegel Online
• Zeitungs-/Zeitschriftenangebote	Spiegel Online, Focus Online, Süddeutsche Online, Heise
• Rundfunk-Angebote	RTL Online, PRO7 Online, BR-online, ARD online
• Fachinformations-/Wissensdatenbanken	Medline, Leo.org, Genois
Organisationspräsentationen	(Spendeninitiativen, Selbsthilfegruppen)
• Angebote von Interessengemeinschaften	Greenpeace, (politische Interessengemeinschaften)
• Vereinsseiten	Löwenfans, DFB
• Behörden-/Verwaltungsseiten	Gemeinde Grasbrunn
Unterhaltungs-Angebote	joecartoon.com, Netzpiloten
• Erotik-Angebote	Orion, Persian Kitty
• Spieleseiten	Zone.com, Gamezone
Suchmaschinen	Altavista, Fireball, Google, Infoseek, Yahoo, Northern Light
Service-Angebote	GMX, Microsoft, Novell, eGroups, Web.de
Privatseiten	(Hobby- und Sammlerseiten, persönliche Steckbriefe)
Communities	Netzpiloten, Ciao, Diraba
Akademisch-wissenschaftliche Angebote	IfKW
• Bildungseinrichtungs-Angebote	LMU Website
• Forschungseinrichtungs-Angebote	NASA, CERN

Tabelle 4.3: Haupt- und Untertypen von Websites mit typischen Vertretern – Ergebnis der von Brandl (2002, S. 87) durchgeführten Experteninterviews

bezüglich des Webdesigns große Unterschiede, ihre Typizität manifestiert sich vornehmlich in einheitlichen Navigationsstrukturen und ähnlichen Themenbereichen (z. B. Informationen über Produkte oder das Unternehmen). Noch ausgeprägter sind die Konventionen, die für den Typ „E-Commerce-Angebote“ genannt werden, denn diese zeichnen sich durch eine „gewisse Gleichförmigkeit“ aus (ebd., S. 93), die „Unterschiede könne man nur an Logo und Farbgebung ausmachen.“ (ebd., S. 94). Die „informationsorientierten Angebote“ werden als „sehr heterogene“ Gruppe bezeichnet (ebd.), die Untertypen bündelt, deren Funktion und Gemeinsamkeit die Vermittlung informationsorientierter Inhalte ist. Bei einigen Untertypen sind ebenfalls Konventionen feststellbar: „Die Befragten waren der Ansicht, dass insbesondere innerhalb der ›Online-Magazine‹ und -›Zeitungen‹ im Hinblick auf die formale Darstellung jeweils große Gleichförmigkeit herrscht.“ (ebd., S. 95).[51] Ein Typ „persönliche Homepage“

[51] Eine bereits angesprochene Konvention (Behme, 2000a, vgl. Indikator 38 auf S. 136) wird von einer Expertin sehr detailliert geschildert: „Beim Typ der ›Online-Zeitung‹ war eine Befragte der Ansicht, dass hier das Seitenraster [...] an das Layout von Print-Zeitungen erinnert. Der Content ist meist mittig [...] platziert, wobei am oberen Seitenrand häufig eine Werbebanner-Leiste angebracht ist. Links und rechts neben dem Haupt-Contentblock findet man meist Navigationsobjekte oder kleinere Werbebanner.“ (Brandl, 2002, S. 95 f.).

wird von den Experten interessanterweise nicht bestimmt. Brandl subsumiert sie unter der Bezeichnung „Privatseiten“ und charakterisiert sie als Angebote von Personen, die ihre Hobbys und Freizeitaktivitäten darstellen: „Das einzige, was Privatseiten inhaltlich gemeinsam haben können ist ein kurzer persönlicher Steckbrief der Person, die sie ins Netz stellt.“ (ebd., S. 105 f.). Die „akademisch-wissenschaftlichen Angebote“ werden primär durch inhaltliche Aspekte definiert, weil sie umfangreiche Informationen in Bezug auf Wissenschaft, Forschung und Lehre zur Verfügung stellen. Form und Darstellung derartiger Websites sind den Befragten zufolge „nüchtern und schlicht“ (ebd., S. 108).

In der zweiten Phase wurden 47 von den Experten genannte prototypische Websites 31 weiteren Probanden vorgelegt, die das WWW häufig nutzen. Die Aufgabe dieser zweiten Gruppe war es, die Websites auf der Grundlage des hierarchischen Sortierens nach ihren Ähnlichkeiten zu ordnen.[52] Brandl (2002, S. 82) geht davon aus, dass „durch den Stimulus einer Reihe hinreichend bekannter Webangebote die Aktivierung von Typen-Schemata bei den Probanden hervorgerufen [wird].“ Durch die hierarchische Sortierung entsteht ein Baum, der Distanzmaße zwischen den Beispielen, d. h. die von den Probanden wahrgenommenen Entfernungen zwischen zwei Websites reflektiert. Die 31 Bäume wurden in Matrizen der Größe 47×47 überführt und einer Analyse unterzogen, die sechs Cluster ergibt, die als die Typen „Service-Portale“, „Firmenwebsites“, „E-Commerce-Angebote“, „Medien-Angebote“, „Organisationspräsentationen“ und „Unterhaltungsangebote“ interpretiert werden (ebd., S. 114 ff.). Die Benennungen dieser Typen basieren auf den am häufigsten als Sortierkriterien genannten Begriffen. Die Ergebnisse der beiden Phasen zusammenfassend berichtet Brandl (2002, S. 140), dass es „erstaunlich [ist], wie stark sich die ermittelten Typen jeweils decken. Es zeigt sich, dass die übergreifenden Typen weitgehend übereinstimmen.“ Neben dieser Korrespondenz zeigt Brandl, dass die Experten zur Charakterisierung der selbstständig definierten Typen insbesondere formale Merkmale verwenden, während sich die Probanden der zweiten Phase an funktionalen Eigenschaften orientieren (ebd., S. 147).

4.4.6 Die Studie von Dewe et al. (1998)

Dewe et al. (1998, siehe auch Karlgren et al., 1998) gehen zur Erstellung einer „genre palette for Internet materials“, die in IR-Experimenten eingesetzt werden soll (vgl. Abschnitt 14.2.3), auf ähnliche Weise wie Brandl vor, verschicken jedoch einen Fragebogen per E-Mail an Angehörige der Universität Stockholm. Es wird argumentiert, dass bezüglich des Einsatzes maschineller Genre-Analysen in Suchmaschinen die Erwartungen der Benutzer in Betracht gezogen werden sollten. Gleichzeitig sollte die „genre palette“ jedoch keine zu umfangreiche Komplexität annehmen, um eine maschinelle Kategorisierung gewährleisten zu können. Von den 648 versendeten Fragebögen trafen 67 teils sehr unterschiedliche Antworten ein: Die meisten von den Angeschriebenen erstellten Listen enthielten nicht Genres, sondern themenspezifische Kategorien (z. B. „tourism“, „sports“, „games“, „economic info“, „culture“

[52] Jeder von den Experten ermittelte Typ war mit mindestens zwei Beispielen enthalten (Brandl, 2002, S. 66 ff.). Eine Liste der Beispiele wurde vorab an alle Teilnehmer geschickt. Die Namen und Adressen dieser Websites wurden zusammen mit einem Bildschirmabzug auf Karteikarten gedruckt, die die Probanden nach begründbaren Kriterien in zwei vergleichbare Stapel sortieren und mit einer Bezeichnung versehen sollten. Die beiden Stapel wurden sukzessive weiter sortiert und benannt. Das Verfahren ist abgeschlossen, sobald der Proband keine weiteren Unterschiede zwischen zwei Beispielen feststellen kann.

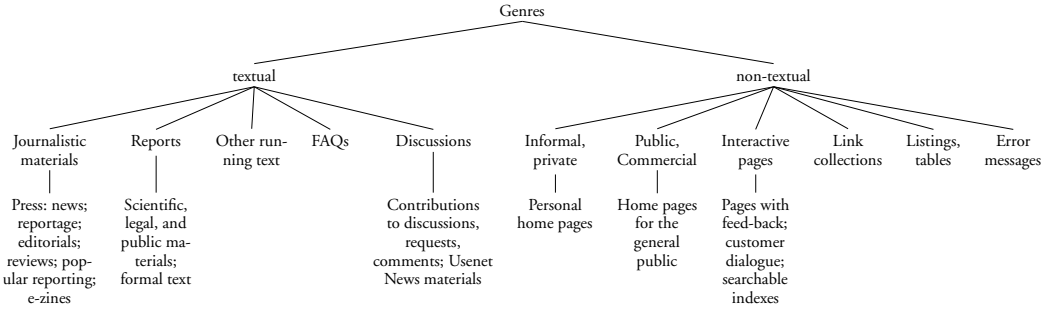


Abbildung 4.7: Die Genre-Hierarchie von Dewe et al. (1998) und Karlgren et al. (1998)

und „science“). Daneben wurden jedoch auch „home pages“, „data bases“, „FAQs“, „search pages“ und „reference materials“ genannt. Weitere Nennungen beziehen sich auf den Status oder die Intentionen der jeweiligen Produzenten, z. B. „here I am“, „sales pitches“, „serious material“, „commercial info“, „public info“ und „non-governmental organization info“. Mit „boring home pages“ wird auch auf die Qualität von Webangeboten eingegangen. Weil die Antworten auf unterschiedlichen Differenzierungskriterien basieren, versuchen Dewe et al., zumindest einen Teil der Nennungen systematisch entlang der von den Anwendern wahrgenommenen Dimensionen zu verorten (vgl. Abbildung 4.7).[53] Die entstandene Liste wurde mit der Frage, ob die enthaltenen Genres verständlich und vollständig seien, an die ursprünglichen Adressaten des Fragebogens versendet. Die 102 Antworten kritisieren unter anderem die Benennungen der Genres und zeigen fehlende Kategorien auf (z. B. „download page“, „ftp database directory“ und „search engine“). Darüber hinaus wird angemerkt, dass einigen der Kategorien im Kontext eines Suchszenarios keine Bedeutung zukomme, schließlich sei es unwahrscheinlich nach „error messages“ oder „interactive pages“ zu recherchieren. Dewe et al. kommen zu der Schlussfolgerung, dass WWW-Benutzer in Bezug auf HTML-Dokumente nur eine vage Vorstellung vom Genre-Konzept besitzen.

4.4.7 Die Studie von Rosso (2005)

Rosso (2005) präsentiert drei Benutzerstudien, von denen zwei der Ermittlung einer „genre palette“ (im Sinne von Dewe et al., 1998) dienen, die für die maschinelle Identifizierung von Web-Genres eingesetzt werden soll. Rosso beruft sich auf Miller (1984) und Swales (1990) und das Kriterium der Diskursgemeinschaft, der ein Genre bekannt ist (diese besitzt ein „shared genre knowledge“) und es in rekurrenten Situationen verwendet. Ein „genre on the web“ wird definiert als „a pragmatic type (with corresponding form and substance), recognized by

[53] Die in Abbildung 4.7 dargestellte Hierarchie wird von Dewe et al. nicht in dieser Form präsentiert. Zunächst stellen sie die „current genre palette“ als Liste dar, anschließend werden 11 Genre-Kategorien aufgeführt, für die aus unterschiedlichen Quellen HTML-Dokumente gesammelt wurden, zuletzt gehen Dewe et al. auf die Differenzierung zwischen „textual“ und „non-textual“ ein, da hierdurch die rudimentär erfolgte Implementierung einer maschinellen Genre-Detektion in der Lage sei, präziser arbeiten zu können (vgl. Abschnitt 14.2.3). Zur Bezeichnung der einzelnen Genres werden an einigen Stellen inkonsistente Termini eingesetzt, zudem werden im Text einige Genres genannt, die in den Listen nicht enthalten sind. Von den zahlreichen problematisch erscheinenden Aspekten seien lediglich zwei genannt: Die Existenz der Kategorie „Other running text“ wird nicht motiviert, ebenfalls wird nicht deutlich, weshalb das Genre „Listings, tables“ angenommen wird.

the genre's user group." (Rosso, 2005, S. 27). Mit Bezug auf Haas und Grams (1998b) wird dieser Definition eine Spezifizierung hinzugefügt: „A web genre without a user group is not a web genre. It is a web page type." (ebd., S. 35), d. h. „web page type" bezeichnet ein übergeordnetes Konzept.[54] Rosso betrachtet „web page types" als a priori postulierte Kategorien, die nicht von Benutzerstudien validiert werden, derartige Klassen sind also „most likely web page types, unless some evidence is provided that a specified user group can recognize them." (ebd.). Web-Genres hingegen können, so Rosso, ausschließlich durch Benutzerstudien ermittelt werden. Damit Web-Genres als Deskriptoren von Dokumenten in den Ergebnislisten von Suchmaschinen eingesetzt werden können, müssen Rosso (2005, S. 61 f.) zufolge drei Kriterien erfüllt sein: Die Anwender müssen ein ausreichendes Wissen über das Web-Genre besitzen, sie müssen die Relevanz eines Web-Genres für einen Informationsbedarf einschätzen können und es muss maschinell identifizierbar sein.[55] Rosso geht somit von der Frage aus, ob es möglich ist, eine „genre palette" zu entwickeln, die sowohl von den Benutzern identifizierbar als auch hinsichtlich der Verwendung von Suchmaschinen hilfreich ist.

Die erste Studie strebt eine Ermittlung der Web-Genres an, die Benutzer wahrnehmen. Rosso nimmt mit Bezug auf Rehm (2002b) eine Stichprobenrestriktion vor, die sich auf Dokumente aus dem Bereich `.edu` bezieht, d. h. es werden die Webauftritte US-amerikanischer Hochschulen und Universitäten untersucht. Durch mehrere Suchanfragen nach hochfrequenten Wörtern des Englischen wird mittels *Google* eine Stichprobe von 102 HTML-Dokumenten erzeugt, die ausgedruckt und von drei Versuchspersonen unter Anleitung des Verfassers hinsichtlich ihrer Genre-bezogenen Ähnlichkeiten sortiert werden (vgl. Brandl, 2002).[56] Anschließend nennen die Probanden das Etikett eines Web-Genres, eine Kurzdefinition und verschiedene Charakteristika. Rosso merkt an, dass die Versuchspersonen bei dieser Aufgabe jeweils unterschiedliche Aspekte fokussierten und somit individuelle Klassifikationen und Benennungen[57] vornahmen; der fehlende Kontext bei- und übergeordneter Dokumente hat die Sortierung ebenfalls erschwert. Von mindestens zwei Versuchspersonen in ein ähnlich benanntes Web-Genre aufgenommene Dokumente erlauben es Rosso, die drei Gruppen von „genre piles", die zwischen 24 und 28 Web-Genres enthalten, auf ein Inventar von 48 Web-Genres abzubilden, deren Etiketten und Definitionen weitestgehend auf den Angaben der Probanden beruhen (vgl. Tabelle 4.4).[58] Die 102 Dokumente wurden anschließend zehn Personen vorgelegt, um ein Inventar von Web-Genres zu produzieren, die HTML-Dokumenten

[54] Eine spezifischere Definition wird von Rosso (2005) *nicht* angegeben. Die Argumentation erscheint fragwürdig: Der initialen Definition von „Web-Genre" fügt Rosso (2005, S. 27) hinzu, dass einige Web-Genres allen, andere jedoch nur einem Teil der WWW-Benutzer bekannt sind. Die Gruppe *aller* WWW-Benutzer konstituiert also, so Rosso, eine Diskursgemeinschaft. Unter dieser Annahme kann jedoch das Konzept „web page type" streng genommen nicht existieren, schließlich kann es sich auf eben diese Gruppe sämtlicher WWW-Anwender beziehen bzw. in dieser Gruppe bekannt sein und verwendet werden.

[55] Das dritte Kriterium wird von Rosso (2005) nicht thematisiert.

[56] Hintergrundgrafiken wurden dabei nicht ausgedruckt. Weiterhin entfernt Rosso (2005, S. 69) „over-abundant page-types in the sample (newsletters/articles, and top-level homepages for schools, for example) [...] to keep the size of the sample down". Zur Webkompetenz der Versuchspersonen werden *keine* Angaben gemacht.

[57] Rosso (2005, S. 78) führt eine private Homepage als Beispiel an, die von den Probanden als „Homepage", „Index/Table of Contents" und „Links/Index" eingestuft wurde. Vergleichbare Beispiele liefert Santini (2005a).

[58] Die von Rosso (2005, S. 81 f.) vorgenommenen Modifikationen beziehen sich unter anderem auf die Änderung von Plural- in Singularbezeichnungen. Außerdem wurden Web-Genres, „which described relatively uncommon types of web pages (e. g., "obituaries")" ohne Angabe von Gründen aus der Liste entfernt (ebd.).

Genre	Description
about	short description of purpose or objectives of an organisation
abstract	title and brief description and reference (one page)
advice	advice on how to deal with a situation
article-1	something about a topic with supporting facts or opinions; tells a story somehow; not conversational
article-2	several pages talking about something (not as time sensitive as enews; more topic focused rather than event focused)
bibliography	page of pointers to books or articles; lists of pointers to papers, books, or other resources (that are usually on other sites)
biography	page primarily about a person
blurb	description page for a place or program (used to find out if you want to go there)
card catalog	reference to a titled work
contact form	for asking questions
conversations, observations, or opinions	opinions, stream of consciousness stuff, just talking
course	classes or programs offered for instructional purposes
course description	what's covered in a course; syllabus
course list	page that lists courses
database	for access to a database (a search engine)
diaries, weblogs, or blogs	a personal narrative or time log of activities (not a biography)
email	form for sending email
enews	online articles
FAQ	frequently asked questions; questions may be links to answers; not interactive like a forum
form	page for entering info
forum/interactive discussion archive	one or more messages and/or responses that are viewable by an audience
full-text index	page pointing to full-text of a book or article or magazine or journal
help	assisting you to perform a task (like an FAQ, but links are topics, not questions)
history	like "reference" but about the past
homepage	mission statement and table of contents for an organization
indices/table of contents/links	page which is primarily a list of links
instructional	recipe to follow a task
job listing	describes one or more jobs that are available
joke	story intended to be funny
navigation	top-level pages with list of links (to same site)
news index	headers for online articles (enews postings)
newsletter	fairly current news for an organization or a group
personal website	page that somebody writes about themselves
picture/photo	page primarily containing a picture with few or no words
poetry	contains a poetry or similar wordplay
product for sale	page from an online store
program description	describing educational programs
publications, bulletins, newsletters	information published about/by various organizations; collection of articles (or links to articles)
reference	detailed facts about a subject
registration	form for registration
resume	for looking for a job
review	short description of literature, art, TV, etc.
search start	place to type-in key words and search
shopping	for purchasing products
speech	text of a speech
story	shorter prose than reference, complete in itself, fiction or non-fiction
syllabus	course syllabus
welcome page	starting page (does not have to be the "top" page in a site)

Tabelle 4.4: Liste von 48 Genres – Ergebnis der ersten Studie von Rosso (2005, S. 193 f.)

von Benutzern mit einer hohen Übereinstimmung zugewiesen werden können. Die Probanden notierten hierzu für jedes Dokument das jeweilige Web-Genre in einer Liste; zusätzlich wurden für 25 der 102 Dokumente insgesamt 31 neue Web-Genres vorgeschlagen. Mehr als die Hälfte der Teilnehmer stimmte in der Zuweisung eines Genres für ein einzelnes Dokument in 60% aller Fälle überein. Auf der Grundlage dieser Ergebnisse reduziert Rosso (2005, S. 94) die Liste von 48 Web-Genres auf 18 Einträge (vgl. Tabelle 4.5): Nur Genres, die eine Übereinstimmung von mindestens 50% besitzen, werden beibehalten; sich in ihrem Skopus überschneidende Genres werden kombiniert.[59] In einem weiteren Experiment zeigt Rosso, dass Benutzer in der Lage sind, die 18 Web-Genres mit einer hohen Übereinstimmung 55 Dokumenten zuzuweisen: 48 Dokumente erreichten einen Konsens von mindestens 50%, die durchschnittliche Übereinstimmung beträgt 71,9%.

[59] Ein Grund für die durchgeführte Reduktion der Liste wird nicht genannt. Unter anderem werden die Web-Genres „joke", „reference" und „story" entfernt.

Genre	Description
article	something about a topic, often with supporting facts or opinions
course description	what's covered in a course; syllabus
course list	page that lists courses
diary, weblog or blog	a personal narrative or time log of activities (not a biography article)
FAQ/Help	frequently asked questions, or assistance in helping you perform a task; questions may be links to answers, or topics may be links to assistance; not interactive like a forum
form	page primarily for entering and submitting information (other than a search engine)
forum/interactive discussion archive	one or more messages and/or responses that are viewable by an audience
index/table of contents/links	a page which is primarily a list of links or text items ordered (usually alphabetically) so that a list item can be found easily, AND the page does not belong to any of the other categories
job listing	describes one or more jobs that are available
other instructional materials	materials (other than a syllabus) used in teaching course, including but not limited to tests, quizzes, assignments, answer keys, etc.
personal website	page (possibly a homepage) that somebody writes about oneself (but not a biographical article)
picture/photo	page primarily containing a picture or pictures with few or no words (other than captions)
poetry	contains poetry or similar wordplay
product for sale/shopping	for purchasing products (not a product review article)
search start	page primarily to enter key words and search a database; a search engine
speech	text of a speech
welcome/homepage	starting page (does not have to be the "top" page in a site); may contain introductory information about a specific organisation, department, program, etc. and a table of contents
NONE OF THE ABOVE	page that definitely does not fit into any of the above categories

Tabelle 4.5: Liste von 18 Genres – Ergebnis der zweiten Studie von Rosso (2005, S. 207)

4.4.8 Fazit – Zum Bedarf einer Restriktion der Untersuchungsdomäne

Die in den vorangegangenen Abschnitten dargestellten Listen von Web-Genres, Cybergenres, Page-Types und Website-Typen bieten im direkten Vergleich ein uneinheitliches Bild. Crowston und Williams ermitteln 48 (1997) bzw. mindestens 64 Genres (2000). Roussinov et al. finden 116 Genres, die auf fünf Kategorien reduziert werden. Eine ebenso kleine Gruppe von sieben „page types" ist das Resultat der Analyse von Haas und Grams, die, wie Roussinov et al., eine automatische Erkennung intendieren. Eine unmittelbare Deckungsgleichheit besteht lediglich in der sehr abstrakten Kategorie „Home page". Prinzipiell verfügt *jede* Website über eine Einstiegsseite, die aufgrund der etablierten Konvention als „Homepage" bezeichnet wird. Insofern handelt es sich bei der Homepage nicht um ein spezifisches Genre, sondern um einen generischen Hypertextknotentyp, für den sich in bestimmten Kontexten spezifischere Genres herausgebildet haben.[60] Brandl reduziert die Ergebnisse ihrer Experteninterviews ebenfalls auf eine kleine Gruppe von 11 Website-Typen. Obwohl sich Brandl nicht auf die einschlägige Literatur bezieht, ist eine erstaunliche Korrespondenz zwischen ihren Website-Typen und den Ergebnissen der zuvor genannten Arbeiten zu verzeichnen. Dies zeigt, dass die Analysen von Crowston und Williams, Shepherd und Watters sowie Haas und Grams durchaus eine gewisse Aussagekraft besitzen, obwohl sie induktiv und introspektiv von den jeweiligen Verfassern durchgeführt wurden und nicht auf den Ergebnissen von Benutzerstudien oder Interviews basieren. Die Ermittlung von Genres durch eine Auswertung per E-Mail eingegangener Fragebögen (Dewe et al., 1998) scheint kein adäquates Verfahren zu sein, da die individuellen Antworten zu inkonsistent sind und die Erstellung einer homogenen Genre-Hierarchie verhindern.[61]

[60] Auch für weitere der von den Verfassern als eigenständige „Genres" aufgelisteten Entitäten kann bezweifelt werden, dass es sich tatsächlich um Genres im eigentlichen Sinne handelt.

[61] Besonders problematisch erscheint die einzige Frage in dem von Dewe et al. (1998) verschickten „genre questionnaire" zu sein: „What genres do you feel you find on the WWW?" Zuvor findet sich eine knappe terminologische Erläuterung, die jedoch nur sehr wenige Beispiele enthält.

Der von Brandl verwendete Ansatz kann zur Aufstellung eines Inventars von Website-Typen eingesetzt werden. Die Studie fasst Websites als von Probanden wahrgenommene Ganzheiten auf und die Ergebnisse belegen, dass die Experten präzisere Resultate erzielen als die Versuchspersonen der zweiten Phase. Es ist davon auszugehen, dass dieses Verfahren für eine Ermittlung von Hypertextknotensorten nicht eingesetzt werden kann, da die korrespondierenden Schemata aller Wahrscheinlichkeit nach weniger eindeutig aus dem Gedächtnis abgerufen und benannt werden können (vgl. Abschnitt 4.3.1). Die Ergebnisse von Brandl zeigen, dass der von Crowston und Kwasnik (2004) präferierte *bottom-up*-Ansatz zum Aufbau einer Genre-Klassifikation auf der Basis von distinktiven Merkmalen, die von Rezipienten wahrgenommen werden, nicht erfolgversprechend ist (vgl. Abschnitt 5.8.2). Zwar wäre es durchaus wünschenswert, von den Benutzern auszugehen, doch sprechen verschiedene Gründe dagegen: Die automatische Verarbeitung muss zwangsläufig auf der Ebene des einzelnen HTML-Dokuments ansetzen; Brandl betrachtet hingegen vollständige Websites als atomare Analyseeinheiten und die von ihr ermittelten Benennungen dürften in Bezug auf einzelne Webseiten eine noch umfangreichere Varianz ergeben, was unter anderem darin begründet ist, dass eine Webseite in der Regel mehrere Funktionen gleichzeitig erfüllt (vgl. Abschnitt 4.5).

Rosso (2005) strebt die Entwicklung eines Inventars von Web-Genres an, das ausschließlich auf den Einschätzungen von Benutzern basiert, so dass diese „genre palette“ im Rahmen einer maschinellen Identifizierung eingesetzt werden kann. Verschiedene Aspekte der von Rosso angewendeten Methodik sind zu kritisieren und zeigen gleichzeitig die Grenzen dieser Vorgehensweise auf: Zunächst handelt es sich um eine extrem kleine Stichprobe von nur 102 Dokumenten, die keinesfalls *alle* im Hochschulbereich existenten Web-Genres abdeckt, so dass die intendierte Anwendung dieser „genre palette“ in maschinellen Verfahren Einschränkungen unterworfen ist. Weiterhin basiert die initiale Liste auf den Einschätzungen von lediglich drei Probanden, zu deren WWW-Kompetenz *keine* Angaben gemacht werden. Es wird nicht deutlich, weshalb diese drei Versuchspersonen „authentischere“ oder „bessere“ Ergebnisse produzieren als eine introspektiv von Experten durchgeführte Stichprobenanalyse (wie z. B. bei Crowston und Williams oder Haas und Grams). Im Gegenteil: Die beiden Listen enthalten zahlreiche undifferenzierte Termini und Definitionen mit einem hohen Abdeckungsgrad. Das zentrale Argument Rossos für diese Vorgehensweise ist die *Erkennbarkeit* des Web-Genres eines gegebenen Dokuments und, hierauf basierend, die Einschätzung der Relevanz eines Web-Genres für einen spezifischen Informationsbedarf. Es wird zwar demonstriert, dass das „shared genre knowledge“ der drei Probanden durchaus beschränkt ist, von wesentlicher Bedeutung ist jedoch im Kontext einer Genre-getriebenen Suchmaschine meiner Ansicht nach die Realisierung der korrespondierenden Benutzerschnittstelle: Wenn es sich hierbei z. B. um ein komfortables grafisches und personalisierbares Interface handelt, kann das Explorieren, Erlernen und individuelle Konfigurieren der unterstützten Web-Genres von der Oberfläche assistiert werden (vgl. Lucas und Topi, 2004). Falls die maschinelle Identifizierung von Web-Genres mit einer sehr hohen Präzision erfolgt und den Benutzern zu effizienteren Recherchen verhilft, spielt es letzten Endes keine Rolle, ob sie das Inventar von Web-Genres in den ersten Sitzungen erlernen mussten. Es können weitere problematische Aspekte beobachtet werden: Hierzu zählt z. B. die isolierte Betrachtung von Einzeldokumenten, die Annahme, dass exakt *ein* Dokument die Instanz *eines* Web-Genres darstellt und die Reduktion der initialen Liste von 48 Web-Genres auf ein Inventar von 18 Kategorien. Abschließend

sei darauf hingewiesen, dass Rosso in den Diskussionen mit den drei Versuchspersonen die Etikettierung der einzelnen „genre piles“ beeinflusst hat. Über eine Probandin wird bezüglich der Kategorisierung von Dokumenten mit HTML-Formularen berichtet: „She did not group forms together, but separated them into singleton piles according to purpose: email, registration, search engine, etc. When the experimenter suggested the word “form”, she said she liked that as an overall term.“ (Rosso, 2005, S. 75). Die initiale Liste mit 48 Einträgen enthält zwar die adäquaten, aber unpräzise etikettierten Web-Genres „email“, „registration“ und „search start“, Rosso reduziert jedoch zwei dieser Einträge in der zweiten Liste zu „form“, so dass die persönlichen Präferenzen des Verfassers in die reduzierte Liste eingeflossen sind.

Die Ansätze zur Sammlung von Hypertextsorten auf der Grundlage von Analysen zufällig zusammengestellter Stichproben zeichnen sich durch die Gemeinsamkeit aus, dass bei der Datenerhebung keine oder nur geringfügige Restriktionen angewendet wurden (vgl. Tabelle 4.6): Crowston und Williams (1997) und Haas und Grams (1998a,b, 2000) beschränken sich auf zufällig ausgewählte, englischsprachige Dokumente, während Shepherd und Watters (1999), Crowston und Williams (2000) und Roussinov et al. (2001) keinerlei Restriktionen anwenden.[62] Den Studien liegen Stichproben zugrunde, die den Datenbeständen von Suchmaschinen entnommen wurden und somit in den unterschiedlichsten Sprachen verfasste und in verschiedenen Kulturen[63] verankerte Instanzen beliebiger Dokumenttypen aus vielfältigen Domänen (private, wissenschaftliche und kommerzielle Angebote etc.) enthalten. Die heterogenen und wenig detaillierten Ergebnisse der Studien sind eine Konsequenz dieser breit gefächerten und ebenfalls heterogenen Stichproben. Die Motivation bestand in der Zu-

[62] Eine Ausnahme stellt die Arbeit von Rosso (2005) dar, der sich, mit Bezug auf Rehm (2002b), ausschließlich auf Dokumente US-amerikanischer Hochschulen bezieht. In Rehm (2002b) wurde die zwingend notwendige Restriktion derartiger Analysen auf eine spezifische Untersuchungsdomäne erstmalig thematisiert.

[63] Jakobs (2003, S. 238) beklagt das Fehlen von Studien, die sich mit „kulturspezifischen Einflüssen“ auf Hypertextsorten beschäftigen (vgl. Beghtol, 2001, Schütte, 2004a, und Schmid-Isler und Oehninger, 2004). Dieser Aspekt wurde vornehmlich für die Websites von E-Commerce-Anbietern untersucht, da er für sie von besonderer Relevanz ist, denn, wie Marcus und Gould (2000, S. 34) anmerken, „different cultures look for different data to make decisions.“ Die Theorie von Hofstede (1980; zitiert nach Marcus und Gould, 2000) sieht fünf Dimensionen vor, hinsichtlich derer sich Kulturen unterscheiden (unter anderem „collectivism vs. individualism“, „femininity vs. masculinity“ und „uncertainty avoidance“). Marcus und Gould (2000) zeigen für diese Dimensionen zahlreiche potenzielle kulturelle Spezifika auf, die an verschiedenen Websites (und an den von Hofstede vorgenommenen Zuordnungen der jeweiligen Länder zu den fünf Dimensionen) exemplifiziert werden. Hofstede (1980) kategorisiert die von ihm untersuchten Nationen basierend auf kulturellen Gemeinsamkeiten in sechs Cluster („Anglo cluster“, „Nordic cluster“, „German cluster“, „Latin cluster“, „Asian cluster“ und „Japan“). Robbins und Stylianou (2003) untersuchen jeweils 15 kommerzielle Websites aus diesen kulturellen Gruppen hinsichtlich ihrer Inhalts- und Design-Merkmale. Die Verfasser stellen fest, dass zwar bezüglich des Inhalts, nicht jedoch hinsichtlich der Gestaltung Unterschiede vorliegen, was durch die zu generelle Methodologie erklärt werden kann (untersucht wurden grobe Merkmale wie „Animation“, „Frames“, „Graphics“ und „Sound“). Cyr und Trevor-Smith (2004) zeigen, dass im WWW sehr wohl kulturelle Spezifika existieren; untersucht werden jeweils 30 Webauftritte von Städten aus den USA, Japan und Deutschland, da diese Länder nach Hofstede (1980) zahlreiche kulturelle Unterschiede aufweisen. Diese manifestieren sich im WWW z. B. in unterschiedlichen Navigationshilfen, Werbebannern, Kontaktmöglichkeiten und Farbräumen. Bucher (2004, S. 158) geht auf kulturelle Spezifika chinesischer Portalseiten ein, die „fast doppelt so viele Linkkategorien (23 Seitenelemente) auf[weisen] wie die internationalen und die sinisierten Portale (14 Seitenelemente), die Anzahl der Links ist mit rund 750 auf der Einstiegsseite um ein mehrfaches höher als auf westlichen Portalseiten (150 bis 180 Links).“ Bucher kommt zu dem Schluss, dass chinesische Portale nach dem Prinzip „soviel wie möglich auf einmal“ gestaltet werden und führt diesen Umstand auf die „Ästhetik der Fülle“ zurück (ebd., S. 159).

	Stichprobe/Datengrundlage	Ergebnisse	Hierarchie	Restriktionen
Crowston und Williams (1997)	100 Dokumente (zufällig ausgewählt aus dem Bestand einer Suchmaschine)	48 Genres	nein	Englisch
Crowston und Williams (2000)	837 Dokumente (zufällig ausgewählt aus dem Bestand einer Suchmaschine)	mindestens 64 Genres	ja	—
Roussinov et al. (2001)	1 234 Dokumente (ermittelt durch Interviews)	5 Genre-Gruppen 116 Genres	ja	—
Shepherd und Watters (1999)	96 Dokumente (zufällig ausgewählt aus dem Bestand einer Suchmaschine)	6 Cybergenres	nein	—
Haas und Grams (1998a, 1998b, 2000)	75 Dokumente (zufällig ausgewählt aus dem Bestand einer Suchmaschine)	7 page types 4 link types	ja	Englisch
Brandl (2002)	(a) Interviews mit 10 Experten; (b) 47 Websites als zu sortierende Beispiele (31 Versuchspersonen)	(a) 11 Typen von Webangeboten; (b) 6 Typen-Cluster	ja nein	— —
Dewe et al. (1998)	67 ausgefüllte Fragebögen	11 Genres; 16 Subkategorien	ja	—
Rosso (2005)	(a) Drei Personen weisen 100 Dokumenten beliebige Web-Genres zu; (b) Zehn Personen verwenden die 48 Web-Genres; Kategorien ohne Übereinstimmung werden entfernt	(a) 48 Web-Genres (b) 18 Web-Genres	nein	Nur `.edu`

Tabelle 4.6: Die Ansätze zur Ermittlung von Hypertexttypen und Hypertextsorten auf der Grundlage von Stichproben im Überblick

sammenstellung möglichst vollständiger Listen von „Web genres“, „Cybergenres“ oder „page types“, die im optimalen Fall alle im WWW existenten Hypertextsorten abdecken sollten. Die Heterogenität der Ergebnisse zeigt, dass eine solche Vorgehensweise Resultate produziert, die im Hinblick auf die Konzeptionierung maschineller Verfahren zur Identifizierung von Hypertextsorten sowohl unvollständig als auch unspezifisch sind (vgl. die von Roussinov et al., 2001, vorgeschlagenen sehr vagen und in dieser Form nicht implementierbaren „recognition indicators“). Roussinov et al. (2001, S. 210) erklären den Umstand, dass innerhalb ihrer Analyse sehr viele Genres ermittelt wurden, mit der Tatsache, dass im WWW zahlreiche Diskursgemeinschaften existieren. Aus diesem Grund beschränkt sich die vorliegende Arbeit auf die deutschsprachigen Dokumente der Webangebote deutscher Universitäten und Hochschulen, da vermutet wird, dass diese Domäne einer relativ homogenen Diskursgemeinschaft zugehörig und somit hinreichend spezifisch ist, um bezüglich der Resultate Homogenität zu gewährleisten, aber dennoch breit genug ist, um eine Typologie der an Hypertexttypen und Hypertextsorten beteiligten Konstituenten zu ermitteln.

4.5 Kerneigenschaften von Hypertextsorten

Hypertextsorten besitzen Eigenschaften, die sie von traditionellen Textsorten unterscheiden. Dieser Abschnitt geht auf diejenigen Aspekte ein, die insbesondere die Konzeptionierung des Hypertextsortenmodells beeinflusst haben (vgl. Kapitel 5). Zunächst wird der Einfluss der Form bei dem Prozess der Erkennung von Hypertextsorten behandelt. Anschließend wird das Phänomen der Einbettung von Hypertextsorten thematisiert (Abschnitt 4.5.2), das eine unmittelbare Relevanz für die beiden anschließend dargestellten Aspekte der Verknüpfung von Knoten (Abschnitt 4.5.3) sowie ihrer hierarchischen Anordnung besitzt (Abschnitt 4.5.4).

4.5.1 Der Einfluss der Form bei der Erkennung von Hypertextsorten

Im WWW fallen die vielfältigen Unterschiede vollständig weg, die Papierdokumente aufweisen können. Unabhängig von ihrer Speicherung, z. B. in einer Datenbank oder im Dateisystem, werden Webseiten letzten Endes immer in HTML präsentiert. Dass jedoch wahrnehmbare Unterschiede existieren, wird anhand der Studie von Crowston und Williams (2000, S. 205) deutlich: Einige zu analysierende Dokumente waren in Sprachen verfasst, die von den Verfassern nicht beherrscht werden. Dennoch konnte in diesen Fällen ein Web-Genre zugewiesen werden, was den Stellenwert der Form, d. h. der Typografie, der Anordnung der Informationsobjekte und Texte, des Textdesigns und des Textstrukturmusters hervorhebt.

Toms und Campbell (1999) untersuchen anhand einer Benutzerstudie die Einflussfaktoren der Form und des Inhalts bei der Erkennung von Textsorte (vgl. auch Toms, 2001). Sie gehen davon aus, dass jedes Genre Eigenschaften besitzt, die eine eindeutige Identifikation gewährleisten, so kann z. B. ein charakteristisches Layout distinktive Signale über den zu erwartenden Inhalt liefern. Es wird argumentiert, dass ein effektiver Umgang mit digitalen Dokumenten davon abhängig ist, wie erfolgreich der Anwender diese Signale erkennen kann, die auf zwei separaten Ebenen anzusiedeln sind. Neben der Form kann der Inhalt entscheidende Hinweise auf die Textsorte liefern, z. B. nach bestimmten Konventionen benannte Überschriften („Zusammenfassung“, „Schlussfolgerungen“, „Leistungsnachweis“ etc.). Toms und

Campbell geht es um die Klärung der Frage, ob Rezipienten in der Lage sind, die Textsorte auf der Basis der Form erkennen zu können, welche Eigenschaften dabei eine Rolle spielen und ob das Medium einen Einfluss auf die Erkennung hat. Zu diesem Zweck wurden Textsorten ermittelt, mit denen Angehörige einer (US-amerikanischen) Universität häufig umgehen: „Journal Article", „Course Reading List", „Departmental Memo", „Dictionary", „Minutes from Meetings" und „Course Calendar". Für jede Textsorte wurde ein Papierdokument und ein digitales Textexemplar aus dem WWW ausgewählt, die in zweierlei Hinsicht modifiziert wurden: Zur Maskierung des Inhalts wurden alle alphanumerischen Zeichen durch „x", „X" und „9" ersetzt; zur Maskierung der Form wurde das gesamte Layout entfernt, so dass der Text als eine fortlaufende und unstrukturierte Sequenz von Wörtern und Zeichen erscheint. Jeweils acht Printdokumente und acht digitale Dokumente wurden 15 Angehörigen einer Universität mit der Bitte um Identifizierung der Textsorte präsentiert; in vier Dokumenten war die Form und in vier weiteren Dokumenten der Inhalt maskiert.

Die Ergebnisse zeigen, dass Formeigenschaften einen erheblichen Anteil an der Identifikation der Textsorten besitzen. Die Probanden konnten durchschnittlich zehn Dokumente (63%) korrekt identifizieren, von denen jeweils etwa die Hälfte aus den Gruppen der Papiertexte und der digitalen Texte stammten, d. h. das Medium hat *keinen* Einfluss auf die Detektion einer Textsorte; die Erkennungssignale sind verlustfrei in digitale Dokumente übertragbar. Die zehn erkannten Dokumente teilen sich weiterhin auf in 5,7 Dokumente, in denen die Form und 4,5 Dokumente, in denen der Inhalt maskiert wurde. Eine Erkennung auf alleiniger Grundlage der Form kann also durchaus erfolgreich durchgeführt werden. Beghtol (2001) merkt an, dass sich diese Erkennung lediglich auf Textsorten mit typischen Formmerkmalen bezieht und unterscheidet „form genres" von „content genres".

Bei der Identifizierung wurden meist Formmerkmale eingesetzt, die eindeutig auf eine Textsorte hinweisen. Mehrdeutige Signale verursachten den Probanden Schwierigkeiten bei der Zuordnung. Ihre Reaktionen zeigten Toms und Campbell zufolge jedoch, dass vermutlich komplexere Prozesse vorliegen als die bloße Abbildung eines Signals auf ein bestimmtes Genre: Viele Versuchspersonen thematisierten zusätzlich zu den Merkmalen das Erscheinungsbild der Texte, zu dem die einzelnen Signale in Relation gesetzt wurden, bevor schließlich der Dokumenttyp benannt wurde. Bei der Maskierung der Form haben die Probanden meist angefangen, den Text zu rezipieren, um beim Vorkommen von Schlüsselwörtern (z. B. „Staff Meeting" und „Present" beim Sitzungsprotokoll) ein Genre zu benennen. Auch hier wurden Formaspekte hinzugezogen (z. B. Häufigkeiten von Zahlenausdrücken). Die Dauer, die zur Identifikation einer Textsorte aufgrund der Form oder des Inhalts benötigt wurde, weist keine Unterschiede auf. Toms und Campbell ziehen hieraus den Schluss, dass die visuellen Signale mit den inhaltlichen Hinweisen in den entscheidenden Sekunden der initialen Präsentation interagieren. Eine weitere Schlussfolgerung betrifft die Genres, die dem Probanden bekannt sind: Beispielsweise hat ein Bibliothekar einen „course calendar", bei dem der Inhalt maskiert wurde, als eine Liste von Zusammenfassungen interpretiert, wohingegen die teilnehmenden Studierenden die Textsorte unmittelbar korrekt benennen konnten. Toms und Campbell gelangen zu der folgenden Schlussfolgerung: „Clearly from this study, document structure can be used as a means of identifying documents. Evident also is the fact that those same cues that make a document immediately identifiable in the paper world are readily transferrable to the digital world." (ebd., S. 7).

4.5.2 Einbettung und Integration von Hypertextsorten

Exemplare von Hypertextsorten erlauben die Einbettung von Instanzen weiterer Hypertextsorten. Bereits Crowston und Williams (1997) weisen darauf hin, dass ihre Stichprobe mehrere Beispiele für dieses Phänomen enthält. Eines der Dokumente wird als „entry in an archive“ aufgefasst, der einen „letter“ enthält, der wiederum „stories for a folklore collection“ beinhaltet. Dies wird als eine Art kaskadierte Einbettung interpretiert: „Each wrapping created a new genre without completely losing the characteristics of the previous instantiation.“ (ebd., S. 37, Crowston und Williams, 2000, S. 207). Da digital verfügbare Texte einfacher manipuliert und in andere Dokumente integriert werden können, kann dieses Phänomen, so Crowston und Williams, im WWW sehr häufig beobachtet werden.[64]

Hyperlinks verschärfen diese Problematik: Crowston und Williams (1997, S. 37) führen das Beispiel einer Webseite über die Filme von Stan Laurel und Oliver Hardy an, die als „filmography“ aufgefasst wird. Jeder beschriebene Film besitzt einen Hyperlink zu den Produktseiten eines Anbieters, bei dem die Filme online bestellt werden können, so dass *zusätzlich* ein Produktkatalog vorliegt.[65] Auch Bittner (2003, S. 127) geht in seiner Analyse von 25 privaten Homepages (vgl. Abschnitt 4.6.3) auf die Eigenschaft dieser Hypertextsorte ein, als „Container“ fungieren zu können, d. h. „andere Textsorten und -formen in sich zu integrieren und dabei funktional und strukturell anzupassen.“ Bittner ermittelt vor allem traditionelle Textsorten wie z. B. „Reisebeschreibungen“, „Weinkunde“, „Kochbuch“ und „wissenschaftliche Texte“, aber auch eine „Fotogalerie“, „Linklisten“ und „E-Mail-Formulare“, die als „medienkontingente Textformen“ bezeichnet werden (ebd.). Der Behauptung Bittners, dass durch das häufige Anbieten von E-Mail-Formularen „die Grenze zwischen Rezeptions- und Kommunikationsorientierung verschwimmt“, ist zuzustimmen. Zu hinterfragen ist jedoch die Ansicht, dass die private Homepage eine Textsorte ist, die „verschiedene – heterogene – Textsorten und -formen zu einem kohärenten Ganzen verbinde[t]“ (ebd.). Kohärenz kann nicht als per definitionem gegebene Eigenschaft aller Exemplare dieser Hypertextsorte postuliert werden, die in jeder Kommunikationssituation und für jeden Rezipienten gilt.

Shepherd und Watters (1999, S. 2) weisen auf das verwandte Problem hin, dass nicht immer eine eindeutige Zuordnung einer Website zu einer Hypertextsorte vorgenommen werden kann: „The boundaries among [...] cybergenres are fuzzy and, in some instances, a Web site may be a composite of two or more of these cybergenres.“[66] Auch Haas und Grams (1998b, S. 104) setzen sich mit diesem Aspekt auseinander (vgl. Haas und Grams, 2000, S. 185):

[64] Crowston und Williams (2000, S. 207) liefern weitere Beispiele: „An *instruction sheet* on how to apply for a loan that included the *eligibility rules* for the program. An *announcement* that included the *Unix man[ual] page* for the software being announced. A *newsletter* that included an *events calendar*. A *press release* that included a *policy statement*.“ (Hervorhebungen hinzugefügt, G. R.).

[65] Crowston und Williams (1997) gehen nicht auf den Status des Autors des ursprünglichen Dokuments ein: Handelt es sich um einen kommerziellen Anbieter, so ist das Dokument als Produktkatalog zu interpretieren. Handelt es sich jedoch um ein privates Webangebot, liegt eine Filmografie vor, in die Hyperlinks zu online bestellbaren Filmen integriert wurden, um möglichst vielen Rezipienten die Gelegenheit zu geben, sich mit dem Œuvre von Laurel und Hardy vertraut zu machen. Auch die Werbung spielt eine wichtige Rolle (vgl. auch Fortanet et al., 1998, 1999, sowie Crijns, 2001): Viele kommerzielle Anbieter vergüten das Verfolgen von Hyperlinks zu ihrem Produktkatalog von privaten Seiten aus mit kleinen Beträgen.

[66] Es wird jedoch nicht deutlich, weshalb Shepherd und Watters (1999, S. 2) unmittelbar im Anschluss an diese Aussage das folgende Beispiel geben: „For instance, a commercial home page may link to a company catalogue.“

> A repeating problem we had while coding the pages concerned the level of granularity of coding [...]. All of the page types [...] occurred as individual Web pages. But there were also many instances where several of these page types occurred on one physical Web page. It is the author's design decision as to how many different functions a single page should accomplish, and it is not at all uncommon to see combinations [...]. Home pages are especially known for including several types. (Haas und Grams, 1998b, S. 104)

Während Shepherd und Watters von „some instances" sprechen, sind Haas und Grams der Ansicht, dass Mischungen in „many instances" vorliegen. Derartige Amalgamierungen *mehrerer* Hypertextknotensorten in *einem* HTML-Dokument könnten sich als Regelfall herausstellen, der die maschinelle Erkennung von Hypertextsorten zusätzlich erschwert. Es sind zwei unterschiedliche Konzeptualisierungen möglich: Entweder werden Hypertextknotensorten als monolithische Einheiten betrachtet (vgl. Haas und Grams, 2000, S. 181) oder sie werden als flexible Typen aufgefasst, die aus einzelnen Bestandteilen – Haas und Grams (1998b, S. 104) sprechen von „primitive building blocks" – zusammengesetzt sind. Bezüglich der zweiten Option ergibt sich die Schwierigkeit, ob einem aus mehreren Bausteinen zusammengesetzten HTML-Dokument im Zuge einer Analyse ein einzelnes oder mehrere Etiketten zugewiesen werden, die den „page type" bzw. die Hypertextknotensorte markieren. Falls dabei ein primitiver Basisdokumenttyp angenommen wird, der verschiedene Bausteine fordert, stellt sich die Frage, welchen Stellenwert der Basistyp für die Benennung besitzt. Es ist denkbar, dass er keinen oder nur einen geringen Beitrag zum Inhalt, zur Form oder zur Funktion des gesamten Dokuments beisteuert, in gewisser Hinsicht also von den integrierten Bausteinen überlagert wird (vgl. Haas und Grams, 2000, S. 185), was sowohl die maschinelle Erkennung als auch die Zuweisung eines eindeutigen Etiketts erschwerte. Haas und Grams verwerfen letzten Endes die Auffassung von „page types" als monolithische Entitäten, denn

> it seems that looking at page types as components of pages, which may occur singly or in combination, would provide the most flexibility in terms of [...] user functions such as searching, or in refining retrieval techniques. The larger questions of who (or what process) determines the page types, and when the classification(s) occurs, are left in need of further research. (Haas und Grams, 2000, S. 186)

Meines Wissens liegen auf die von Haas und Grams angesprochenen Forschungsfragen bislang keine Antworten vor.[67] Das in Kapitel 5 eingeführte Hypertextsortenmodell setzt an eben diesen sowie den nachfolgend thematisierten Fragen an.

4.5.3 Der Einfluss der Verknüpfung von Dokumenten

Die Verknüpfung einzelner HTML-Dokumente ist für die Bestimmung der zugrunde liegenden Hypertextsorte relevant und sie muss simultan in die Analyse der Bestandteile eines Dokuments einfließen. Haas und Grams (1998b, S. 101) geben ein Beispiel: „One author may design a page to fulfill several functions, such as providing an index to a document collection as well as an order form and purchase information, where another author would put these onto separate pages. Home pages provide good examples of the many different ways in

[67] Relevante Ansätze werden in Arbeiten diskutiert, die die Identifizierung von Strukturschemata in Webseiten anstreben (z. B. Carchiolo et al., 2003). Dieses Thema wird in Kapitel 14 ausführlich diskutiert.

which authors may divide essentially the same information into different configurations of actual pages." Es ist dem Autor eines Hypertextes überlassen, in welcher Form die Bestandteile realisiert werden, ob sie als Baustein in ein übergeordnetes Dokument integriert oder in eine separate, untergeordnete Datei ausgelagert und verknüpft werden. Diese Verwendung von Hyperlinks hat nicht notwendigerweise eine Veränderung der zugrunde liegenden Hypertextsorte zur Folge, weshalb Crowston und Williams (2000, S. 208) von einem „multipage document" sprechen. Hinzu kommen Verweise, die externe Hypertexte referenzieren.[68] Da HTML keine typisierten Hyperlinks unterstützt, kann nur näherungsweise bestimmt werden, welchen Status ein verknüpfter Knoten besitzt. Die Identifikation der Grenzen eines Hypertextes ist mit automatischen Verfahren sehr schwer zu erreichen, dies gilt umso mehr, wenn auf einem Webserver mehrere Hypertexte gleichberechtigt nebeneinander gepflegt werden und sich zur Ablage der Dateien Verzeichnisse teilen (vgl. Abschnitt 14.6.1).

Die bislang vorliegenden Analysen von HTML-Dokumenten, in denen mehrere Web-Genres ermittelt oder untersucht werden, betrachten die Ebene der einzelnen Datei. Die unterhalb dieser Ebene befindliche Schicht der eingebetteten Bestandteile wird meist ebenso vernachlässigt wie die abstraktere Ebene des Hypertextes, da sie die Komplexität einer Studie drastisch erhöhen. Dennoch ist es zwingend notwendig, beide Ebenen zu berücksichtigen. Crowston und Williams (2000, S. 210) schlagen vor, Web-Genres anhand korrespondierender Verknüpfungsmuster zu spezifizieren (vgl. die Indikatoren 21 und 22, S. 131): Beispielsweise wird die Hotlist definiert als Liste von Hyperlinks zu anderen Websites (vgl. Abschnitt 4.6.6). Ein online verfügbares Buch umfasst in der Regel eine sequenziell angeordnete Liste von Kapiteln, wobei jedes Kapitel einen Verweis auf den nachfolgenden Teil enthält.[69] Ausgehend von diesen Beispielen konstatieren Crowston und Williams (2000, S. 210): „Thus, the genre of a hyperdocument might be determined in part by examining how its component parts are linked together." Ein Hypertext kann also aus Knoten bestehen, die in typischer Weise miteinander verknüpft sind und gemeinsam die Instanz einer Hypertextsorte bilden und es können weitere Instanzen zusätzlicher Hypertextsorten integriert sein. Es können Abhängigkeiten zwischen der übergeordneten und den untergeordneten Hypertextsorten existieren: Der *Webauftritt einer Universität* enthält z. B. eingebettete Instanzen der Hypertextsorte *Webauftritt eines Fachbereichs*, während die Einbettung dieser Hypertextsorte innerhalb einer Online-Zeitung keinesfalls zu erwarten ist.[70]

Crowston und Williams (1999) beschäftigen sich mit den Zusammenhängen zwischen der Verlinkung und dem Web-Genre, wobei 70 aus dem Bestand von *Yahoo!* gesammelte

[68] Haas und Grams (1998b, S. 101) vergleichen in diesem Zusammenhang professionelle Hypertextanwendungen mit dem WWW: „Pages on the Web [...] share no such unity of purpose or style. An author may link with almost any other page (etiquette issues aside), regardless of the differences between the source and target page in terms of genre, style, intended audience, or even language or culture."

[69] Neben vollständigen Büchern existieren auch sehr viele „companion websites", die eine Buchpublikation begleiten und unter anderem der Werbung dienen (vgl. Miles-Board et al., 2004).

[70] Schönefeld (2001, S. 95 f.) postuliert „neue Textsorten" (vgl. Fußnote 34, S. 174) und merkt an, dass sie sich „nur auf die Gesamttextsorten der untersuchten Hypertexte beziehen. Es ist zu vermuten, daß diese Gesamttextsorte die Anzahl und die Art der Textsorten, die in verschiedenen Bereichen eines Hypertextes auftreten können, restringiert, ähnlich der Textsorte Fernsehnachrichten, die die weiteren enthaltenen Textsorten eingegrenzt [sic] (unter anderem Kommentar, Wetterbericht und Katastrophenwarnung)." Es existieren also hierarchische Abstufungen („Gesamttextsorte", „Textsorte", „enthaltene Textsorten"). Das Hypertextsortenmodell (vgl.Kapitel 5) stellt eine Terminologie zur Differenzierung dieser Konstituenten zur Verfügung.

FAQ-Dokumente als Stichprobe zur Untersuchung der These dienen, dass nur diejenigen Verknüpfungen, die die Funktion eines Dokuments ändern, auch das Genre des Dokuments beeinflussen, denn „merely dividing a document into pages does not, any more than routine repagination affects the genre of a paper document. Stated alternately, we argue that there is a class of changes to form, namely those related to pagination rather than purpose, that do not affect genre." (Crowston und Williams, 1999, S. 2).[71] Die Textsorte FAQ wurde ausgewählt, um zu überprüfen, welchen Einfluss die Verknüpfung auf ein Genre besitzt (vgl. Fußnote 153, S. 133). In 63 FAQs werden durchschnittlich 40 Hyperlinks pro Dokument verwendet; die verbleibenden sieben FAQs stammen ursprünglich aus dem Usenet und wurden ohne jegliche Modifikation im WWW publiziert.[72] Zur Ermittlung von Hyperlinkmustern wurde eine hierarchische Cluster-Analyse auf der Grundlage der Linkfrequenzen durchgeführt, die vier Gruppen von Dokumenten aufgedeckt hat: (i) „no links at all" (sieben Dokumente), (ii) „links primarily on the same page" (19 Dokumente), (iii) „links primarily to URLs with the same host name" (33 Dokumente) und (iv) „links primarily to URLs with different host names" (11 Dokumente). Die Dokumente des zweiten Clusters werden als eine Erweiterung des Genres FAQ aufgefasst, die es dem Rezipienten erleichert, die Antwort auf eine Frage zu lokalisieren (vgl. Fußnote 72).[73] Die im dritten Cluster enthaltenen FAQs bestehen aus mehreren Dateien, umfangreiche Dokumente werden also in meist nach thematischen Gesichtspunkten geordnete Teile zerlegt. Hierdurch findet nach Crowston und Williams eine Erweiterung des Genres FAQ statt, weil es an die Bedürfnisse des Mediums WWW angepasst wird. Die FAQs des vierten Clusters sind in den meisten Fällen jeweils nur ein einziges physikalisches HTML-Dokument und verwenden Hyperlinks, um Informationen auf anderen Websites zu referenzieren. Dies ist Crowston und Williams zufolge als Beibehaltung der Form und Erweiterung der Funktion zu interpretieren, denn „the documents have been transformed from simple FAQs to a hybrid of FAQ and hotlist". Die genannten Beispiele zeigen, dass die Verknüpfung einzelner HTML-Dokumente zu einem als Ganzheit fungierenden Hypertext im Kontext der Erkennung von Hypertextsorten als zentral zu betrachten ist.

[71] Crowston und Williams (1999, S. 2) führen zwei Beispiele an: Ein als Hypertext realisiertes Referenzhandbuch, das in separate Kapitel aufgeteilt ist, die nur in streng sequenzieller Form rezipierbar sind, verfehlt seinen Zweck, da es unmöglich ist, beliebige Informationen nachzuschlagen. Durch das Hinzufügen von Hyperlinks kann zusätzliche Funktionalität entstehen: Ein Roman kann mit einem Index versehen werden, der den Zugriff auf Szenen erlaubt, in denen bestimmte Charaktere agieren, wodurch den Verfassern zufolge „something more than a traditional novel" entsteht (ebd.). In gleicher Weise könnten jedoch z. B. auch farbliche Markierungen in die Papierversion dieses Romans geklebt werden, die ebenfalls einen derartigen Zugriff gestatteten. Es geht nicht darum, zu diskutieren, ob durch derartige peritextuelle Hyperlinks tatsächlich *neue* Textsorten entstehen, sondern primär um den Einfluss der Verlinkung auf zusammengehörige Dokumente.

[72] Alle Dokumente, in denen Hyperlinks eingesetzt werden, enthalten Verknüpfungen zu anderen HTML-Dateien auf dem gleichen Webserver (z. B. um einen sehr großen Bestand von Fragen und zugehörigen Antworten in mehrere Dateien aufzuteilen). Verweise innerhalb eines Dokuments existieren in 30 Dokumenten (z. B. um am oberen Seitenrand die Fragen darzustellen, die mit den im unteren Teil des Dokuments befindlichen Antworten verlinkt sind). Etwa 75% der Dokumente enthalten Verweise auf andere Webserver.

[73] Crowston und Williams (1999) weisen darauf hin, dass es nicht gerechtfertigt ist, bereits bei einer derartig marginalen funktionalen Erweiterung von einem anderem oder einem neuen Genre zu sprechen: „We contend that if the document still serves primarily as a repository of questions and answers that have come up in a [newsgroup], it remains an FAQ. However, it is unclear to us at what point such navigational aids allow a sufficiently new purpose to be served, giving rise to a novel genre."

4.5.4 Hierarchien und Typologien von Hypertextsorten

Textklassen werden in Typologien angeordnet, um Zusammenhänge zwischen ihnen zu verdeutlichen (vgl. Abschnitt 2.3.3). In der Diskussion ihrer Benutzerstudie geben Toms und Campbell (1999, S. 7) an, dass die Antworten der Probanden auf die Existenz eines „concept of taxonomic families of documents based on document structure" hindeuten: Einige Teilnehmer haben z. B. eingerückte Zeilen in fragmentarischen Absätzen als eine Reihe von Literatureinträgen interpretiert, wohingegen andere Probanden das Dokument unmittelbar als Literaturliste bezeichnet haben. Wieder andere Teilnehmer gingen von einer noch abstrakteren Ebene aus und zogen diejenigen Dokumenttypen in Betracht, die Literaturlisten beinhalten, z. B. wissenschaftliche Artikel. Aus dieser Beobachtung wird gefolgert, dass digitale Dokumente in Hierarchien gruppiert werden sollten. Unklar sei jedoch, ob die verwendeten Relationen partitiver (*has-part*) oder subsumierender (*is-a*) Natur sein sollten. In jedem Fall ist es jedoch nach Toms und Campbell (1999, S. 7) von zentraler Bedeutung, die Ebene der einfachen Auflistung von Genres zu verlassen.

Crowston und Williams (1997, S. 31) beschäftigen sich ebenfalls mit Genre-Hierarchien und geben ein intuitives Beispiel an: Das „social science paper" ist eine Subklasse des Genres „research paper", das wiederum unterhalb von „paper" angeordnet ist. Andere Subklassen von „research paper" verwenden sowohl identische (z. B. Nennung der Autoren, Angabe von Literaturverweisen) als auch variierte Merkmale (etwa nach fachspezifischen Konventionen benannte Abschnittsüberschriften). Crowston und Williams gehen von einer *is-a*-Relation aus, die spezifischere Subklassen beschreibt („social science paper" *is-a* „research paper" *is-a* „paper"), die die Eigenschaften ihrer Superklassen erben, wohingegen das von Toms und Campbell beschriebene Beispiel durch eine partitive Relation zu repräsentieren ist („wissenschaftlicher Artikel" *has-part* „Literaturliste" *has-part* „Literatureintrag"). Dies gilt auch für die Repräsentation von „index" und „table of contents", die der von Crowston und Williams (2000, S. 213) angenommenen Kategorie „partial documents" angehören, weil es sich um *Teile* von Textsorten wie z. B. „technischer Bericht" oder „Lehrbuch" handelt.

4.5.5 Fazit – Bezug zur maschinellen Analyse von Hypertextsorten

Trotz einer gewissen, von der Allgegenwärtigkeit der *Hypertext Markup Language* hervorgerufenen Gleichförmigkeit aller Webseiten, können unterschiedliche Typen von Hypertexten und Hypertextknoten differenziert werden, deren Spezifika sich auf einer abstrakten Ebene als unterschiedliche Merkmale von Inhalt, Form und Funktion beschreiben lassen. Bei einigen Textexemplaren ist eine alleinige Analyse der Ebene der Form zur Bestimmung oder Eingrenzung der korrespondierenden Hypertextsorte ausreichend (Toms und Campbell, 1999). Eine derartige Formanalyse muss eng mit der Erkennung der in ein Dokument integrierten und von der zugrunde liegenden Hypertextsorte vorausgesetzten Bausteine verbunden werden. Als Ansatzpunkt für derartige maschinelle Verfahren kann die Baumstruktur dienen, die innerhalb einer Webseite durch die enthaltenen HTML-Elemente aufgespannt wird. Eine Analyse der Linkstrukturen eines gegebenen Hypertextes ist ebenfalls von essenzieller Bedeutung, schließlich können sich die von der Hypertextsorte verlangten Bausteine unmittelbar in einem Dokument befinden oder in separate Dateien ausgelagert sein. Zur Modellierung derartiger Beziehungen bietet sich eine partitive Relation an, die innerhalb einer Hierarchie

von Hypertextsorten die Zusammengehörigkeit einzelner Bestandteile markiert. Gleichzeitig muss eine derartige Hierarchie jedoch auch die Spezialisierung von Hypertextsorten zulassen, so dass Subsumptionsrelationen unterschiedliche Granularitäts- bzw. Spezialisierungsstufen von Hypertextsorten modellieren können. Eine Genre-Hierarchie muss also mindestens diese beiden Ebenen in Betracht ziehen. Mit der Ebene der von Hyperlinks realisierten Verknüpfungen kommt eine weitere Ebene hinzu, die Graphenstrukturen beschreibt. Die nachfolgenden Kapitel gehen in detaillierter Form auf dieses Thema ein.

4.6 Charakterisierungen von Hypertextsorten

Neben den in Abschnitt 4.4 angesprochenen Studien zur Ermittlung von Web-Genres beschäftigen sich zahlreiche Analysen mit ausgewählten Hypertextsorten. Im Einzelnen handelt es sich dabei um verschiedene Ausprägungen der institutionellen (Abschnitt 4.6.2) und privaten bzw. persönlichen Homepage (Abschnitt 4.6.3). Während die Online-Zeitung ebenfalls ausführlich untersucht wurde (Abschnitt 4.6.4), liegen zur Online-Enzyklopädie (Abschnitt 4.6.5) bislang erst wenige Erkenntnisse vor. Die Hotlist wird in verschiedenen Beiträgen nur am Rande thematisiert, Abschnitt 4.6.6 fasst die wesentlichen Aspekte zusammen. Abschnitt 4.6.7 stellt das Weblog als eine Hypertextsorte vor, die der asynchronen computervermittelten Kommunikation sehr nahe steht. Dies gilt auch für das Gästebuch (Abschnitt 4.6.8). Abschnitt 4.6.9 geht auf verschiedene Hypertextsorten ein, die Aspekte der Interaktivität betonen, woraufhin Abschnitt 4.6.10 abschließend Hypertextsorten darstellt, die sich auf einer peritextuellen Ebene auf die Website selbst beziehen.

4.6.1 Vorbemerkungen – Das Konzept „Homepage“

Mit dem Begriff „Homepage“ wird im Allgemeinen die Einstiegsseite eines Informationsangebots im WWW bezeichnet, die dem Rezipienten simultan als Übersicht, Navigationshilfe und Kurzzusammenfassung der Inhalte dient. Obwohl der Begriff in allen aktuellen Wörterbüchern mit einander ähnlichen Bedeutungsparaphrasen aufgeführt wird (vgl. Schütte, 2004a, S. 198, für Beispiele), ist ihm eine gewisse Ambiguität inhärent (vgl. Rehm, 2002b): Nielsen (1995b, S. 183 f.) bezeichnet Homepages als „top nodes of an *organization's hypertext*“, die einen „high-level overview of an *organization* or an *individual user* with links to more indepth information“ präsentieren; Schlobinski (2000b, S. 812) definiert die Homepage als „*Titelseite einer Web-Site* [...] oder [...] *Web-Site einer Person*“ (Hervorhebungen hinzugefügt, G. R.). Dürscheid (2000, S. 63) führt verschiedene Synonyme an: „Die Seite, von der alle Verbindungen innerhalb des Textes ausgehen, ist die Startseite, die auch Begrüßungsseite, Einstiegsseite oder Homepage genannt wird.“[74] Die synonymen Ausdrücke (vgl. Storrer, 1999a, S. 39) belegen die Multifunktionalität: Die Homepage begrüßt den Leser, dient als Einstiegspunkt und fungiert als Titelseite. Bucher (1999, S. 14) nennt sie zusätzlich

[74] Im Anschluss erläutertDürscheid (2000, S. 63): „Diese Seite ist von jeder anderen Seite des Hypertextes aus anwählbar, möglicherweise aber auch von anderen, externen Webseiten aus.“ Dieses Zitat ist nur ein Beleg für den Umstand, dass oftmals die Ebenen der Charakterisierung der realen Gegebenheiten mit Gestaltungsempfehlungen vermengt werden: Es ist nicht zutreffend, dass die Einstiegsseite gleichsam per definitionem „von jeder anderen Seite des Hypertextes aus anwählbar [ist]“ (ebd.).

einen „Advanced Organizer“, da sie den Rezipienten in die Lage versetzen soll, sich einen vollständigen Überblick über die Inhalte und Funktionen der einzelnen Teile eines Hypertextes machen zu können. Auch Storrer (2001c, S. 193) merkt an, dass die Homepage den Umfang und die Struktur eines Hypertextes verdeutlichen sollte, wobei sie „als Einstieg, Wegweiser und als Dreh- und Angelpunkt beim Herumstöbern im Informationsangebot“ fungiert. Schütte (2004a, S. 195–200) geht im Rahmen „nominationstheoretischer Vorüberlegungen“ auf die Bedeutungen der Lexeme „Homepage“, „Einstiegsseite“ und „Begrüßungsseite“ ein, wobei „die bislang kaum erfolgte lexikographische Erfassung der Benennungen auf ein gegenwärtig noch schwach konturiertes Bedeutungsprofil [... hindeutet]“ (Schütte, 2004a, S. 197). Dieses manifestiert sich unter anderem in der Tatsache, dass der Begriff „Homepage“ in Webangeboten, auf Visitenkarten, in Anzeigen und auch umgangssprachlich oftmals zur Bezeichnung *vollständiger* Websites verwendet wird (vgl. z. B. Bittner, 2003, S. 125, und Jakobs, 2003, S. 236): Da die Homepage die Einstiegsseite darstellt, besitzt ihre URL in gewisser Hinsicht den Status der Adresse des gesamten Hypertextes.[75] Diese Auffassung der Extension von „Homepage“ entspricht zwar nicht dem etablierten Sprachgebrauch (vgl. Reiss, 2000, S. 10), sie entbehrt aber auch nicht jeglicher Plausibilität.

Schütte (2004a, S. 195) zufolge handelt es sich bei der Homepage um „eine für das World Wide Web spezifische Textsorte“, deren „wesenhaft bestimmende Parameter“ sind, dass sie „die erste Seite [...] eines Angebots“ darstellt, eine Informationsfunktion besitzt und Personen, Firmen oder Institutionen als Emittenten fungieren können (ebd., S. 200). Der Geltungsbereich dieser Parameter ist sehr umfassend, es liegt ein sehr hoher Abstraktionsgrad vor und es werden keine Merkmale genannt, die zwingend ausgeprägt sein müssen. Bezüglich der in Abschnitt 2.3.2 dargestellten Terminologie handelt es sich bei dem Konzept Homepage nicht um eine Textsorte, sondern um einen Texttyp.[76] Die Homepage ist ein inhärenter Bestandteil beinahe *aller* WWW-basierten Hypertexte, der – im Vergleich zu gedruckten Büchern – unter anderem die Funktionen der Titelseite, des Inhaltsverzeichnisses und des Index in sich vereinen kann (vgl. Abschnitt 3.5.2) und somit kein unmittelbares Pendant in der Klasse gedruckter Texte besitzt (vgl. Askehave und Nielsen, 2005). Inhaltliche, gestalterische oder z. B. auch strukturelle Konventionen können ausschließlich in unterschiedlichen Typen von Homepages ermittelt werden. Für *sämtliche* im WWW verfügbaren Homepages in globaler Weise geltende Konventionen können naturgemäß nicht über die eingangs dargestellten allgemeinen Eigenschaften hinausgehen. Dieser Umstand kann z. B. durch einen

[75] Schütte (2004a, S. 138) findet in ihrem Korpus von Unternehmenshomepages Belege, in denen die URL eines Webservers als „Seite“ bezeichnet wird und führt diesen Umstand auf eine Verwechslung von „Site“ und „Seite“ zurück. Eine alternative bzw. zusätzliche Erklärung betrifft die Möglichkeit, dass „Seite“ in diesem Zusammenhang als Synonym von „Homepage“ benutzt wird.

[76] Innerhalb des Hypertextsortenmodells (Kapitel 5) besitzt die Homepage den Status eines Hypertextknotentyps. Androutsopoulos und Kraft (2003, S. 3) vertreten die Ansicht, dass der „textsortenlinguistische Status der Homepage [...] bislang nicht geklärt ist“, weshalb sie – mit Bezug auf den angesprochenen sehr hohen Abstraktionsgrad – „die Homepage als Kommunikationsform [auffassen], auf deren Basis in einzelnen Domänen des World Wide Web Textsorten entstehen können.“ Die Verfasser gehen jedoch nicht auf die Frage ein, welchen Status die eingebetteten HTML-Dokumente einer Website besitzen, es wird also nicht deutlich, ob sie ebenfalls als (untergeordnete) Kommunikationsformen oder Bestandteil der „Kommunikationsform Homepage“ aufzufassen sind. Bezüglich dieser Problematik ist die von Schütte (2004a) vorgeschlagene Konzeptualisierung einer vollständigen Website als Kommunikationsform adäquater (vgl. Abschnitt 4.2.4).

Vergleich rudimentärer studentischer Homepages mit den Webangeboten international tätiger Unternehmen gezeigt werden, für deren professionelle Realisierung in aller Regel Budgets in Millionenhöhe zur Verfügung stehen.

Die Charakteristika von Homepages können wie folgt zusammengefasst werden (vgl. Storrer, 1999b): Sie dienen als Einstiegspunkte in Hypertexte und als generelle Orientierungshilfen, sie werden umgestaltet und umstrukturiert, um neu hinzugekommene Knoten in die Hypertextbasis einzubinden, vermitteln einen Eindruck über die Größe und den Umfang einer Website und sie werden in vielen Fällen regelmäßig inhaltlich aktualisiert. Angaben zum Erstellungsdatum sowie zur letzten Modifikation reflektieren die Aktualität, die als wichtiges Qualitätskriterium gilt (vgl. Aladwani und Palvia, 2002). Aktualisierungen können einerseits manuell erfolgen, andererseits binden viele Homepages, die einen Portalcharakter besitzen, die unterschiedlichsten Informationen dynamisch ein, z. B. Nachrichtenticker, Börsenkurse oder Schlagzeilen.[77] Die Produzenten von Homepages wünschen sich, dass diese von möglichst vielen Menschen wiederholt besucht werden, häufig wird den Rezipienten auch die Möglichkeit zur Kontaktaufnahme gegeben (z. B. per E-Mail); viele Anbieter fordern dazu auf, Anregungen oder Hinweise auf thematisch verwandte Angebote einzusenden.

Storrer (1999b) betrachtet fünf Typen von Homepages (gemeint sind Typen von Webangeboten) unter dem Aspekt ihrer kommunikativen Funktion: (i) Private Homepages werden von Privatpersonen veröffentlicht und schildern deren Interessen, während (ii) persönliche Homepages die berufliche Rolle einer Person präsentieren (vgl. Abschnitt 4.6.3). (iii) Institutionelle Homepages befinden sich nach Storrer auf den Websites von Universitäten, Forschungseinrichtungen, Behörden, Parteien und weiteren Organisationen des öffentlichen Lebens. Derartige Websites werden üblicherweise mit einer spezifischen Fragestellung besucht, weshalb Storrer den Aspekt der Navigation und effizienten Auffindung von Informationen in diesen meist komplexen Angeboten hervorhebt. Zur dritten Kategorie gehören auch die Homepages von Organisationen und Institutionen sowie Stadt- und Regionalinformationen. In den Webangeboten von Organisationen befinden sich häufig Mitarbeiterverzeichnisse, Telefon- und Adresslisten sowie Wegbeschreibungen zur Institution. In sprachlicher Hinsicht ist ein „nüchterner, geschäftsmäßiger Tonfall“ vorherrschend (ebd., S. 6), Begrüßungsfloskeln werden vermieden und die Benutzer werden im Falle einer direkten Ansprache gesiezt. Stadt- und Regionalinformationen umfassen Verzeichnisse von Hotels und Gaststätten, Veranstaltungskalender, Informationen zum öffentlichen Nahverkehr und „kommen meist etwas bunter und flotter daher“ (ebd.). Die Produzenten intendieren in aller Regel eine möglichst effiziente Abwicklung der unterschiedlichen Anliegen der Rezipienten. (iv) Themenbezogene Homepages stellen Informations- und Kommunikationsangebote zu spezifischen Themen dar und sind meist in umfassendere Websites eingebettet (z. B. von Verlagen oder Online-Zeitungen). Als Qualitätsmerkmale für diesen Typ gelten nach Storrer die Verständlichkeit und Attraktivität der Themenbehandlung, die Vollständigkeit und Verlässlichkeit der Informationen sowie ihre Aktualität, weshalb sie von Interessierten meist regelmäßig besucht werden. Ein „typischer Bestandteil“ sind Storrer (1999b, S. 7) zufolge Listen von Hyperlinks zu

[77] Mit RSS (*Really Simple Syndication*) existiert ein XML-Datenformat, das sowohl auf Produzentenseite die *syndication*, die Verteilung derartiger Informationsströme erleichtert, als auch auf Rezipientenseite ihren Empfang und die Konvertierung in HTML-Fragmente vereinfacht. Das Format wird auch häufig in Weblog-Plattformen (vgl. Abschnitt 4.6.7) zur Verteilung von Inhalten eingesetzt (vgl. Cayzer, 2004).

themenrelevanten Angeboten im WWW, häufig bestehen themenbezogene Homepages ausschließlich aus kommentierten Linklisten (vgl. Abschnitt 4.6.6). Ein zweiter hochfrequenter Bestandteil sind FAQ-Dokumente (ebd.). (v) Kommerzielle Homepages dienen der Werbung, der Kundenbetreuung und als elektronisches Kaufhaus (vgl. Abbildung 3.7, S. 151). Die Funktionen und Gestaltungen derartiger Websites variieren erheblich und sind von dem beworbenen Produkt abhängig. Große Firmen behandeln ihre Websites als Prestigeobjekt, „das vor allem prunkvoll und teuer wirken muss“ (ebd., S. 8). Als hochfrequente thematische Bereiche werden Produktinformationen, Kontaktadressen, Kundendienst und Angaben zur Firmenstruktur genannt, die Hersteller von Software haben die Möglichkeit, Demonstrationsversionen ihrer Produkte zum kostenlosen Download anzubieten. Während gerade in dieser Branche der eindeutige Mehrwert eines Webauftritts identifiziert werden kann, ist dies bei *low interest*-Produkten wie Nahrungsmitteln, Süßigkeiten oder dem allgemeinen Haushaltsbedarf weniger der Fall. Storrer (1999b, S. 8) ist der Auffassung, dass aus diesem Grunde auf den Websites dieser Unternehmen vornehmlich unterhaltende Angebote, Spiele und Rätsel vertreten sind, in die das jeweilige Produkt oder die Marke eingebunden sind, weshalb diese Webpräsenzen auch als „Adutainment“ bezeichnet werden. Storrer (1999b, S. 8) beobachtet eine direkte, informelle Ansprache des Rezipienten, profesionell formulierte Texte und einen „jugendliche[n] Sprachstil“. Diese Websites sollen den Rezipienten dazu anregen, möglichst viel Zeit in dem Angebot zu verbringen, ihn an das Produkt binden und zu dessen Kauf bewegen (vgl. Runkehl et al., 1998, S. 174 f.). Folglich existieren dort auch nur selten Linklisten zu externen Angeboten, die den Leser von der werbenden Website ablenken könnten.[78] Es wird deutlich, dass eine Generalisierung kommerzieller Homepages als digitale Werbeanzeigen (vgl. z. B. Singh und Dalal, 1999) zu kurz greift.

4.6.2 Die institutionelle Homepage

Dieser Abschnitt geht auf verschiedene Eigenschaften institutioneller Homepages ein. Zunächst wird eine Usability-Studie der Einstiegsseiten von 50 Webpräsenzen kommerzieller Unternehmen thematisiert und ihr Zusammenhang zu Hypertextsorten diskutiert. Kontrastiert wird diese Studie mit einer textlinguistischen Analyse von 105 Homepages der Webpräsenzen russischer und deutscher Unternehmen sowie einer Betrachtung der rhetorisch-funktionalen Struktur kommerziell ausgerichteter Homepages. Darauf wird der Aufbau von Konferenz-Websites erläutert. Abschließend wird eine Analyse vorgestellt, die die Entwicklung der Homepages der 50 US-Bundesstaaten über einen Zeitraum von mehreren Jahren aufzeigt. Abschnitt 4.6.3 geht ausführlich auf persönliche bzw. private Homepages ein.

Die kommerzielle Homepage

Nielsen und Tahir (2002, S. 37–53) nehmen eine empirische Analyse von 50 kommerziellen Homepages vor.[79] Die Ergebnisse der Analyse richten sich an Webdesigner. Nielsen und

[78] Online-Zeitungen verfahren ähnlich (vgl. Riley et al., 1998). Mit Bezug auf die Allgegenwärtigkeit von Hotlists auf privaten Homepages findet es Bittner (2003, S. 101 f.) „[p]aradox […], daß sie genau das Gegenteil der vom Autor intendierten Absicht bewirken, nämlich, daß der Rezipient die Homepage wieder »verläßt«, wenn er dem Hyperlink folgt.“ Die „intendierte Absicht“ wird in Abschnitt 4.6.3 genauer thematisiert.

[79] Es handelt sich um die Einstiegsseiten der Webauftritte von z. B. Disney, Ford, Microsoft und Amazon.

Tahir merken an, dass die Benutzer mehr Zeit auf anderen Websites als auf der eigenen Website verbringen (vgl. hierzu S. 145). Den Verfassern zufolge impliziert dies

> the need [...] to follow design conventions – users will be hard pressed to remember any special interaction tricks from one visit to the next, given the amount of time they'll spend on other sites between the two visits. [... We] measure the extent to which such conventions exist on the Web today. Over time, we expect even more design conventions to emerge. (Nielsen und Tahir, 2002, S. 37)

In Bezug auf den zentralen Aspekt der Benutzerfreundlichkeit raten Nielsen und Tahir, die ermittelten Konventionen anzuwenden, um die Eingewöhnungszeit für die Benutzer so gering wie möglich zu halten.[80] Es wird antizipiert, dass Webdesigner eine Vereinheitlichung von Gestaltungsprinzipien befürchten und diesen Ratschlag ablehnen könnten. Für die Grundlosigkeit dieser Befürchtung wird folgender Beleg angegeben:

> [The] design conventions don't at all mean that all homepages will look the same. Almost all magazines follow the convention of placing page numbers in the corner of the pages, displaying headlines in type larger than that of the body text, having the table of contents at the beginning, and many, many more principles that are common for the simple reason that they make magazines easy to read. This doesn't mean *Vogue* looks the same as *Sports Illustrated* or the *Far Eastern Economic Review*. Similarly, homepages that address different audiences or represent different companies will look different, even if they all promote ease of use by sticking to the conventions. (Nielsen und Tahir, 2002, S. 38)

Die ermittelten Konventionen beziehen sich nur sekundär auf die Ebene der grafischen Gestaltung. Primär geht es um die *Existenz* spezifischer Bestandteile einer Homepage und ihre kanonische Positionierung. Nielsen und Tahir zufolge sind diese Elemente zur bestmöglichen Unterstützung der Les- und Benutzbarkeit einer Website so essenziell wie Seitenzahlen in gedruckten Texten.[81] Diesbezüglich existiert kein Bezug zum Inhalt: Websites, die die unterschiedlichsten Themen diskutieren, können simultan den Konventionen folgen, d. h. in textlinguistischer Hinsicht beziehen sich Nielsen und Tahir primär auf peritextuelle Merkmale, die nur mittelbar einer Textsorte zugeordnet werden können.[82]

Tabelle 4.7 fasst die Ergebnisse der Analyse von Nielsen und Tahir (2002, S. 37–53) zusammen.[83] In Bezug auf „fundamental page design elements" diskutieren Nielsen und Tahir

[80] Es existieren vergleichbare Studien, die sich der Thematik aus betriebswirtschaftlicher Perspektive nähern, insbesondere unter den Gesichtspunkten des Direktmarketings (vgl. z. B. Liu et al., 1997, Palmer und Griffith, 1998, Huizingh, 2000, Zhang et al., 2000, und Aladwani und Palvia, 2002).

[81] Mehrere Arbeiten diskutieren den Stellenwert, den Klassifikationen von Web-Genres bzw. Hypertexttypen und -sorten für das Webdesign besäßen. Existierte eine derartige Klassifikation, könnte für ein spezifisches Projekt gezielt eine Hypertextsorte ausgewählt und gewinnbringend eingesetzt werden (Ryan et al., 2003, S. 407), um den kommunikativen Zweck zu betonen (Crowston und Williams, 2000, S. 211). Haas und Grams (1998b, S. 106) streben langfristig die Entwicklung von Empfehlungen zur Gestaltung benutzerfreundlicher Webseiten an. Crowston und Williams (2000, S. 211) raten, auf der Ebene des einzelnen HTML-Dokuments etablierte Genres zu reflektieren, da Anwender von Suchmaschinen häufig nicht zur Einstiegsseite, sondern auf interne Dokumente geführt werden, d. h. „the purpose and form of even a single page should be evident." (ebd.).

[82] Dieser Aspekt wird, wenngleich nur implizit, auch von Haas und Grams (2000, S. 185) angesprochen: „Organization/corporate home pages especially are starting to coalesce into a recognizable genre, with expected components, such as a site map (or partial map), search tool, and various text and multimedia pieces."

[83] Die prozentualen Angaben in der rechten Spalte von Tabelle 4.7 beziehen sich auf die Gesamtvorkommen.

Kategorie	Merkmal	Submerkmal bzw. Hyperlinkanzeiger	Proz.
Elemente der Seitengestaltung	Logo	Links oben: 84%, rechts oben: 6%, oben in der Mitte: 6%	100
	Suchmöglichkeiten	Suchmöglichkeiten gesamt	86
		Suche mittels Suchbox unmittelbar auf der Homepage	81
		Rechts oben: 35%, links oben: 30%, oben in der Mitte: 14%	
		Suche mittels separatem Dokument	20
		Weißer Hintergrund des Suchformulars	97
		„Search“ (42%), „Go“ (40%), „Find“ (9%), „Find It“ (5%)	
Navigationselemente	Primäre Navigationshilfen	Navigationsleiste am linken Seitenrand	30
		Reiter-Metapher („tabs“) am oberen Seitenrand	30
		Hyperlinkliste am oberen Seitenrand	18
		Link-Kategorien im Seitenzentrum	12
		Pull-Down-Menüs	10
	Sitemap	„Site Map“ (63%), „Site Index“ (13%), „Site Guide“ (8%)	48
	Splash-Seite		6
	Navigationselemente in der Fußzeile		80
Häufig eingesetzte Module	Benutzeranmeldung	„Your Account“ (19%), „Login“ (19%), „Sign In“ (15%)	52
	Firmeninformationen	„About *Firma*“ (55%), „About Us“ (21%)	84
	Kontaktinformationen	„Contact Us“ (89%), „Contact *Firma*“ (4%)	90
	Datenschutzrichtlinien	„Privacy Policy“ (47%), „Privacy“ (19%)	86
	Stellenanzeigen	„Careers“ (18%), „Jobs at *Firma*“ (16%), „Jobs“ (13%)	74
	Hilfeseiten		54
Werbung	Anzeigen (eigene Produkte)	Median: 4,5	84
	Anzeigen (Fremdprodukte)	Median: 3	46
Typografie	Schwarze Schrift		72
	Blaue Schrift		8
	Graue Schrift		8
	Weißer Hintergrund		84
	Weiße Schrift auf schwarzem Hintergrund		4
	Hyperlinks unterstrichen		80
	Darstellung von Hyperlinks in blauer Farbe		60
	Darstellung von Hyperlinks in schwarzer Farbe		12

Tabelle 4.7: Charakterisierung des Hypertexttyps kommerzielle Homepage (nach Nielsen und Tahir, 2002, S. 37–53)

(2002, S. 41 f.) die Bestandteile Logo und Suchformular. Alle 50 untersuchten kommerziellen Homepages besitzen ein Logo, das sich in 84% aller Fälle in der linken oberen Ecke der Seite befindet, so dass von einer Konvention gesprochen werden kann. Suchmöglichkeiten sind in 86% der Homepages enthalten. In 81% der Seiten kann der Anwender unmittelbar ein kleinformatiges Suchformular verwenden („search box“), wohingegen in 20% der Homepages ein separates HTML-Dokument angesteuert werden muss – einige Webseiten bieten beide Möglichkeiten. Die Suchbox ist in fast allen kommerziellen Homepages verfügbar, bezüglich ihrer Platzierung hat sich jedoch noch keine Konvention etabliert.

In der Kategorie „navigation“ analysieren Nielsen und Tahir (2002, S. 43 f.) Navigationshilfen, Fußzeilen, Sitemaps und Splash-Seiten. Bezüglich der Anordnung und Gestaltung der primären Navigationshilfe dominieren am linken Rand angebrachte Navigationsleisten („left-hand navigation rails“) sowie am oberen Rand befindliche Reiter (jeweils 30%).[84] Eine Auflistung der Hyperlinks am oberen Seitenrand wird in 18% der Dokumente beobachtet.

[84] Bucher (2004, S. 154) kommt im Rahmen einer Rezeptionsstudie kommerzieller Webangebote zu ähnlichen Schlussfolgerungen. Die Ergebnisse zeigen, dass „dass diejenigen Angebote, ganz unabhängig von der Branche, als nutzerfreundlicher gewertet werden, deren Seiten nach einem Prototypen gestaltet sind, der auf der linken Seite eine vertikale, am Kopf der Seite eine horizontale Navigation verwendet, dessen *content*-Teil die rechten beiden Drittel der Seite umfasst, und der über eine Suchoption verfügt.“

Zusätzlich existiert die Anordnung von Links im Seitenzentrum (12%), Pull-Down-Menüs (10%) und sonstige Typen. Navigationselemente in der Fußzeile kommen zwar in 80% der Dokumente vor, aber „there is no agreement about what to include in this navigation list" (Nielsen und Tahir, 2002, S. 43). Einige Fußzeilen enthalten Hyperlinks, die am ehesten als Fußnoten bezeichnet werden können, z. B. Copyright- und Kontaktinformationen: „Because of the lack of a standard, users don't know what to expect in the footer, and this makes the area less useful than it could be." (ebd.).[85] Einen Link zu einer Sitemap enthalten 48% der Homepages. Splash-Seiten werden in 6% der analysierten Dokumente verwendet. Aus Sicht der Benutzerfreundlichkeit sind sie abzulehnen (vgl. Abschnitt 3.6.5).

Abstrakte funktionale und thematische Bereiche werden von Nielsen und Tahir (2002, S. 45 ff.) als „frequent features" diskutiert. Über eine Anmeldeprozedur können sich Benutzer in 52% der untersuchten Homepages dem Webserver gegenüber identifizieren, um z. B. Personalisierungsfunktionen zu nutzen.[86] Hyperlinks zu Informationen über den Anbieter der Website sind in 84% der Homepages enthalten. Kontaktinformationen bieten 90% der Homepages an, wobei 60% einen mit „Contact Us" beschrifteten Hyperlink in der Einstiegsseite präsentieren, während sich derartige Links auch auf der „About Us"-Seite (22%) und in Bereichen wie „Help" oder „Customer Service" befinden (14%). Viele kommerzielle Websites sammeln Informationen über die Benutzer (z. B. ihre E-Mail-Adressen), weshalb 86% der Homepages Angaben zu den Datenschutzrichtlinien eines Unternehmens machen (vgl. Vila et al., 2003).[87] Als weitere Konvention gelten Hyperlinks zu Stellenanzeigen, die in 74% der Homepages enthalten sind und in 53% aller Fälle das Wort „Jobs" enthalten. Links zu Hilfeseiten besitzen einen weniger ausgeprägten Grad der Konventionalisierung, sie werden in 54% der Websites angeboten, wobei Positionierungen in der rechten oberen (41%) und der linken unteren Ecke (19%) dominieren. Nielsen und Tahir (2002, S. 45) gehen davon aus, dass sich in Zukunft weitere Konventionen bilden werden: „As the Web evolves, it's likely that even more conventional features will become established, especially for sites in specific genres such as corporate sites and goverment sites." (vgl. Teil III).[88]

[85] Nielsen und Tahir (2002, S. 44) empfehlen eine „footnote-style navigation", wobei die primären Navigationsbestandteile nicht wiederholt werden sollten. Ein generelles Usability-Prinzip besagt, die Duplikation von Hyperlinks, d. h. Redundanz, zu vermeiden.

[86] Es handelt sich um ein Merkmal, das sich fast ausschließlich auf kommerziellen sowie kommunikationsorientierten Websites (Diskussionsforen) findet. Die Varianz in der Benennung der Hyperlinkanzeiger belegt, dass keine standardisierte Terminologie existiert (Nielsen und Tahir, 2002, S. 45).

[87] Jensen und Potts (2004) gehen auf *privacy policies* ein und zeigen, dass in einer Stichprobe von Websites in 77% aller Fälle Datenschutzrichtlinien angeboten werden. In 86% der Homepages befindet sich der Link in der Fußzeile, drei Websites (5%) stellen den Link in einer Navigationshilfe am linken Seitenrand dar und zwei Einstiegsseiten (3%) enthalten den Hyperlink am oberen Seitenrand. Die Verfasser stellen fest: „Even if one assumes that companies sincerely follow practices that comply with their posted policies, the form, location and legal context of policies make them essentially unusable as decision-making aids for a user concerned about privacy." (ebd., S. 477). Die Privacy-Policy besitzt Pendants in den traditionellen Textsorten, die primär bei Preisausschreiben, Umfragen und Benutzertests eingesetzt werden.

[88] Reiss (2000) diskutiert unterschiedliche Typen von Websites (z. B. „Newsletter sites", „Image sites", „Tile sites", „Traditional sites" und „Search sites") und konstatiert: „In the years to come, new styles will evolve as more sophisticated web conventions are established. For the time being, though, very few navigational devices can be taken for granted. That's why the good sites have a tendency to resemble each other: visitors stand a better chance of recognizing and understanding navigation and structure if they've already seen it somewhere else. That's also why traffic signals are red/yellow/green from Toledo to Tashkent." (ebd., S. 81).

Abschließend gehen Nielsen und Tahir (2002, S. 50 f.) auf die Bereiche Werbung und Typografie ein. Anzeigen für die Produkte anderer Firmen sind in 46% der Homepages enthalten, Werbung für eigene Produkte wird in 84% der Einstiegsseiten festgestellt. In 72% der Homepages wird schwarze Schrift benutzt, in 84% wird ein weißer Hintergrund verwendet.[89] Links werden in 80% der Seiten unterstrichen dargestellt.

Die Analyse von Nielsen und Tahir (2002, S. 52 f.) mündet in einem Empfehlungskatalog, der Richtlinien für 40 Kriterien und Gestaltungsmerkmale liefert, wobei die „strength" einer Empfehlung „default", „strong" oder „essential" sein kann.[90] Auf die Frage, ab welcher Frequenz eine Konvention vorliegt, gehen Nielsen und Tahir nicht ein, es lassen sich jedoch Rückschlüsse aus den prozentualen Angaben ziehen (vgl. Tabelle 4.7): Viele der essenziellen Empfehlungen besitzen sehr hohe (z. B. Platzierung des Logos in der linken oberen Ecke oder Angaben von Firmeninformationen) oder sehr niedrige Frequenzangaben (Vermeidung von Splash-Seiten, Animationen oder eingebetteter Musik). Nielsen und Tahir scheinen sich an dem Wert von 50% für essenzielle Empfehlungen zu orientieren.

Die Homepages der Webpräsenzen russischer und deutscher Unternehmen

Schütte (2004a, S. 135) nimmt eine textlinguistische Verortung der „mediumsspezifischen Textsorte Homepage" anhand eines Korpus von 105 im Jahr 2000 archivierten Einstiegsseiten der Webauftritte russischer und deutscher Unternehmen vor.[91] Die Textsortenkonventionen werden – wie in fast allen in diesem Kapitel vorgestellten Arbeiten – mit einem empirisch-induktiven, *bottom-up*-basierten Vorgehen untersucht. Die Homepages der deutschen Unternehmen werden nach formalen und funktionalen Kriterien drei Typen zugeordnet, die als „Leitseite", „Leit-Inhalts-Seite" und „Pre-Homepage" bezeichnet werden (ebd., S. 154; vgl. Tabelle 4.8). Die Funktion des Typs „Leitseite" betrifft die Vermittlung von Übersicht über die Inhalte einer Website; es liegen vier Subtypen vor: (i) Inhaltsübersicht (15 Dokumente), (ii) Inhaltsübersicht + Neues/Aktuelles (12 Dokumente), (iii) Inhaltsübersicht + Unternehmenskurzvorstellung (4 Dokumente), (iv) Inhaltsübersicht + Aktuelles + Unternehmenskurzvorstellung (4 Dokumente). Der abgeleitete Typ „Leit-Inhalts-Seite" ist „signifikant der Inhaltspräsentation verpflichtet" (ebd., S. 155) und umfasst primär aktuelle Informationen (ebd., S. 168). Die „Pre-Homepage" besteht aus der „Pre-Page und einer nachgeschalteten Leitseite oder Leit-Inhalts-Seite" (ebd., S. 157).[92] Falls eine Pre-Page verwendet wird,

[89] Dieser Trend wird von Ryan et al. (2003, S. 418) bestätigt (vgl. auch Murayama et al., 2004).

[90] Nielsen und Tahir (2002, S. 52) betonen erneut den Stellenwert von Konventionen: „[. . .] sites work best when they follow the conventions users know from other sites. [. . .] Even when a convention may be suboptimal from a theoretical perspective, in practice it will work well because users will *know* how it works."

[91] Mit Bezug auf Bucher (1999) ermittelt Schütte (2004a, S. 153) einander widersprechende Empfehlungen, die für praktische alle Homepages gelten: „Dem für Übersichtsseiten typischen Prinzip der ökonomischen Gestaltung steht das des reichhaltigen Angebots, dem der Selektion das Prinzip der Vollständigkeit und dem Prinzip der optischen Auffälligkeit das der Übersichtlichkeit gegenüber."

[92] Der unübliche Begriff „Pre-Page" (d. h. Splash-Seite) ist an Runkehl et al. (1998, S. 173) angelehnt. Es wird nicht deutlich, weshalb Schütte annimmt, dass eine „Pre-Homepage" aus zwei separaten Dokumenten besteht. Die Verfasserin präsentiert die absolute und prozentuale Verteilung der drei bzw. vier Typen anhand zweier Tabellen (Schütte, 2004a, S. 157, S. 177), wobei diejenigen „Leitseiten" und „Leit-Inhalts-Seiten", die „Pre-Pages" nachgeschaltet sind, nicht in die Vorkommen dieser Typen einfließen, wodurch eine leichte Verzerrung entsteht. Tabelle 4.8 enthält sowohl die von Schütte gewählte Präsentation als auch die absolute Verteilung.

Typen	*Deutsches Korpus* Anzahl	%	Anzahl *mit Pre-Homepages*	%	*Russisches Korpus* Anzahl	%	Anzahl *mit Pre-Homepages*	%
Leitseiten	35	64	49	89	22	44	31	62
Leit-Inhalts-Seiten	3	5	6	11	9	18	15	30
Inhaltsseiten					3	6	4	8
Pre-Homepages	17	31			16	32		
Pre-Page + Leitseite	14				9			
Pre-Page + Leit-Inhalts-Seite	3				6			
Pre-Page + Inhaltsseite					1			
Homepages gesamt	55	100	55	100	50	100	50	100

Tabelle 4.8: Verteilung der Homepage-Typen in den beiden von Schütte (2004a, S. 157, S. 177) untersuchten Korpora

übernimmt diese die Titel- und Begrüßungsfunktion, wohingegen die nachfolgende Webseite die Leitfunktion besitzt. Neben der Nennung des Firmennamens ist die Präsentation unterschiedlicher Sprachoptionen die Hauptfunktion der Pre-Page (ebd., S. 171). Der Zugang zur eigentlichen Website erfolgt beim Einsatz einer Pre-Page durch mindestens eine der folgenden Arten: Automatische Weiterleitung, Sprachauswahl, Firmenname und/oder Logo bzw. Substantiv „Start“ als (möglicherweise einziger) Einstiegslink. Bei den russischen Homepages wird die Einführung des zusätzlichen Typs „Inhaltsseite“ notwendig (enthält keine Inhaltsübersicht), weil die korrespondierende Website lediglich aus einem oder – beim Einsatz einer Splash-Seite – zwei Dokumenten besteht. Schütte (2004a, S. 186) schreibt derartigen Webangeboten den Status einer „elektronischen Visitenkarte“ zu, die als obligatorische Elemente einen Firmenschriftzug, ein Logo, eine Kurzvorstellung sowie Kontaktinformationen enthält, so dass es sich lediglich „um ein frühes Entwicklungsstadium, gleichsam die Vorform einer Website“ handelt (ebd., S. 187). Die textlinguistischen Analysen erfolgen auf der Basis eines Kriterienkatalogs, der unter anderem die Kommunikationssituation, das Thema, die Makrostruktur, Verweisstrukturen und sprachliche Spezifika umfasst, wobei Schütte (2004a, S. 203 f.) darauf hinweist, dass „eine umfassende, geschweige denn vollständige Erhebung aller textsortenrelevanter Merkmale [...] kaum zu leisten [ist].“

Von besonderer Relevanz für die vorliegende Arbeit ist die Analyse der Makrostruktur, die Schütte als inhaltlich-formale sowie funktional bedingte Textstruktur versteht, deren Bestandteile visuell voneinander abgegrenzt werden und somit Cluster bilden (Weingarten, 1997b). Schütte (2004a, S. 214) erhebt für die deutschen Unternehmenshomepages zehn makrostrukturelle Konstituenten (vgl. Tabelle 4.9), von denen der Firmenname und die Inhaltsübersicht aufgrund ihrer Frequenz als „obligatorische Bestandteile“ aufgefasst werden.[93]

[93] Ein Vergleich dieser Komponenten (vgl. Tabelle 4.9) mit den von Nielsen und Tahir (2002) diskutierten Merkmalen (vgl. Tabelle 4.7) deckt nur wenige Gemeinsamkeiten auf: Übereinstimmung herrscht lediglich in Bezug auf das Logo, das in beiden Studien mit einer Frequenz von 100% angegeben wird. Während Nielsen und Tahir fünf Typen „primärer Navigationshilfen“ ansetzen, geht Schütte von dem Typ „Inhaltsübersicht“ aus. Nielsen und Tahir beobachten Splash-Seiten in 6% der untersuchten Homepages, wohingegen 31% der deutschen Homepages eine „Pre-Page“ verwenden. Die „funktionalen Komponenten“ in Tabelle 4.9 zeigen, dass Schütte unter einem Terminus verschiedene Komponenten subsumiert, für die Nielsen und Tahir separate Kategorien angeben. Die sehr hohe Frequenz von 84% für das Merkmal „Anzeigen für eigene Produkte“ bei

Komponente	*Deutsches Korpus* Anzahl	%	Prototyp	*Russisches Korpus* Anzahl	%	Prototyp
Hyperlinkfunktionalität						
interhypertextuelle Links	55	100	✓	48	96	✓
extrahypertextuelle Links	29	53	✓	33	66	✓
intrahypertextuelle Links	2	4		4	8	
Firmenname (Schriftzug, Logo)	55	100	✓	50	100	✓
Inhaltsübersicht	55	100	✓	46	92	✓
Abbildungen	53	96	✓	43	86	✓
Funktionale Komponenten (E-Mail, Sitemap, Suche etc.)	49	89	✓	41	82	✓
Kontaktsequenz (primär Begrüßung)	39	71	✓	11	22	
Metainformationen	31	56	✓	38	76	✓
Aktuelle Meldungen	28	51	✓	12	24	
Slogan	25	45		9	18	
Unternehmensvorstellende Kurztexte	13	24		22	44	
Adresse (Unternehmenssitz)	8	14		12	24	

Tabelle 4.9: Makrostrukturelle Komponenten von russischen und deutschen Unternehmenshomepages (nach Schütte, 2004a, S. 214, S. 223, S. 258, S. 262)

Des Weiteren gibt Schütte (2004a, S. 222) die Bestandteile einer prototypischen Unternehmenshomepage an, wobei angenommen wird, dass diejenigen Komponenten für den Prototypen konstitutiv sind, deren Frequenz mindestens 50% beträgt. Den makrostrukturellen Komponenten lassen sich dabei die primären kommunikativen Funktionen der Homepage zuordnen: Identifikation (Firmenname), Titelfunktion (Abbildung, Firmenname, Begrüßung), Kontaktherstellung (Begrüßung), Überblick über die Inhalte (Inhaltsübersicht) sowie Informationsankündigung und -verweisung (Inhaltsübersicht, aktuelle Meldungen; ebd., S. 223). Bezüglich der Makrostrukturen der russischen Homepages ermittelt Schütte (2004a, S. 257) „keine nennenswerten Abweichungen“, jedoch ist deren Textsortenprofil insgesamt „schwächer konturiert“ und enthält weniger klare Muster als die deutschen Homepages (ebd., S. 342). Zusammenfassend charakterisiert Schütte Unternehmenshomepages als den tatsächlichen Inhalten einer Website vorgeschaltete Übersichtsseite, die Einstiegs- und Orientierungsfunktion besitzt, wobei verschiedene Teilaufgaben existieren (ebd., S. 329 f.): Als obligatorisch werden die Titelfunktion (Identifikation des Emittenten und des Gegenstandes), die Übersicht über die Inhalte und die Kontaktfunktion aufgefasst. Fakultative Teilaufgaben sind die Vermittlung initialer Informationen zum Kommunikationsgegenstand (einschließlich Werbung) sowie die Kommunikation der *corporate identity*.

Die rhetorisch-funktionale Struktur einer kommerziellen Homepage

Askehave und Nielsen (2005) betrachten die Einstiegsseite des Webauftritts eines dänischen Unternehmens und verorten sie als Genre. Die Verfasser stützen sich auf Swales (1990), der davon ausgeht, dass ein Genre eine kommunikative Funktion besitzt, die durch funktional

Nielsen und Tahir legt nahe, dass es sich bei den untersuchten Homepages primär um „Leit-Inhalts-Seiten“ im Sinne von Schütte handelt. Diese sind jedoch im deutschen Korpus mit einer Frequenz von lediglich 11% und im russischen Korpus mit 30% verzeichnet, weil vornehmlich „corporate sites“ – im Gegensatz zu „brand sites“ – enthalten sind (Schütte, 2004a, S. 147).

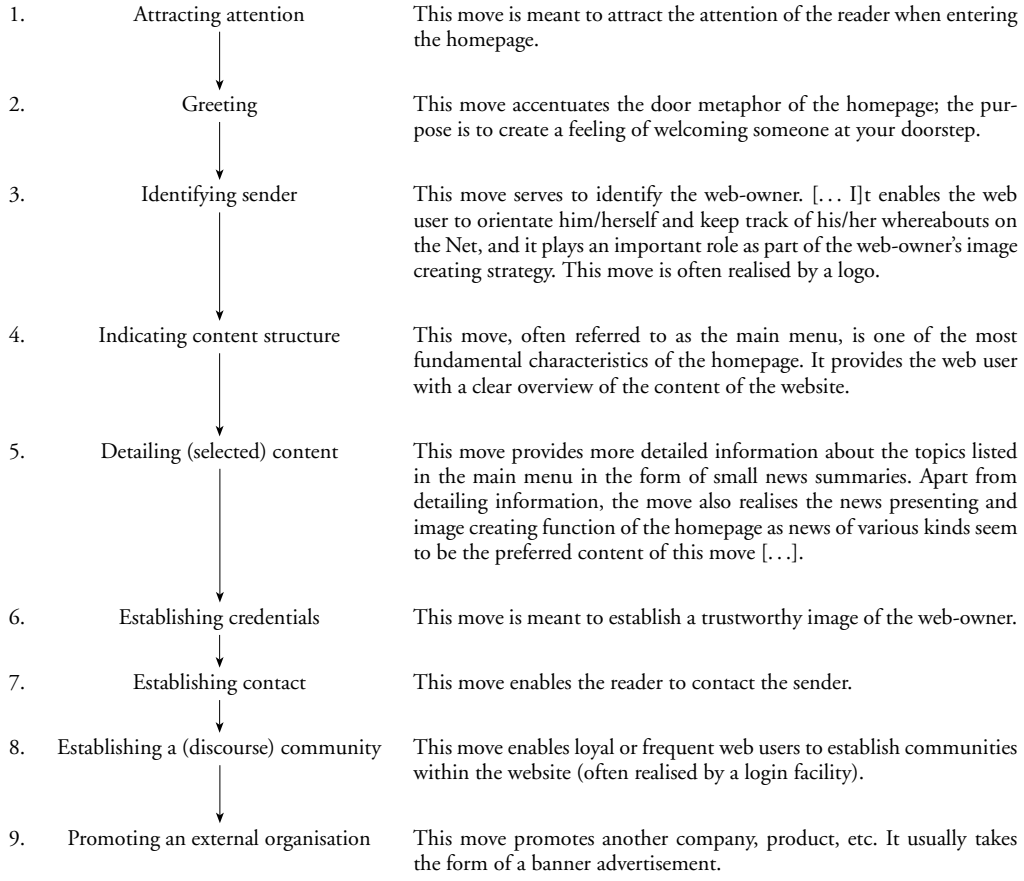


Abbildung 4.8: Die „move structure“ der Homepage des Webauftritts eines Unternehmens (nach Askehave und Nielsen, 2005, S. 5)

markierte Kommunikationsschritte (Swales spricht von „moves“) realisiert wird, die sich wiederum auf spezifische rhetorische Strategien sprachlicher und visueller Natur in einer konventionalisierten Textstruktur beziehen (vgl. Abschnitt 2.3.7).

Abbildung 4.8 stellt eine Analyse der „move structure“ dar, die der betrachteten Einstiegsseite zugrunde liegt. Zunächst soll die Aufmerksamkeit des Rezipienten erregt werden, woraufhin dieser in dem Webangebot begrüßt wird. Nach der Identifikation des Emittenten durch ein Logo oder einen Schriftzug werden die angebotenen Inhalte in Form einer strukturierten Navigationshilfe präsentiert, woraufhin spezifische Inhalte hervorgehoben werden, häufig handelt es sich hierbei Askehave und Nielsen zufolge um nationale oder internationale Meldungen, Pressemitteilungen oder neue Produkte. Anschließend ist der Emittent bestrebt, sich als vertrauenswürdiges Unternehmen zu positionieren, woraufhin Kontaktinformationen dargestellt werden. Viele kommerzielle Einstiegsseiten geben dem Rezipienten darüber hinaus die Möglichkeit, sich gegenüber der Website zu authentifizieren, um z. B. Diskussionsforen benutzen zu können. Abschließend enthalten viele derartige Einstiegsseiten Werbung für die Produkte anderer Unternehmen (vgl. auch Nielsen und Tahir, 2002).

Die Website einer Konferenz

Die Hypertextsorte Webauftritt einer Konferenz hat in den vergangenen Jahren etablierte Konventionen hervorgebracht. Im Gegensatz zur persönlichen Homepage, die nur fragmentarisch auf Textsorten basiert, die aus anderen Medien bekannt sind, beruht die Website einer Konferenz primär auf traditionellen Textsorten. Hierzu gehört unter anderem die generelle Ankündigung einer Konferenz, die in der Vergangenheit in Rundschreiben, Zeitschriften, Newsgroups und per E-Mail publiziert wurde, sowie der *Call for Papers*, der ebenfalls über den Titel der Konferenz, den Veranstalter und den Veranstaltungsort sowie das Datum informiert und eine Liste möglicher Themenschwerpunkte präsentiert und dazu einlädt, Beiträge einzureichen und möglicherweise auch Workshops zu veranstalten.

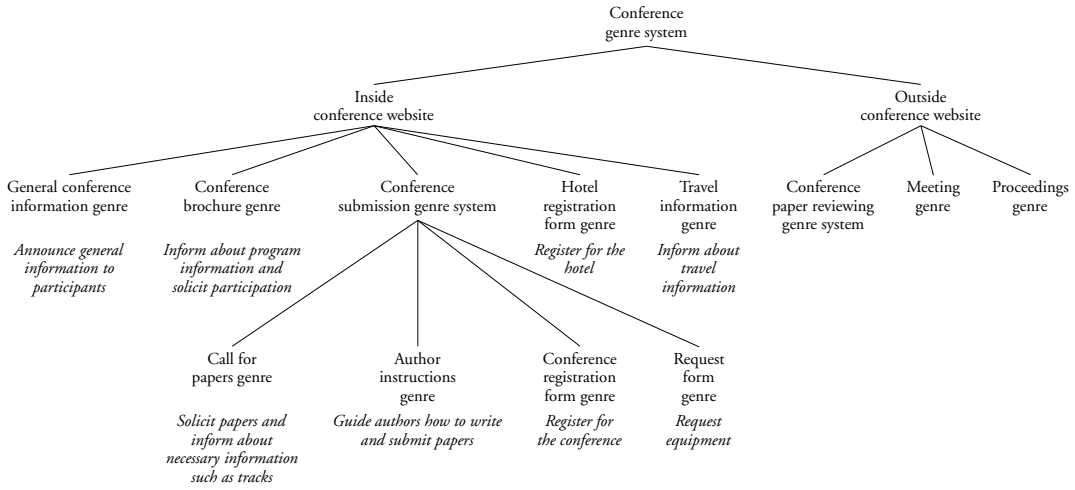


Abbildung 4.9: Die Genres und Genre-Systeme der Website einer Konferenz (nach Yoshioka und Herman, 2000, S. 4 f.)

Yoshioka und Herman (2000) analysieren diese Hypertextsorte am Beispiel der Website der 32. *Hawai'i International Conference on System Sciences* (`http://www.hicss.org`).[94] Abbildung 4.9 zeigt die von Yoshioka und Herman aufgrund ihrer kommunikativen Funktion (in der Grafik in kursiver Schrift dargestellt) ermittelten Genres, in deren Zentrum das „conference submission genre system" steht. Dieses umfasst vier Genres, die die kommunikativen Handlungen zur Einreichung von Beiträgen koordinieren. Potenzielle Autoren rezipieren zunächst den *Call for Papers* und entschließen sich, einen Artikel einzureichen. Ein Exemplar des Genres „author instructions" erläutert die thematische Ausrichtung der verschiedenen Tracks und Minitracks und dass eine Zusammenfassung des geplanten Beitrags bis zu einem gewissen Stichtag an den Leiter eines Minitracks zu senden ist. Daraufhin findet ein Begutachtungsprozess statt, über dessen Ergebnis der Autor von der Leitung des Minitracks infor-

[94] Yoshioka und Herman (2000) untersuchen die Genres aus Sicht der „Coordination Theory", die zur Modellierung von Geschäftsprozessen entwickelt wurde (vgl. Fußnote 71, S. 61). Mit den Websites von Konferenzen beschäftigen sich auch Mehler et al. (2004), die jedoch die maschinellen Erkennung der involvierten Bausteine anstreben (vgl. Abschnitt 14.6.3).

miert wird.[95] Daraufhin muss sich der Beiträger bis zu einem bestimmten Datum für die Konferenz registrieren („conference registration form genre“) und die benötigte technische Ausrüstung (Videoprojektor etc.) spezifizieren („request form genre“). Nach Yoshioka und Herman (2000, S. 5) handelt es sich um ein Genre-System, weil die erfolgreiche Durchführung einer mit einem bestimmten Genre assoziierten kommunikativen Handlung eine Voraussetzung für die Durchführung der kommunikativen Handlung des nachfolgenden Genres ist.[96] Yoshioka und Herman (2000, S. 5) geben an, dass die auf der Website der HICSS ermittelten Genres ein Bestandteil des „conference genre systems“ sind, das unter anderem das „meeting genre“, das „conference paper reviewing genre system“ und das „proceedings genre“ enthält (vgl. Abbildung 4.9). Nicht thematisiert wird der Aspekt, welche der gefundenen Genres tatsächlich als konventionalisierte Textsorten im engeren Sinne gelten können. Meines Erachtens sind zunächst mindestens drei unterschiedliche Hypertexttypen zu unterscheiden: Die Websites kleinerer Arbeitstreffen und Workshops bestehen in der Regel nur aus einer sehr kleinen Zahl meist schlicht gestalteter Dokumente, die über die thematische Ausrichtung, Ort und Zeit sowie die Veranstalter informieren. Größere, aber primär die nationale Ebene fokussierende Konferenzen enthalten weiterführende Informationen, z. B. über die Zusammensetzung des Programmkomitees. Sehr große internationale Konferenzen besitzen immer häufiger sehr umfangreiche Webpräsenzen, die zahlreiche Zusatzinformationen und -funktionen bieten, die in den meisten Fällen nur durch eine finanzielle Unterstützung aus der Privatwirtschaft ermöglicht werden.[97]

Die Homepage eines Bundesstaates

Ryan et al. (2003) beschäftigen sich mit den Dimensionen anhand derer unterschiedliche Webangebote von Rezipienten differenziert werden. Als Untersuchungsdomäne dienen die insgesamt etwa 300 Einstiegsseiten der Webauftritte der 50 US-Bundesstaaten aus den Jahren 1997 bis 2002.[98] Ryan et al. haben die im Korpus enthaltenen Dokumente ausgedruckt, um sie von insgesamt 180 Probanden (30 pro Jahr) nach beliebigen (aber begründbaren) Kriterien gruppieren zu lassen.[99] Die von den Probanden gebildeten Gruppen werden in Ma-

[95] Der Begutachtungsprozess ist nicht Teil dieses nur auf die Website bezogenen Genre-Systems. Einige Konferenz-Websites unterstützen den Prozess mit spezialisierten Datenbank-Systemen, in die potenzielle Beiträger digitale Versionen ihrer Papiere hochladen können, die anschließend online bewertet werden. Die Benachrichtigung erfolgt per E-Mail und befindet sich somit ebenfalls außerhalb des Genre-Systems.

[96] Dies widerspricht der Definition von „systems of genre“ (Bazerman, 1994, S. 99, vgl. auch Fußnote 64, S. 58). Die vier Genres bilden vielmehr ein „genre set“, weil von den auf der Website angegebenen Genres lediglich die kommunikativen Handlungen des potenziellen Beiträgers spezifiziert werden.

[97] Ein gutes Beispiel hierfür sind die Websites der Reihe „International World Wide Web Conference“, die seit einigen Jahren jeweils eigene Domains besitzen (`http://www2002.org`, `http://www2003.org` etc.) und das WWW dazu einsetzen, der aus den Teilnehmern bestehenden Diskursgemeinschaft die Möglichkeit zu geben, miteinander in Kontakt zu treten (vgl. z. B. `http://w3photo.org/photos/www2004/`).

[98] Die Daten wurden dem „Web Archive“ entnommen (`http://www.archive.org`, vgl. Abschnitt 7.5.3). Da einige Dokumente nicht verfügbar waren, sind lediglich 293 Homepages in die Studie eingeflossen.

[99] Diese Vorgehensweise entspricht der zweiten Phase der Analyse von Brandl (2002) und der ersten Studie von Rosso (2005). Während Ryan et al. (2003) die „US state government home page“ vorgeben, ermittelt Brandl Hypertexttypen anhand von Experteninterviews (vgl. Abschnitt 4.4.5) und die Versuchspersonen von Rosso sortieren 102 Dokumente einer Stichprobe. Die Probanden von Brandl erstellen sukzessive einen Sortierbaum, während die Teilnehmer von Ryan et al. keine hierarchische Untergliederung vornehmen (wie bei Rosso).

trizen überführt, die die Ähnlichkeit einzelner Dokumente repräsentieren und einer Cluster-Analyse unterzogen. Ryan et al. (2003, S. 414 ff.) ermitteln mehrere Dimensionen der Differenzierung: „navigation support" bezieht sich auf Dokumente, die primär die Navigation zwischen verknüpften Webseiten fokussieren, „layout" bezieht sich auf die Anordnung der Informationsobjekte in einem Dokument, und „information density" charakterisiert die Menge von Informationen in einem HTML-Dokument. Interessant ist die Gewichtung dieser Dimensionen innerhalb der fünf Jahrgänge: Während „navigation" für die Jahre 1997 und 1998 diejenige Dimension mit der breitesten Varianz war, wurde für die aus dem Jahr 1999 stammenden Dokumente die Dimension „layout" zur Sortierung eingesetzt, die 2000/2001 von „information density" abgelöst wurde. Ryan et al. (2003, S. 418) folgern hieraus, dass bezüglich der Einstiegsseiten der untersuchten Domäne ein Trend zu „high information density 'portal' pages" und „low information density 'entry' pages" auszumachen ist. Weiterhin ermitteln die Verfasser sechs grundlegende Typen von Homepages: Der „long list of text links type", der sich an Linklisten orientiert, konnte für das Jahr 2002 nicht mehr festgestellt werden (1997: 7, 1998: 11, 1999: 11, 2000: 6), er ist offensichtlich nicht mehr zeitgemäß. Der „simple rectangle type" bezieht sich auf Dokumente mit wenigen Grafiken, einigen inhaltlichen Kategorien und einem Layout, das für eine bestimmte Bildschirmauflösung und Größe des Browser-Fensters konzipiert wurde. Dieser Typ hatte in den Jahren 2000 und 2001 die meisten Mitglieder. Der „short 'l' type" hat ein ähnliches Äußeres wie das einfache Rechteck, besitzt jedoch eindeutig erkennbare Navigationsleisten am oberen und linken Rand, die sich in der oberen linken Ecke treffen. Für das Jahr 2002 konnte dieser Typ nicht mehr eindeutig festgestellt werden und wurde vom „high density/long 'l' type" abgelöst, der mehr Hyperlinks und Bilder beinhaltet und etwas umfangreicher ist. Erstmals im Jahr 2001 konnte der „portal type" nachgewiesen werden, der am oberen, linken und rechten Rand Navigationshilfen einsetzt; ein Jahr später hatte dieser Typ bereits 15 Mitglieder (vgl. Indikator 38, S. 136). Der „boxes type" versammelt unterschiedlich große, aber eindeutig visuell hervorgehobene rechteckige Bereiche mit Informationen und Hyperlinks und wurde nur im letzten Jahrgang gefunden. Nach Ansicht von Ryan et al. (2003, S. 424) ist nicht abzusehen, ob sich diese neue Variation langfristig durchsetzen wird oder ob es sich um einen experimentellen Typ handelt.[100] Neben diesen dominanten Typen konnten weitere, weniger umfangreiche Cluster ermittelt werden, z. B. „big buttons", „scattered buttons" und „high color" (ebd., S. 425), die in den Folgejahren jedoch nicht wieder auftraten:

> [A]ny communication medium will have experiments of new types and usage patterns. Some will fail and some will succeed, to be repeated in the future. The home page is evolving and changing. We saw early experimentation, but now see more agreement on the usage of only a few types. (Ryan et al., 2003, S. 426)

Obwohl sich Ryan et al. ausschließlich auf das Webdesign beziehen, gelingt ihnen der Nachweis unterschiedlicher evolutionärer Entwicklungsformen von Homepages (vgl. Abbildung 4.4, S. 171). In ihren Schlussfolgerungen gehen Ryan et al. (2003, S. 428) auf Arbeiten

[100] Ryan et al. gehen nicht auf die Herstellung der Dokumente ein (vgl. Abschnitt 3.3.6). Vermutlich ist der erstmals für das Jahr 2002 festgestellte „boxes type" auf CMS-Software zurückzuführen. Derartige Systeme sind umfangreich konfigurierbar, die korrespondierenden Inhaltsbestandteile (Kalender, Linklisten, Suchfunktionen etc.) werden typischerweise als rechteckige Bereiche dargestellt.

zu Web-Genres ein: „[G]enre theory may not be the most appropriate lens for viewing home page evolution. While some researchers consider the home page to be a genre [...], many would not count varieties of home pages as different genres." Es wird nicht pauschal gegen Genre-basierte Ansätze argumentiert, lediglich zur Darstellung unterschiedlicher Entwicklungsstufen von Homepages seien sie nicht geeignet. Ein zweiter von Ryan et al. thematisierter Aspekt betrifft die in der *North American Genre Theory* vernachlässigte Frage nach der Granularität einer Typologie von Textklassen (vgl. Abschnitt 2.3.7): Ist diese grob, können Homepages mit ihrer Funktion als Einstieg in eine Website konzeptualisiert werden (Auffassung der Homepage als einer von vielen Hypertextknotentypen), ist sie jedoch fein, können unterschiedliche Varianten von Homepages als Hypertextknotensorten oder Varianten von Hypertextknotensorten betrachtet werden, die sich mehrere Eigenschaften teilen (etwa die in diesem Abschnitt genannten) und bezüglich anderer Merkmale differieren. Alternativ kann z. B. die kommerzielle Homepage auf der Ebene des Hypertextknotentyps angesetzt werden, um, etwa auf der Ebene der Branche (Ho, 1997), auf der nachfolgenden Ebene unterschiedliche Hypertextknotensorten und daraufhin, etwa basierend auf verschiedenen Sprach- oder Kulturkreisen (vgl. Fußnote 63, S. 190), unterschiedliche Varianten dieser Hypertextknotensorten zu differenzieren (vgl. die Kapitel 11 bis 13). Eine systematische Untersuchung von Web-Genres erfordert nach Ryan et al. (2003, S. 428) die Konzeptionierung von Methoden „to clarify the purposes, content, form and functionality of home pages, as well the connection between home pages and the sites they introduce."

4.6.3 Die persönliche Homepage

Die persönliche Homepage besitzt kein unmittelbares Pendant unter den Textsorten, deren Textexemplare üblicherweise auf Papier gedruckt werden (vgl. Dillon und Gushrowski, 2000), was den Umstand erklärt, dass ihre zahlreichen Facetten Gegenstand einer Vielzahl von Publikationen sind (vgl. Döring, 2001a, 2002, für einen umfassenden Überblick).[101] In persönlichen Homepages stellen sich Personen als Individuen und Netzbenutzer vor, wobei zwei grundlegende Typen zu unterscheiden sind: Private Homepages werden von Privatpersonen veröffentlicht und enthalten in der Regel Informationen über ihre Hobbys und Interessen (sie dienen vornehmlich der Selbstdarstellung). Berufliche Homepages (z. B. von Hochschullehrern) stellen dagegen die Rolle einer Person in ihrem beruflichen Kontext vor.[102]

[101] Der Begriff „persönliche Homepage" ist die geläufige und – im Gegensatz zum „Personalcomputer" (*personal computer*, PC) – korrekte Übersetzung des englischen Ausdrucks *personal home page*. Der Terminus bezeichnet nicht nur die Einstiegsseite, sondern das gesamte Webangebot, so dass Döring (2001a, S. 328) zufolge eigentlich von der „Homesite" die Rede sein müsste. Dass die „home"-Metapher, wie Bates und Lu (1997, S. 334) vermuten, aus dem Bereich des Bergsteigens („home base") oder des Baseballs („home plate") stammt, muss bezweifelt werden, schließlich wurde diese Metapher bereits in dem von (dem Briten) Tim Berners-Lee entwickelten ersten grafischen Browser namens „WorldWideWeb" verwendet.

[102] Haas und Grams (1998b, S. 103) bezeichnen berufliche Homepages als „professional home pages", die das „professional life" einer Person thematisieren. Obwohl eine Differenzierung zwischen privaten und beruflichen Homepages offensichtlich erscheint, werden diese beiden Kategorien häufig unter dem generischen Etikett der „home page" subsumiert, z. B. von Crowston und Williams (2000, S. 208), die auch institutionelle Einstiegsseiten einbeziehen: „We defined a home page as personal or organizational information plus links to other pages reflecting the subject's interests that are intended to introduce the person or organization to the world and to facilitate further contact." (vgl. Abschnitt 4.6.1).

Die berufliche Homepage

Nach Storrer (1999b, S. 6) dienen berufliche Homepages dazu, „Mitarbeiter von Institutionen und Firmen in den zugehörigen Sites vorzustellen, Funktionen, Kompetenzen und Zuständigkeiten abzustecken und die betreffenden Personen in institutionellen und betrieblichen Hierarchien einzuordnen."[103] Ihre Gestaltung ist dabei Storrer zufolge in vielen Fällen weniger eine Manifestation des persönlichen Ausdrucks, sondern beruht – zur Erzeugung von Kohärenz und optischer Konsistenz – auf dem Design der einbettenden Website als übergeordnetem Ganzen. Zu diesem Zweck werden häufig vorgefertigte Schablonen zur Verfügung gestellt, die von der darzustellenden Person mit den persönlichen Informationen bestückt werden (vgl. auch Döring, 2001a, S. 327). Die Anfertigung der Homepage kann jedoch auch im Auftrag der darzustellenden Person erfolgen, z. B. von einer Hilfs- oder Schreibkraft oder einer Abteilung für Öffentlichkeitsarbeit (vgl. Abschnitt 3.5.3). Crowston und Williams (1997) finden z. B. in der von ihnen untersuchten Stichprobe mehrere persönliche Homepages (vgl. Abschnitt 4.4.1), die nicht von der dargestellten Person, sondern von der Organisation bzw. Institution veröffentlicht wurden, für die der vermeintliche Autor tätig ist. Derartige Seiten erinnern Crowston und Williams (1997, S. 37) zufolge an das „faculty profiles book" US-amerikanischer Universitäten. Falls die von der Schablone vorgegebenen Möglichkeiten nicht ausreichend sein sollten, kann eine Verknüpfung zur privaten Homepage integriert werden, auf der weiterführende Informationen zu finden sind.[104] Storrer (1999b, S. 6) gibt zwei Beispiele an: Auf der Homepage einer Professorin befinden sich in der Regel Angaben über Forschungsschwerpunkte, Projekte, Publikationen, Vorträge, Lehrveranstaltungen und Sprechstunden. Die Homepage eines Verwaltungsangestellten beinhaltet seinen Zuständigkeitsbereich und Sprechzeiten.

Die private Homepage

In den meisten Arbeiten zu persönlichen Homepages wird die private Homepage fokussiert, die von Haas und Grams (1998b, S. 103) als „true "personal" page" bezeichnet wird, weil sie Informationen enthalte, die nur für jemanden von Interesse seien, der den Autor kenne oder kennen lernen möchte. Storrer (1999b, S. 5) bemerkt in privaten Homepages ein „Bemühen um Authentizität und Spontaneität, teilweise zu Lasten von Rechtschreibung, Syntax und Textplanung" verbunden mit einer Hinwendung zum Rezipienten, die durch die „Orientierung an der mündlichen Sprache und an dialogischen Kommunikationsformen" deutlich wird und listet verschiedene Merkmale privater Homepages auf: Ihre Inhalte sind meist selbstgemacht (z. B. Fotos, Artikel, Bilder, Gedichte, Geschichten, Kompositionen, Software etc., vgl. auch Erickson, 1996, S. 15), sie besitzen einen authentischen Ausdruck (vgl. Bittner,

[103] Storrer (1999b, S. 6) bezeichnet berufliche Homepages als „persönliche Homepages". Döring (2001a, S. 327) weist darauf hin, dass dies nicht dem üblichen Sprachgebrauch entspricht.

[104] Siehe hierzu auch Sandbothe (1997, S. 68): „Unsere Netzpersönlichkeit setzt sich aus einem Geflecht unterschiedlicher Rollen, Identitäten und Funktionen zusammen, die wir streng voneinander isolieren oder aber bewußt miteinander vernetzen können. Das World Wide Web ermöglicht uns auf diese Weise eine artistische Ausgestaltung und ästhetische Dramatisierung derjenigen pluralen Identitätsnetze, die zwar auch unser »reales Leben« bestimmen, aber IRL [in real life, G. R.] nicht immer ausagiert werden können." Bahl (1997) liefert diesbezüglich für die IRC- und MUD-Kommunikation verschiedene Fallbeispiele.

2003, S. 116, sowie Machilek et al., 2004) und eine individuelle Gestaltung, wodurch ein informelles und persönliches Ambiente verbreitet und Offenheit sowie Gesprächsbereitschaft gegenüber den dargestellten Inhalten signalisiert werden. Außerdem bieten sie, ebenso wie berufliche Homepages, mehrere Möglichkeiten der Kontaktaufnahme an.

Zur Funktion der persönlichen Homepage

Berufliche und private Homepages können zwar häufig in Reinform beobachtet werden, ebenso oft sind jedoch Mischungen zu verzeichnen. Erickson (1996) liefert eine prägnante Charakterisierung der wesentlichen Funktion dieses Hypertexttyps:

> I believe this seemingly frivolous blending of the professional and the personal is the key to why the Web is becoming a fundamentally different creature from the systems of information servers that preceeded it. Personal pages and the Web are not being used to "publish information"; they are being used to construct identity – useful information is just a side effect. A personal page is a carefully constructed portrayal of a person. (Erickson, 1996, S. 15)

Aus diesem Grund bezeichnet Erickson das WWW auch als „sozialen Hypertext", der unter anderem Knoten umfasst, die nicht etwa spezifische Themen diskutieren, sondern „representations of people" sind (ebd.).[105] Die Ordnung dieser Knoten basiert Erickson (1996, S. 16) zufolge auf einer „sozialen Logik", die es den Rezipienten ermöglicht, sich einen Überblick über das Netzwerk von Freunden, Arbeitskollegen und persönlichen Interessen zu verschaffen, indem sie sich durch „social navigation" (ebd.) von Individuum zu Individuum hangeln (vgl. auch Thelwall, 2003). Der Wille, eine explizite Darstellung des Selbst konstruieren zu wollen, ist nach Erickson ein zentraler menschlicher Wesenszug und ein wesentlicher Faktor der permanenten Erweiterung des WWW. Die Aufgabe der persönlichen Homepage, die eigene Identität im *World Wide Web* zu (re)kreieren, wird in zahlreichen Arbeiten thematisiert (z. B. von Furuta und Marshall, 1996, Chandler, 1998, und Döring, 2001a).

Miller (1995) nähert sich persönlichen Homepages ebenfalls aus soziologischer Perspektive und vermutet: „[P]eople feel a desire to establish their selves on the Web."[106] Die explizite Verschriftlichung derartiger Selbstdarstellungen ist jedoch oft mit Problemen verbunden, weil viele Autoren von persönlichen Homepages keine Erfahrung damit haben: „For most people, though, it is difficult to establish yourself as a whole person through a self-description: it feels like an extended lonely-hearts advert." (Miller, 1995, vgl. auch Eckkrammer, 2001, S. 63, sowie Bittner, 2003, S. 87). Daher weichen viele Produzenten auf eine alternative Strategie aus, die Miller sehr treffend als „show me what your links are, and I'll tell you what kind of person you are" umschreibt (Chandler und Roberts-Young, 1999, liefern einige Beispiele); Roberts (1998, S. 80) bezeichnet Hyperlinks in diesem Kontext als „a type of orientation clause that helps to define the essence of the author." Die gerade in den Anfangsjahren

[105] Erickson (1996, S. 15) geht davon aus, dass persönliche Homepages nützliche Informationen enthalten. Die Studie von Roussinov et al. (2001) zeigt jedoch, dass sie in den meisten Fällen zur Beantwortung von Suchanfragen als irrelevant eingestuft werden (vgl. Fußnote 40, S. 177). Groth (1998) weist – ebenfalls anhand einer Benutzerstudie – nach, dass sich Wissenschaftler häufig auf den persönlichen Homepages anderer Wissenschaftler über eine Person informieren (besonders in Bezug auf Kontaktinformationen und Publikationen).

[106] Roberts (1998, S. 79) betrachtet persönliche Homepages als „billboards that say, "Look at me!" [A personal home page] claims a piece of territory in cyberspace."

des WWW auf praktisch allen persönlichen Homepages allgegenwärtigen Listen von Hyperlinks hatten bzw. haben also nicht nur den Zweck, den Rezipienten auf andere interessante Websites hinzuweisen, sondern auch, die anbietende Person selbst mittelbar als Individuum mit spezifischen Vorlieben und Abneigungen zu charakterisieren. Nach Ansicht von Bittner (2003, S. 101) stellen Produzenten mit Linklisten ihre „Netzweltläufigkeit" unter Beweis. Die Individualisierung einer persönlichen Homepage erfolgt auch durch ihr Text- bzw. Webdesign, das z. B. spezielle horizontale Trennlinien, besonders ästhetische oder humorvolle (animierte) Icons, Hintergrundgrafiken und verspielte Schriftarten umfasst (vgl. Roberts, 1998, S. 79, und Walker, 2000, S. 110).[107] Furuta und Marshall (1996) weisen darauf hin, dass die persönliche Homepage auch eingesetzt wird, um die Zugehörigkeit zu einer oder mehreren Gruppen zu signalisieren. So legen z. B. Wissenschaftler auf ihren Homepages häufig Hyperlinks zu ihren Arbeitsgruppen, Instituten, Fachbereichen und Universitäten und auch zu den Websites von Fachverbänden, denen sie angehören. Die aktive Zugehörigkeit kann ebenfalls explizit markiert werden, z. B. durch das Anbieten von Veröffentlichungen, die den Autor innerhalb einer Forschungs-Community als aktives und beitragendes Mitglied positionieren.

Funktionsorientierte Typologien persönlicher Homepages

Miller (1995) stellt eine vorläufige Klassifikation persönlicher Homepages auf, die sich auf soziale Rollen und Eigenschaften bezieht: (1) Webseiten der Kategorie „Hi, this is me (as an individual)" besitzen einzig und allein selbstdarstellende Funktion (vgl. Schütz et al., 2003) und werden Miller zufolge vornehmlich von männlichen Studierenden veröffentlicht. Sie enthalten Informationen über die Herkunft einer Person, ihren Studiengang, ihre Hobbys, ein Foto und Links zu den Homepages von Freunden. Eine Variante betrifft Webseiten, auf denen umfangreiche Informationen zu einem spezifischen Interesse des Autors angeboten werden.[108] (2) Als Beispiel für Webseiten des Typs „Hi, this is me (as a member of an organisation)" gibt Miller die Homepages von Wissenschaftlern an. Die zugehörigen Autoren brechen Miller zufolge bewusst Konventionen, falls zur Erstellung vorgefertigte Schablonen einzusetzen sind, um ihren Homepages eine persönliche Note zu geben. Diese beiden ersten Kategorien entsprechen der Unterscheidung privater und beruflicher Homepages. (3) Den

[107] Furuta und Marshall (1996) vermuten, dass die Popularität des WWW insbesondere mit der einfachen Erlernbarkeit von HTML zusammenhängt, die es Autoren erleichtert, etablierte Konventionen zu übernehmen: „Like school uniforms worn to promote student equality, Web capabilities help make every author's pages look quite similar – straightforward use of [HTML] produces like-appearing documents. And like barrettes in the hair of uniform-wearing students, relatively minor differences serve to personalize the standard, and indeed are embraced as a means of expressing individuality." (vgl. auch den Titel der Arbeit von Amitay, 1997, die private Homepages untersucht: *Hypertext – The Importance of being Different*). Nach Ansicht von Chandler und Roberts-Young (1999) sind die privaten Homepages von Jugendlichen mit ihren Zimmern vergleichbar, die ebenfalls Poster ihrer Lieblingsstars aus Sport und Musik, Fotos von Freunden und Familienangehörigen und diverse Andenken enthalten, um ihnen eine persönliche Note zu verleihen.

[108] Haas und Grams (1998b, S. 103) bezeichnen diesen Typ, der z. B. Hintergrundinformationen zum Lieblingsschauspieler des Autors enthält, als „avocational home page", die einen „service to the public" darstellen (vgl. Walters, 1996). Döring (2001a, S. 340) fasst den Typ als „instrumentelle Homepage" auf, der von der „expressiven Homepage" differenziert wird. Eine Analyse von 279 zufällig ausgewählten studentischen Homepages zeigt, dass 42% als expressiv und 18% als instrumentell aufzufassen sind. Neben diesen „realisierten Homepages" sind 29% „projektierte Homepages" zu verzeichnen (sie bestehen nur aus einer Vorankündigung). Parallel zu diesen „faktischen Homepages" ermittelt Döring 11% „nominelle Homepages" (nicht abrufbar).

dritten Typ nennt Miller „Hi, this is us“. Hiermit sind Homepages gemeint, die Familien vorstellen, d. h. es es geht nicht primär um die Darstellung eines Individuums, sondern vielmehr um die Geschichte und gemeinsamen Interessen des dargestellten Paares, die „corporate identity of the family“ (Miller, 1995).[109] (4) Die vierte Kategorie, „This is what I think is cool“, wurde bereits angesprochen (vgl. Fußnote 108). Sie enthält kaum persönliche Informationen, sondern vielmehr Links zu Themen, Produkten, Bands oder sonstigen Websites, die dem Autor gefallen oder gerade nicht gefallen. (5) Die Kategorie „An advertisement for myself“ umfasst drei Gruppen: (5 a) Webseiten vom Typ „Cool style“ zeichnen sich durch eine technisch fundierte und visuell beeindruckende Aufbereitung aus, sie sollen Medienkompetenz signalisieren. (5 b) Persönliche Homepages der Kategorie „The electronic curriculum vitae“ interpretiert Miller als ehrlichen und sehr direkten Versuch, eine Anstellung zu finden, indem die eigenen Qualifikationen präsentiert werden. (5 c) Der Typ „An advertisement for the service I can provide“ betrifft Webangebote von Personen, die eine Dienstleistung anbieten, wobei Miller Überlappungen mit den Typen (5 a) und (5 b) feststellt.[110]

Walker (2000) beschäftigt sich mit der Erzeugung von Identität in persönlichen Homepages (vgl. Makhfi, 2002). Die Studie basiert auf der Analyse mehrerer hundert Homepages; zusätzlich wurden Fragebögen an die Autoren einhundert weiterer Homepages verschickt, aus denen Walker (2000, S. 101) schließt, dass ein Großteil der Homepages von Studierenden und in der Computerbranche Tätigen angeboten wird (Buten, 1996). Es werden drei Typen ermittelt, die sich zum Teil mit den Kategorien von Miller (1995) decken: (1) Persönliche Homepages der ersten Kategorie umfassen nur rudimentäre Informationen und enthalten den Namen des Autors, Alter und Wohnort, den ausgeübten Beruf und möglicherweise Angaben zu Hobbys. Walker (2000, S. 102) charakterisiert diesen Typ wie folgt: „There is a sameness to these announcements; their format is “This is my name, this is where I live, this is what I do.” [...] Some pages never move past this initial formulation.“ Die konstatierte Gleichförmigkeit deutet an, dass es sich um eine etablierte, jedoch nicht spezifisch etikettierte Hypertextsorte handelt. Zur Erzeugung von Identität und Individualität werden insbesondere relationale Kategorien eingesetzt, die einen Produzenten z. B. als begeisterten Kinogänger oder Liebhaber klassischer Musik zu erkennen geben.[111] (2) Der zweite Typ ist als eine Erweiterung aufzufassen, in der das Leben des Produzenten als Fließtext geschildert wird. Walker (2000, S. 103) stellt in diesem Typ eine große Bandbreite fest: „There is a wide variety in how revealing autobiographical narratives on home pages are. Some tell small stories, while

[109] Schick (2003, S. 2 f.) geht ausführlich auf Familien-Homepages ein, wobei es jedoch schwierig ist, die Grenze zu persönlichen Homepages zu ziehen, da auch letztere in vielen Fällen Inhalte umfassen, die sich auf die eigene Familie beziehen. Außerdem werden viele Familien-Homepages, die die Mitglieder einer Familie einzeln vorstellen, tatsächlich nur von einer einzelnen Person angefertigt (vgl. Chandler und Roberts-Young, 1999). Döring (2001a, S. 326) subsumiert die Webangebote informeller Kleingruppen (Freundes- oder Liebespaar, Familie, Clique, Wohngemeinschaft) unter der Bezeichnung der „kollektiven persönlichen Homepage“.

[110] Bates und Lu (1997, S. 333 f.) schätzen die Funktionen einer Stichprobe von 114 persönlichen Homepages wie folgt ein: „Present own professional capability and experience“ (44,7%), „Play with system capability“ (28,9%), „Announce products or upcoming performance“ (20,2%), „Political campaign“ (2,6%), „Looking for dates“ (2,6%), „Miscellaneous and unknown“ (4,4%).

[111] Den Einsatz relationaler Kategorien betrachtet Walker (2000, S. 104) als logische Konsequenz, denn „the technology behind the [WWW], through this system of hyperlinks, invites people to represent themselves through relational categories.“ (vgl. auch Fußnote 108 sowie Abschnitt 4.6.6, S. 243 ff.).

others give entire life histories." (3) Homepages des dritten Typs konzentrieren sich auf ein bestimmtes Interesse des Autors, biografische Informationen werden nur in rudimentärer Form angegeben. Über die genannten Typen hinaus sind auch Mischformen möglich, deren Bestandteile verschiedene Aspekte der Persönlichkeit bzw. unterschiedliche Rollen reflektieren (ebd., S. 105). Alle drei Typen besitzen jedoch die Gemeinsamkeit, dass sie – bewusst oder unbewusst – Identität offenbaren. Im Hinblick auf die von den Produzenten verfolgte Intention unterscheidet Walker hinsichtlich der Zielgruppe zwischen intrinsischen und extrinsischen Homepages: Die Autoren intrinsischer Homepages richten sich an eine unbekannte Leserschaft, was durch den Inhalt der Dokumente reflektiert wird, die meist generelle und einleitende Angaben zur Identität umfassen (Walker, 2000, S. 108, vgl. auch de Saint-Georges, 1998, S. 71). Bei den Produzenten liegt der Wunsch vor, mit anderen Menschen in Kontakt zu treten, Freundschaften zu schließen und die eigene Meinung zu meist kontroversen Themen zu äußern. Extrinsische Homepages richten sich hingegen an Leser, die den Autoren bekannt sind. Sie dienen der Aufrechterhaltung bestehender Freund- und Bekanntschaften durch die regelmäßige Aktualisierung mit Neuigkeiten aus dem Leben des Autors (Walker, 2000, S. 107, spricht vom „news flash", vgl. auch Chandler und Roberts-Young, 1999). Weiterhin werden Homepages auch mit dem Zweck erstellt, als zentrale Sammelstelle für Hyperlinks zu dienen, auf die der Autor häufig zugreift (vgl. Groth, 1998).[112] Die Produzenten extrinsischer Homepages gehen davon aus, dass Rezipienten ihr Webangebot *nicht* als „identity statement" interpretieren (weil es nicht als solches intendiert ist) und dass die Fremdwahrnehmung ihrer Freunde und Bekannte nicht durch ihre Homepage beeinflusst wird. Walker (2000, S. 111) merkt jedoch zu Recht an: „[L]ack of intent does not translate into lack of identity statement."[113] Trotz der unterschiedlichen Intentionen gelangt Walker (2000, S. 110) zu der Schlussfolgerung, dass sich intrinsische und extrinsische Homepages bezüglich ihrer Gestaltung sehr ähneln. Beide Typen enthalten Zugriffszähler und E-Mail-Adressen, jedoch werden Walker zufolge nur in intrinsischen Homepages Gästebücher eingesetzt (vgl. Abschnitt 4.6.8, S. 247 ff.). Der Einfluss des WWW auf die unterschiedlichen Typen von Homepages und Identitätsangaben ist also nur sehr gering. Das Medium ist zwar an ihrer Ausprägung beteiligt, es determiniert sie jedoch nicht.[114]

[112] In diesen Fällen ist damit zu rechnen, dass der Produzent die Homepage als Startseite seines Browsers konfiguriert hat, d. h. einige persönliche Homepages übernehmen in der Tat die ursprünglich von Berners-Lee et al. (1992) intendierte Funktion als persönliche Einstiegsseite ins *World Wide Web*, die permanent gepflegt und um interessante Hyperlinks ergänzt wird.

[113] Bates und Lu (1997, S. 335) sind diesbezüglich anderer Ansicht: „Making a home page available to the virtual community can both reveal and *hide* the person." (Hervorhebung hinzugefügt, G. R.).

[114] Buten (1996) hat 121 Antworten auf einen an 316 Betreiber von Websites verschickten Fragebogen erhalten und vermutet, dass in den USA ca. 600 000 aktiv gepflegte persönliche Homepages existieren, von denen ca. 75% von Studierenden angefertigt wurden (Döring, 2001b, S. 344, schätzt, dass etwa 10% aller WWW-Benutzer eine persönliche Homepage besitzen). Als Gründe für ihre Veröffentlichung werden die Selbstdarstellung (49%), das Erlernen von HTML (48%, vgl. Berker, 2001, S. 222), das Bereitstellen von Informationen für Freunde (43%) und Netzbekanntschaften (34%) sowie das Publizieren von Linklisten genannt (32%). Buten zufolge ist den Autoren bewusst, dass ihre Website von einem breit gefächerten Publikum rezipiert werden kann. Die Dokumente werden sowohl für ihre Familien, Freunde und Bekannte als auch für Rezipienten angefertigt, die zufällig auf ihre Seite stoßen. Dass die Homepage eine akkurate Darstellung der eigenen Person ist, geben 91% der Befragten an. Nur 78% gehen davon aus, dass auch andere Autoren authentische Darstellungen publizieren. Walker (2000) kommt bezüglich der Intentionen zu übereinstimmenden Ergebnissen.

Die persönliche Homepage aus der Sicht von Usability-Ratgebern

Webdesign- und Usability-Ratgeber geben verschiedene Hinweise zur Gestaltung persönlicher Homepages. Nielsen (1999, S. 66) empfiehlt bezüglich der Verknüpfung von Dokumenten, den Namen einer Person nur dann als Linkanzeiger zu verwenden, wenn der Hyperlink zu einer Seite führt, die eine Biografie der Person enthält.[115] Ebenfalls sollte unmittelbar ein „traditional portrait photo" integriert werden, das zur Gewährleistung kurzer Downloadzeiten kleinformatig sein sollte. Weitere Fotos, die die Person in unterschiedlichen Umgebungen zeigen, sollten auf zusätzliche Webseiten ausgelagert werden. Die persönliche Homepage sollte kurz auf den Hintergrund einer Person eingehen und gegebenenfalls Links zu weiterführenden Informationen anbieten. Wünschenswert wäre nach Nielsen neben der traditionellen Publikationsliste auch eine Liste aller Webseiten, die von der dargestellten Person auf der zugehörigen Website erstellt wurden oder gepflegt werden. Abschließend sollten diejenigen Kontaktinformationen aufgeführt werden, die die Person bereit ist, einer breiten Öffentlichkeit zur Verfügung zu stellen; mindestens sollte dies die als `mailto:`-Link realisierte E-Mail-Adresse sein. Die nachfolgend thematisierten Studien zeigen, dass Nielsens Forderungen von realen persönlichen Homepages nur partiell reflektiert werden.

Empirische Analysen von persönlichen Homepages

Nachfolgend werden sieben empirische Analysen privater Homepages vorgestellt, die auf Stichproben im Umfang von 25 bis 114 Dokumenten basieren. In fünf dieser Studien wird der Standpunkt vertreten, dass es sich bei der privaten Homepage um ein Web-Genre bzw. eine Textsorte handelt, die kein Pendant in den traditionellen Medien besitzt.

Walters (1996) analysiert eine Stichprobe von 100 persönlichen Homepages mit dem Ziel, eines oder mehrere Genres zu ermitteln. Die Verfasserin orientiert sich an den Arbeiten von Yates und Orlikowski (1992), weshalb besonderes Augenmerk auf die jeweiligen Funktionen und Formen der Dokumente gelegt wird. Walters verwendet ein Klassifikationsschema mit mehr als 70 Merkmalen wie z. B. „identifying information", „purpose", „overall form", „graphics", „other form features" und „language".[116] Walters betont, dass die Zuweisung derartiger Kategorien hochgradig subjektiver Natur ist (insbesondere in Bezug auf den Zweck eines Dokuments), weshalb die Auswertung mit einer bestmöglichen Konsistenz vorgenommen wurde: Fast alle Homepages (96%) enthalten persönliche Informationen, 6% beschäftigen sich mit einem bestimmten Interessenbereich, 7% enthalten berufliche Informationen. In 80% der Dokumente wird dem Leser die Möglichkeit gegeben, eine Rückmeldung an den Autor zu senden. Umgangssprache wird in 91% der Dokumente gefunden. Die Varianz der analysierten Dokumente ist sehr groß. Walters beobachtet sowohl spartanisch anmutende als auch sehr aufwändig und mühevoll gestaltete Homepages mit animierten Grafiken und Hintergrundbilder. Bezüglich des Ziels der Ermittlung von Genres gelangt Walters zu folgendem

[115] Zur Begründung: „I recommend against making a person's name into a link to email that person. Doing so violates expectations on the Web because a link normally takes you to information *about* the thing you clicked to rather than making you communicate *with* the thing. Also, it is jarring to click on a normal hypertext link and be transferred into an email application." (Nielsen, 1999, S. 66).

[116] Walters (1996) weist, wie einige andere Autoren vergleichbarer Studien, explizit darauf hin, dass sich die Arbeit mit dem Klassifikationsschema als sehr komplex, schwierig und aufwändig herausgestellt hat.

Schluss: „I found that in practice, few homepages actually have a specific purpose. [... Many] are just floundering and trying to figure out what to put on their pages." (vgl. Berker, 2001, S. 218 f.). Aufgrund dessen vermutet die Verfasserin, dass es möglicherweise verfrüht sei, eine entsprechende Untersuchung von persönlichen Homepages vorzunehmen, da sich dieses Genre erst zu einem späteren Zeitpunkt entwickle. Walters schätzt diese Möglichkeit jedoch kritisch ein: „I tend to believe that they may never do so."

Groth (1998) hat im Sommer 1996 Interviews mit acht an einer Universität und sechs in einem Forschungsinstitut angestellten Wissenschaftlern sowie mit acht in einem Telekommunikationsunternehmen beschäftigten Ingenieuren durchgeführt. Bis auf eine Ausnahme geben alle Autoren auf ihren Homepages verschiedene Kontaktmöglichkeiten an, einige nennen auch die Privatadresse und ihre Telefonnummer. Angaben zu aktuellen Forschungs- oder Entwicklungsprojekten werden von 18 Produzenten veröffentlicht, 15 Personen geben rudimentäre Informationen zu ihren beruflichen Interessengebieten. Weitere hochfrequente Elemente sind Fotos, Publikationslisten, digitale Dokumente und private Interessen. Angaben zum persönlichen Hintergrund (z. B. eine Biografie) machen nur sechs Personen. Hyperlinks zu anderen Websites werden in fast allen Homepages angeboten.

Dillon und Gushrowski (2000, S. 203) untersuchen 100 persönliche Homepages und ermitteln zahlreiche Gemeinsamkeiten, die den Verfassern zufolge in ihrer Gesamtheit als „digital genre" aufgefasst werden können. Tabelle 4.10 zeigt die Liste der Bestandteile; die fünfte Spalte enthält die von Dillon und Gushrowski ermittelten Frequenzen.[117] Anhand fünf häufiger („E-mail address", „External links", „Welcome message", „One to four graphics", „Brief bio") und fünf seltener Elemente („Table of contents", „Text-only option", „Guestbook", „Frames", „Ten or more graphics") wurden insgesamt acht neue persönliche Homepages hergestellt, die jeweils die gleiche Anzahl Elemente besitzen, bezüglich ihrer beobachteten Frequenz jedoch auf einem Kontinuum zwischen „typisch" und „untypisch" liegen. Diese Dokumente wurden ausgedruckt und 57 Studierenden vorgelegt, die sie bezüglich ihrer „preference as a personal home page design" (Dillon und Gushrowski, 2000, S. 204) sortieren sollten. Die Verfasser ermitteln eine Übereinstimmung zwischen den Bewertungen der Dokumente, die auf den Erfahrungen und Erwartungshaltungen der Versuchspersonen basieren, und den enthaltenen häufigen und seltenen Elementen. Zusätzlich wurde den Probanden die Liste häufiger Elemente mit der Bitte präsentiert, diejenigen Elemente auszuwählen (oder neue hinzuzufügen), die in einer typischen persönlichen Homepage enthalten sein sollten.

[117] Die Datensammlung erfolgte auf Basis der folgenden Definition: „Home pages were defined as belonging to a named individual who was not advertising or selling a business or service and whose information content primarily related to him/herself." (Dillon und Gushrowski, 2000, S. 203). Bezüglich der Liste häufiger Elemente fallen verschiedene Aspekte auf: Zunächst ist unklar, weshalb die Verfasser in der Lage sind, das Element „Advertisement" identifizieren zu können, da Webseiten, die Werbung enthalten, bei der Datensammlung ausgeschlossen wurden (vgl. die Definition). Insbesondere von der Perspektive der Benutzernennungen (siehe im Folgenden) ist die Annahme des Elements „Table" problematisch, da auch unsichtbare Tabellen existieren können, die für Layout-Zwecke eingesetzt werden. Mit HTML vertraute Probanden könnten dieses Element also – im Gegensatz zu technisch weniger versierten Personen – durchaus als typisch ansehen.

Die Tabellen 4.10 bis 4.12 stellen die Ergebnisse von Dillon und Gushrowski (2000), Bates und Lu (1997) und Bittner (2003) dar und enthalten in den vier rechten Spalten die Resultate der jeweils anderen zwei Studien; es werden nur unmittelbare Übereinstimmungen aufgeführt. Die in den Tabellen 4.11 und 4.12 genannten Resultate von Dillon und Gushrowski beziehen sich auf die Angaben der Probanden. Die Namen der Elemente wurden nicht übersetzt, um die von den Autoren gewählten Bezeichnungen nicht zu verfälschen.

	Element	Total (of 57)	Total (%)	Initial sampling (%)	B. & L. (1997) Rang	B. & L. (1997) Proz.	B. (2003) Rang	B. (2003) Proz.
1	Title	55	96	71	—	—	—	—
2	E-mail address	49	86	82	1	92,1	1	100
3	Update date	48	84	39	8	43,0	—	—
4	Table of contents	42	74	11	—	—	—	—
5	Create date	41	72	20	49	1,8	—	—
6	External links	39	68	72	4	56,1	—	—
7	Welcome message	38	67	51	—	—	—	—
8	One to four graphics	34	60	52	3	78,9	—	—
9	Photographs	32	56	42	11	35,1	4	72
10	Brief bio	32	56	49	14	31,6	—	—
11	Text-only option	26	46	2	—	—	—	—
12	Five to nine graphics	22	39	31	3	78,9	—	—
13	Site map	14	25	4	—	—	—	—
14	Guestbook	11	19	16	20	17,5	—	—
15	Lists	9	1	33	—	—	—	—
16	Animation	8	14	37	12	33,3	—	—
17	Tables	7	12	37	—	—	—	—
18	Frames	7	12	11	—	—	—	—
19	Sound	7	12	5	37	5,3	—	—
20	Image map	5	9	4	—	—	—	—
21	Counter	2	4	39	5	52,6	—	—
22	Advertisements	0	0	33	—	—	—	—
23	Ten or more graphics	0	0	17	3	78,9	—	—
24	Back to top button	1	—	—	—	—	—	—
25	Thumbnails of images	1	—	—	—	—	—	—

Tabelle 4.10: Bestandteile privater Homepages (nach Dillon und Gushrowski, 2000, S. 203)

Tabelle 4.10 stellt die genannten Elemente in der dritten und vierten Spalte dar. Dillon und Gushrowski ermitteln ebenfalls eine signifikante Korrelation zwischen den von den Versuchspersonen als typisch erachteten Elementen und ihren Frequenzen in der Stichprobe: „From this we may conclude that users' ideas of what a personal home page should contain are broadly agreed upon, and that these ratings reflect existing home page contents on the web." (ebd., S. 204). Basierend auf diesen Ergebnissen wird die persönliche Homepage als „first unique digital information genre" bezeichnet (ebd., S. 205).

Bates und Lu (1997, S. 332) wollen mit ihrer Analyse von 114 persönlichen Homepages, die im Juni 1996 von einem Verzeichnis privater Webangebote gesammelt wurden, ein vorläufiges Profil dieser „new social form" entwickeln. Tabelle 4.11 zeigt die Ergebnisse in gebündelter Form: In nur 90 Homepages (79%) wird der vollständige Name des Autors angegeben, in 17 Webseiten (15%) wird ein Spitzname verwendet. Eine E-Mail-Adresse wird jedoch in 105 Dokumenten (92%) aufgeführt, als zusätzliche Möglichkeit der Kontaktaufnahme wird in 18 Homepages (16%) eine Telefonnummer genannt. Diese hier nur ausschnittsweise angesprochenen Elemente werden von Bates und Lu als „personal information" aufgefasst, die der Produzent auf seiner Homepage preisgibt. Insgesamt existieren durchschnittlich 5,9 dieser 32 Elemente auf einer Seite. Bezüglich der Strukturierung werden drei Typen differenziert: Der erste Typ wird als „menu home page" bezeichnet und umfasst ausschließlich Listen von Hyperlinks zu weiterführenden Seiten des Webangebots (9% der Dokumente). Der zweite Typ besteht beinahe vollständig aus Fließtext mit nur wenigen Links (14%). Der letzte, die Stichprobe dominierende Typ enthält viele Hyperlinks und zu Beginn der Seite eine Art Inhaltsverzeichnis (77%). Andere Homepages werden in 107 Dokumenten (94%) referenziert, 41 Homepages (36%) bieten Links zu Suchmaschinen an, und 42 Seiten (37%) verweisen auf frei zugängliche Software-Pakete. Neben den Inhaltselementen vom Typ „personal information" führen Bates und Lu die Gruppe „Miscellaneous" auf, die sich auf die Interaktivität und

	Element	Type	Occur.	Perc.	D. & G. (2000) Rang	D. & G. (2000) Proz.	B. (2003) Rang	B. (2003) Proz.
1	E-mail address	Pers. inform.	105	92,1	2	86	1	100
2	Name	Pers. inform.	90	78,9	—	—	2	100
3	Icons or artwork	Visual/audio	90	78,9	8	60	—	—
4	Favourite Web sites	Pers. inform.	64	56,1	6	68	—	—
5	Number of hits to home page	Misc.	60	52,6	21	4	—	—
6	Background: Patterned design – any colour	Visual/audio	52	45,6	—	—	—	—
7	Background: White or grey	Visual/audio	44	38,6	—	—	—	—
8	Last update date	Misc.	49	43,0	3	84	—	—
9	Gender	Pers. inform.	44	38,6	—	—	—	—
10	'Under construction' and its variants	Misc.	41	36,0	—	—	—	—
11	Photo	Pers. inform.	40	35,1	9	56	4	72
12	Motion picture	Visual/audio	38	33,3	16	14	—	—
13	Current work	Pers. inform.	37	32,5	—	—	6	64
14	Educational background	Pers. inform.	36	31,6	10	56	—	—
15	Hobbies, interests	Pers. inform.	35	30,7	—	—	5	68
16	Motion picture: On screen	Visual/audio	33	28,9	—	—	—	—
17	Free speech blue ribbon	Misc.	25	21,9	—	—	—	—
18	Address	Pers. inform.	22	19,3	—	—	8	32
19	Past work	Pers. inform.	22	19,3	—	—	—	—
20	Sign guestbook	Misc.	20	17,5	14	19	—	—
21	Phone or fax number	Pers. inform.	18	15,8	—	—	9	28
22	Background: Colours other than white or grey	Visual/audio	18	15,8	—	—	—	—
23	Alias or nickname	Pers. inform.	17	14,9	—	—	—	—
24	Personal history	Pers. inform.	15	13,2	—	—	—	—
25	Work capacity	Pers. inform.	15	13,2	—	—	—	—
26	Copyright notice	Misc.	15	13,2	—	—	—	—
27	Icon 'new'	Misc.	15	13,2	—	—	—	—
28	Resume (formally labelled)	Pers. inform.	13	11,4	—	—	—	—
29	Age	Pers. inform.	12	10,5	—	—	7	64
30	Favourite book, CD, food, etc.	Pers. inform.	12	10,5	—	—	—	—
31	Publication list	Pers. inform.	11	9,7	—	—	—	—
32	Geographical region	Pers. inform.	10	8,8	—	—	3	72
33	Friends' home pages	Pers. inform.	8	7,0	—	—	—	—
34	Family background	Pers. inform.	7	6,1	—	—	—	—
35	Honour or award received	Pers. inform.	7	6,1	—	—	—	—
36	Marriage status (explicitly stated)	Pers. inform.	7	6,1	—	—	—	—
37	Sound (click to pull up the feature)	Visual/audio	6	5,3	19	12	—	—
38	Cultural background	Pers. inform.	5	4,4	—	—	—	—
39	Research areas	Pers. inform.	5	4,4	—	—	—	—
40	Disclaimer	Misc.	5	4,4	—	—	—	—
41	Motion picture: Click to pull up	Visual/audio	5	4,4	—	—	—	—
42	Military service	Pers. inform.	4	3,5	—	—	—	—
43	Religious beliefs	Pers. inform.	3	2,6	—	—	—	—
44	Clock or calendar	Misc.	4	3,5	—	—	—	—
45	Directories where indexed	Misc.	3	2,6	—	—	—	—
46	Current Web project	Pers. inform.	2	1,8	—	—	—	—
47	Personal seal	Pers. inform.	2	1,8	—	—	—	—
48	Member of HTML Writer's Guild	Misc.	2	1,8	—	—	—	—
49	Time of creation	Misc.	2	1,8	5	72	—	—
50	Favourite people	Pers. inform.	1	0,9	—	—	—	—
51	List of courses taught	Pers. inform.	1	0,9	—	—	—	—
52	News groups sponsored by	Pers. inform.	1	0,9	—	—	—	—
53	Awards for Web site	Misc.	1	0,9	—	—	—	—

Tabelle 4.11: Bestandteile privater Homepages (nach Bates und Lu, 1997, S. 336–338)

fortwährende Veränderung des WWW bezieht. Der Zugriffszähler wird als „potential source of pride“ bezeichnet und reflektiert die Popularität einer Homepage (ebd., S. 337), in gleicher Weise erfüllt ein prall gefülltes Gästebuch eine Prestigefunktion (Bittner, 2003, S. 91). Darüber hinaus signalisieren Counter und Gästebuch Medienkompetenz, da die Realisierung technisch aufwändig sein kann, sofern nicht externe Dienstleister in Anspruch genommen werden. Besonders auffällig ist die hohe Anzahl von „under construction“-Hinweisen, die in 36% aller Angebote ermittelt wurden. Bates und Lu zufolge zeigen sie dem Rezipienten, dass der Autor seine Homepage aktiv pflegt, sie können jedoch auch prophylaktisch platziert werden, um zu testen, ob eine neue und bereits vollständig vorliegende Homepage von den Rezipienten akzeptiert wird.[118] Als dritte Gruppe werden „physical features“ von persönlichen Homepages (vgl. „Visual/audio“ in Tabelle 4.11) untersucht. Bates und Lu kommen zu dem Schluss, dass eine sehr große Bandbreite von Inhalten zu verzeichnen ist, die einer individuellen Gestaltung unterliegen; keines der ermittelten Elemente wird in jeder Homepage verwendet. Bezüglich des Grades der Standardisierung konstatieren die Verfasser: „In sum [...] it appears that the form and content of personal home pages on the [WWW] is still quite open and various. Though certain popular features may be found in it, the public social form known as the ‘personal home page’ has not yet fully developed a fully standardised character and social role, recognised by all.“ (ebd., S. 339).

De Saint-Georges (1998) untersucht lokale, personale und temporale Deiktika in 38 persönlichen Homepages von Studierenden.[119] Es werden zwei Varianten der Konstruktion eines deiktischen Zentrums ermittelt: Beispiele wie „Welcome to my very own home-on-the web page“, „Welcome to my tiny little corner of cyberspace“, „Welcome to my fast-paced life“ und „Come on in and make yourself at home“ zeigen, dass die Autoren von einem abstrakten oder konkreten, in jedem Fall jedoch begrenzten und ihnen zugehörigen Raum ausgehen, in den der Rezipient eingeladen wird. Die zweite Variante bezeichnet de Saint-Georges als „office home page“. Die Präsenz des Erzählers wird nur implizit deutlich, es werden kaum Pronomina verwendet, der Inhalt des Dokuments, das in den meisten Fällen die Tätigkeit des Autors thematisiert, ist faktenorientierter und es findet keine explizite Einladung statt. Sehr häufig werden ana- und katadeiktische Ausdrücke verwendet (z. B. „top“, „up“, „below“ und „down“), die nur aufgrund ihrer Positionierung korrekt interpretiert werden können. Sie deuten eine Bewegung an, die vom aktuell rezipierten Satz wegführt. De Saint-Georges (1998, S. 70) weist darauf hin, dass eine Verbindung zwischen Deiktika, die häufig als Zeigegesten aufgefasst werden, und Hyperlinks besteht, denn „they behave so much like pointers

[118] Ebenso wie Verweise auf Suchmaschinen werden die ehemals allgegenwärtigen, meist an US-amerikanische Straßenschilder angelehnten Grafiken der Typen „under construction“ und „men at work“ mittlerweile kaum noch eingesetzt. Meiner Einschätzung nach haben die Benutzer (von denen einige auch als Autoren tätig sind) mittlerweile erkannt, dass derartige Hinweise redundant sind, denn von nahezu jeder Website wird verlangt, dass sie permanent gepflegt und aktualisiert wird, somit ist *jede* Website *immer* „under construction“, und diese Tatsache muss nicht notwendigerweise explizit betont werden (vgl. Abschnitt 4.6.10). Rosso (2005, S. 87) ist der Ansicht, dass „under construction“ zwar als eigenständiges Genre erkannt wird, „[u]nfortunately, it is probably not a genre that would be helpful for use in searching.“ Falls ein korrespondierendes Exemplar vorliegt, könnte jedoch bei Suchanfragen eine niedrigere Gewichtung vorgenommen werden.

[119] Als Kriterien zur Aufnahme eines Dokuments in das Korpus musste ein Dokument die explizite Angabe enthalten, dass es sich um eine „home page“ handelt, es musste persönliche Informationen umfassen und sich auf nur eine einzige Person beziehen (de Saint-Georges, 1998, S. 68 f.).

that deictics seem to serve naturally as hyperlinks." (vgl. auch die Angaben zur Kohäsion in Abschnitt 3.5.3, sowie Schütte, 2004a, S. 226). Ein Hyperlink kann bezüglich des deiktischen Raumes in zweierlei Hinsicht konzeptualisiert werden: Entweder verweist die Verknüpfung (ana- oder katadeiktisch) auf den Text selbst, oder es wird exophorisch auf ein anderes Dokument verwiesen. Die personale Deixis kodiert die Rolle der Individuen, die an einer kommunikativen Handlung beteiligt sind. De Saint-Georges (1998, S. 71) fasst persönliche Homepages als narrative Texte mit einem Ich-Erzähler auf, die persönliche oder berufliche Informationen enthalten, oftmals beginnen sie mit dem Namen des Autors. Der Name fungiert als Titel, der dem Rezipient mitteilt, in wessen „universe of discourse" er oder sie im Begriff ist einzutreten.[120] Der Verfasserin zufolge ist es ist es gerade der eindeutig identifizierbare Produzentenname (entweder eingeführt als Titel oder Überschrift oder als informelle Begrüßung wie z. B. „Hi, my name is ..."), der eine Homepage letzten Endes *persönlich* macht. Die Rezipientenansprache erfolgt in den meisten Fällen informell, der Autor stellt sich zu Beginn der Homepage vor („Howdy-do!"), und am Ende verabschiedet er sich, was als Einladung zu einem späteren Besuch oder zur Kontaktaufnahme interpretiert werden kann („Thanks for visiting!"). Obwohl sie an eine unbestimmte und öffentliche Zielgruppe gerichtet sind, ähneln persönliches Homepages also folglich eher einem Brief als einem publizierten Text. In vielen der untersuchten Dokumente beobachtet de Saint-Georges Textteile, die Lebensläufen ähneln. Oftmals benutzen die Autoren diesbezüglich nicht die erste Person Singular, sondern verwenden das etablierte (US-amerikanische) Textmuster, das Pronomina verbietet; der Rezipient wird, so de Saint-Georges, vom Autor als potenzieller Arbeitgeber aufgefasst.[121] Bezüglich einer Auswertung der temporalen Deiktika gibt de Saint-Georges an, dass das Präsens deutlich dominiert, an zweiter Stelle folgt das Präteritum (Bayerl, 2002, und Bittner, 2003, bestätigen diese Befunde). Ein Hinweis auf den Entstehungs- oder Aktualisierungszeitpunkt ist – meist als explizites Datum – in 13 Homepages enthalten. Derartige Angaben sind notwendig, um implizite Aussagen wie „This semester's class ..." interpretieren zu können. Temporale Deiktika beziehen sich auf zeitliche Nähe oder Gleichzeitigkeit („this past fall ...", „over the past few years ...", „this semester ..."). Ebenso bezieht sich das Futur meist auf eine sehr nahe Zukunft, wobei de Saint-Georges von dem seitens des Autors geplanten Zeitpunkt der nächsten Aktualisierung ausgeht. Imperative werden in Verbindung mit Hyperlinks (vgl. den Titel der Studie) eingesetzt. Die temporalen Deiktika zeigen nach de Saint-Georges, dass persönliche Homepages keinen finalen Zustand besit-

[120] Roberts (1998) analysiert narrative Sätze und Nebensätze in 36 studentischen Homepages. Explizit narrative Sätze sind Roberts zufolge temporal verbunden, einem Lebenslauf ähnliche Textstrukturen, Listen und insbesondere Hyperlinks fungieren als implizit narrative Sätze. Die Ergebnisse dieser Arbeit werden an dieser Stelle nicht näher thematisiert, da die Vorgehensweise meines Erachtens nicht schlüssig ist. Unter anderem wird keine Definition von „implicit narrative clause" vorgenommen, d. h. es wird nicht deutlich, ob *alle* Hyperlinks in die Analyse eingeflossen sind, und es werden ebenfalls keine Beispiele angeführt, die andeuten könnten, auf welche Weise zwischen „explicit narrative clauses" und „non-narrative clauses" differenziert wurde. Roberts (1998) zeigt, dass Homepages, die große Mengen zusammenhängenden Textes enthalten, eher wenige Hyperlinks besitzen, und Dokumente, die viele Hyperlinks beinhalten, eher weniger Fließtext umfassen.

[121] In einigen Homepages werden auch Freunde und Bekannte des Produzenten direkt angesprochen, was die von Walker (2000) ermittelten Intentionen bestätigt. Die Untersuchung von Bittner (2003, S. 87) verifiziert diesen Befund, den er als „hybride Öffentlichkeit" bezeichnet. Die Produzenten der von Bittner untersuchten privaten Homepages richten sich „zwar stets an einen großen, unbestimmten Rezipientenkreis, in vielen Fällen wird aber an weniger herausragender Stelle explizit Bezug auf Freunde und Bekannte genommen."

zen, laufende Aktivitäten oder Situationen beschreiben und dass angenommen wird, dass sie im Falle neuer Ereignisse mit entsprechenden Darstellungen aktualisiert werden (vgl. Abschnitt 4.6.7). De Saint-Georges (1998, S. 76) fügt hinzu, dass persönliche Homepages, „unlike any other form of writing, [. . .] intrinsically ephemeral" sind.[122] Aus den Ergebnissen entwickelt de Saint-Georges (1998, S. 76) eine nach eigenen Angaben vorläufige Definition:

> Personal home page: presentation of the self in digital (hypertextual) form, authored by one individual, and which (i) emphasizes a person (minimally, by a picture or a name); and/or (ii) a person's current activities; and/or (iii) professional experience; and/or (iv) displays a person's interests (in the body of the text and/or through hyperlinks to other sites). (de Saint-Georges, 1998, S. 76)

De Saint-Georges (1998, S. 77) geht explizit auf den Umstand ein, dass es vermieden werden sollte, disjunktiv verknüpfte Aussagen in Definitionen zu verwenden. Dies ist jedoch ihrer Ansicht nach ein deutliches Indiz dafür, dass das Definiendum nicht präzise charakterisiert werden konnte. Dennoch umschließt die angegebene Definition nach eigener Aussage das breite Spektrum der im Korpus vorhandenen Homepages.[123]

Bayerl (2002) untersucht eine Stichprobe von 40 studentischen Homepages, die mit Hilfe der in Kapitel 7 vorgestellten Korpusdatenbank analysiert wurden und führt eine Arbeitsdefinition dieser Hypertextsorte ein, die sich an de Saint-Georges (1998) anlehnt:

> Studentische persönliche Homepages präsentieren Inhalte in digitaler (hypertextueller) Form, die (i) von Studenten erstellt werden und (ii) der Darstellung ihrer Person (mindestens durch ein Bild oder einen Namen) und/oder (ii) [sic] ihrer Interessen und/oder (iii) ihrer studentischer Aktivitäten dienen und darüberhinaus (iv) Dienstleistungen (in textueller Form oder als Links auf externe Seiten) unabhängig von der Person anbieten können. (Bayerl, 2002, S. 16)

Die Studie von Bayerl deckt sich in großen Teilen mit den bereits dargestellten Analysen: Die Kontaktfunktion ist Bayerl zufolge die primäre mit studentischen Homepages verbundene Intention, da der Leser in 90% der Dokumente explizit oder implizit gebeten wird, mit dem Autor in Kontakt zu treten. Das Erscheinungsbild[124] der Webangebote erweist sich als sehr heterogen, es ist, so Bayerl (2002, S. 22 f.), „nur abhängig vom Können und vom Ehrgeiz des Homepagebesitzers." Als gemeinsame Formeigenschaft wird der Seitenaufbau ermittelt (Kopf-, Mittel- und Fußteil). Überschriften werden – ebenso wie Fußzeilen mit dem Erstellungs- oder Aktualisierungsdatum – in 87,5% aller Einstiegsseiten eingesetzt, 65% davon enthalten das Wort „Homepage", das oft von einer Begrüßung und dem Namen des

[122] Diese Darstellung gilt jedoch nur für Homepages, die *tatsächlich* in regelmäßigen Abständen aktualisiert werden. Legt man eine großzügige Extension der Phrase „any other form of writing" zugrunde, können auch Einkaufs- und Notizzettel als „kurzlebig" und „flüchtig" aufgefasst werden.

[123] Das Spektrum der oben dargestellten Varianten bzw. Funktionen persönlicher Homepages wird von dieser Definition ebenfalls zu großen Teilen abgedeckt, wobei eine Differenzierung, die auf der Befragung der Produzenten basiert, naturgemäß deutlich präzisere Ergebnisse liefert (vgl. z. B. Walker, 2000).

[124] Bayerl (2002, S. 31 ff.) gibt zusätzlich an, dass 22,5% der Dokumente einen weißen Hintergrund verwenden, in 2,5% der Fälle wird die Hintergrundfarbe nicht modifiziert, andersfarbige Hintergründe sind in 60% der Homepages zu finden und 15% verwenden eine Hintergrundgrafik. Neben Text und Bild werden in 27,5% der Homepages andere Medien verwendet, z. B. Audiodateien (7,5%).

Autors flankiert wird. Die Homepages besitzen eine große Varianz bezüglich ihres Umfangs, neben der Einzelseite ist eine Homepage mit insgesamt 173 HTML-Dokumenten im Korpus enthalten (Durchschnitt: 16,7; 30% der Homepages enthalten Frames). Obwohl das durchschnittliche Alter der Dokumente 20,7 Monate beträgt (Minimum: 2, Maximum: 64), ist in 22,5% ein Hinweis auf den fortwährenden Umgestaltungsprozess enthalten. In Bezug auf Themen ermittelt Bayerl die Gruppen Selbstdarstellung, Kontakt zum Leser, Dienstleistungen, Unterhaltung und Metainformation. Elemente der Selbstdarstellung sind in 39 der 40 Dokumente vorhanden, wobei grob zwischen den Gruppen Person und privates Umfeld, Freizeitaktivitäten (62,5%) und Angaben zur Rolle als Student (22,5%) differenziert wird.[125] Die Autoren versuchen insbesondere durch umgangssprachliche Begrüßungen („Hi, Leute!" „Howdy Folks!", insgesamt 57,5%) ein persönliches Verhältnis zum Leser herzustellen, Verabschiedungen kann Bayerl nur in wenigen Fällen belegen („Viel Spaß dabei!"). Neben expliziten Aufforderungen zur Kontaktaufnahme („Schreib mir!") enthalten einige Homepages auch Gästebücher (10%). Vom Autor verfasste Texte (z. B. Anleitungen oder Erlebnisberichte) oder Hyperlinks zu weiterführenden Websites, die einen spezifischen Informations- oder Wissensbedarf des Rezipienten abdecken könnten, fasst Bayerl als Dienstleistungen auf, die häufig als „Nützliches" oder „Auskünfte aller Art" überschrieben sind. Die Verfasserin vermutet, dass Informationen dieser Rubrik, die einen potenziellen Mehrwert darstellen, den Rezipienten zum Verweilen auf der Homepage animieren sollen, wenn kein unmittelbares Interesse an der Person des Produzenten vorliegt. Unterhaltende Elemente betonen den informellen Charakter zusätzlich und sind in 22,5% der Homepages enthalten. Als Metainformationen werden Angaben zum Stand der Homepage und zu ihrer Nutzung aufgefasst.[126] Neben diesen formalen Aspekten geht Bayerl (2002, S. 62–108) auf sprachliche Besonderheiten ein, wobei insbesondere zahlreiche Phänomene beobachtet werden, die kommunikative Nähe hervorrufen sollen und der konzeptionellen Mündlichkeit zugerechnet werden können. Hierzu gehören z. B. Assimilationen („wen's interessiert", „Hier geht's weiter", „Allzuviel gibts hier aber auch nicht"), Elisionen („Hab gehoert es soll [...]", „fühl' Dich wie zu Hause!") und Tilgungen von Silben („schreib mir bitte ne mail", „Ist ja auch nur 'ne Homepage").[127] Bezüglich der verwendeten Lexik ermittelt Bayerl Anleihen bei Jugend- („superschnelle Neuentwicklung", „tonnenweise Bilder", „ganz ganz toll", „mal checken, was so laeuft", „ziemlich

[125] Angaben zur Person und zum privaten Umfeld beziehen sich in der Regel auf den Namen, das Studienfach, Alter, Wohnort und Freunde. Der Name des Autors wird in 12,5% der Homepages *nicht* erwähnt (Bayerl, 2002, S. 38). Die vollständige Form mit Vor- und Nachname wird in 77,5% der Dokumente angegeben, wohingegen sich 5% der Seiten auf eine Nennung des Vor- oder Spitznamens beschränken. Der Name des Autors wird in der Regel in der Überschrift genannt. Selbstdarstellungen, die über den Namen und ein Foto hinausgehen, beziehen sich meist auf einen Werdegang, Links zu den Homepages von Freunden (25%) oder Nebentätigkeiten. Offizielle Lebensläufe werden Bayerl zufolge höchstens zusätzlich angeboten (5%).

[126] Das Datum der Erstellung oder Änderung (72,5%) wird „stets am Ende einer Einzelseite" angegeben (Bayerl, 2002, S. 55). Einen Zugriffszähler enthalten 30%, technische Informationen zur Darstellung umfassen 12,5% der Homepages. Ebenfalls ermittelt werden Haftungsausschlüsse (10%) und die Verwendung des Copyright-Zeichens (©, 20%). Eine englischsprachige Version der Homepage wird in 10% aller Fälle bereitgestellt.

[127] Bayerl (2002, S. 81) berichtet, dass diese Mündlichkeitsmerkmale nur in spezifischen thematischen Bereichen der Homepage eingesetzt werden: „Wechsel zwischen dem Auftreten und Fehlen solcher Charakteristika an thematisch unterscheidbaren Stellen einer Homepage (offizieller Lebenslauf vs. Bericht von einem Segeltörn) deuten zudem darauf hin, daß die Autoren den unterschiedlichen Formalisierungsanforderungen bzw. den Ansprüchen verschiedener Zielgruppen gerecht zu werden suchen."

	Element	Vork.	Proz.	B. & L. (1997) Rang	B. & L. (1997) Proz.	D. & G. (2000) Rang	D. & G. (2000) Proz.
1	E-Mail-Adresse	25	100	1	92,1	2	86
2	Name	25	100	2	78,9	—	—
3	Wohnort	18	72	32	8,8	—	—
4	Bild	18	72	11	35,1	9	56
5	Hobbys	17	68	15	30,7	—	—
6	Beruf	16	64	13	32,5	—	—
7	Alter	16	64	29	10,5	—	—
8	Anschrift	8	32	18	19,3	—	—
9	Telefonnummer	7	28	21	15,8	—	—

Tabelle 4.12: Bestandteile privater Homepages (nach Bittner, 2003, S. 98)

fetzig“, „ganz krass“), Computer-, Fach- und Fremdsprachen („Sorry, Baustelle!“, „Das muss directly korrigiert werden.“). Ebenfalls gefunden wurden Sprecher- und Hörersignale wie z. B. „Mhhhhhh, nicht einfach“ und „Naja“. Bezüglich der Syntax orientieren sich die Autoren an mündlichen Ausdrucksformen, was die zahlreichen Ellipsen und Aposiopesen belegen. Para- und nonverbale Ausdrucksmittel, die für die Chat-Kommunikation charakteristisch sind (vgl. etwa Haase et al., 1997, und Runkehl et al., 1998, sowie Abschnitt 8.3), werden ebenfalls in den studentischen Homepages eingesetzt. Hierzu zählen Akzent und Emphase, die durch Veränderung der Schriftgröße und konstante Großschreibung ausgedrückt werden, die stimmliche Dehnung durch Reduplikation von Vokalen („meeeeeeeehr Homepages“), Pausen (realisiert durch Auslassungspunkte), Mimik (Smileys, die meist am Ende einer Äußerung eingesetzt werden und typografisch markierte Verbstämme wie „∗freu∗“, „∗lach∗“, „∗breitgrins∗“) sowie Gestik („∗riesenknuddel∗“).

Bittner (2003) geht auf die „Textsorte“ private Homepage (ebd., S. 75) ein, die anhand einer Stichprobe von 25 Dokumenten charakterisiert wird. Als konstitutiv für ihren Objekt- und Gegenstandsbereich wird die „selbst verfaßte Darstellung einer Person auf einer oder mehreren Internetseiten“ angenommen (ebd., S. 79), die sich implizit oder explizit an Freunde und Bekannte richtet. Daher kann auch eine Tendenz zur Erzeugung kommunikativer Nähe und Vertrautheit festgestellt werden, die sich in einer „tendenziell geringeren Normgerechtigkeit sprachlicher Formen“ und erhöhter Expressivität manifestiert, weil der breite, multimediale Elemente umfassende Spielraum zur Gestaltung der Homepage den „Versprachlichungszwang“ relativiert (ebd., S. 87 f.). Ein weiteres Merkmal des Strebens nach kommunikativer Nähe betrifft die große Kooperationsbereitschaft, die durch direkte Ansprachen und Kommunikationsappelle deutlich wird. Die bereits angesprochene Individualisierung privater Homepages fasst Bittner (2003) als eine durch „Satzzeichen, Farben oder grafische Elemente“ sowie „Exklamationen, Imperative, eine entsprechende Lexik, aber auch Textauszeichnungen und Emoticons, Soundfiles und animierten Text“ hervorgerufene „Emotionalisierung“ auf, die sich „vor allem para- und nichtsprachlicher Elemente bedient“ (ebd., S. 88).[128] Bezüglich der Kommunikationssituation und den Kommunikationsbedingungen privater Homepages kommt Bittner (2003, S. 97) zu dem Schluss, dass deutliche Unterschiede zu etablierten Textsorten vorliegen, die insbesondere auf das „hybride Öffentlich-

[128] An anderer Stelle fügt Bittner (2003, S. 123) die Merkmale „Abkürzungen“ und „Verbstammbildungen“ hinzu, die bereits in Analysen der computervermittelten Kommunikation als hochfrequent berichtet wurden (vgl. Haase et al., 1997, Bayerl, 2002, sowie Kapitel 8, S. 367 ff.).

keitsverhältnis“ zurückgeführt werden. Zu den Inhalten der 25 Homepages wird angegeben, dass sich „ein gewisser Konsens über bestimmte »Kerninhalte« der Textsorte etabliert“ (ebd., S. 93). Gleichwohl zeigen die Dokumente jedoch eine große Themenvielfalt abseits des inhaltlichen Kerns, der sich aus den Komponenten „persönliche Informationen“, „Interessen und Hobbies“ [sic], „Berufliches“, „Nützlichkeit“ und „Kommunikationsplattform“ zusammensetzt und sich mit den Ergebnissen anderer Studien weitestgehend deckt. Hinsichtlich der persönlichen Informationen ermittelt Bittner (2003, S. 98) die in Tabelle 4.12 dargestellten Frequenzen. Interessen und Hobbys nehmen einen sehr großen Raum innerhalb der privaten Homepages ein, wobei insbesondere Themen aus dem Computerbereich vorherrschend sind, die meist auf Folgeseiten diskutiert werden.[129] Eine besonders interessante Beobachtung Bittners (2003, S. 99 f.) betrifft den Umstand, dass nur zwei der 25 untersuchten Homepages „Manifestationen persönlicher, politischer oder weltanschaulicher Überzeugungen“ enthalten. Gerade derartige Themen führt Walker (2000) als Kennzeichen intrinsischer Homepages an, mit denen sich die Autoren anderen Benutzern vorstellen möchten, um mit ihnen bezüglich der kontroversen, politischen oder philosophischen Themen ins Gespräch zu kommen.[130] Auch zu beruflichen Themen gibt Bittner (2003, S. 100) an, dass sie auf den Homepages in unterschiedlicher Breite thematisiert werden; es überwiegen jedoch Berufe, die mit Computern oder dem Internet zu tun haben oder solche, „für die die Homepage ein geeignetes Dokumentations- und Publikationsmittel ihrer Arbeit bzw. von Ergebnissen derselben darstellt.“ (vgl. auch Storrer, 1999b, S. 5).[131] Ebenfalls als Teil des Kernbereichs wird die „Nützlichkeit“ von Informationen angesehen, die in fast allen Homepages explizit hervorgehoben wird (Dienstleistungen im Sinne von Bayerl, 2002). Der abschließende Bestandteil „Kommunikationsplattform“ manifestiert sich durch explizite Kommunikationsappelle (68%) und `mailto:`-Links (96%). Im Hinblick auf textstrukturierende Ordnungsmuster gibt Bittner an, dass die Homepages einen unterschiedlich ausgeprägten Gebrauch von den Möglichkeiten des WWW machen. Es wird – mit Bezug auf Freisler (1994) – insbesondere eine „Artikulation des Textes in funktionale Subtexte“ (die durch Hyperlinks verknüpft werden) und eine „Auflösung“ des Textes, („im Sinne einer komplex strukturierten, kommunikativ, konzeptuell und thematisch gegliederten Einheit, die vorrangig aus Sätzen besteht“) beobach-

[129] Bezüglich der in der Einstiegsseite verwendeten Lexik ermittelt Bittner (2003, S. 122) viele Fachbegriffe aus dem Computer- und Netzwerkbereich (z. B. „Farbtiefe“, „Browser“, „Surfen“, „Pixel“, „Webcam“, „einloggen“ etc.) und nimmt an, „daß ein Teil dieser Termini auch dauerhaft in den Alltagswortschatz eingehen wird, weil die bezeichneten Objekte zur alltäglichen kommunikativen Praxis gehören werden.“ Ihr Verständnis sei „teilweise sinnvoll, teilweise unabdingbar für die Nutzung des Internet als Kommunikationsmittel“ (ebd.). Bittner kann hier nur bedingt zugestimmt werden, da Begriffe wie „Browser“, „Download“, „Link“ und „Datei“ zwar so eng mit dem WWW verbunden sind wie „Inhaltsverzeichnis“, „Seite“, „Kapitel“ und „Seitenzahl“ mit gedruckten Texten. Termini wie „Farbtiefe“, „Pixel“, „Script“, „dynamische IP-Adresse“ oder „komprimierte Dateien“ (ebd., S. 122) können jedoch tatsächlich als Fachvokabular im engeren Sinne eingestuft werden, die keinesfalls „unabdingbar“ für die Nutzung des WWW sind (vgl. auch Dürscheid, 2004, S. 143 ff.). Vielen Personen, die gelegentlich in Internet-Cafés auf das WWW zugreifen, dürften diese Begriffe nicht bekannt sein.

[130] Zur Erklärung der Abwesenheit derartiger Homepages in Bittners Korpus können zwei Möglichkeiten herangezogen werden: Entweder ist die untersuchte Stichprobe zu klein (d. h. es liegen tatsächlich nur zwei intrinsische Homepages vor), oder es handelt sich bei der Diskussion kontroverser Themen um ein kulturelles Spezifikum (vgl. Fußnote 63, S. 190). Dieses Thema wird nachfolgend erneut aufgegriffen.

[131] Homepages im universitären Bereich schreibt Bittner (2003, S. 101) zwar eine besondere Rolle zu, zwischen privaten und beruflichen Homepages wird jedoch nicht differenziert.

tet (Bittner, 2003, S. 113, vgl. auch Abschnitt 3.5.6). Hinsichtlich der sprachlichen Wohlgeformtheit und Darstellungsweise merkt Bittner an, dass die grammatische und orthografische Normgerechtigkeit keinesfalls die Regel ist (ebenso Bayerl, 2002). Die privaten Homepages werden in die vier Gruppen „hoher Elabortationsgrad, stärkerer Integrationscharakter, längere Texte“ (4 Dokumente), „mittlerer Elaborationsgrad, immer noch integriert, kürzere Texte“ (3 Dokumente), „geringer Elaborationsgrad, nur sehr kurze Texte bzw. einzelne Sätze“ (15 Dokumente) und „Fehlen von Texten (nur Überschriften/Listen), stark aggregativer Charakter“ (3 Dokumente) eingeteilt, d. h. nur vier Homepages weisen Merkmale auf, die für viele Exemplare traditioneller Textsorten typisch sind. Die Angaben zur Syntax sind entsprechend: Sie wird als „stark aggregativ“ (ebd., S. 123) bezeichnet, im Extremfall besteht der Text lediglich aus Überschriften, einigen Stichpunkten und Listen. In elaborierteren Texten bilden Überschriften und Schlagwörter eine grobe Makrostruktur, in die einzelne Sätze eingebettet werden, die meist kurz und parataktisch angeordnet sind. Die in Einzelfällen sehr umfangreichen Webseiten, die von der Homepage aus verlinkt sind, weisen dagegen eine elaboriertere Syntax und interessanterweise auch korrespondierende Textsortenbezeichnungen auf: „Reisebericht, Tourenbeschreibung, Aufsatz, Plädoyer, Artikel, Leserbrief, Tagebuchauszug, »My story«, Kochbuch, Weinkunde, Reisebericht, Witz der Woche.“[132] Insgesamt gelangt Bittner (2003, S. 125) zu dem Schluss, dass sich mit privaten Homepages „eine besondere Form schriftlicher Umgangs- und Alltagssprache entwickelt und etabliert, die die computerspezifischen Bedingungen von Produktion und Rezeption reflektiert.“

Ursprünge der persönlichen Homepage

Es steht außer Zweifel, dass die persönliche Homepage kein Pendant in der Klasse gedruckter Texte besitzt. Dies bemerkt auch Bittner (2003, S. 97) und argumentiert, dass die Produzenten aus eben diesem Grund „nicht auf ein präfabriziertes Modell zurückgreifen [können], das ihnen als Orientierung im Produktionsprozeß dienen könnte.“ Zum einen können Autoren, wie bereits mehrfach angesprochen, durchaus direkt oder indirket auf „präfabrizierte Modelle“ zurückgreifen, allerdings auf diejenigen anderer Produzenten, die bereits eigene Homepages erstellt haben (vgl. Abschnitt 4.3.2).[133] Zum anderen übersieht Bittner den Umstand, dass zwischen einem *unmittelbaren* Pendant und *verwandten* Textsorten zu unterscheiden ist, aus denen Text- und Formulierungsmuster übernommen werden, die durchaus mögliche Ursprünge verschiedener spezifischer Facetten der persönliche Homepage erklären können.[134]

[132] Bittner (2003, S. 124) führt weiter aus, dass diese Bezeichnungen bei den Produzenten Textmuster etablierter Textsorten aktivieren, die auf die digitalen Fassungen angewendet werden: „Diese Texte sind [...] als digitale Versionen traditioneller Textsorten und weniger als *genuine* digitale Texte anzusehen.“ (vgl. Abschnitt 4.5.2).

[133] Buten (1996) berichtet (vgl. Fußnote 114, S. 218), dass 95% der Befragten inhaltliche oder gestalterische Elemente aus anderen Webangeboten für die eigenen Homepages übernommen haben, d. h. sie orientieren sich sowohl inhaltlich auch als formal aneinander, was bis zur unmittelbaren Kopie eines Angebots führen kann (vgl. Döring, 1997, S. 324, und Döring, 2001a, S. 337).

[134] Diese Einflussfaktoren führt Bittner (2003, S. 127) selbst an: „Dabei greifen die Homepage-Autoren [...] auf Textmuster verschiedener Textsorten zurück, die im weitesten Sinne als »biographisch« zu bezeichnen sind oder der Beschreibung oder Kennzeichnung von Individuen dienen.“ Und: „Vielfach, teilweise sogar explizit, lehnen sich die Präsentationen persönlicher Informationen an tabellarischen Darstellungen wie dem »Lebenslauf« oder dem »Steckbrief« an.“ (Bittner, 2003, S. 98). Des Weiteren geht Bittner (2003, S. 96 f.) auf das Etikett „private Homepage“ ein: „Der Textsortenname an sich zielt zwar über das beschreibende Adjektiv und

Miller (1995) gibt für die insgesamt sieben von ihm ermittelten Typen persönlicher Homepages Textsorten an, die den jeweiligen Typ möglicherweise beeinflusst haben oder diesem in Grundzügen entsprechen: Der Typ „Hi, this is me (as an individual)" ist Miller zufolge mit dem (ersten) Brief an einen Brieffreund vergleichbar, und der Typ „Hi, this is me (as a member of an organisation)" ähnelt – zumindest in der Ausprägung der studentischen Homepage oder der Website eines Jugendlichen – dem „entry in a student handbook" oder der Vorstellung eines Schülers in einer Abiturzeitung (vgl. Miller, 1999). Die Familien-Homepage („Hi, this is us") vergleicht Miller mit einem „company report" und dem insbesondere in den USA bekannten „Annual Family Circular", der zu Weihnachten an Freunde und Verwandte geschickt wird, um über die wichtigsten Ereignisse des ausklingenden Jahres zu berichten. Eine direkte Zuordnung wird beim Typ „This is what I think is cool" nicht vorgenommen, hier erfolgt eher eine Inferenzziehung, d. h. der Rezipient schließt über die präsentierten und kommentierten Hyperlinks auf die Vorlieben und Abneigungen des Autors. Zu den als Werbung fungierenden Typen zählen zunächst Homepages der Klasse „Cool style". Miller vergleicht sie mit Arbeitsproben eines Kunstschaffenden sowie den Objekten und Materialien, die Lehrende in den Bereichen Werbung, Gestaltung und Design in ihren Dienstzimmern aufstellen, um Besuchern und Studierenden mitzuteilen: „see what a cool person I am". Der Typ „Electronic curriculum vitae" besitzt im traditionellen Lebenslauf das unmittelbarste Pendant. Der abschließende Typ „An advertisement for the service I can provide" entspricht Miller zufolge Flugzetteln (vgl. Androutsopoulos, 2000), Demo-Bändern und CDs von Nachwuchsbands, „or the people who stop you in the street in Edinburgh at Festival time to charm you into coming to their Fringe performances."

Auch de Saint-Georges (1998, S. 71) geht der Frage nach, welchem Genre persönliche Homepages angehören: „They are not autobiography, because they are written by the ordinary citizen, and they are not journals and diaries because they are designed for a potentially wide audience like a book would be." Die mündlich realisierte Lebensgeschichte kann ebenfalls als unmittelbares Pendant ausgeschlossen werden. Bittner (2003, S. 126) ist ähnlicher Meinung: „Die private Homepage ist kein Lebenslauf, sie ist keine (Auto-)Biographie, und sie ist auch kein Lexikoneintrag oder ein Artikel aus dem »Who is who«." Die von Walker (2000, S. 107 f.) befragten Autoren intrinsischer Homepages vergleichen ihre Dokumente häufig mit traditionellen Textsorten und Objekten, die in Gesprächssituationen verwendet werden, z. B. „signatures", „greeting card", „calling cards", „journals" und „letters of introduc-

den englischen Begriff des »home« [...] auf den privaten Bereich. Diese Bestimmung ist jedoch ausreichend vage, so daß die Produzenten einen großen Freiraum für die individuelle Ausgestaltung ihrer Homepage besitzen und diesen auch nutzen." Bittner argumentiert, dass sich die Autoren von der (lexikalischen) Semantik der Bezeichnung „private Homepage" leiten lassen, dass die inhaltliche Varianz in gewisser Weise auf die Vagheit des *Terminus* zurückzuführen ist. Zunächst müsste jedoch untersucht werden, ob den Autoren der Begriff „private Homepage" überhaupt bekannt ist (vgl. Schütz und Machilek, 2003). Die Produzenten werden sich vielmehr unter anderem an der eigenen Intuition und denjenigen Textexemplaren orientieren, die sie als Benutzer des Mediums WWW rezipiert haben (vgl. Abschnitt 4.3.2). Es ist bemerkenswert, dass Bittner (2003, S. 98) dienen Aspekt selbst thematisiert: „Das etablierte analoge Textmuster [der Lebenslauf, G. R.] dient dabei nicht nur als formale, sondern vor allem auch als inhaltliche Orientierung." Die angesprochene Intuition bezieht sich auf die Textsorten, die dem Produzenten bekannt sind und die für eine spezifische Kommunikationssituation (hier: Darstellung der eigenen Person) bewusst oder unbewusst als adäquat erachtet werden. Aus diesem Grund erinnern einige private Homapages nur vage an einen Lebenslauf, während andere vollständig ausgeprägte tabellarische Lebensläufe besitzen (siehe im Folgenden).

tion“. Die in diesen Nennungen vorhandene Bandbreite verdeutlicht Walker (2000, S. 108) zufolge, dass den Autoren durchaus implizit bewusst ist, dass sie unterschiedliche Typen privater Homepages betreiben, denn „a greeting card reveals less than a journal.“ Erickson (1996, S. 15) vergleicht diese Hypertextsorte mit dem Lebenslauf: „Personal pages are similar to informal resumes, except that in addition to professional material they often contain personal information.“ Dieser Aspekt wird im nachfolgenden Abschnitt genauer thematisiert.

Auch technische Spezifika haben einen Einfluss auf persönliche Homepages. Furuta und Marshall (1996) weisen darauf hin, dass in ihnen häufig detaillierte Informationen zum Tagesablauf oder zu wichtigen bevorstehenden Terminen enthalten sind, für eben diesen Zweck wurden auf UNIX-Systemen bis in die neunziger Jahre hinein `.plan`-Dateien eingesetzt (vgl. auch Crowston und Williams, 2000, S. 208).[135] Einen Einfluss besitzen auch die verschiedenen Kommunikationsdienste, die das Internet anbietet. Die empirischen Analysen haben deutlich gemacht, dass in privaten Homepages häufig Smileys und auch typografisch markierte Verbstämme verwendet werden, um kommunikative Nähe zu erzeugen. Dass auch Merkmale der E-Mail-Kommunikation einen Einfluss auf die Gestaltung und Strukturierung persönlicher Homepages besitzen, wird Kapitel 9 verdeutlichen.

Die möglichen Ursprünge der persönlichen Homepages können nur unter Beachtung der individualisierenden Metaphern betrachtet werden (vgl. Abschnitt 3.5.4). Häufig wird die Metapher des virtuellen Zuhauses benutzt. Bates und Lu (1997, S. 334) finden folgende Belege: „Aaron's *Home Sweet Home* Page“, „Spartan's *Haven*. You are *guest* number ... to *visit my house*“ (Hervorhebungen hinzugefügt, G. R.). Sie können Bates und Lu zufolge mit meist selbst angefertigten Namens- bzw. Haustürschildern verglichen werden, die den Gast freundlich einladen (vgl. Miller, 1999), die Familienmitglieder namentlich vorstellen, eine „proud social unit“ markieren, und sie können als moderne Version des Familienwappens aufgefasst werden. Weiterhin werden abstraktere Begriffe gefunden, die sich auf Lokationen beziehen: „quadrant“, „world“, „space“, „spot“, „domain“ und „inferno“. Insgesamt merken Bates und Lu jedoch an, dass sich weit weniger Homepages auf Metaphern des virtuellen Raumes oder Zuhauses als auf das Modell des Lebenslaufes stützen.

Die Veröffentlichung von Lebensläufen in persönlichen Homepages

Lebensläufe bzw. an den Lebenslauf angelehnte Texte sind hochfrequente Bestandteile persönlicher Homepages. Der Lebenslauf ist eine hinsichtlich Inhalt und Form normierte Textsorte. Wesentliche Prämissen zur erfolgreichen Realisierung seines kommunikativen Ziels sind eine strikte Befolgung der Regeln und Konventionen sowie orthografische und grammatikalische Korrektheit (vgl. auch Abschnitt 2.3.7).

[135] Ein Merkmal von UNIX-Betriebssystemen betrifft die Tatsache, dass mehrere Benutzer gleichzeitig an einem Rechner angemeldet sein können bzw. die Benutzerverzeichnisse über das lokale Netzwerk verteilt werden und simultan auf allen Maschinen zur Verfügung stehen. Jeder Benutzer hat – sofern diese Funktion nicht aus Sicherheitsgründen gesperrt wurde – die Möglichkeit, in seinem Verzeichnis eine Datei namens `.plan` anzulegen, in die beliebige Informationen eingetragen werden können. Über das Kommando `finger` *`benutzername`* können andere Anwender diese Daten, die häufig Kontaktinformationen oder Angaben über bevorstehende Termine (und damit verbundene Abwesenheiten) enthalten, abrufen. Die technischen Grundlagen des WWW wurden auf UNIX-Systemen entwickelt und auch die WWW-Anwender der ersten Jahre waren ausschließlich UNIX-Benutzer, weil für andere Betriebssysteme noch keine Browser existierten.

Bates und Lu (1997) ermitteln eine kleine Zahl von Dokumenten, deren Aufbau an das „resume“ angelehnt ist (und in einigen Fällen auch eine gleichlautende Überschrift besitzt) oder Komponenten beinhaltet, die sich daran orientieren. Es wird angemerkt, dass die korrespondierenden Regeln nur vereinzelt befolgt werden: „[...] even in those pages with the most resume-type organisation, there are often substantial deviations from the standard format. The creators [...] appear to enjoy experimenting with the more flexible possibilities of Web documents and include photographs, icons and other format deviations from conventional resumes.“ (Bates und Lu, 1997, S. 334). Die Autoren experimentieren also mit der Textsorte, sie brechen die existenten Regeln, weil es sich um eine Kommunikationssituation handelt, in der die Sanktionierung von Verstößen ausgeschlossen werden kann. Es entstehen Dokumente, die zwar oberflächlich an einen Lebenslauf erinnern, tatsächlich aber „quite freely and variously structured“ sind (Bates und Lu, 1997, S. 339). Nach de Saint-Georges (1998, S. 75) wird das Präteritum verwendet, wenn sich die Produzenten in narrativer Weise selbst als Person darstellen und Ereignisse, Gedanken, Anekdoten und private oder berufliche Aktivitäten schildern, die in den umgebenden Text eingebettet werden und einem Lebenslauf ähneln (z. B. „Below are a few notes about me that I scribbled on a paper bag on the bus on my way here“, „I studied at Simon Fraser University“, „I was born in Easton, PA“).

Heiber (2001) analysiert 50 willkürlich zusammengestellte studentische Homepages der Ruhr-Universität sowie der Fachhochschule Bochum. Bezüglich der Gestaltung beobachtet sie mit Abstand am häufigsten Fließtext, einen an Lebensläufe in Bewerbungen angelehnten, rein tabellarischen Aufbau und Mischformen; ein Autor bietet seinen Lebenslauf als Fließtext und zusätzlich in tabellarischer Form an. Die Lebensläufe sind mit durchschnittlich 176 Wörtern relativ kurz, die Extrembeispiele umfassen 33 bzw. 1 326 Wörter. Insgesamt sind 47 orthografische und 40 Interpunktionsfehler in den 50 Homepages enthalten. Das auffälligste Merkmal ist Heiber (2001, S. 17) zufolge der „unbekümmerte, schon fast lässige Einsatz der Sprachmittel“, der sich durch verschiedene Eigenschaften auszeichnet, die konzeptionelle Mündlichkeit ausdrücken. Die Syntax ist kaum durchkomponiert, es treten Anakoluthe und Ellipsen auf, weiterhin werden Ausklammerungen und Parenthesen benutzt. Die Lexik ist umgangssprachlich und dialektal gefärbt und es werden Füllwörter und fremdsprachliche Termini eingesetzt. Heiber beobachtet einen narrativen und häufig mit Humor und Ironie durchsetzten Stil, der die konzeptionelle Mündlichkeit zusätzlich betont. Dass es sich bei den untersuchten Textexemplaren nicht um Lebensläufe im eigentlichen Sinne handelt, zeigen die Überschriften, in denen dieser Begriff nur selten verwendet wird: „Was bisher geschah“, „About me“, „Persönliche Daten“, „Infos über meine Person“, „Ich“, „Ich über mich“, „Just me, myself and I“.[136] Heiber (2001, S. 24) vertritt diesbezüglich den folgenden Standpunkt: „Dass der Leser die Texte als Lebenslauf erkennt, kann nur an seinem Vorwissen liegen [...]. Der Autor muss [...] daher nicht unbedingt explizit angeben, dass

[136] Auch hier können meines Erachtens Parallelen zu Abiturzeitungen gezogen werden, denn dies dürfte die einzige Gelegenheit sein, in der alle Studierenden sich selbst (oder eine andere Person) in knapper und möglichst authentischer und humorvoller Weise schriftlich charakterisieren mussten. Die Überschrift „Was bisher geschah“ erinnert an die von einem Off-Sprecher eingeführte Zusammenfassung wöchentlich ausgestrahlter US-amerikanischer Fernsehserien („Previously on ...“), und die Überschrift „Just me, myself and I“ bezieht sich auf den 1989 erschienenen Titel „Me, myself and I“ der Hip-Hop-Formation „De La Soul“. Bayerl (2002, S. 89) findet ebenfalls ein Vorkommen: „ME, MYSELF AND I...“.

es sich um den Textabschnitt ›Lebenslauf‹ [...] handelt.“ Da der Lebenslauf eine sehr stark normierte Textsorte ist, stellt sich die Frage, ob es sich bei den in den Homepages beobachteten Textexemplaren tatsächlich um Instanzen der Textsorte Lebenslauf oder daran angelehnte Texte handelt. Die Kapitel 9 und 10 werden zeigen, dass auch Exemplare des traditionellen Lebenslaufes in persönlichen Homepages angeboten werden.

Killoran (2003a,b, 2004) geht auf die Frage ein, inwieweit sich das traditionelle Genre „resume“ verändert, wenn es im WWW eingesetzt wird. Mit Hilfe einer Suchmaschine wurden persönliche Homepages lokalisiert, die ein *resume* enthalten, woraufhin die Autoren per E-Mail angeschrieben und zur Bearbeitung eines Fragebogens gebeten wurden. Den 100 Rückmeldungen entnimmt Killoran (2004), dass die Befragten ihren Lebenslauf vor durchschnittlich vier Jahren im WWW veröffentlicht haben und dass er einige Male pro Jahr aktualisiert wird (auch bezüglich des Webdesigns). Meist werden zwei Lebensläufe gepflegt, einer für Stellenbewerbungen, und einer für das WWW, letzterer enthält zusätzliche Informationen, d. h. der für den Ausdruck gedachte Lebenslauf wird gerade nicht ohne jegliche Modifikation übernommen. Im Gegenteil: Einige der Befragten konzipierten ihren Lebenslauf im Hinblick auf eine Veröffentlichung im WWW und aktualisieren ihn regelmäßig, so dass der für Bewerbungen eingesetzte Lebenslauf per *Copy & Paste* auf der Grundlage der Webseite zusammengesetzt werden kann. Killoran (2004) bezeichnet den im WWW publizierten Lebenslauf daher auch als „über-resource“. Es werden unter anderem die folgenden Intentionen ermittelt: „To seek new employment [...]“ (56%), „To inform people [...] about who I am“ (46%), „To seek new clients for my self-employment“ (45%), „To enhance my profile among colleagues [...]“ (42%), „To inform visitors to the rest of my Web site about who made this site“ (32%), „To make myself part of the new medium of the [WWW]“ (29%), „To showcase my Web design skills“ (23%) und „To practice how to make a Web page“ (21%). Diese Bandbreite zeigt, dass mit der Veröffentlichung eines Lebenslaufes im WWW ein umfangreicheres Inventar von Intentionen verbunden ist als dies mit dem traditionellen Pendant je der Fall sein könnte. Es sind also, so Killoran (2004), nicht nur die technologischen Möglichkeiten für die beobachteten Änderungen verantwortlich zu machen, vielmehr nutzen und modifizieren die befragten Autoren das Genre in individueller und unerwarteter Weise. Lebensläufe im WWW sind also flexibler und idiosynkratischer und es können Erweiterungen hinsichtlich der Form und der Funktion beobachtet werden (Killoran, 2003a). Etwa ein Drittel der Befragten bietet einen Lebenslauf in mehreren Formaten an (z. B. HTML, PDF und ASCII). Einige Produzenten bieten sogar *mehrere* Lebensläufe an, die unterschiedliche Karrierebahnen reflektieren. Die WWW-Version des Lebenslaufes ist in fast allen Fällen Teil einer privaten oder beruflichen Homepage und enthält üblicherweise Links zu anderen Websites.

Zur Privatheit der privaten Homepage

Eine im WWW publizierte private Homepage ist, sofern kein Zugriffsschutz existiert, von jedem Benutzer weltweit abrufbar, weshalb verschiedentlich das veränderte Spannungsverhältnis zwischen Öffentlichkeit und Privatheit diskutiert wird. Den Autoren privater Homepages ist dieser Umstand in den meisten Fällen durchaus bewusst, denn sie bieten sowohl Informationen an, die üblicherweise auf Visitenkarten zu finden und somit für eine eher unbekannte Öffentlichkeit gedacht sind, als auch Daten privater Natur, z. B. Verknüpfun-

gen zu den Homepages von Freunden oder von den Kindern gemalte Bilder. Furuta und Marshall (1996) weisen darauf hin, dass die Produzenten vermutlich aufgrund der anonymen Leserschaft und der Passivität des Mediums Materialien einer weltweiten Öffentlichkeit zugänglich machen, die sie weder an eine Mailing-Liste schicken noch in einer Newsgroup posten würden.[137] Nach Ansicht von Bittner (2003, S. 71) macht der Begriff „Homepage" auf das „neue, ein wenig paradox anmutende Verhältnis zwischen Öffentlichtkeit und Privatheit, zwischen öffentlichem und privatem Raum aufmerksam. Denn das *Heim* ist ja eigentlich der genuine Ort der Privatheit und mithin das Gegenteil von Öffentlichkeit, und die Homepage nun propagiert exakt einen öffentlichen privaten Raum." (vgl. Boardman, 2005, S. 48). Sowohl Heiber (2001, S. 6) als auch Bittner (2003, S. 72) merken an, dass es sich um Tendenzen handelt, die nicht nur auf das WWW beschränkt sind, schließlich rücken TV-Formate vom Typ „Big Brother", Casting Shows und täglich ausgestrahlte Talkshows zunehmend die Privatsphäre einzelner Personen für einen ca. 15 Minuten umfassenden oder auch sehr langen Zeitraum ins Licht der Öffentlichkeit. Bittner (2003, S. 86) zufolge deuten Indizien in den von ihm untersuchten privaten Homepages darauf hin, dass die Autoren sich an eine prinzipiell unbekannte Zielgruppe richten, jedoch finden sich ebenfalls Hinweise darauf, dass ein restringierter Rezipientenkreis vermutet wird. Aus diesem Grund konstituiert sich laut Bittner (2003, S. 86) in privaten Homepages aus Sicht der Produzenten ein „hybrides Öffentlichkeitsverhältnis, das gekennzeichnet ist vom Widerspruch zwischen einer theoretisch totalen, faktisch aber sehr eingeschränkten Öffentlichkeit." Den Autoren ist der Umstand, dass ihre Homepage nur von einem Bruchteil der WWW-Benutzer rezipiert wird, zwar durchaus bewusst, aber dennoch führt die „Unentscheidbarkeit zwischen Privatheit und Öffentlichkeit", so Bittner, zu „Unsicherheiten und Uneinheitlichkeiten der Anrede, die zwischen Nähe und Distanz schwankt". Dieses Dilemma führt Bittner (2003, S. 87) zufolge zu einem Kommunikationsmodus, der weder vollständig privat noch vollständig öffentlich ist, „sondern sich in die Richtung einer »öffentlichkeitstauglichen« (Pseudo-)Privatheit bewegt."

Bittner (2003) findet in seiner Stichprobe von privaten Homepages lediglich zwei Dokumente, die politische oder weltanschauliche Themen diskutieren. Es wird eine Verschiebung der sichtbaren „Grenze zwischen öffentlichem und privatem Raum in die Richtung des letzteren" beobachtet, wobei gleichzeitig „der Begriff dieses privaten Raums und dessen, was als »privat« gilt, re-definiert" wird, so dass „das »wahrhaft Private« hinter eine öffentlich zugängliche Privatheit zurückweicht" (ebd., S. 100). Zur Begründung führt Bittner an, „daß die Explizierung oder Zurschaustellung privater und persönlicher Überzeugungen (insbesondere politischer und weltanschaulicher Art), die auf Kritik und Widerspruch stoßen kann, gerade im Rahmen einer [...] Textsorte wie der privaten Homepage deplaziert wirken kann." (ebd.). Bittner (2003, S. 126) argumentiert bezüglich der von ihm als „paradox" empfundenen Bezeichnung „private Homepage", dass zwar die Produktionsweise, der Inhalt und die Gestaltung dem Begriff des Privaten entsprechen mögen, „durch die Publikation jedoch der »private« Bereich dem öffentlichen Zugriff freigegeben wird und somit die Homepage auch in den öffentlichen Bereich übergeht. Damit [...] unterliegt nicht nur das Verhältnis zwischen Privatheit und Öffentlichkeit, sondern in letzter Instanz auch der Begriff des Privaten einer

[137] Diese Beobachtung ist nicht zutreffend, denn es existieren tausende von Mailing-Listen und Newsgroups, in denen sich die Mitglieder über teils sehr private und extrem kontroverse Themen austauschen.

grundlegenden Veränderung“. Diese Argumentation ist aus zwei Gründen zurückzuweisen: Zunächst ist die Stichprobe von 25 Dokumenten zu klein, um eine „grundlegende Veränderung“ des Begriffs der Privatheit ausmachen zu können. Weiterhin orientiert sich Bittner zu sehr an dem eingedeutschten Begriff der „*privaten* Homepage“, der ursprünglich auf den Terminus „personal home page“ zurückgeht. Bittner geht in seiner Argumentation von einer zu intimen und zu stark in der Privatsphäre verhafteten und geradezu absoluten Lesart aus, statt der von Bittner angesprochenen Privatheit geht es vielmehr um das Privatleben des Autors. Eine derartige Unterscheidung führt Bittner mit „dem wahrhaft Privaten“ und der „öffentlich zugänglichen Privatheit“ selbst ein.[138] Letztere entspricht den Informationen, die Autoren privater Homepages bereit sind, der Öffentlichkeit anzubieten, die sich, wie die in diesem Abschnitt vorgestellten Arbeiten zeigen, aus Sicht der Produzenten zu einem Großteil aus Freunden und Bekannten konstituiert; viele Autoren berücksichtigen bei der Planung und Anfertigung ihrer Webangebote eben diese Zielgruppe. Einige Personen dieser Zielgruppe interagieren aber auch abseits des Netzes mit dem Autoren, sie haben also nicht nur Einblick in sein Privatleben, sie sind ein *Teil* seines Privatlebens. Aus Sicht des Autors spricht offenbar nichts dagegen, Aspekte des Privatlebens, von denen seine Freunde Kenntnis haben, auch auf der privaten Homepage zur Sprache zu bringen, schließlich geht er davon aus, dass vornehmlich Freunde und Bekannte diese rezipieren. Ähnlich verhält es sich mit politischen oder weltanschaulichen Themen, die nach Ansicht Bittners nicht auf privaten Homepages diskutiert werden, weil sie auf „Kritik und Widerspruch stoßen könnten“. Viele Personen, die sich z. B. mit einer bestimmten politischen Überzeugung vollständig identifizieren und auch in der netzunabhängigen Öffentlichkeit – somit auch in ihrem Privatleben – entsprechend agieren, werden weder Scheu noch Hemmungen haben, dieses Thema auch auf ihren privaten Homepages zur Sprache zu bringen, um die damit verbundenen, vermeintlich positiven Aspekte in ein gutes Licht zu rücken. Falls jedoch lediglich eine geringe Sympathie für diese oder jene von der Öffentlichkeit als kontrovers eingestufte politische Neigung vorhanden ist, von der auch der engeren Freundeskreis keine Kenntnis besitzt, wird diese nur in den seltensten Fällen Gegenstand der privaten (d. h. namentlich gekennzeichneten) Homepage sein. Insofern ist die Bezeichnung „private Homepage“ keinesfalls paradox, sie ist vielmehr durchaus treffend, da sie denjenigen Personen, die ihr Privatleben einer größeren Öffentlichkeit zugänglich machen wollen, eine entsprechende Plattform bietet. Als Gründe hierfür können unter anderem Geltungssucht, Exhibitionismus oder Narzissmus angeführt werden, insofern bezieht sich, wie Heiber (2001, S. 6) zutreffend anmerkt, der Aspekt der Privatheit privater Homepages lediglich darauf, „dass der Autor der Homepage nur das veröffentlicht, was er gerne möchte, ohne dass er Zwängen unterliegt. Wo die Grenzen zum Öffentlichen liegen, bleibt selbstbestimmt.“ Die Auffassung von Öffentlichkeit und Privatheit ist demnach immer subjektiver Natur, kann jedoch durchaus von äußeren Faktoren beeinflusst werden (vgl. Chandler und Roberts-Young, 1999).

[138] De facto existieren verschiedene Abstufungen von „Privatheit“, die sich z. B. auf Aspekte beziehen, die nur dem Partner, dem besten Freund, dem engeren Freundeskreis, dem weiteren Freundeskreis, flüchtigen Bekannten oder Fremden im Rahmen eines Smalltalks mitgeteilt werden. Unter gewissen Umständen können alle hier aufgeführten Personengruppen bzw. Rollen von einigen privaten Themen ausgeschlossen sein, die ein Individuum z. B. nur einem Arzt, einem Therapeuten oder einem Geistlichen anvertraut.

Fazit

Walters (1996) gelingt es nicht, in ihrer Stichprobe Tendenzen zur allmählichen Herausbildung eines eigenen Genres zu ermitteln. Bates und Lu (1997, S. 335) decken zwar verschiedene Gemeinsamkeiten auf, äußern sich bezüglich dieser Frage jedoch ebenfalls zurückhaltend: „If there is a standardised format for the home page as document, however, we have been unable to identify it yet." Dillon und Gushrowski (2000), die den Aspekt der Rezipientenwahrnehmung berücksichtigen, kommen hingegen zu einem eindeutigen Ergebnis:

> [T]he first unique digital genre has already arrived in the form of the personal home page. [... H]ome pages have no obvious paper equivalent. As such, the personal home page might be thought of as highly idiosyncratic and lacking common form. Yet it appears that the exact opposite is really the case. As the growth of home pages has increased, the genre characteristics of the form have begun to take shape. (Dillon und Gushrowski, 2000, S. 203)

Die in diesem Abschnitt dargestellten Analysen belegen den allmählichen Prozess der Konventionalisierung verschiedener Charakteristika dieser Hypertextsorte. Zugleich zeigen sie jedoch auch verschiedene Probleme auf, die insbesondere die Ebene ihrer Granularität betreffen: Während Dillon und Gushrowski (2000) 25 Merkmale erheben, die sich vornehmlich auf Aspekte der Form beziehen, untersuchen Bates und Lu (1997) 53 Merkmale, die zusätzlich inhaltliche Aspekte beinhalten. Bittner (2003, S. 98) stellt in einer Abbildung neun Eigenschaften des „Kernbereichs" vor. Dass die von den jeweiligen Verfassern gewählten Merkmale nur partiell aufeinander abgebildet werden können, zeigen die Tabellen 4.10 bis 4.12. Es soll an dieser Stelle nicht diskutiert werden, welche nun die *tatsächlich* typischen Eigenschaften persönlicher Homepages sind. Diese Frage entspricht in etwa der Problematik, die typischen Eigenschaften von Werbeanzeigen oder wissenschaftlichen Veröffentlichungen zu benennen. Die Ursachen, die die Beantwortung derartiger Fragen erschweren, sind in der Existenz subgenerischer Varietäten begründet, die in Teil III diskutiert werden.

4.6.4 Die Online-Zeitung

Neben der persönlichen Homepage als Paradebeispiel einer Hypertextsorte, die in den traditionellen Medien kein unmittelbares Pendant besitzt, stehen die Webpräsenzen von Tageszeitungen im Mittelpunkt der Forschungen zu digitalen Genres, weil sie ein Gegenstück in den gedruckten Medien besitzen, das einen direkten Vergleich der Online- und der Papierausgabe erlaubt (vgl. Bucher, 1999, 2000, 2001, Schmitz, 2001, Storrer, 2001c, und Seibold, 2001).

Watters und Shepherd (1997b, S. 23 f.) beschäftigen sich mit der Motivation, die Benutzer dazu veranlasst, digitale Zeitungen zu lesen. Einerseits kann ein spezifischer Bedarf vorliegen, sich über ein bestimmtes Thema zu informieren oder die neuesten Entwicklungen eines Ereignisses zu verfolgen. Diese Art des Zeitunglesens ist in Bezug auf digitale Medien mit der Recherche in Datenbanken und Artikelarchiven verbunden.[139] Online-Zeitungen sind demnach als IR-Systeme aufzufassen, in denen Meldungen und Artikel als eigenständige, recherchierbare Texte vorliegen. Neben der zielgerichteten Lektüre existiert das gewohnheitsmäßige

[139] Eine Alternative stellen Medienbeobachtungsdienste (*clipping services*) dar, die insbesondere von Unternehmen beauftragt werden, um die Reichweite, d. h. den Verbreitungsgrad von Pressemitteilungen zu ermitteln und Informationen über Mitbewerber und deren Produkte einzuholen. Ein ähnlicher Dienst, der sich vornehmlich

Überfliegen einer Zeitung. Auch dieser Modus sollte nach Watters und Shepherd (1997b, S. 24) von Online-Zeitungen bestmöglich unterstützt werden, z. B. durch eine geschickte Anordnung vieler Nachrichten in einem HTML-Dokument und einfache Navigationsmöglichkeiten zwischen Rubriken, so dass das grobe Überfliegen oder die ausführliche Lektüre eines zufällig gefundenen Artikels zu einem „enjoyable part of the day“ wird.[140]

Mehrere schwedische Arbeitsgruppen haben sich ausführlich mit Eigenschaften der Webpräsenzen von Zeitungen beschäftigt. Von diesen Studien werden nachfolgend die wichtigsten in chronologischer Reihenfolge dargestellt, da sie die allmähliche Entwicklung des Hypertexttyps *Online-Zeitung* aufzeigen. Eriksen und Sørgaard (1996) untersuchen in einer Pilotstudie Herstellungsaspekte der Webpräsenzen dreier skandinavischer Tageszeitungen anhand von Interviews mit den beteiligten Redakteuren und Beobachtungen ihres Arbeitsalltags (vgl. Eriksen, 1997).[141] Die Webauftritte der drei Zeitungen gingen 1995 in Produktion, die Befragungen wurden einige Monate später durchgeführt. Während bei den jeweiligen Intentionen zur Einrichtung einer Online-Zeitung durchaus Unterschiede[142] festzustellen sind, ähneln sich die technologischen Bedingungen und Zielvorstellungen über das Aussehen des Produkts sehr: Die online zu publizierenden Artikel aller drei Zeitungen wurden bei einer morgendlichen Redaktionssitzung ausgewählt, in sehr aufwändigen Prozessen aus der verwendeten Software zur Texterfassung nach HTML konvertiert, manuell korrigiert und per FTP auf den Webserver kopiert. Hyperlinks zur Einbindung der neuen Artikel wurden manuell angelegt. Die befragten Mitarbeiter der Zeitungen gaben an, dass das Layout des Webauftritts dem der gedruckten Ausgabe möglichst präzise entsprechen sollte. Die online veröffentlichten Artikel waren also in dieser frühen Phase lediglich digitale Versionen der in der Papierausgabe erschienenen Artikel (vgl. Neuberger et al., 1998). Eriksen (1997, S. 95) stellt fest, dass eine große Diskrepanz zwischen dem technischen Potenzial des Mediums WWW

an Privatpersonen richtet, ist *Google News* (`http://news.google.com`): Vollständig auf automatischen Verfahren basierend werden permanent etwa 4 500 Websites von Agenturen und Zeitungen beobachtet, Nachrichten und unterschiedliche Meldungen zu einer Nachricht identifiziert, gruppiert und dem Leser präsentiert. Stichwortrecherchen können gespeichert werden, so dass bei der Identifikation einer neuen Nachricht zu einem spezifizierten Stichwort eine automatische Benachrichtigung per E-Mail stattfinden kann.

[140] Ausgehend von den Bedürfnissen der Rezipienten haben Bellamy et al. (1999) ein digitales Genre als eine auf das Wesentlichste reduzierte Variante der Online-Zeitung entwickelt, das von drei unterschiedlichen Prototypen dargestellt werden kann. Seine Textexemplare gestatten sowohl ein beiläufiges Überfliegen als auch die ausführliche Lektüre einer Nachricht. Das Genre wird ausschließlich in der Arbeitsgruppe der Verfasser eingesetzt (unter anderem auf der internen Website sowie per Videoprojektor im Aufenthaltsraum der Gruppe) und basiert auf computerlinguistischen Analysen eingehender Nachrichtenströme, die die wesentlichsten Inhalte extrahieren (vgl. Boguraev et al., 1999) und mit Hilfe einer „temporal typography“ genannten Darstellungstechnik in unterschiedlichen Abstraktionsstufen präsentieren.

[141] Sparks (2003) beschäftigt sich mit traditionellen Zeitungen, die Online-Auftritte betreiben und stellt eine Typologie auf, die von Typ 0 (keine Website), Typ 1 (die Website enthält sehr viel weniger Inhalte als die gedruckte Ausgabe) bis zu Typ 6 (Website mit vollkommen unterschiedlichen Inhalten) reicht. Runkehl et al. (1998, S. 145–154) gehen auf deutsche Online-Zeitungen und -Zeitschriften ein.

[142] Bei den untersuchten Zeitungen handelt es sich um *Jyllands-Posten*, eine Tageszeitung aus Dänemark, die Regionalzeitung *Göteborgs-Posten* sowie *Dagbladet*, eine norwegische Boulevardzeitung. Der Webauftritt von *Jyllands-Posten* wurde initiiert, um „serious news articles“ zu publizieren und mittelfristig Einnahmen durch einen Subskriptionsdienst zu generieren (Eriksen, 1997, S. 96). Die Intention des Boulevardblatts *Dagbladet* war es, die Leser zu unterhalten und Werbung für die Printausgabe zu betreiben. Den Verantwortlichen der *Göteborgs-Posten* ging es primär darum, mit den Möglichkeiten der Technologien zu experimentieren.

und seiner Nutzung durch die Zeitungen besteht.[143] In nur einem Fall wurden Artikel speziell für die Online-Ausgabe geschrieben, zwei Webangebote umfassten Zusatzfunktionen wie z. B. rudimentäre Diskussionsforen. Insgesamt charakterisiert Eriksen (1997, S. 94) die drei Online-Zeitungen als das Ergebnis einer initialen Experimentierphase mit den Möglichkeiten des Internet-basierten Publizierens durch das bloße Reproduzieren der Printausgaben.

Eriksen und Ihlström (1999, 2000) haben die Redaktionen dieser Zeitungen drei Jahre später erneut besucht, um die Entwicklung des „web news genre" genauer beschreiben zu können. Aufgrund der gestiegenen Konkurrenzsituation herrschen in allen drei Fällen Trends zur Diversifizierung und zur Besetzung von Nischen. Um auf dem umkämpften Markt bestehen zu können, ist es notwendig, sehr häufig Aktualisierungen vorzunehmen, d. h. neu eingetroffene Agenturmeldungen so schnell wie möglich redaktionell aufzubereiten und auf der Einstiegsseite zu platzieren. Die Online-Zeitung wird nun nicht mehr als eine digitale und im Umfang reduzierte Kopie der Printversion, sondern als eigenständiges Produkt aufgefasst, das das Intervall zwischen der Veröffentlichung zweier Printausgaben überbrücken soll. Weiterhin werden auf den Ebenen der Präsentation und der Konzeption verschiedene gemeinsame Merkmale der für Online-Zeitungen typischen „design language" ermittelt. Die Präsentation von Inhalten basiert auf vier möglichen Elementen: Der „article" stellt eine eigenständige Nachricht dar und ist mit einem Printartikel vergleichbar. Die beiden Elemente „hard composite" und „soft composite" bilden die Eckpunkte eines Kontinuums, das die Aggregation *mehrerer* Artikel in *einem* Präsentationsobjekt beschreibt: Beim „hard composite" werden viele Artikel in einem räumlich begrenzten Bereich präsentiert, beim „soft composite" sind es hingegen nur wenige Artikel.[144] Das Objekt „structure" enthält Navigationselemente. Die fünf Elemente der konzeptionellen Ebene sind nach Eriksen und Ihlström als unterschiedliche „views" auf den Artikelbestand zu betrachten: Das Element „headlines" umfasst die Schlagzeilen der Nachrichten, die von der Redaktion als interessant und für eine möglichst große Zielgruppe relevant erachtet und – wie bei einer Papierausgabe – als „soft composite" auf der Einstiegsseite dargestellt werden. Der „news stream" umfasst Hyperlinks zu Artikeln in umgekehrt chronologischer Reihenfolge und wird üblicherweise als „soft composite" realisiert; ein traditionelles Pendant existiert nicht. Bezüglich der inhaltlichen Darstellung von Rubriken werden „thematic hard news" (spezifischen Zeitpunkten zugeordnete Nachrichten aus den klassischen Rubriken) und „thematic soft news" („zeitunabhängige" Meldungen) unterschieden. Abschließend ist das „archive" als fester Bestandteil einer Online-Zeitung aufzufassen, der ebenfalls kein traditionelles Pendant besitzt. Ursprünglich kon-

[143] Auch Dillon und Gushrowski (2000, S. 202 f.) weisen auf dieses Problem hin, das sie durch den Einsatz korrespondierender Metaphern erklären: „In the digital domain many existing paper-based conventions are adopted in the hope that such familiarity of form will leverage user comprehensions, i. e., web versions of such paper formats as newspapers and magazines frequently adhere so closely to type that users can rely on their experiences with physical newspapers [. . .]." Die Orientierung an Metaphern, die auf etablierten Textsorten basieren, kann Dillon und Gushrowski zufolge eine restriktive Wirkung besitzen, da diese Konventionen möglicherweise den Anforderungen des Mediums widersprechen (z. B. bezüglich des Textdesigns), und das bloße Nachahmen gedruckten Papiers die Entwicklung innovativer Informationsstrukturen verhindert.

[144] Es handelt sich um unterschiedliche Typen von Hyperlinklisten (vgl. Abschnitt 4.6.6, S. 243 ff.): Je mehr Artikel in einem Objekt präsentiert werden sollen, desto weniger Platz bleibt für die genauere Beschreibung der jeweiligen Nachricht. Bei einem „soft composite" könnte z. B. die Überschrift als Hyperlink realisiert und von einem Zeitstempel und dem Anreißer flankiert sein (Eriksen und Ihlström, 1999, S. 303).

zipiert als digitale Version der gedruckten Ausgabe, beobachten Eriksen und Ihlström (2000, S. 9) eine zunehmende Eigenständigkeit von Online-Zeitungen, die sich in einem „"live" scheme of news reporting" manifestiert, d. h. die konzeptionelle Struktur und die Präsentationselemente bleiben fixiert, während immer wieder neue Artikel und Nachrichten durch die Website geschleust[145] und sukzessive um neue Fakten ergänzt werden, damit permanent das Kriterium der Aktualität gewährleistet ist. Gerade weil Online-Zeitungen im Gegensatz zu Printpublikationen aktualisiert werden können, hat sich die neuartige Form der chronologischen Präsentation von Nachrichten etabliert. Im Vergleich zu gedruckten Zeitungen fällt ein entscheidenes Merkmal weg: Der Inhalt eines Artikels kann sich – ebenso wie seine Adresse – ändern, das Konzept der „Ausgabe einer Online-Zeitung", die sich auf ein bestimmtes Datum bezieht, existiert nicht. Folglich entsteht für den Leser ein Problem: Wenn nicht explizit bekannt ist, in welchen Intervallen eine Online-Zeitung aktualisiert wird, könnte der Fall eintreten, dass Nachrichten, die für einen Leser von Interesse sind, nicht bemerkt werden, weil sie beim nächsten Besuch schon wieder von der Einstiegsseite entfernt worden sind (Eriksen und Ihlström, 2000, S. 10), was jedoch dem vom Benutzer wahrgenommenen Konzept einer Online-Zeitung als „digitale Version einer gedruckten Zeitung" widerspräche.[146]

Ihlström und Lundberg (2003) zeigen anhand von Interviews mit 153 Lesern von neun schwedischen Online-Zeitungen und einer parallelen Inhaltsanalyse der Websites, dass sich zwar gemeinsame Merkmale etabliert haben, die als Genre-spezifische Abkehr von traditionellen Zeitungen betrachtet werden können, aber dennoch fasst etwa die Hälfte der Rezipienten die Online-Zeitung – insbesondere hinsichtlich struktureller Aspekte – als digitale Replik der gedruckten Zeitung auf.[147] Aus diesem Grund bezeichnen Ihlström und Lundberg (2003, S. 9) die Online-Zeitung als „variant genre" (vgl. Abschnitt 4.3.2). Ihlström und Lundberg schlagen ein „repertoire of genre elements" vor, das eine Charakterisierung auf zwei Ebenen zulässt: Von der Navigationsperspektive aus betrachtet, werden „navigation elements" (Menüs, die Titelzeile, die URL etc.) und „landmarks" (z. B. das Logo eines Unternehmens, vgl. auch die Abschnitte 3.6.1 und 4.6.2 sowie Nielsen und Tahir, 2002) unterschieden. Als Genre-spezifische Elemente werden „news stream", „headlines", „search/archive" und „advertisements" genannt (vgl. Eriksen und Ihlström, 2000). Die Inhaltsanalyse deckt die Verwendung fünf unterschiedlicher Navigationselemente auf: „menus", „bars", „tabs", „banners" und „breadcrump trails". Ihlström und Lundberg (2003, S. 5) gehen insbesondere auf die Positionierung dieser Elemente ein, die in einigen Fällen kanonisch erfolgt: Das Logo befindet sich bei allen Online-Zeitungen am oberen Seitenrand (links oder in der Mitte); dort wird auch ein Werbebanner dargestellt. Fast alle Websites enthalten am linken Rand eine Navigationshilfe und bieten im mittleren Bereich die Schlagzeilen an.

[145] Eriksen und Ihlström (2000, S. 9) geben hierfür ein plastisches Beispiel an: „An article on a revolutionary chip design may be present in the news stream for 6 hours, while at the same time spending 12 hours in the headlines, and 36 hours in a hard news section. From there it may migrate to the IT soft section as background information for half a year, where after [sic] it is replaced by a new story on a revolutionary chip design."

[146] Es ist zu vermuten, dass sich in den kommenden Jahren weitere Merkmale dieses Genres etablieren werden, um diesem Effekt entgegenzuwirken, z. B. Personalisierungsfunktionen mit individuellen Beobachtungsdiensten, wie sie bereits in Fußnote 139 (S. 236) angesprochen wurden.

[147] Zu diesem Schluss kommen auch Neuberger et al. (1998), die die Websites deutscher Online-Zeitungen analysiert und Befragungen der Redaktionen und Rezipienten durchgeführt haben und deren Ergebnisse sich mit den Analysen von Eriksen und Sørgaard (1996) und Eriksen (1997) weitgehend decken.

Ihlström und Åkesson (2004) fokussieren ebenfalls das Webdesign und stellen eine Methodik zur Ermittlung der Eigenschaften spezifischer Web-Genres vor, die anhand einer Analyse der Einstiegsseiten von 85 schwedischen Online-Zeitungen exemplifziert wird (vgl. auch Åkesson, 2003). Die Verfasserinnen beziehen sich auf Shepherd und Watters (1998, 1999) und gehen davon aus, das ein Web-Genre durch das Quadrupel „<content, form, functionality, positioning>" beschrieben werden kann. Das vierte Merkmal bezieht sich auf die Positionierung von Inhaltselementen, die in einem fünf Zeilen und vier Spalten umfassenden Gitternetz verortet werden.[148] Diese Zuordnung wurde für alle 85 Einstiegsseiten durchgeführt, wobei auch „form" und „functionality" bestimmt wurden, um zu einer Liste der „genre characteristics specific for online newspapers" zu gelangen (Ihlström und Åkesson, 2004, S. 4). Für jedes Merkmal wurde bestimmt, ob es von gedruckten Zeitungen übernommen wurde. Die Ergebnisse werden der Taxonomie von Cybergenres (Shepherd und Watters, 1998) zugeordnet (vgl. Abschnitt 4.3.2), wobei aber keine der Online-Zeitungen als „emergent" (d. h. „novel") klassifiziert wird, da sich die beobachteten Formen noch zu sehr an gedruckten Zeitungen orientieren. Die ermittelte Bandbreite ist derart umfangreich, dass neben den Entwicklungsstufen „replicated" und „variant" (unter dem Knoten „extant") eine dritte Stufe namens „progressed" eingeführt wird. Die Zuordnung ergibt, dass 13 der 85 Online-Zeitungen replizierte Genres besitzen, 49 Einstiegsseiten sind Varianten gedruckter Zeitungen und 23 Websites basieren auf einer fortgeschrittenen Adaption dieses Genres. Der wesentlichste Unterschied zwischen den drei Gruppen bezieht sich auf die Spaltenanzahl: „the more evolved [...] the more columns [...]. Most online newspapers have an additional column to the right for advertisements, 43 have it as their forth [sic] column and 8 as their fifth." (Ihlström und Åkesson, 2004, S. 6). Die Verfasserinnen kommen zu dem Schluss, dass sich Online-Zeitungen bezüglich ihres Layouts wieder vermehrt an gedruckten Zeitungen orientieren, was sich insbesondere in der Anordnung in mehreren Spalten widerspiegelt.[149]

Bucher et al. (2004) führen eine Rezeptionsstudie der E-Paper-Ausgabe der *Rhein-Zeitung* durch, die eine über das WWW zugängliche Variante der Zeitung darstellt, denn sie umfasst faksimilierte Seiten der gedruckten Ausgabe. Während sich Instanzen des Hypertexttyps *Online-Zeitung* in zunehmender Weise an der Gestaltung ihrer gedruckten Pendants orientieren, sehen viele Redaktionen traditioneller Zeitungen einen Bedarf für die alternative Publikation einer digital reproduzierten Variante der Druckausgabe. Die Reproduktion erfolgt entweder durch die Präsentation einer PDF-Version oder durch die Integration einer Einzelseite als eingebettete Grafikdatei in einen mit unterschiedlichen Mitteln navigierbaren HTML- bzw. XML-basierten Hypertext. Bucher et al. (2004, S. 48) zeigen, dass die Rezipienten sowohl für die Tageszeitung als auch für die Online-Zeitung jeweils charakteristische „Nutzungs-Skripts" entwickelt haben, die jedoch zur Aufmerksamkeits- und Rezeptionssteuerung der E-Paper-Version noch nicht geeignet sind. Die Verfasser kommen zu dem

[148] Die erste Zeile („A") wird als „header of front page" bezeichnet (Ihlström und Åkesson, 2004, S. 3), die zweite Zeile entspricht dem Bildschirminhalt bei einer Auflösung von 1 280 × 1 024 Punkten (eingeteilt in eine obere, „Bt[op]", und eine untere Hälfte, „Bb[ottom]"). Die dritte Zeile („C") ist der „rest of front page" [sic], und die vierte Zeile ist ihr „footer" („D").

[149] Hierfür ist meines Erachtens insbesondere die zunehmende Ablösung nicht mehr aktueller Hardware verantwortlich, denn die Akzeptanz eines derartigen Layouts basiert auf der Prämisse, dass die Mehrzahl der Rezipienten hochauflösende Bildschirme einsetzt, mit deren Hilfe die Websites der Online-Zeitungen in voller Breite, d. h. ohne im Browser Rollbalken zu erzwingen, dargestellt werden können (vgl. Fußnote 29, S. 172).

Schluss, dass der Vorteil der E-Paper-Variante in der Komplementärfunktion zur gedruckten Ausgabe besteht, der sich am besten durch den Einsatz des kurz vor der Marktreife stehenden „digitalen Papiers“ als Trägermedium ausspielen ließe. Zusätzlich wird eine Anreicherung der E-Paper-Version durch multimediale Elemente sowie weiterführende Navigations- und Suchmöglichkeiten vorgeschlagen, weil die Nutzer nicht bereit sind, „auf den internetspezifischen Bedienungskomfort [... zu] verzichten“ (ebd.).

4.6.5 Die Online-Enzyklopädie

Im WWW stehen verschiedene Online-Enzyklopädien zur Verfügung, von denen einige kommerzielle Produkte darstellen, weshalb eine Subskriptionsgebühr zu entrichten ist. Neben diesen qualitativ hochwertigen Enzyklopädien, die in aller Regel auf der jeweiligen Printausgabe basieren, existieren jedoch auch Angebote, die von Freiwilligen gepflegt und um neue Artikel ergänzt werden. Der prominenteste Vertreter ist die Wikipedia (`http://www.wikipedia.org`), deren Inhalte nach dem Prinzip des Wikis erstellt werden (vgl. Kleinz, 2004). In einem Wiki besitzt *jeder* Leser die Möglichkeit, ein HTML-Dokument zu ändern, um z. B. neue Informationen beizutragen oder fehlerhafte Angaben zu entfernen – das Wiki-Prinzip realisiert also die von nahezu allen Hypertexttheoretikern beschriebene Konvergenz von Autor und Leser im wortwörtlichen Sinne (vgl. auch Abschnitt 3.7).[150] Die Wikipedia fußt auf dem Gedanken, dass interessierte Leser Einträge rezipieren und zu einem Schlagwort gegebenenfalls zusätzliche Hintergrundinformationen eintragen oder neue Artikel erstellen, in denen mit Hilfe einer speziellen Notation Hyperlinks zu anderen Schlagwörtern hinterlegt werden. Auf der Basis dieses hochgradig dynamischen Prozesses sind mittlerweile Wikipedia-Versionen in zahlreichen Sprachen entstanden, die englische Wikipedia (`http://en.wikipedia.org`) umfasst z. B. etwa 577 000 Artikel (Stand: 31.05.2005), die im Mai 2001 ins Leben gerufene deutschsprachige Variante zählt etwa 238 000 Einträge (`http://de.wikipedia.org`), die von mehr als 40 000 Freiwilligen gepflegt werden. Sie besitzen einen sehr großen Abdeckungsgrad und eine inhaltliche Qualität, die über vergleichbare kommerzielle Produkte hinausgeht (vgl. Kurzidim, 2004).[151] Auf der Grundlage dieser immensen Anzahl unsequenziert vernetzter Knoten kann die Wikipedia als eines der wenigen Beispiele für prototypische Hypertexte im

[150] Ein Wiki (von „wikiwiki“, Hawaiianisch für „schnell“) ist in technischer Hinsicht mit sehr einfachen Mitteln zu realisieren, z. B. mit Hilfe eines *Perl*-Skriptes. Da hierdurch jedoch die Sicherheit des Webservers beeinträchtigt werden kann, operieren die meisten Wiki-Plattformen auf relationalen Datenbanken, in denen Revisionen gespeichert werden, so dass fehlerhafte oder mit böswilliger Absicht durchgeführte Modifikationen rückgängig gemacht werden können. Ein Wiki kann somit als sehr spezielle Ausprägung eines WWW-basierten CMS betrachtet werden: Jedes Dokument enthält einen Hyperlink wie z. B. „edit page“, der zu einer Eingabemaske führt, in der die Inhalte bearbeitet werden können. Das erste Wiki wurde von Ward Cunningham in *Perl* implementiert, am 25. März 1995 in Betrieb genommen und diente als Werkzeug zum Wissensmanagement (vgl. `http://c2.com/cgi/wiki?WikiHistory`).

[151] Da jeder Leser Änderungen vornehmen kann, existieren Fälle von Vandalismus (Artikel werden z. B. gelöscht, mit anstößigen Inhalten versehen oder um irrelevante Inhalte ergänzt), die jedoch durch Schutzvorrichtungen schnell erkannt und behoben werden können (vgl. Viégas et al., 2004). Ein weiteres Phänomen bezieht sich auf Artikel, die kontroverse Themen behandeln: Häufig entstehen „edit wars“ (Viégas et al., 2004, S. 580), in denen Anhänger unterschiedlicher Lager einen Eintrag jeweils zugunsten ihrer politischen oder ethischen Weltanschauung modifizieren. In besonders schweren Fällen wird ein Artikel vom Administrator auf unbestimmte Zeit „eingefroren“, so dass keine weiteren Modifikationen mehr möglich sind.

WWW gelten. Ein besonderes Merkmal aller Wikipedia-Enzyklopädien betrifft den rechtlichen Status der Inhalte, die der *GNU Free Documentation License* (GFDL) unterliegen.[152]

Gedruckte oder kommerziell auf CD ROM vertriebene Enzyklopädien werden von Experten erstellt und unterliegen einer Qualitätskontrolle, um eine bestmögliche Aktualität und eine umfassende thematische Abdeckung zu gewährleisten. Die Wikipedia-Angebote basieren hingegen auf dem Prinzip der kollaborativen Erstellung von Inhalten (vgl. Kuhlen, 2004). Es existiert keine zentrale Kontrollinstanz. Auch bei den Autoren handelt es sich vermutlich nur in den wenigsten Fällen um Personen, die in beruflicher Hinsicht mit der Erstellung von Enzyklopädieartikeln zu tun haben. Ein Vergleichstest hat gezeigt, dass die Wikipedia bezüglich ihrer Inhalte und insbesondere hinsichtlich der Aktualität den meisten kommerziellen Produkten überlegen ist (jedoch nicht im Bereich der Multimedialität; vgl. Kurzidim, 2004). Es stellt sich jedoch in diesem Zusammenhang die Frage, welche stilistischen Phänomene in den Artikeln beobachtet werden können, ob also Merkmale, die für gedruckte Enzyklopädien beschrieben wurden, auch in der Wikipedia ermittelt werden können. Dieser Frage gehen Emigh und Herring (2005) nach, die jeweils 15 Artikel aus der englischsprachigen Wikipedia, der Website Everything2 und der traditionellen, jedoch auch online verfügbaren *Columbia Encyclopedia* analysieren (z. B. Einträge zu „British Empire“, „Karl Marx“, „Mind the Gap“ und „Sing Sing“).[153] Die Untersuchung stützt sich auf das von Biber (1988) vorgeschlagene Verfahren, die „formality“ bzw. „informality“ eines Textkorpus durch die Erhebung der Frequenzen zugehöriger linguistischer Merkmale zu ermitteln (z. B. Assimilationen und Personalpronomina).[154] Emigh und Herring zeigen, dass zwischen der Wikipedia und der *Columbia Encyclopedia* keine Unterschiede existieren. Die online von Freiwilligen erstellte Enzyklopädie besitzt demnach einen „degree of formality“, der dem traditionellen, von Experten angefertigten Pendant entspricht. Im Gegensatz dazu sind die aus Everything2 stammenden Artikel eher informeller Natur. Eine qualitative Analyse der Beiträge belegt, dass die Wikipedia keine umgangssprachlichen oder informellen Ausdrücke enthält, von stilistischer Homogenität geprägt ist und sich auf die Erläuterung eines Schlagwortes in Bezug auf eine spezifische Lesart beschränkt; weiterhin liegt ein einheitliches und standardisiert erscheinendes Gestaltungsformat vor (z. B. konventionalisierte Überschriften). Die Einträge in Everything2 enthalten dagegen sehr viele informelle und auch bewertende Ausdrücke, die sich auf unterschiedliche Lesarten eines Konzeptes beziehen.

Von einem einheitlichen Genre der Online-Enzyklopädie kann also Emigh und Herring (2005, S. 9) zufolge nicht gesprochen werden. Die Wikipedia und Everything2 weisen zwar

[152] Die *GNU Free Documentation License* wurde ihrerseits von der GPL abgeleitet, die vielen Software-Paketen, die als freie bzw. Open-Source-Software vertrieben werden, zugrunde liegt (vgl. Rehm und Lobin, 2003).

[153] Everything2 ist nach eigenen Angaben eine „online community with a focus to write, publish and edit a quality database of information, art and humor.“ (`http://www.everything2.com`). Diese Website wendet andere Erstellungsprinzipien: Es können nur registrierte Benutzer neue Knoten anlegen, und nur der Autor eines Knotens ist berechtigt, Änderungen vornehmen zu können. Zusätzlich existiert ein Ranking-System, das auf Punkten basiert, die beispielsweise bei der Anlegung eines neuen Knotens oder der Bewertung eines existierenden Knotens vergeben werden (vgl. Emigh und Herring, 2005, S. 3). Während die Wikipedia der neutralen Darstellung von Sachverhalten verpflichtet ist, schildern die Everything2-Autoren („noders“, in Anlehnung an „coders“) in vielen Fällen – und oftmals in humoristischer Form – ihre subjektive Meinung.

[154] Die Unterscheidung „formality vs. informality“ entspricht in etwa der Differenzierung zwischen konzeptioneller Schriftlichkeit und Mündlichkeit (vgl. Abschnitt 2.2.7). Ein ähnliches Verfahren wird in der ersten Stichprobenanalyse angewendet (vgl. Kapitel 8, S. 367 ff.).

Gemeinsamkeiten auf – beide streben die Herstellung einer möglichst umfangreichen Sammlung von Allgemeinwissen an, sind online verfügbar, werden von zahlreichen Privatpersonen erstellt, sind recherchierbar und verwenden Hyperlinks. Dennoch rechtfertigen es diese Korrespondenzen nicht, von zwei Instanzen *eines* Genres (im Sinne von Yates und Orlikowski, 1992) zu sprechen. Emigh und Herring erklären die stilistischen Diskrepanzen zwischen Everything2 und Wikipedia sowie die auffälligen Gemeinsamkeiten zwischen Wikipedia und der *Columbia Encyclopedia* durch die unterschiedlichen Arten der Erstellung von Inhalten, so dass ein gemeinsames übergeordnetes Genre (d. h. ein Hypertexttyp) „online knowledge repository" angenommen werden kann, wobei jedoch unterschiedliche Ausprägungen als „online collaborative authoring environment" vorliegen (Emigh und Herring, 2005, S. 9). Bereits Yates und Orlikowski (1992) haben darauf hingewiesen, dass Eigenschaften des Mediums die Entwicklung von Konventionen beeinflussen können. Im vorliegenden Fall handelt es sich dabei um die redaktionellen Mechanismen, die in Everything2 und Wikipedia eingesetzt werden. Die sehr offenen, liberalen und demokratischen Prinzipien der kollaborativen Erstellung von Inhalten mit Hilfe einer Wiki-Plattform führen in der Wikipedia zur Reproduktion von Normen, die aus dem Printbereich bekannt sind. Als Ursache werden zwei soziale Faktoren genannt: Einerseits orientiert sich der Großteil der Autoren an den Normen, die sie einer Enzyklopädie intuitiv zuschreiben, andererseits werden diese Normen zusätzlich durch eine kleine Gruppe sehr engagierter Benutzer forciert, die Abweichungen eigenständig an die Konventionen anpassen.[155] Im Falle der Wikipedia liegt also Emigh und Herring zufolge die Reproduktion eines Genres vor, während Everything2 als „a blend of discussion forum and knowledge repository" und somit als „emergent genre" charakterisiert werden kann (ebd.).

4.6.6 Die Hotlist

Hotlists – in der Regel von Autoren persönlicher Homepages zusammengestellte, sehr umfangreiche Listen von Hyperlinks zu interessanten, nützlichen und auch amüsanten Websites – waren im WWW bis etwa 1998 allgegenwärtig. Diese Hypertextsorte hat sich vermutlich auf Basis der thematischen Auflistung von FTP-Servern und Usenet-Newsgroups gebildet, die seit den achtziger Jahren verwendet wurden, um das unstrukturierte und kaum überblickbare Angebot von Ressourcen zu gliedern, die in dieser Zeit ausschließlich per FTP und NNTP erreichbar waren (vgl. auch Crowston und Williams, 2000, S. 208). Mittlerweile hat die Verbreitung von Hotlists deutlich abgenommen, was verschiedene Ursachen hat: Zunächst haben sich Suchmaschinen (und zu einem gewissen Grad auch Webkataloge, die die Aufgabe von Hotlists übernommen haben, vgl. Brandl, 2002, S. 96) als alltägliches Handwerkszeug etabliert, so dass Benutzer nicht mehr auf externe Listen von Angeboten (oder gar gedruckte Sammlungen von URLs mit der Stärke eines Telefonbuchs, z. B. Maxwell und Grycz, 1994) angewiesen waren, um das WWW zu erkunden. Parallel zum Erfolg des WWW wurden auch andere Internet-Dienste immer häufiger eingesetzt, insbesondere die elektronische Post, über die vermehrt URLs an Arbeitskollegen, Freunde und Bekannte verschickt werden, insbesondere diejenigen der eingangs zuletzt genannten Kategorie. Des Weiteren fällt die „nüchterne, schmucklose Darstellung" (Brandl, 2002, S. 96) von Hotlists im Ver-

[155] Der erste Aspekt dieses Befundes kann als weiterer Indikator für die Validität des in Abschnitt 4.3.2 dargestellten Modells der Entwicklung von Hypertextsorten betrachtet werden.

gleich zu anderen Webangeboten deutlich ab, sie gelten schlicht als nicht mehr zeitgemäß und werden nur noch auf themenspezifischen Websites eingesetzt, um weiterführende Links zu einem speziellen Themenbereich anzubieten.[156] Die ursprüngliche Funktion der Hotlist als implizite Darstellung der eigenen Interessen auf der persönlichen Homepage wird seit einigen Jahren von Weblogs übernommen (vgl. Abschnitt 4.6.7).

Die Hypertextsorte Hotlist wird in den meisten Arbeiten zu Web-Genres nur am Rande thematisiert. Crowston und Williams (2000, S. 208) definieren sie als eine Liste von Hyperlinks zu weiterführenden Webangeboten (zu einem oder mehreren Themen), die nicht vom Autor der Hotlist selbst kontrolliert werden. Eine Hotlist enthält somit – im Gegensatz zu einer Einstiegsseite, einem Index oder einem Inhaltsverzeichnis – sehr viele abgehende und nur wenige eingehende Hyperlinks. Furuta und Marshall (1996) gehen davon aus, dass Hotlists, die meist auf persönlichen Homepages veröffentlicht werden, den Status von Annotationen in experimentellen bzw. kommerziellen Hypertextsystemen besitzen (vgl. Fußnote 5, S. 68): Die Aufnahme der Adresse einer Website in eine Liste wie „Die besten Sites im Web“ oder „Die langweiligsten Sites im Web“ fungiert als eine Art Annotation, weil sie die Meinung des Rezipienten reflektiert und vom Produzenten nicht unmittelbar eingesehen werden kann (vgl. Walker, 2000, S. 104).[157] Crowston und Williams finden in ihrer Stichprobe von 837 Dokumenten 26 Hotlists zu Themenbereichen wie „music, HTML, nanotechnology, films, environmental organizations, computer stores, and presidential candidates.“ (ebd.). Als eine Art Vorläufer der Hotlist gelten Bookmark-Dateien, die die vom Benutzer im Browser gepflegte Liste von Lesezeichen umfassen (vgl. Abschnitt A.4.3, S. 754 ff.). Crowston und Williams finden in ihrer Stichprobe 22 derartige Dateien, die jedoch weniger strukturiert als Hotlists und meist nicht thematisch sortiert sind. Weiterhin identifizieren Crowston und Williams (2000, S. 208) ein weiteres Web-Genre, das sie als „topical home page“ bezeichnen. Diese ist vergleichbar mit der Hotlist, da es ihre primäre Funktion ist, Hyperlinks zu bestimmten Themen zu präsentieren, der Unterschied besteht jedoch in der Menge der angebotenen Informationen: Hotlists bestehen ausschließlich aus Hyperlinks und sortierenden Überschriften, „topical home pages“ bieten zusätzlich einen Themenüberblick und möglicherweise auch Erläuterungen oder Kommentare zu Verknüpfungen an. Da hierdurch für den Rezipienten ein Mehrwert entsteht, gehen Crowston und Williams (2000, S. 209) davon aus, dass „topical home pages“ Hotlists langfristig verdrängen werden.

4.6.7 Das Weblog

Weblogs (kurz: Blogs) werden von einzelnen Personen, Arbeitsgruppen oder Interessengemeinschaften – in jüngerer Zeit auch von Firmen, Politikern, Künstlern, Wissenschaftlern und Journalisten (vgl. Rosenbloom, 2004, S. 32) – betrieben und enthalten Einträge, die in umgekehrt chronologischer Reihenfolge dargestellt werden. Die Einträge selbst sind in

[156] Eine nach thematischen Gesichtspunkten sortierte und gegebenenfalls kommentierte Hotlist kann als Zusammenstellung relevanter Hypertexte aufgefasst werden. Im Printbereich übernehmen (kommentierte) Bibliografien diese Funktion, die zumeist Fachliteratur zu einem bestimmten Themenkomplex bündeln.

[157] HTTP übermittelt bei der Aktivierung eines Hyperlinks jedoch den *referrer* (vgl. RFC 2616), d. h. auf dem Webserver, der den Zielknoten bereitstellt, kann die URL desjenigen Knotens protokolliert werden, der den aktivierten Hyperlink enthält. Auf diese Weise können mit etwas technischem Aufwand Listen der Adressen erstellt werden, die auf eine bestimmte Seite verweisen.

der Regel kurze Textabschnitte zu tagesaktuellen Themen oder spezifischen Themenbereichen und verweisen auf andere Websites oder sie beschreiben Ereignisse und Anekdoten aus dem Leben des Autors. Sie fungieren somit tatsächlich als eine Art Web-Logbuch, da neue Einträge im Regelfall mehrmals pro Tag hinzugefügt werden, weshalb Killoran (2002) sie als Nischenmedium charakterisiert, das einen „guerilla-like style of Web publishing" ermöglicht. Der Begriff „Weblog" geht auf Jorn Barger, den Betreiber von „Robot Wisdom" zurück, das als eines der ersten Blogs gilt (`http://www.robotwisdom.com`). Barger definierte „Weblog" 1997 als „[a] Web page where a Web logger 'logs' all other Web pages she finds interesting." (Blood, 2004, S. 54). Ursprünglich fokussierten die Betreiber von Weblogs die Kommentierung externer Hyperlinks, weshalb sie vom Standpunkt der Hypertexttheorie als eine weitere WWW-spezifische Art der Realisierung von Annotationen einzelner Knoten betrachtet werden können (vgl. Fußnote 5, S. 68).[158] Weblogs werden meist mit einer Software gepflegt, die das Schreiben neuer Artikel komfortabel gestaltet, eine Aktualisierung kann somit sehr schnell erfolgen. Zusätzlich ermöglichen es diese Pakete den Rezipienten, an Einträgen Kommentare zu hinterlassen – falls der Autor eines Blogs vornehmlich individuelle Annotationen externer Ressourcen anbietet, können von den Lesern daher diese Annotationen zusätzlich annotiert werden. Neben privaten Blogs existieren weitere Typen, z. B. themenspezifische oder auf eine bestimmte Benutzergruppe ausgerichtete Weblogs.[159]

Herring et al. (2004, S. 1) charakterisieren das „emergent blog genre" anhand einer Inhaltsanalyse, die auf 203 Weblogs basiert.[160] Die Verfasser haben insgesamt 44 Merkmale betrachtet, unter anderem Angaben zum Autor eines Weblogs, seinen Zweck (z. B. „filter" oder „personal journal"), die Anzahl enthaltener Hyperlinks und die verwendete Softwa-

[158] Hotlists (vgl. Abschnitt 4.6.6) besitzen ebenfalls diese Funktion, falls sie mit Kommentaren versehen werden. Im Gegensatz zu Blogs werden Hotlists jedoch weder in regelmäßigen Abständen aktualisiert noch in der Reihenfolge der Eintragung sortiert. Neben Hyperlinks können die Einträge eines Blogs beliebige weitere Inhalte umfassen. Herring et al. (2005) untersuchen die Vernetzung von Weblogs untereinander und zeigen, dass eine Klasse von etwa 100 „A-list blogs" existiert, auf die überproportional häufig von den „B-list blogs" verwiesen wird. Im Gegensatz dazu verknüpfen „A-list blogs" in der Regel Einträge in anderen „A-list blogs".

[159] Die von mehreren Redakteuren gepflegte Website „Slashdot" (`http://slashdot.org`) ist ebenfalls eine Art Weblog, allerdings werden dort veröffentlichte Neuigkeiten in der Regel von den Besuchern an die Betreiber geschickt, die daraufhin über die Publikation entscheiden. Der eigentliche Mehrwert entsteht durch die Kommentare und zusätzlichen Informationen der zahlreichen Stammgäste, die bei sehr brisanten Themen oder Ereignissen deutlich mehr als 1 000 Einträge hinterlassen. In dieser ursprünglich auf die Themen *Linux* und freie Software ausgerichteten Community-Site hat sich eine feste Terminologie entwickelt. Besonders bekannt ist der „Slashdot effect" (vgl. Halavais, 2001): Innerhalb von Sekunden greifen mehrere tausend Slashdot-Leser gleichzeitig auf einen Webserver zu, der in einer neu veröffentlichten Nachricht genannt wird, wodurch der Server – falls es sich nicht um einen Hochleistungsrechner mit hervorragender Netzanbindung handelt – unter der Last der eingehenden Verbindungen zusammenbricht.

[160] Analysiert wurde jeweils die Einstiegsseite sowie der jüngste Eintrag. Die Weblogs wurden im Mai 2003 mit Hilfe einer Zufallsfunktion von der Website `http://blo.gs` gesammelt. Nach Herring et al. (2004, S. 3) waren dort am 26.09.2003 insgesamt 710 755 Weblogs gelistet. Mittlerweile ist die Zahl auf 15 780 225 angestiegen (09.08.2005). Blood (2004, S. 55) kommentiert die Popularität von Weblogs: „When I began blogging, I imagined that someday there might be hundreds of Weblogs, with tens of thousands of readers. [...] Instead of dozens of Weblogs with a million readers, there are now well over four million Weblogs worldwide – most with only a few dozen readers [...]." Kommunikationswissenschaftliche und soziolinguistische Betrachtungen des Genres Weblog werden in dem Online-Journal *Into the Blogosphere* publiziert (vgl. `http://blog.lib.umn.edu/blogosphere/`). Der Terminus „Blogosphere" bezieht sich auf die Gesamtheit aller Blogs (Rosenbloom, 2004), die häufig auch als „Blogspace" (Kumar et al., 2004) bezeichnet wird.

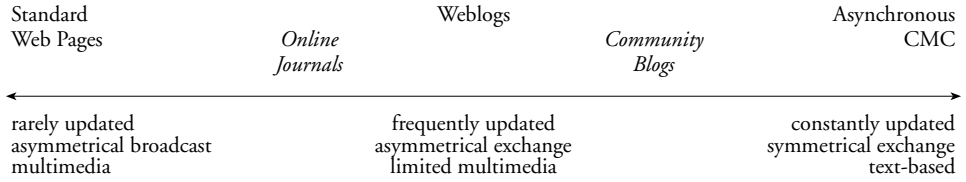


Abbildung 4.10: Weblogs auf einem Kontinuum zwischen HTML-Dokumenten und asynchroner computervermittelter Kommunikation (nach Herring et al., 2004)

re. Etwa 90% der analysierten Blogs werden von nur einem Verfasser gepflegt; in 55% der Blogs machen die Verfasser Angaben zu ihrer Tätigkeit: 58% der Autoren sind Studierende, 19% arbeiten als Systemadministratoren, Web-Entwickler oder Programmierer.[161] Die meisten Blogs gehören dem „personal journal type" an (70%), „in which authors report on their lives and inner thoughts and feelings" (Herring et al., 2004, S. 6). Die Darstellung einer subjektiven und gelegentlich auch intimen Perspektive zu individuellen Interessengebieten und Ereignissen des Alltags identifizieren Herring et al. als gemeinsamen Zweck aller untersuchten Weblogs.[162] Bezüglich struktureller Charakteristika berichten Herring et al., dass die meisten Weblogs deutlich weniger Gästebücher, Suchfunktionen oder Werbeanzeigen enthalten als private Homepages. Gleichzeitig umfassen sie aber auch inhaltliche Bereiche, die in private Homepages üblicherweise nicht vorhanden sind, z. B. Archive, also Links zu älteren Einträgen (74%); derartige Funktionen gehen unmittelbar auf die Möglichkeiten der Software zurück, die einem Weblog zugrunde liegt. Weiterhin kommen Herring et al. zu dem Schluss, dass Weblogs im Schnitt alle 2,2 Tage aktualisiert werden, wobei die Rezipienten an einem Eintrag durchschnittlich 0,3 Kommentare hinterlassen. Ein Eintrag enthält im Schnitt 210 Wörter bzw. 16 Sätze und 3,5 Absätze und wird fast immer von einer Überschrift sowie einem Zeitstempel begleitet, der Datum und Uhrzeit markiert.

Die unterschiedlichen Typen von Weblogs fußen Herring et al. (2004, S. 10) zufolge nicht auf einem spezifischen Genre, sondern werden als Hybrid aufgefasst, der Eigenschaften von Tagebüchern, Leitartikeln, Leserbriefen, Notizzetteln, Reiseberichten, Fotoalben, E-Mails und Presseschau in sich vereint (vgl. auch Simanowski, 2004, S. 209 ff., sowie Bucher, 2004, S. 162). Zusätzlich teilen sich Blogs funktionelle Gemeinsamkeiten mit privaten bzw. persönlichen Homepages, die ursprünglich die präferiertere Methode darstellten, eigene Interessen und Ansichten im WWW zu präsentieren. Des Weiteren werden Gemeinsamkeiten mit Medien der *Computer-Mediated Communication*, insbesondere asynchronen Diskussionsforen ermittelt (vgl. Efimova und de Moor, 2005). Nach Ansicht von Herring et al. (2004,

[161] Kumar et al. (2004) liefern weitere demografische Angaben und Analysen gemeinsamer Interessen von Blog-Autoren. Diese Studie basiert auf der Untersuchung der von ca. 1,3 Mio. Autoren verfassten persönlichen Profile, die auf der Website `http://www.livejournal.com` verfügbar sind: Weblogs werden auf allen Kontinenten gepflegt, die meisten Autoren sind zwischen 16 und 24 Jahren alt.

[162] Nardi et al. (2004, S. 43) haben Interviews mit 23 Autoren durchgeführt und extrahieren fünf grundlegende und einander häufig überlappende Motivationen zur Pflege von Weblogs. Hierzu zählt die Dokumentation des eigenen Lebens, das Präsentieren von Kommentaren und kritischen Meinungen, das Ausdrücken von Emotionen, das Artikulieren und Verbalisieren von Ideen sowie das Aufbauen und Pflegen eines Diskussionsforums für spezifische Interessengruppen.

S. 10) überbrücken Weblogs die technologische Kluft, die noch immer zwischen textbasierten CMC-Medien und traditionellen HTML-Dokumenten besteht. Diese Überbrückung wird auf den Ebenen der Aktualisierungsfrequenz, Symmetrie der Kommunikation und Multimedialität angesiedelt. Aus diesem Grunde verorten Herring et al. (2004) Weblogs auf einem Kontinuum – aufgespannt zwischen den Polen asynchroner CMC und traditionellen Webseiten – in der Mitte (vgl. Abbildung 4.10). Die ehemals eindeutige Trennung dieser Pole wird durch das sehr flexible und in zahlreichen unterschiedlichen Kommunikationssituationen einsetzbare Genre Weblog abgeschwächt, wodurch sich die „genre ecology" nach Ansicht von Herring et al. (2004, S. 11) grundlegend ändern könnte. Abschließend kann festgehalten werden, dass Hypertexttypen wie z. B. das Weblog oder die unterschiedlichen Arten von Diskussionsforen die ursprüngliche Funktionalität des WWW als Medium zur Distribution von Informationen und Dokumenten drastisch erweitern, wodurch das Medium *World Wide Web* in die Nähe der asynchronen *Computer-Mediated Communication* rückt.

4.6.8 Das Gästebuch

Digitale Gästebücher werden von unterschiedlichen Anbietern betrieben, z. B. Privatpersonen, Schulen, Vereinen und Fanclubs. Kommerzielle Anbieter wie etwa Verlage, kleinere Museen oder Firmen verzichten in ihren Webauftritten üblicherweise auf Gästebücher.[163] Diekmannshenke (2000) beschäftigt sich mit traditionellen und digitalen Gästebüchern und geht der Frage nach, weshalb sich Gästebücher im WWW einer großen Beliebtheit erfreuen. Bei traditionellen Gästebüchern liegt Diekmannshenke (2000, S. 132 f.) zufolge eine dominante Kontakt- und Erinnerungsfunktion vor, die Kommunikation wird vom Gastgeber initiiert, aber nur vom Gast realisiert, der sich in einem Eintrag z. B. verabschiedet, Freude zum Ausdruck bringt oder dem Gastgeber dankt und ihn lobt. Bezüglich digitaler Gästebücher kann eine Verschiebung der Kommunikationsfunktion beobachtet werden.

Ein Eintrag wird durch das Ausfüllen eines HTML-Formulars hinzugefügt, das auch meist spezifische Felder für den Namen des Absenders, seine E-Mail-Adresse und die URL seiner Homepage bereitstellt (vgl. auch Walker, 2000, S. 106), viele Einträge stammen jedoch von Personen, die Pseudonyme und Spitznamen verwenden, um ihre wahre Identität zu verschleiern (Diekmannshenke, 2000, S. 143). Die Gästebucheinträge werden in ihrer Gesamtheit auf einer einzelnen oder – z. B. als Gruppe von 20 Einträgen – auf mehreren Webseiten entweder in chronologischer oder umgekehrt chronologischer Reihenfolge dargestellt (wie bei Weblogs). Aufgrund der Gleichförmigkeit der Einträge, denen wegen technischer

[163] Der Anbieterkreis wird von Diekmannshenke (2000, S. 137) wie folgt umrissen: „Die technischen Möglichkeiten des Internet und der dort erhältlichen Software führen zumindest innerhalb einer gewissen ›Szene‹ zur Herausbildung bestimmter Standards für die Gestaltung der eigenen Homepage, wozu offensichtlich auch ein elektronisches Gästebuch gehört." Es wird jedoch nicht thematisiert, um welche „bestimmten Standards" oder um welche „Szene" es geht. Im Kontext der vorliegenden Arbeit bedeutet dies jedoch, dass einige Hypertexttypen in der Regel keine Gästebücher enthalten, andere dagegen sehr wohl. Meiner Ansicht nach handelt es sich dabei primär um Angebote, die von Privatpersonen (private Homepage), Freiwilligen (Vereine) oder Institutionen angeboten werden, die eine treue Fangemeinde besitzen (Musikbands, Sportvereine). Kommerzielle Websites verzichten auf Gästebücher, weil sie vermutlich verhindern möchten, dass unzufriedene Kunden in öffentlicher Weise (noch dazu auf der *eigenen* Website) drastische Kritik äußern, was den Ruf des Unternehmens gefährden könnte. Weiterhin können Gästebucheinträge verschiedene rechtliche Probleme verursachen (vgl. hierzu auch Diekmannshenke, 2000, S. 144).

Einschränkungen ein Fehlen des „persönlichen Fingerabdrucks“ inhärent ist, der ja gerade die einzelnen Einträge traditioneller Gästebücher auszeichnet, beobachtet Diekmannshenke (2000, S. 141) in digitalen Gästebüchern einen „Formular- und Listencharakter“. Die Individualisierung kann nur durch den Eintragstext erfolgen. Es überrascht daher auch nicht, dass Diekmannshenke (2000, S. 140 ff.) in den von ihm untersuchten Gästebüchern zahlreiche Merkmale für konzeptionelle Mündlichkeit und kommunikative Nähe findet, die bereits für die Kommunikationsdienste E-Mail, Newsgroups und IRC berichtet wurden (vgl. Haase et al., 1997). Hierzu gehören typografisch markierte Verbstämme („∗liebknuddel∗“), spezielle Abkürzungen („ROTFLMAO“) und Smileys. Gästebücher im WWW sind nach Diekmannshenke (2000, S. 136) ein „Ort der Selbstdarstellung, sowie eines weitgehend unspezifisch adressierten Klatsches und Tratsches, an dem persönliche Kontakte in vergleichsweise unverbindlicher Weise angeboten, angebahnt und gepflegt werden, an dem – für das Internet typisch [...] – Privates öffentlich präsentiert und der allgemeinen Kommentierung preisgegeben wird.“ (vgl. hierzu auch Abschnitt 4.6.3 ab S. 233).

Traditionelle und digitale Gästebücher unterscheiden sich in vielerlei Hinsicht: Gästebücher im WWW sind „generell öffentlich“, der Zugang ist jederzeit möglich (ebd., S. 137). Es besteht keine Notwendigkeit für eine persönliche Bekanntheit der Kommunikationspartner (ebd., S. 139). Eine Befragung von Gästebuch-Benutzern hat ergeben, dass diese eher zufällig auf ein Gästebuch treffen und dann einen Eintrag hinterlassen, weshalb Diekmannshenke (2000, S. 139) die Ansicht äußert, dass derartige Gästebucheinträge eher dem „Graffito und dem ›Klospruch‹“ ähneln. Viele Anbieter eines Gästebuchs sind offenbar bestrebt, „möglichst viele Einträge von möglichst vielen Menschen aufweisen zu können“ (ebd., S. 138). Hieraus ergibt sich das Phänomen, dass dem Eintrag in ein Gästebuch häufig der Eintrag des Besitzers eben dieses Gästebuchs in das eigene folgt (ebd.) – viele Personen, die sich in Gästebücher eintragen, scheinen also selbst ein Gästebuch anzubieten. Vermutlich erfolgen viele Eintragungen nur aus dem Grund, einen Gegeneintrag zu erwirken. Weiterhin beobachtet Diekmannshenke eine häufige Benutzung von Anreden und Verabschiedungen und viele Einträge von einzelnen Personen, die in kurzen Abständen erfolgen. Auf die Frage, was denn die Benutzer in Gästebüchern eigentlich machen, gibt Diekmannshenke (2000, S. 145) eine prägnante Antwort: „Alles.“ Wesentliche Intentionen der Einträger sind Selbstdarstellung, Kontaktpflege sowie Klatschen und Tratschen, die gruppen- und beziehungskonstituierend wirken, d. h. Gästebücher werden zur Verfolgung der eigenen sozialen und kommunikativen Interessen eingesetzt.[164] Die Nachrichten sind häufig intendiert als unverbindliches Angebot zur Kommunikation und Kontaktaufnahme (Diekmannshenke, 2000, S. 152), weshalb der Verfasser „analog zum Pausen- oder Kaffeeklatsch“ einen „neuen Klatschtypus“, d. h. den „Gästebuchklatsch“ konstatiert (ebd., S. 153). Gästebücher im WWW werden also in vielen

[164] Bittner (2003, S. 92) äußert im Rahmen einer Betrachtung der Dialogizität (vgl. Fußnote 78, S. 94) privater Homepages die Ansicht, dass Gästebücher der „Validierung“ einer Homepage dienen, die Einträge verdeutlichen, dass eine Homepage auch tatsächlich besucht wird. Weiterhin seien Gästebücher nicht als „dialogisch“ zu betrachten, „weil sie nur einen einseitigen Informationsfluß von Seiten des Rezipienten zulassen.“ Dabei übersieht Bittner die Tatsache, dass der Anbieter auch durchaus als Rezipient in Erscheinung treten kann, um – dann wiederum als Produzent im Hinblick auf das Gästebuch – Gästebucheinträge mit eigenen Einträgen zu kommentieren (dies ähnelt der Kommentierung eines Weblog-Eintrags durch Benutzer und dem Antworten auf diese Kommentare durch den Autor des Weblogs). Software-Pakete zur Realisierung von Gästebüchern heben Einträge des Besitzers typografisch hervor.

Fällen als Diskussionsforen zweckentfremdet.[165] Diekmannshenke (2000, S. 153) gelangt zu folgender Schlussfolgerung: „Die elektronischen Gästebücher bewegen sich [...] noch im Spannungsfeld zwischen tradierten Formen und Mustern und der kreativen Veränderung dieser Traditionen innerhalb spezifischer Domänen. [... W]ir sind in der glücklichen Lage, das Entstehen und den Wandel einer Textsorte und einer Kommunikationsform unmittelbar erleben zu können." (vgl. Fußnote 165).

4.6.9 Weitere interaktive Hypertextsorten

Der Aspekt der Interaktivität kann zur Differenzierung zwischen unterschiedlichen Hypertextsorten eingesetzt werden. Dabei geht es insbesondere um Möglichkeiten zur Interaktion mit dem HTML-Dokument, dem Hypertext oder einer entfernten Anwendung.[166] Crowston und Williams (2000, S. 209 f.) nennen verschiedene in ihrer Stichprobe enthaltene Dokumente, die eher Benutzerschnittstellen zu Programmen als Texten ähneln. Die erste Gruppe von Webseiten betrifft die Infrastruktur des WWW, z. B. vom Rezipienten mit seiner E-Mail-Adresse ausfüllbare HTML-Formulare, die ihn automatisch über Änderungen an einem bestimmten Dokument informieren, Formulare, die die Anmeldung einer spezifischen URL an einer Suchmaschine erlauben, Eingabemasken von Suchmaschinen oder Dokumente mit den Resultaten von Datenbankrecherchen. Die zweite Gruppe von Dokumenten ermöglicht asynchrone Kommunikationsprozesse, die zuvor ausschließlich auf andere Dienste beschränkt waren (z. B. E-Mail oder Newsgroups), beispielsweise Formulare zur Einsendung

[165] Es kann tatsächlich von einer Zweckentfremdung gesprochen werden (kritisch hierzu Diekmannshenke, 2000, S. 135): Digitale Gästebücher sind in technischer Hinsicht vergleichsweise einfach zu realisieren, weshalb sie schon 1994/1995 häufig eingesetzt wuden. Da auf Seiten der Benutzer offenbar ein Bedarf zur asynchronen Kommunikation herrschte, wurden für diesen Zweck die bereits in zahlreichen Ausprägungen verfügbaren Gästebücher verwendet. Erst sehr viel später wurden Software-Pakete entwickelt, die die Funktionalität von Newsgroups im WWW nachbilden und dort „discussion board", „Diskussionsforum" oder kurz „Forum" genannt werden. Meiner Vermutung nach zeichnet sich ab, dass das Gästebuch im WWW zu seinen ursprünglichen Textsortenwurzeln zurückfinden wird, weil die komfortableren Diskussionsforen ihnen in technischer Hinsicht weit überlegen und komfortabler zu bedienen sind. Da letztere in aller Regel eine Registrierung des Benutzers voraussetzen, können die Anwender noch deutlicher Gruppenzugehörigkeit signalisieren, allerdings wird das Verfassen einer Nachricht durch die in der Regel obligatorische Authentifizierung erschwert.

[166] Auf eine detaillierte Diskussion des Interaktivitätsbegriffs wird an dieser Stelle verzichtet (vgl. Bucher, 2001, 2004, sowie die weiteren Beiträge in Bieber und Leggewie, 2004). Stattdessen verdeutlicht ein medienübergreifender Vergleich, was im Kontext des WWW mit Interaktivität gemeint ist: Der Rezeption eines Papierdokuments, z. B. eines Buches, sind verschiedene interaktionale Handlungen inhärent. Dies betrifft unter anderem das Festhalten des Buches, das Umblättern, das Aufsuchen des Inhaltsverzeichnisses und das anschließende Lokalisieren einer bestimmten Stelle anhand der Seitenzahlen, die als Navigationshilfe fungieren (vgl. Haack, 1997). Die Rezeption eines oder mehrerer Hypertexte im WWW ist von der Verwendung eines Browsers abhängig, wobei die Verfolgung von Hyperlinks bzw. das Betätigen des Rollbalkens (oder der Leertaste oder der Bild-ab- und Bild-auf-Tasten) dem Umblättern in gedruckten Dokumenten entspricht (Jakobs, 2003, S. 245, fasst dies als „Interaktivität im Sinne des Reagierens der Computerumgebung auf menschliche Eingaben" auf und betrachtet Interaktivität konsequenterweise als konstitutives Merkmal von Hypertext). Alle Handlungen, die über das Auswählen einer Verknüpfung oder der Navigation innerhalb eines Knotens hinausgehen, können – in unterschiedlichen Abstufungen (vgl. Goertz, 1995) – als interaktiver Umgang mit dem System aufgefasst werden (Boardman, 2005, S. 19), z. B. die Benutzung einer Suchmaschine, bei der eine Anfrage in mehreren Schritten verfeinert wird, um eine überschaubare Menge möglichst präziser Treffer zu erhalten (vgl. z. B. Spink et al., 2000, Jansen et al., 2000, und Strube und Hölscher, 2000) oder das Hinterlassen eines Eintrags in einem Weblog oder Gästebuch (vgl. auch Bittner, 2003, S. 112 f.).

von Kommentaren zu einer Website oder Web-basierte Diskussionsforen, die das Absenden neuer Nachrichten und den Zugriff auf ein Archiv erlauben. Die dritte Gruppe betrifft E-Commerce-Anwendungen, z. B. virtuelle Einkaufswagen oder Bestellformulare, die zwar traditionellen Bestellformularen sehr ähneln; die jeweiligen Produkte können jedoch unmittelbar eingetragen und per Mausklick in Auftrag gegeben werden. Derartige Einkaufssysteme können eine sehr umfangreiche Funktionalität besitzen, der Benutzer interagiert hier primär mit der extrem komplexen Prozesslogik einer Datenbankanwendung, die sich als dynamisch generierte HTML-Dokumente präsentiert. Weitere Anwendungen, die eine interaktive Benutzung über das WWW gestatten, sind z. B. Routenplaner oder HTML-Editoren.

Eckkrammer (2001) geht im Rahmen einer kontrastiven, multilingualen und medienübergreifenden Korpusanalyse der Gebrauchstextsorten Kontaktanzeige, Stellenanzeige und Kochrezept auf Transformationsprozesse ein, die bei korrespondierenden Textexemplaren im WWW beobachtet werden können.[167] Untersucht wird, inwiefern sich textsortengebundene Konventionen, d. h. „natürlich innerhalb der Kultur gewachsenes Wissen um die Gepflogenheiten bei der Versprachlichung von Textsorten“ (ebd., S. 48) in digitalen Textexemplaren wiederfindet. Die Ergebnisse der Studie belegen, dass die Autoren bei der Anfertigung von Texten im WWW vornehmlich durch die „traditionelle Schriftlichkeit und ihre Produkte“ geprägt sind, das traditionelle Textwissen wird also im WWW „perpetuiert“ (ebd., S. 51; vgl. Abschnitt 4.3.2). Festgestellt werden bei allen drei Textsorten Tendenzen zum Formularcharakter, zum direkten Adressatenbezug und zu einer Anlehnung an die konzeptionnelle Mündlichkeit. Die digitalen Textexemplare aus den Subkorpora der Kontakt- und Stellenanzeigen sind deutlich umfangreicher, was Eckkrammer durch den Wegfall des „sprachökonomischen Drucks“ erklärt, d. h. längere Anzeigen im WWW verursachen den Produzenten nicht notwendigerweise höhere Kosten – meist sind derartige Dienste kostenfrei nutzbar. Neben Texten, die vollständig auf den traditionellen Konventionen und Normen der jeweiligen Textsorte basieren, findet Eckkrammer (2001, S. 51) Beispiele, die „massive Auflösungstendenzen“ belegen. Diese sind jedoch in der Minderheit, vornehmlich orientieren sich die Produzenten thematisch und in Bezug auf Stil und Lexik an den traditionellen Textsorten (Eckkrammer, 2001, S. 53). Modifikationen dieser etablierten Muster resultieren Eckkrammer zufolge aus Textproduktionsvorgaben wie HTML-Formularen und „intertextuellen Phänomenen, da das Medium dem Textproduzenten eine starke Anlehnung an die Vertextungsstrategien der Textsorte ›E-Mail‹ suggeriert.“ (ebd.). Insgesamt stellt Eckkrammer Phänomene fest, die ebenfalls in Untersuchungen privater Homepages berichtet werden (vgl. Abschnitt 4.6.3): Die digitalen Textexemplare sind persönlicher, sprechen die Kommunikationspartner direkt an, enthalten zahlreiche Fehler (z. B. inflationäre oder fehlerhafte Benutzung von Interpunktion, allgemeine Rechtschreibmängel etc.) sowie ein umfangreiches Inventar von Merkmalen für konzeptionelle Mündlichkeit, wodurch kommunikative Nähe signalisiert wird (z. B. Smileys, geringe Elaboriertheit, grammatische Mängel, Auslassungen, Anakoluthe, geringe strukturelle Planung mit Nachträgen und Post-Scriptum-ähnlichen Zusätzen etc.). In den Kontaktanzeigen treten diese Merkmale hochfrequent auf, in den Kochrezepten werden sie eher selten beobachtet; die Stellenanzeigen sind primär in der traditionellen

[167] Diese Studie basiert auf einem Korpus von 600 Kontaktanzeigen, 800 Stellenanzeigen und 400 Kochrezepten, die jeweils zur Hälfte aus traditionellen Printmedien und dem *World Wide Web* stammen.

Textsorte verankert. Die digitalen Textexemplare der drei Textsorten zeichnen sich zusätzlich durch einen vermehrten Einsatz humoristischer Komponenten und metadiskursiver Bemerkungen aus, die die erwarteten Inhalte der jeweiligen Textsorte thematisieren. Eckkrammer (2001, S. 56) generalisiert diesen Befund als „Durchmischung traditioneller Vertextungskonventionen mit neuen Parametern der Textproduktion, die mit Versprachlichungstendenzen konzeptioneller Mündlichkeit einhergehen“, die wiederum insbesondere auf die ambivalente Produktionssituation zurückzuführen ist, in der sich der Autor befindet: Einerseits legt das Medium WWW eine größere Freiheit und Nähe zum Kommunikationspartner nahe, andererseits existiert aufgrund der unterschiedlich ausgeprägten HTML-Formulare eine „extern determinierte Einengung des textproduktiven Rahmens“ (ebd., S. 60), wodurch eine „Überforderung“ der Produzenten hervorgerufen werden kann (ebd., S. 63).

4.6.10 Hypertext- und Webserver-bezogene Hypertextsorten

Verschiedene Hypertextsorten beziehen sich auf einer Metaebene auf die Website bzw. den Webserver selbst, ihre dem Peritext zugehörigen Instanzen vermitteln den Rezipienten also Informationen *über* den Hypertext, die sich unter anderem auf technische Funktionen beziehen. Diese Hypertextsorten lassen sich aufgrund der Unterschiede, die bei der Herstellung korrespondierender Textexemplare festgestellt werden können, in die Gruppen manuell und automatisch erstellter Dokumente einteilen (vgl. Abbildung 4.11).

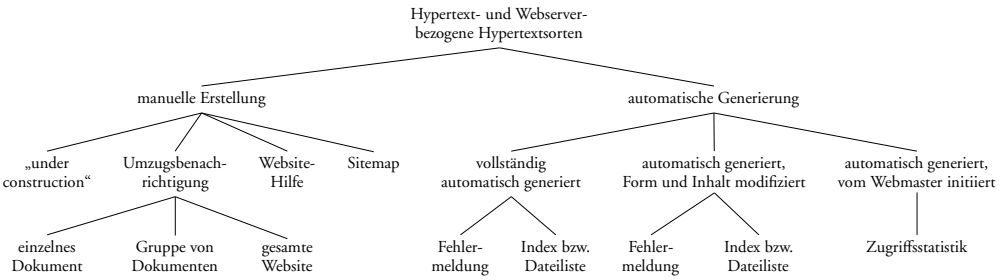


Abbildung 4.11: Hypertext- und Webserver-bezogene Hypertextsorten (Ausschnitt)

Zu den manuell angelegten Dokumenten dieser Kategorie gehören die insbesondere in den Anfangsjahren des WWW allgegenwärtigen „This site is under construction“-Hinweise, die signalisieren, dass sich ein Webangebot noch im Aufbau befindet und dass noch nicht alle geplanten Inhalte tatsächlich verfügbar sind. Korrespondierende Textinstanzen wurden bzw. werden häufig flankiert von einem – gelegentlich animierten – Icon, das einen stilisierten Bauarbeiter zeigt und auf ein US-amerikanisches Straßenschild („men at work“) zurück zu führen ist (vgl. Fußnote 118, S. 223). Crowston und Williams (2000, S. 209) ermitteln in ihrer Stichprobe 16 Vorkommen dieser Hypertextsorte, die als Platzhalter für Inhalte fungiert. Die Gründe, die zu ihrer Entwicklung und schließlich zu ihrem Rückgang geführt haben, wurden bislang nicht untersucht, können aber durch eine Art kollektiven Lernprozess erklärt werden: Gerade wissenschaftliche Institutionen und in der Computer-Branche tätige Unternehmen waren ab etwa 1993/1994 aus den verschiedensten Gründen bestrebt, Webangebote aufzubauen. Diese Websites wurden jedoch nicht, wie dies heute der Regelfall ist, zunächst

am Reissbrett konzipiert, sondern auf dem unmittelbar von jedem Benutzer zugänglichen „heißen" Webserver angelegt und sukzessive erweitert. Aufgrund der einfachen Erlernbarkeit von HTML war man vielfach der Ansicht, dass eine Website in sehr kurzer Zeit realisiert werden kann. Falls dem Webmaster bewusst war, dass in das Angebot eine bestimmte Themenrubrik integriert werden sollte, für die aber noch keine Inhalte oder schlicht keine Zeit zur Erstellung vorlagen, wurde stattdessen ein entsprechender Hinweis aufgenommen. Dieser sollte den Rezipienten mitteilen, dass sehr wohl intendiert sei, ein bestimmtes Dokument in Zukunft mit Inhalt zu füllen, dies aber momentan aus logistischen Gründen nicht geschehen könne. In diesem von einer gewissen Naivität geprägten, frühen Stadium des WWW war es wichtig, möglichst umfangreiche Angebote zu präsentieren – falls die Erarbeitung umfangreicher Angebote nicht realisiert werden konnte, wurde stattdessen zumindest eine entsprechende Absichtserklärung publiziert, die meist auch sehr unspezifisch war, denn in aller Regel wurde *nicht* thematisiert, welche konkreten Themen zu welchem spezifischen Zeitpunkt verfügbar sein werden. Der zur Entwicklung und insbesondere zur fortwährenden Pflege einer Website benötigte zeitliche und personelle Aufwand wurde also massiv unterschätzt, denn viele Hinweise vom Typ „Dieser Teil unserer Website befindet sich momentan noch im Aufbau" wurden nie durch die tatsächlich geplanten Inhalte ersetzt, weshalb diese Hinweise auf Seiten der Rezipienten nach und nach negativ konnotiert wurden.[168] Ein weiterer Grund für ihren Rückgang war, dass derartige Hinweise für den Rezipienten keinen Mehrwert besaßen: Dieser benötigt in einer bestimmten Situation gewisse Informationen und wenn sie auf einer spezifischen Website zwar erwartet werden konnten, jedoch nicht verfügbar waren, so spielte es letzten Endes keine Rolle, ob der Produzent einen „under construction"-Hinweis angeboten hat oder nicht. Diese Hypertextsorte hat in sehr kurzer Zeit einen Wandel erfahren: Lag sie vor einigen Jahren noch regelrecht im Trend, ist es mittlerweile geradezu verpönt, den Rezipienten auf einen unspezifischen Zeitpunkt zu vertrösten: Entweder möchte der Produzent bestimmte Inhalte anbieten oder eben nicht – eine entsprechende Absichtserklärung, so argumentiert die Ratgeberliteratur, soll nur dann publiziert werden, wenn eindeutig abzusehen ist, dass ein spezifisches Angebot an einem bestimmten Datum publiziert werden wird, alles andere wirke nicht professionell und bewirke das Gegenteil: Der Rezipient wird sich die Website gerade *nicht* für einen weiteren Besuch vormerken, sondern sie fortan ignorieren.[169] Die ursprünglich Verwendungsintention kehrte sich also allmählich in das Gegenteil um.

Eine weitere Hypertextsorte aus der Anfangszeit des WWW bezeichnen Crowston und Williams (2000, S. 209) als „Web site has moved". Hiermit sind HTML-Dokumente gemeint, die den Besuchern mitteilen, dass entweder ein einzelnes Dokument, eine Gruppe von Dokumenten oder die gesamte Website unter einer neuen Adresse verfügbar ist. Diese

[168] Crowston und Williams (2000, S. 209) bezeichnen die in diesem Abschnitt vorgestellten Hypertextsorten als „novel genres". Zumindest „under construction"-Hinweise besitzen aber durchaus ein Pendant: Entwurfsversionen längerer Texte (z. B. wissenschaftliche Qualifikationsarbeiten), die auch häufig explizit als solche gekennzeichnet sind („Entwurf", „Draft"), besitzen eine Gliederung und vor der eigentlichen Ausarbeitung notieren sich Autoren häufig in den einzelnen Abschnitten, welche Aspekte dort zu thematisieren sind. Eine Variante stellen Preprints wissenschaftlicher Veröffentlichungen dar, die in der Regel ebenfalls explizit als solche markiert werden und im WWW häufig als Postscript- oder PDF-Dateien verfügbar sind.

[169] Reiss (2000, S. 90) merkt in seinem Ratgeber an: „In principle, websites should always be under construction, but there's rarely a reason to advertise the fact! Most visitors interpret these signs to mean "launched and forgotten" – and never return." (vgl. Indikator 49, S. 145).

Hypertextsorte ist indirekt aus der Konvention entstanden, möglichst mnemonische Adressen zu verwenden: Die ersten universitären Websites wurden in der Regel von Mitarbeitern der jeweiligen Rechenzentren aufgebaut, die die Potenziale des Mediums WWW erkannten. Bevor sich die Konvention etabliert hat, einen Webserver im *Domain Name System* (DNS) `www` zu nennen, hatten diese Maschinen meist kryptischere Bezeichnungen. Nach dem Umzug des Angebots auf einen Rechner mit einem sprechenderen Namen (oder der Eintragung eines Alias im DNS) mussten die Benutzer, die ein Lesezeichen auf die mittlerweile veraltete Adresse angelegt hatten, umgeleitet werden, wodurch ein Bedarf für Hinweise entstanden ist, die auf die neue Adresse verweisen. Interessant ist, dass aus technischer Sicht kein Grund besteht, explizit derartige Hinweise zu geben: Webserver können so konfiguriert werden, dass der Benutzer beim Aufruf der alten Adresse automatisch zur neuen URL weitergeleitet wird. Entweder war bzw. ist diese Möglichkeit vielen Produzenten nicht bekannt, oder sie messen der neuen Adresse, die möglicherweise auch mit einem neuen Design des Angebots verbunden ist, so viel Gewicht zu, dass der Benutzer explizit auf den Umzug hingewiesen wird. Musste der Benutzer früher den verlinkten Hyperlink selbst aktivieren, findet heutzutage meist nach wenigen Sekunden eine automatische Weiterleitung („Redirect“) statt, auf die in dem Textexemplar üblicherweise ebenfalls explizit hingewiesen wird, weil sie – abhängig von ihrer technischen Realisierung und den Einstellungen des verwendeten Browsers – möglicherweise nicht funktioniert. Zu den manuell erstellten Dokumenten, die sich auf einen Hypertext selbst beziehen, gehören auch Webseiten, die z. B. die technische Infrastruktur des Webservers und verwendete Software-Pakete thematisieren, das Impressum und Hilfeseiten, die z. B. die einzelnen Bestandteile der Navigationsleisten erläutern (vgl. auch Abbildung 3.8).

Zur Gruppe der automatisch erzeugten Dokumente gehören zunächst diejenigen Webseiten, die der Webserver beim Aufruf einer nicht existenten Adresse dynamisch generiert (vgl. Abschnitt A.3.8, S. 733 ff.). Diese Dokumente beruhen in der Regel auf den optisch sehr schlichten Schablonen, die der Hersteller der Webserver-Software voreingestellt hat und die meist lediglich darüber Auskunft geben, dass die angeforderte Adresse ungültig ist, wobei üblicherweise auch der Name des Webservers, die E-Mail-Adresse des Verwalters und ein Zeitstempel eingeblendet werden. Alle verbreiteten Webserver-Pakete gestatten es jedoch dem Administrator, diese Dokumentschablonen durch eigene zu ersetzen. Gerade kommerzielle Anbieter machen hiervon Gebrauch: Da die traditionelle und sehr schmucklose „Error 404: Document not found“-Fehlermeldung auf grauem Hintergrund beim Rezipienten häufig den Eindruck erweckt, dass der Produzent sein Angebot nicht sorgfältig pflegt, wird stattdessen meist ein farbenfrohes Webdesign verwendet, wobei auch häufig auf die Nennung des Status-Codes 404 verzichtet wird.[170] Neben der Vermeidung negativer Dissonanzen ist ein zweiter Grund, dass auch die Fehlermeldungen in dem übergreifenden Design des Angebots erscheinen sollen, wobei meist auch eine Suchfunktion angeboten wird, so dass der Benutzer das ge-

[170] Prinzipiell sind derartige Fehlermeldungen, die aus „dangling links“ resultieren, also Verweisen, die auf nicht mehr existente Adressen zeigen (vgl. Boardman, 2005, S. 77 ff.), aus technischer Sicht nicht zu tolerieren: Es existieren verschiedene Werkzeuge, die – z. B. einmal pro Nacht – für jeden in einer Website enthaltenen Hyperlink automatisch prüfen, ob der Zielknoten noch erreichbar ist. Jeder nicht mehr erreichbare Knoten wird in einen Bericht aufgenommen, der automatisch per E-Mail and den Webmaster geschickt wird, der daraufhin die betroffenen Hyperlinks entfernen oder aktualisieren kann (vgl. Benbow, 1998). Insofern sind die auf vielen Websites befindlichen Hinweise, der Leser möge dem Webmaster nicht mehr aktuelle Hyperlinks per E-Mail melden, zumindest aus technischer Sicht nicht nachvollziehbar.

wünschte Dokument eventuell auf diese Weise finden kann (Crowston und Williams, 2000, S. 209, finden 11 Dokumente dieser Hypertextsorte, die sie „Custom 404“ nennen). Neben den Fehlermeldungen des voreingestellten Typs werden auch Auflistungen von Dateien, die – ähnlich einem FTP-Client – den Inhalt von Verzeichnissen darstellen, automatisch vom Webserver generiert, können aber vom Benutzer zumindest teilweise modifiziert werden, indem zu Beginn oder am Ende Dokumentfragmente, die spezielle Dateinamen besitzen müssen, integriert werden (z. B. zur Aufnahme einer Überschrift, eines Erläuterungstextes, einer Fuß- oder einer Navigationszeile).[171] Schließlich existiert noch eine Hypertextsorte, deren Textexemplare Statistiken über den benutzten Webserver präsentieren, z. B. die Anzahl der Zugriffe und das Datentransfervolumen in einem bestimmten Zeitraum. Diese Dokumente werden zu bestimmten Zeiten (z. B. einmal pro Tag oder einmal pro Woche) automatisch von speziellen Werkzeugen erzeugt, die die Protokolldateien der Webserver-Software analysieren und in Form von HTML-Dokumenten aufbereiten, deren Aussehen primär vom verwendeten Werkzeug abhängt. Crowston und Williams (2000, S. 209) finden neun Dokumente vom Typ „server statistics“ und sind der Auffassung, dass dieses Web-Genre beim Rezipienten Verwirrung hervorrufe, da sein kommunikativer Zweck unklar sei. Zahlreiche Werkzeuge erlauben die Aufbereitung der Protokolldateien, so dass es in technischer Hinsicht geradezu trivial ist, Zugriffsstatistiken anzubieten. Diese Informationen sind jedoch primär für den Verwalter der Website von Interesse, weshalb mittlerweile die Konvention existiert, den Verweis zu einer solchen Statistik entweder nicht besonders prominent zu platzieren oder gar nicht erst in das öffentlich zugängliche Angebot aufzunehmen.

4.7 Zusammenfassung

Dieses Kapitel umfasst eine ausführliche und kritische Darstellung des aktuellen Forschungsstandes zum Thema Hypertextsorten. Ausgehend von verschiedenen Arbeiten zu Digital Genres wurde deutlich, dass sich in den digitalen Informations- und Kommunikationsmedien neue Textsorten bilden oder traditionelle Textsorten in unterschiedlichen Ausprägungen eingesetzt werden. Für das Hypertextsystem *World Wide Web* bieten sich diesbezüglich im Deutschen die Termini Hypertexttyp bzw. Hypertextsorte sowie Hypertextknotensorte an, um zwischen der funktionalen Ganzheit und einzelnen, von ihr determinierten Knoten differenzieren zu können. Ein zentraler Unterschied von Text- und Hypertextsorten betrifft ihre Etikettierung.[172] Es ist Konsens in der Textlinguistik, dass es sich bei konventionalisierten Textstrukturen nur dann um Textsorten handeln kann, wenn sie innerhalb der verwenden-

[171] Dem Rezipienten werden derartige Listen präsentiert, wenn die aufgerufene URL ein Verzeichnis referenziert, ein dieser Adresse zugeordnetes Dokument, das üblicherweise `index.html` oder `welcome.html` heißt (diese voreingestellten Dateinamen können durch entsprechende Direktiven innerhalb der Konfiguration der Server-Software modifiziert werden), jedoch nicht existiert. Die Auflistung des Verzeichnisinhalts kann den Zugriff auf sensitive Daten ermöglichen, weshalb diese Funktion von den meisten Administratoren deaktiviert wird.

[172] Ein weiterer Unterschied zu traditionellen Textsorten, der erneut hervorzuheben ist, betrifft die optische Gleichförmigkeit aller Webseiten, die durch das Medium WWW, speziell die Beschreibungssprache HTML, hervorgerufen wird. Damit einher geht die „tendenzielle Dominanz des Bildes gegenüber der Schrift, des Flächigen und Virtuell-Räumlichen gegenüber dem Linearen sowie die Zerstückelung der Information in kleine Portionen“, die, wie Schmitz (2003, S. 270) sehr zutreffend anmerkt, „universal neuartige Mitteilungsformen [erzeugen], die herkömmliche regionale, soziale und thematische Unterschiede tendenziell überlagern.“

den Sprachgemeinschaft ein Etikett besitzen. Im *World Wide Web* hingegen existieren diverse Hypertextsorten, die *keine* derartigen Etiketten besitzen, weil weder bei den Rezipienten noch den meisten Produzenten ein konkreter Bedarf besteht, diese zu etikettieren. Hingegen sind in Bezug auf andere wesentliche Eigenschaften von Textsorten bei Hypertextsorten und in Teilen auch bei Hypertextknotensorten deutliche Übereinstimmungen festzustellen (vgl. zur Begründung Abschnitt 4.4): Sie können als Alltagsklassifikationen der Sprach- und Textbenutzer gelten und stellen ein Reservoir an Kenntnissen dar (vgl. Heinemann und Viehweger, 1991, S. 144) und sie können als Operator für Zuordnungsoperationen der Individuen dienen (vgl. Heinemann und Heinemann, 2002, S. 140). Bezüglich der Untersuchung von Hypertextsorten existieren zwei grundsätzliche Forschungsinteressen: Entweder wird eine spezifische Hypertextsorte ausgewählt, aus Verzeichnissen eine Stichprobe erzeugt und diese analysiert oder es wird versucht, in einer zufällig zusammengestellten Stichprobe beliebiger Webseiten iterativ und induktiv die korrespondierenden Hypertextsorten zu ermitteln. Die in Abschnitt 4.4 vorgestellten Studien zeigen, dass das *gesamte* WWW als Basis einer derartigen zufallsgetriebenen Zusammenstellung ungeeignet ist, weil die umfangreiche Varianz der beobachtbaren Hypertextsorten die Erarbeitung homogener Ergebnisse verhindert.[173] Aus diesem Grund ist die Domäne der zu untersuchenden Dokumente einzuschränken, was in den nachfolgenden Kapiteln mit einem Bezug auf die deutschsprachigen Dokumente geschieht, die die Webserver deutscher Hochschulen anbieten (vgl. hierzu auch die große Zahl unterschiedlicher Hypertexttypen in Abbildung 3.7, S. 151). Bezüglich der Kernmerkmale von Hypertextsorten sind zwei Aspekte hervorzuheben: Instanzen einer Hypertextsorte können Instanzen weiterer Hypertextsorten umfassen. Dies bezieht sich zunächst auf die Ebene der Hypertextsorte, die eine spezifische Gruppe eingebetteter Hypertextsorten oder Hypertextknotensorten vorsieht (die persönliche Homepage kann z. B. einen Lebenslauf und ein Gästebuch enthalten); betroffen sind aber auch Hypertextknotensorten, die Objekte untergeordneter Ebenen einbetten können. Mit dieser Problematik eng verbunden ist der Aspekt der Verlinkung einzelner Knoten. Das nachfolgende Kapitel 5 führt ein Hypertextsortenmodell zur Beschreibung dieser und weiterer Ebenen ein.[174]

4.8 Fazit

Die in Teil III dargestellten Analysen beziehen sich auf die deutschsprachigen HTML-Dokumente, die auf den Webservern deutscher Hochschulen angeboten werden (vgl. hierzu Kapitel 6). Die vorliegende Arbeit konzentriert sich also auf universitäre Websites, in denen viele der in diesem Kapitel vorgestellten Hypertextsorten ebenfalls verwendet werden. Hierzu

[173] Ein zusätzlicher Aspekt betrifft die Tatsache, dass die bislang vorgelegten Arbeiten in Bezug auf ihren Beitrag zur Konzeptionierung und Implementierung eines Systems zur automatischen Identifizierung von Hypertextsorten entweder – im Falle der Untersuchung bestimmter Hypertextsorten – zu spezifische oder zu grobe Ergebnisse liefern, falls weltweite Stichproben untersucht werden, die in heterogenen und nicht miteinander in Einklang zu bringenden Mengen arbiträrer Hypertextsorten resultieren. Dies dürfte auch der wesentliche Grund sein, weshalb Roussinov et al. (2001) lediglich die maschinelle Erkennung von fünf sehr groben Hypertexttypen statt der ermittelten 116 spezifischen Genres anstreben (vgl. Rehm, 2002b).

[174] Auf eine Anordnung der zahlreichen in den Abschnitten 4.4 sowie 4.6 dargestellten Hypertexttypen und Hypertextknotentypen in Form zweier Typologien, wie sie sich am Ende von Kapitel 3 befinden, wird an dieser Stelle aus Komplexitäts-, Platz- und Darstellungsgründen verzichtet.

gehören zunächst berufliche und private Homepages, aber auch Gästebücher und Hotlists. Online-Zeitungen als Pendant gedruckter Zeitungen beziehen sich in dieser Domäne unter anderem auf wissenschaftliche Zeitschriften bzw. auf ausschließlich online verfügbare Journale. Kommerziell ausgerichtete Webauftritte dürften nur in Ausnahmefällen zu finden sein (z. B. die Websites von Firmen, von Abteilungen für den Wissens- und Technologietransfer oder Ausgründungen von Forschungsprojekten, d. h. *Spin-Offs*). Die Analysen charakterisieren das Inventar von Hypertextsorten, die in den Websites von Universitäten eingesetzt werden, unter anderem wird dabei die Einstiegsseite des Webauftritts einer Universität thematisiert. Die Abschnitte 4.6.2 und 4.6.3 haben gezeigt, dass bei kommerziellen Homepages ein Trend zur Angleichung der enthaltenen Bausteine vorliegt. Dieser Trend ist bei privaten Homepages weniger ausgeprägt, denn die Exemplare dieser Hypertextsorte zeichnen sich insbesondere durch Individualität und eine persönliche Note aus. In diesem Zusammenhang stellt sich die Frage, welche Trends bei den Einstiegsseiten universitärer Websites zu verzeichnen sind, schließlich befinden sich diese in gewisser Hinsicht in einem Spannungsverhältnis zwischen Kommerzialität und Privatheit (vgl. Abschnitt 6.3, S. 306 ff.).

In textlinguistischer Hinsicht ist die Frage, ob Hypertextsorten mit dem vorhandenen Beschreibungsinventar (vgl. Kapitel 2) erfasst werden können, von entscheidender Bedeutung. Das Analysemodell für Hypertexte von Huber (2002, vgl. Abschnitt 3.5.7) sowie die ausführliche Analyse von Schütte (2004a, vgl. Abschnitt 4.6.2) zeigen, dass die traditionellen Kriterien zwar eine Grundlage darstellen, jedoch verschiedene Erweiterungen notwendig sind, die sich insbesondere auf die Einbettung von Hypertextsorten und die Ebene der Verlinkung beziehen müssen (vgl. die Abschnitte 4.5.2 und 4.5.3). Die traditionellen Methoden sind vornehmlich auf die Beschreibung prototypischer, ›genuiner‹ Langtexte ausgelegt, weshalb sie – wie dieses und das vorherige Kapitel gezeigt haben – nicht auf Webseiten angewendet werden können. „Multimodale Kommunikationsformen“ sind, wie Bucher (2000, S. 689) betont, „für die Linguistik deshalb eine Herausforderung, die traditionell gewachsene Beschränkung auf den Textmodus aufzugeben.“ (vgl. auch Schmitz, 2003, S. 259). Doch bei aller, gelegentlich ein wenig polemisch anmutender Kritik an „Schnipsel-“ oder „Mosaiktexten“ stellen diese zweifelsohne die realen Gegebenheiten im *World Wide Web* dar, weshalb der textlinguistische Beschreibungsapparat entsprechend zu ergänzen ist. Die Erweiterungen, die in den nachfolgenden Kapiteln vorgeschlagen werden, zielen insbesondere auf ein maschinenlesbares Repräsentationsformat sowie eine automatische Verarbeitung ab, die auch als hilfreiche Werkzeuge für die traditionelle textlinguistische Arbeit intendiert sind, denn gerade eine derartige maschinelle Unterstützung bei der rein deskriptiven Analyse scheint angesichts der zahlreichen beteiligten Ebenen (vgl. Huber, 2002, siehe auch die Tabellen 3.1 und 3.2, S. 117 f.) notwendig zu sein.

Teil II

Das Rahmenmodell und die Methodologie

Überblick

Kapitel 5 stellt zunächst das Hypertextsortenmodell vor, auf das sich die Ausführungen in den Teilen III und IV beziehen. Das Modell besitzt eine textlinguistische Ausrichtung, es stellt einen Beschreibungsapparat für Hypertextsorten zur Verfügung. Zugleich ist es für einen gewinnbringenden Einsatz in sprach- und texttechnologischen Systemen ausgelegt, der in Teil IV diskutiert wird. Innerhalb des Modells werden drei grundlegende und hierarchisch gestufte Beschreibungsebenen angenommen, die sich auf Hypertexttypen bzw. Hypertextsorten, Hypertextknotentypen bzw. Hypertextknotensorten und Hypertextsortenmodule beziehen. Die Ausrichtung des Modells betrifft beliebige Gebrauchshypertextsorten. Die in Kapitel 4 diskutierten Studien zur Sammlung von Hypertextsorten beziehen sich nahezu ausschließlich auf die Untersuchung von Stichproben, die den Datenbeständen von Suchmaschinen entnommen wurden und somit *beliebige* Webseiten umfassen – entsprechend heterogen und teilweise widersprüchlich fallen ihre Ergebnisse aus. In der vorliegenden Arbeit wird das Hypertextsortenmodell anhand der Untersuchungsdomäne der universitären Webangebote exemplifiziert, was in der Hypothese begründet ist, dass eine derartige Restriktion homogenere und detailliertere Resultate ermöglicht, die sich auch auf die Problematik der Typologisierung positiv auswirken. In Kapitel 6 werden die Charakteristika dieser Untersuchungsdomäne erläutert. Die Darstellung geht von den traditionellen Textsorten des Kommunikationsbereiches Hochschule und Wissenschaft aus, woraufhin verschiedene Kerneigenschaften universitärer Webauftritte aufgezeigt werden, die sich unter anderem auf situative, strukturierende und gestalterische Faktoren beziehen. Im Anschluss stellt Kapitel 7 die Korpusdatenbank vor, mit deren Hilfe die in Teil III präsentierten Analysen durchgeführt wurden. Das der Arbeit zugrunde liegende Korpus umfasst die etwa vier Millionen deutschsprachigen Dokumente der Webauftritte von 100 deutschen Universitäten und Hochschulen. Das Kapitel erläutert neben der Datensammlung, an der ein automatischen Sprachenidentifizierer beteiligt ist, die Funktionsweise der Korpusdatenbank. Diese besitzt eine Web-Oberfläche, die die Generierung und Analyse von Stichproben erlaubt.

5 Das Hypertextsortenmodell

5.1 Einleitung

Dieses Kapitel geht auf die textlinguistische und texttechnologische Ausrichtung eines Hypertextsortenmodells ein, das in Bezug auf zwei wesentliche Anwendungsszenarien entwickelt wurde. Zunächst soll das Hypertextsortenmodell linguistische und textlinguistische Analysen von Hypertextexemplaren ermöglichen, deren Ergebnis empirisch fundierte Profile von Hypertextsorten sind. Zusätzlich bildet das Hypertextsortenmodell die konzeptionelle Grundlage einer sprachtechnologischen Architektur, die die automatische Identifizierung von Hypertextsorten, d. h. die Klassifikation von Hypertexten in ihre korrespondierenden Hypertextsorten und verwandte Verarbeitungsprozesse zum Ziel hat.

Während die Kapitel 8 bis 12 die Anwendung des Hypertextsortenmodells auf die Untersuchungsdomäne der universitären Webangebote darstellen, wird im Folgenden dessen theoretische Ausrichtung erläutert. Abschnitt 5.2 geht auf Aspekte der Charakterisierung von Textsorten ein, wobei die Kerneigenschaften textlinguistischer und texttechnologischer Beschreibungen dargestellt und mit den korrespondierenden Arbeiten zur linguistischen Modellierung von Hypertextsorten kontrastiert werden. Abschnitt 5.3 diskutiert – ausgehend von den texttechnologischen Spezifika der *Hypertext Markup Language* – die Ausrichtung des Hypertextsortenmodells in Bezug auf die Texttechnologie und die Textlinguistik, geht auf die Problematik traditioneller Textdefinitionen im Hinblick auf das WWW ein und thematisiert die unterschiedlichen Granularitätsstufen der an einer Hypertextsortenanalyse beteiligten Entitäten. Das Hypertextsortenmodell basiert auf drei Schichten: Hypertexttypen bzw. Hypertextsorten stellen die oberste Abstraktionsebene dar (Abschnitt 5.4). Eine Hypertextsorte wird durch die Menge der in einem korrespondierenden Hypertextexemplar enthaltenen Knoten instanziiert, die ihrerseits Hypertextknotentypen bzw. Hypertextknotensorten zugehörig sind. Instanzen von Hypertextknotensorten stellen somit die spezifische Ausprägung einer Hypertextsorte dar (Abschnitt 5.5). Da die Aufteilung von Inhalten in einen oder

mehrere Knoten vollständig arbiträrer Natur und dem Produzenten eines Hypertextexemplars freigestellt ist, wird die konzeptionelle Ebene der Hypertextsortenmodule eingeführt, die als flexible und primär inhaltlich-thematisch markierte Makrostrukturbausteine fungieren (Abschnitt 5.6). Verschiedene Merkmalsausprägungen, die unter anderem die kommunikative Funktion und die Dekoration betreffen, werden bei der Analyse einer Instanz auf dieser untersten Ebene erfasst und über den Knoten bis zur obersten Ebene propagiert, so dass Varietäten von Hypertextsorten beschrieben werden können. Abschnitt 5.7 geht auf die Interaktion der Ebenen ein, woraufhin Abschnitt 5.8 weitere Aspekte des Hypertextsortenmodells darstellt. Dabei handelt es sich um eine Betrachtung von Hypertextsorten als Prototypen und die Methodologie der Sammlung und Identifizierung von Hypertextsorten.

5.2 Zur Charakterisierung von Text- und Hypertextsorten

Das in Kapitel 13 thematisierte und anhand der Ergebnisse aus Teil III exemplifizierte Repräsentationsformat für Hypertextsorten ist als zentrale Ressource für computerlinguistische Anwendungen des in diesem Kapitel eingeführten Hypertextsortenmodells ausgelegt, weshalb im Folgenden diejenigen korrespondierenden Ansätze vorgestellt werden, die das Modell maßgeblich beeinflusst haben und für das Repräsentationsformat von Relevanz sind. Abschnitt 5.2.1 geht auf Verfahren zur Beschreibung von Textsorten ein, woraufhin Abschnitt 5.2.2 Ansätze zur Repräsentation von Hypertextsorten diskutiert.

5.2.1 Zur Repräsentation von Textsorten

Dieser Abschnitt thematisiert zunächst ausgewählte Verfahren zur Charakterisierung und Repräsentation von Textsorten aus einer textlinguistischen Perspektive, woraufhin die etablierte texttechnologische Methodik zur Erstellung von Dokumentgrammatiken skizziert wird.

Textlinguistische Ansätze zur Charakterisierung von Textsorten

Im Allgemeinen bestehen textlinguistische Charakterisierungen von Textsorten aus natürlichsprachlichen Beschreibungen, die aus den Analysen eines Korpus repräsentativer Textexemplare gewonnen werden; oftmals beziehen sich diese Charakterisierungen auf die Untersuchung lediglich eines Exemplars, um die generellen Parameter einer Textsorte spezifizieren zu können. Die Analysen basieren wiederum auf einem Katalog von Untersuchungskriterien und Merkmalen, deren Zusammenstellung vom Erkenntnisinteresse abhängig ist und deren individuelle Ausprägungen in den Textexemplaren ermittelt werden. Auf diese Weise können Generalisierungen in Form von Textsortenprofilen konstruiert werden. Hinzu kommen typische kontextuelle Faktoren der textuell realisierten kommunikativen Handlung, die sich z. B. auf die Kommunikationssituation und das Verhältnis von Produzent und Rezipient beziehen. Sobald mehr als eine Textsorte untersucht wird, werden in der Regel Aspekte der Typologisierung der charakterisierten Textklassen diskutiert, um die zwischen ihnen geltenden Unterschiede und Gemeinsamkeiten zu beschreiben (vgl. ausführlich hierzu Kapitel 2).

Im Kontext der maschinellen Anwendung ist der Ansatz von Sandig (1972) von zentraler Bedeutung: Trotz der häufig geäußerten Kritik bezüglich einer mangelnden theoretischen

Fundierung hinsichtlich der Gewinnung und Ausrichtung der Attribute entspricht die von Sandig (1972) beschriebene Methodologie zur merkmalbasierten Klassifikation von Textsorten (vgl. Abschnitt 2.3.5) in ihren Grundzügen den bislang vorgelegten Verfahren zur maschinellen Identifizierung von Textsorten (vgl. ausführlich hierzu Abschnitt 14.2, S. 628 ff.). Diese basieren auf der Berechnung verschiedener linguistischer Merkmale eines gegebenen Textexemplars, die mit manuell kategorisierten Daten verglichen werden. Diese werden wiederum empirisch auf der Grundlage eines Korpus von Texten verschiedener Textsorten erhoben, so dass dem unbekannten Text die Textsorte mit der größten Übereinstimmung bezüglich der beobachteten Merkmalsausprägungen zugewiesen werden kann.

Texttechnologische Ansätze zur Repräsentation von Textsorten

Dokumentgrammatiken sind sowohl die historische als auch die konzeptionelle Basis der Texttechnologie (vgl. Lobin und Lemnitzer, 2004, für einen Überblick), die die linguistisch motivierte Informationsanreicherung und Verarbeitung digital verfügbarer Texte mit standardisierten Auszeichnungssprachen fokussiert (Rehm, 2004e, S. 138). Mit Hilfe der Metasprachen SGML (ISO 8879) und XML (Bray et al., 2004b) definierte Auszeichnungssprachen werden als explizite Element- und Attributdeklarationen in Dokumentgrammatiken (DTDs, Dokumenttyp-Definitionen) gebündelt (vgl. Lobin, 2001a, und Rehm, 2004e). Die Strukturierung, Hierarchisierung und Etikettierung von Informationseinheiten in derartigen Auszeichnungssprachen unterliegt keinerlei Beschränkungen, doch orientieren sich diese Prozesse – falls sie auf eine spezifische Textklasse angewendet werden – in aller Regel an den generischen Textstrukturmustern derjenigen Textsorte, für die eine DTD erstellt werden soll (vgl. Lobin, 2004, S. 53).[1] Die Anfertigung einer solchen Auszeichnungssprache basiert üblicherweise auf mehreren repräsentativen Textexemplaren der zu modellierenden Textsorte, die im Hinblick auf rekurrente Muster sowie übergreifende hierarchische Strukturen und Komponenten analysiert werden (vgl. ausführlich hierzu Maler und Andaloussi, 1996). Die beobachteten Regularitäten manifestieren sich in der Spezifikation einer Dokumentgrammatik (z. B. in Form einer DTD), die die zulässigen Strukturen annotierter Dokumentinstanzen beschreibt und z. B. bei der Texterfassung sowie der Transformation in unterschiedliche Ausgabeformate eine immense Erleichterung darstellen kann (vgl. Lobin, 2000).

SGML- oder XML-basierte Dokumenttyp-Definitionen bzw. die von ihnen konstituierten Auszeichnungssprachen spezifizieren jedoch nicht notwendigerweise ein Höchstmaß an Struktur: Ein Element wie `address` kann einerseits einer hochgradigen Binnendifferenzierung unterliegen (mit untergeordneten Elementen wie z. B. `street`, `postbox`, `zip_code`, `city` etc.), andererseits nur Textinhalt umfassen (wie das gleichnamige Element in HTML). Die Modellierung einer Textsorte als DTD gewährleistet also nicht unmittelbar für jeden informationellen Bestandteil eine maximale Strukturierung, vielmehr ist es die Aufgabe des Autors der DTD, in den Textexemplaren, die als repräsentative Vertreter fungieren, möglichst umfassende Gemeinsamkeiten zu isolieren, die sich als maximal restriktive, aber zugleich erfolgreich auf die Menge der zu erwartenden Textexemplare anwendbare Regeln in der DTD

[1] Der angesprochene Mangel an Beschränkungen bezieht sich nicht auf die rigiden syntaktischen Vorgaben der Metasprachen SGML und XML, sondern auf die Methodik, die zur Erstellung einer DTD verwendet wird, auf die Granularität der angenommenen Informationseinheiten, auf ihre Etikettierung sowie ihre Verschachtelung.

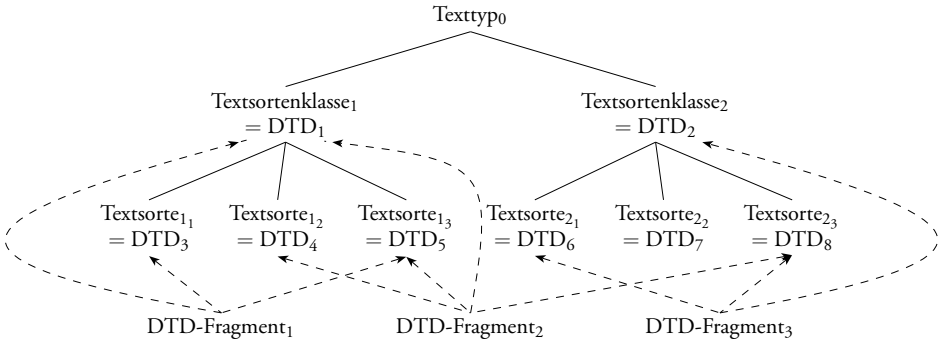


Abbildung 5.1: Die Repräsentation der in einer Texttypologie enthaltenen Textklassen als SGML- bzw. XML-Dokumentgrammatiken (vgl. Abbildung 2.2, S. 40)

manifestieren sollten. In vielen Fällen können derartige Gemeinsamkeiten jedoch – eine optimale, d. h. maximal ausführliche Analyse bezüglich der beobachtbaren Textstrukturmuster vorausgesetzt – nicht ermittelt werden, weshalb in impliziter Weise unterschiedliche Grade von Strukturiertheit in DTDs existieren.[2] Je spezifischer die enthaltenen Element- und Attributdeklarationen sind, desto restriktiver sind zugleich die erlaubten Strukturierungen und desto härter (im Sinne von Schultze und Boland, 1997, vgl. Abschnitt 4.2.3) ist somit auch die DTD bzw. das von ihr modellierte Textstrukturmuster der korrespondierenden Textsorte.

Wenn eine DTD als Spezifizierung der gültigen Textstrukturmuster einer bestimmten Textsorte aufgefasst wird, stellt sich unmittelbar die Frage nach der Repräsentation von Texttypologien. Der Funktionsumfang der syntaktischen Konstrukte der Metasprachen SGML (ISO 8879) und XML (Bray et al., 2004b) zeigt, dass diese Formalismen auf die ausschließliche Modellierung *einzelner* Textsorten ausgelegt sind, d. h. die unter textlinguistischen Gesichtspunkten zentrale Anwendung der Repräsentation von Beziehungen zwischen mehreren verwandten Textsorten kann nicht bzw. nur sehr eingeschränkt erfolgen. Abbildung 5.1 verdeutlicht die Problematik: Zwischen den Textklassen, die in einer Typologie angeordnet sind, existieren notwendigerweise verschiedene Relationen (vgl. auch Abbildung 2.2, S. 40). Diese Relationen können jedoch durch Dokumentgrammatiken nicht ausgedrückt werden. Die Strukturmuster der beiden Textsortenklassen sowie der drei jeweils untergeordneten Textsorten resultieren zwangsläufig in acht voneinander unabhängigen DTDs. Die drei DTD-Fragmente deuten an, dass übergreifende, d. h. identische Strukturfragmente zwar mit Hilfe

[2] Ein weiteres Beispiel soll dies verdeutlichen: Falls ein Element wie z. B. `name` die beiden obligatorischen, untergeordneten Elemente `firstname` und `surname` mit einem optional zwischen ihnen vorkommenden Element `middlename` sowie einem weiteren optionalen initialen Element `title` erwartet, liegt eine explizitere Strukturierung vor als bei einem Element `name`, als dessen Inhalt beliebiger Text (`#PCDATA`) vorgesehen ist, so dass bei der Anfertigung von Dokumentinstanzen Zeichenketten wie z. B. „Dr. Petra K. Schmidt“, „Schmidt, Petra Klara“, „Frau Dr. P. K. Schmidt“ und „Dr. Schmidt“ zu erwarten sind, weil die in der DTD existierende und nicht weiter binnenstrukturierte Vorgabe `name` von Anwendern der DTD unterschiedlich aufgefasst werden kann. Dieses Beispiel zeigt darüber hinaus, dass der Strukturiertheitsgrad einer DTD nur für ihren Einsatzzweck eruiert werden kann: Lobin (2001a, S. 181) unterscheidet diesbezüglich zwischen darstellungsbezogenen DTDs (z. B. HTML), standardisierten DTDs (z. B. die TEI-Richtlinien), zentralen proprietären DTDs (z. B. zur Strukturierung des vorhandenen und absehbaren Datenaufkommens eines Unternehmens) und proprietären Benutzer-DTDs (zur Gewährleistung der Konsistenz bei der Erfassung neuer Daten).

des Einsatzes von Parameterentitäten parallel in mehreren Dokumentgrammatiken verwendet werden können (z. B. zur Beschreibung der generischen Struktur einer Adresse), doch handelt es sich hierbei nur um sehr eingeschränkte Möglichkeiten, da die korrespondierenden Fragmente in separaten Dateien hinterlegt und in die eigentliche DTD importiert werden.[3] Durch die Aufnahme eines derartigen Fragments in mehreren DTDs werden keine Beziehungen zwischen einzelnen Textklassen ausgedrückt. Eine für diesen Zweck notwendige zusätzliche Beschreibungsebene, die in struktureller Hinsicht oberhalb der einzelnen Dokumentgrammatiken anzusiedeln wäre, um Beziehungen zwischen mehreren DTDs zu modellieren, existiert weder in SGML noch in XML.

DTDs stellen den ursprünglich von SGML eingeführten und in XML übernommenen Mechanismus zur Kodierung von Dokumentgrammatiken dar. Da DTDs zahlreichen Einschränkungen unterliegen, wurden verschiedene alternative Metasprachen zur Spezifikation von Dokumentgrammatiken entwickelt (vgl. Lobin, 2004, für einen Überblick).[4] XML Schema (Fallside et al., 2001) ist eine vollständig in XML realisierte Schemasprache, die viele der in Bezug auf DTDs häufig geäußerten Probleme löst. In Anlehnung an das Paradigma der objektorientierten Programmierung erlaubt XML Schema z. B. die Deklaration abstrakter Elementtypen, die durch konkrete Elemente instanziiert und erweitert werden können. Diese Elemente fungieren somit als Objekte, die die Eigenschaften ihrer zugehörigen Klassen (die Elementtypen) erben. Lobin (2004, S. 68) spricht daher auch von einer „objektorientierten Schema-Spezifikation“. Dieses Verfahren erleichtert die Realisierung in unterschiedlicher Weise spezifizierter Dokumentgrammatiken, die auf beliebig umfangreichen Inventaren abstrakter Elementtypen basieren. Duckett et al. (2001) gehen ausführlich auf die hier nur grob angedeutete Modularisierung von Schemata ein.

Der Ansatz, spezifische Elemente auf abstrakte Elemente zu beziehen, besitzt seinen konzeptionellen Ursprung in den *Architectural Forms*, die als Bestandteil der *SGML Extended Facilities* im Anhang des HyTime-Standards veröffentlicht wurden (ISO 10744). Eine herkömmliche DTD kann demnach als Meta-DTD, d. h. als eine abstrakte Vorlage fungieren, auf deren Element- und Attributtypen sich verschiedene Client-DTDs über entsprechende Referenzen innerhalb der nicht vererbbaren, sondern obligatorisch lokal vorzunehmenden Deklarationen beziehen können (vgl. ausführlich hierzu Lobin, 2001a). Client-DTDs können somit als konkrete Ausprägungen derjenigen Architektur aufgefasst werden, die die Meta-DTD vor-

[3] Ein Beispiel für diesen Mechanismus zur Importierung von DTD-Fragmenten sowie Hilfs-DTDs stellt die von der *Text Encoding Initiative* erarbeitete Dokumentgrammatik dar (Sperberg-McQueen und Burnard, 2002): Diese besteht aus zwei obligatorischen *core tag sets*, die allgemeine Auszeichnungselemente (Textabschnitte, Zitate, Listen etc.) und einen Satz von Metadaten zur Verfügung stellen. Zusätzlich muss eines der *base tag sets* eingebunden werden, um Textexemplare eines spezifischen Texttyps annotieren zu können. Die Version P4 der TEI-Richtlinien enthält für diesen Zweck DTDs für Erzählungen, Gedichte, Dramen, Transkriptionen gesprochener Sprache, Wörterbücher und Terminologiedatenbanken. Beliebige Mischungen dieser Typen sind ebenfalls möglich. Abschließend können verschiedene *additional tag sets* importiert werden, die z. B. sehr detaillierte Strukturierungsmöglichkeiten für Verknüpfungen, kritische Apparate, Namen und Datumsausdrücke, Korpora und Graphen zur Verfügung stellen.

[4] Die genannten Einschränkungen beziehen sich zunächst auf den Umstand, dass XML-DTDs nicht in einem XML-basierten Format repräsentiert werden und somit nicht Gegenstand XML-basierter Verarbeitungsprozesse werden können (z. B. zur Transformation der DTD durch ein XSLT-Stylesheet). Weiterhin gestattet die Syntax von XML-DTDs nur sehr rudimentäre Möglichkeiten der Spezifizierung von Inhaltsmodellen. Abschließend existiert in DTDs kein Mechanismus zur Verwendung spezifischer Datentypen.

gibt. Umgekehrt ist es auch möglich, z. B. hinsichtlich ihrer Semantik identische Elemente unterschiedlicher DTDs auf eine übergeordnete Meta-DTD zu beziehen (vgl. Sasaki, 2004). Das Konzept der Architekturen konnte sich aufgrund der Problematik, dass alle Vererbungen manuell kodiert werden müssen und keine Richtlinien zur Validierung der architektonischen Zuordnung existieren, nicht durchsetzen (vgl. Lobin, 2004, S. 69).

5.2.2 Zur linguistischen Beschreibung von Hypertextsorten

Nur wenige Arbeiten beschäftigen sich auf einer abstrakten Ebene mit der Beschreibung und Repräsentation von Hypertextsorten. Diese Ansätze werden im Folgenden diskutiert.

Merkmalbasierte Verfahren zur Repräsentation von Hypertextsorten

Crowston und Kwasnik (2004) schlagen zur Repräsentation von Genres einen Ansatz vor, der auf den von Rezipienten wahrgenommenen unterschiedlichen Facetten (*facets*) beruht, die als klassifizierende Beschreibungsdimensionen von Textsorten aufgefasst werden können. Der Klassifikationsprozess beruht also im Vergleich zu Texttypologien nicht auf einer gemeinsamen Einordnungsinstanz, sondern simultan auf voneinander unabhängigen Merkmalen, so dass eine auf diese Weise vorgenommene Repräsentation konkreter Textsorten in einer Matrix resultiert, die die unterschiedlichen Ausprägungen aller Facetten im Hinblick auf jede charakterisierte Textsorte aufzählt; dieses Verfahren entspricht prinzipiell der von Sandig (1972) vorgeschlagenen merkmalsbasierten Klassifikation von Gebrauchstextsorten (vgl. Abschnitt 2.3.5, insbesondere Tabelle 2.2, S. 50). Der wesentliche Unterschied betrifft jedoch die Tatsache, dass die von Crowston und Kwasnik beschriebenen Facetten keine binären bzw. ternären Merkmale sind, sondern autarke Klassifikationsstrukturen darstellen, die intern z. B. als Zeitleiste, Hierarchie, Teil-Ganzes-Taxonomie oder auch als binäres Merkmal realisiert sein können: „The process of facet analysis is to view the object from all its angles – same object, but seen from different perspectives." (Crowston und Kwasnik, 2004, S. 5). Die Ermittlung und die interne Strukturierung von Facetten wie z. B. „Communication act", „Source", „Structure", „Layout", „Content", „Language level" oder „Length" sollte Crowston und Kwasnik zufolge auf denjenigen Kriterien basieren, die Rezipienten von Textexemplaren zur Differenzierung der jeweiligen Textsorten einsetzen. Auf diese Weise sollten zusätzlich formale oder inhaltliche Hinweise erhoben werden, die der maschinellen Bestimmung eines Merkmals dienen können. Crowston und Kwasnik nennen eine Reihe von Vorteilen, die mit einem derartigen Verfahren zur Repräsentation von Genres verbunden wären: Die Klassifikation auf der Basis von Facetten wird als „relativ gastfreundlich" beschrieben, d. h. neue Textsorten sind verhältnismäßig einfach zu integrieren, indem sie auf der Grundlage der vorhandenen Merkmale charakterisiert werden. Weiterhin ist es durch die Verwendung unabhängiger Attribute flexibel. Der vermutlich größte Vorteil besteht in der beliebigen Binnenstrukturierung der einzelnen Facetten, die zur Spezifizierung der jeweiligen Beschreibungsdimensionen gleichsam maßgeschneidert werden können. Zusätzlich wird keine zugrunde liegende, globale Theorie benötigt, weil die einzelnen Facetten als Klassifikationsstrukturen fungieren. Diese sollten jedoch ihrerseits über eine theoretische Fundierung verfügen, denn „[t]he study of genre draws from many disparate disciplines, which could not easily be accommodated un-

der the umbrella of a single classificatory scheme." (Crowston und Kwasnik, 2004, S. 7). Neben den Vorteilen dieses Ansatzes existieren den Verfassern zufolge jedoch auch verschiedene Nachteile: Zunächst ist es schwierig, ohne detailliertes Wissen über die zu charakterisierende Domäne die relevanten distinktiven Merkmale zu bestimmen, was ebenfalls für die Mehrzahl der für diese Aufgabe zu befragenden Nutzer gelten dürfte, denn „people may not be aware of what allows them to recognize a given genre, and thus the determination of an adequate set of fundamental categories will be a challenge." (Crowston und Kwasnik, 2004, S. 7). Weiterhin existieren keinerlei Beziehungen zwischen einzelnen Merkmalen, d. h. Wechselwirkungen oder spezifische Relationen zwischen Facetten können nicht erfasst werden. Auch die Visualisierung einer derartigen Klassifikation wird durch die multiplen Klassifikationen in Form individuell strukturierter Facetten erschwert, so dass ein derartiges Schema also jeweils nur aus einer oder maximal zwei Perspektiven gleichzeitig betrachtet werden kann. Ein zusätzliches Problem betrifft die Aufnahme von Beschreibungsebenen, die für eine maschinelle Verarbeitung ausgelegt sind – Crowston und Kwasnik (2004, S. 8) äußern sich diesbezüglich kritisch und stellen eine derartige Realisierung in Frage (vgl. Abschnitt 14.2, S. 628 ff.).

Das in diesem Kapitel vorgestellte Hypertextsortenmodell sowie die in Teil III präsentierten Analysen bauen hinsichtlich verschiedener Aspekte auf diesen Überlegungen auf: Crowston und Kwasnik (2004) beziehen sich zwar wiederholt auf die Potenziale einer maschinellen Anwendung der „facetted classification", gehen jedoch nicht auf die Frage ein, wie sie praktisch bzw. technisch realisiert werden kann. Das Hypertextsortenmodell sowie die in Kapitel 13 vorgestellte Hypertextsortenontologie können diesbezüglich als eine Konkretisierung der Vorüberlegungen von Crowston und Kwasnik (2004) aufgefasst werden, die zwei von den Verfassern angesprochene Probleme – die Visualisierung sowie die Integration maschinell verwertbarer Analyseinformationen – zumindest partiell löst. Crowston und Kwasnik weisen mehrfach darauf hin, dass sich eine Klassifikation von Genres primär an den Verwendern korrespondierender Textexemplare orientieren sollte, so dass die von ihnen wahrgenommenen Differenzierungskriterien aufgegriffen werden können (vgl. Toms und Campbell, 1999, Brandl, 2002, und Ryan et al., 2003). Obwohl eine derartige Vorgehensweise zur Erhebung der realen Gegebenheiten zu präferieren ist (vgl. auch Haas und Grams, 2000, S. 190), sind mit ihr jedoch, wie die Diskussion der Arbeit von Rosso (2005) gezeigt hat (vgl. Abschnitt 4.4.8), zahlreiche Probleme verbunden (vgl. auch Crowston und Kwasnik, 2004, S. 7). Daher basieren die in Teil III vorgestellten Analysen sowie die Hypertextsortenontologie ausschließlich auf von dem Verfasser induktiv und introspektiv durchgeführten Untersuchungen innerhalb der Domäne universitärer Webangebote (vgl. Abschnitt 5.8.2).

An der Textlinguistik orientierte Beschreibungen von Hypertextsorten

Kapitel 4 ist ausführlich auf verschiedene Arbeiten zur Charakterisierung konkreter Hypertextsorten eingegangen (vgl. insbesondere die Abschnitte 4.6.2 und 4.6.3). Schütte (2004a) verwendet z. B. einen stringenten Katalog von Kriterien zur natürlichsprachlichen Beschreibung zweier Korpora, die aus den Homepages der Webauftritte russischer und deutscher Unternehmen bestehen. Neben der Kommunikationssituation geht die Verfasserin auf den Kommunikationsgegenstand (das Thema) ein, woraufhin eine Analyse der Makrostruktur der Textexemplare erfolgt, die in einer Untersuchung der Teiltext- und Verweisstrukturen

mündet. Dabei beschäftigt sich Schütte insbesondere mit den hochfrequenten Komponenten „Inhaltsübersicht“, „aktuelle Meldungen“ und „unternehmensvorstellende Kurztexte“.

Jakobs (2003, S. 238) vertritt den Standpunkt, dass zur Charakterisierung von Hypertextsorten keine „per se neuen Modelle notwendig [sind]“ (ebenso Storrer, 2000b, Huber, 2002, und Schütte, 2004a, vgl. auch die Abschnitte 3.5.7 und 3.5.8) und modifiziert das pragmatisch-funktionale Textmustermodell von Sandig (1997) zu einem „Beschreibungsrahmen für Hypertextsorten“ (vgl. zu Sandigs Modell Abschnitt 2.3.5).[5] Im Hinblick auf die nichtsprachlichen Rahmenbedingungen geht Jakobs (2003, S. 239) davon aus, dass der Zweck den zentralen Faktor des Handlungstyps darstellt, wodurch Hypertexte als „funktional-thematisch bestimmte Ganzheiten“ aufgefasst werden können. Zusätzliche Faktoren wie Situationseigenschaften (z. B. die Institution oder das Medium) oder Situationsbeteiligte „kontextualisieren das Handlungsmittel und nehmen Einfluss auf dessen Ausstattung im Sinne von Gestaltungsinventaren, -freiheiten und -restriktionen.“ (ebd.). Bezüglich des sprachlichen Handlungsmittels Hypertextsorte nimmt Jakobs verschiedene Modifikationen an Sandigs Modell vor. Hierzu zählen zunächst neue Typen innerhalb der Handlungshierarchie, die sich durch spezifische Eigenschaften des Mediums Hypertext ergeben, wobei Jakobs (2003, S. 240 f.) „sprachliche“ und „metasprachliche Äußerungen mit Hyperlinkfunktion“ als Beispiele nennt. Zusätzlich wird die bei Sandig in die Handlungshierarchie eingebettete Themenhierarchie von Jakobs (2003, S. 241) als eigenständige und übergeordnete Kategorie betrachtet, weil Hypertexte in vielen Fällen „nach thematischen Gesichtspunkten“ strukturiert sind und die „Darstellung von Inhalt in thematisch geschlossenen Einheiten oder Knoten“ ein „Grundprinzip“ von Hypertext ist.[6] Weiterhin ersetzt Jakobs die primär für linear organisierte Texte intendierte Kategorie Sequenzmuster durch die allgemeinere Kategorie Strukturierungsmuster, um der multilinearen Anordnung von Knoten Rechnung zu tragen. Als Beispiele werden unter anderem die von Storrer (2000b) thematisierten Sequenziertheitsgrade genannt (vgl. hierzu Abschnitt 3.5.1), die von verschiedenen Größen abhängig sind. Jakobs (2003, S. 242) diskutiert in diesem Zusammenhang insbesondere die einfache Konversion von Printtexten, deren Resultat von der Verfasserin als „mediale Variante ihres Printpendants“ und nicht als Exemplar einer Hypertextsorte aufgefasst wird.[7] Eine weitere Möglichkeit der Organisation von Informationseinheiten besteht in einer Strukturierung, die

[5] Abbildung 2.6 (S. 52) zeigt das Textmustermodell von Sandig (1997) in schematischer Form. Jakobs (2003) modifiziert einige der in dem Modell enthaltenen Kategorien, um es zur Charakterisierung von Hypertextsorten zu adaptieren, gibt jedoch keine vergleichbare Abbildung an.

[6] Zur Begründung dieser Modifikation führt Jakobs (2003) einen Bildschirmabzug an, der die thematisch strukturierte Einstiegsseite eines universitären Webauftritts zeigt.

[7] Da die einfache Konversion vorhandener Texte für eine Zweitverwertung im WWW meist mit weit verbreiteten Werkzeugen durchgeführt wird, beinhalten die entstehenden HTML-Dokumente ebenfalls konventionalisierte Textstrukturen, insbesondere hinsichtlich der Ebene des Peritextes. Deshalb wird in der vorliegenden Arbeit die Auffassung vertreten, dass diese „medialen Varianten“ (Jakobs, 2003, S. 242) durchaus eigenständigen Hypertextsorten zugerechnet werden können (vgl. Abschnitt 4.3.2). Jakobs (2003, S. 242) geht implizit davon aus, dass einige Textexemplare im WWW auf Hypertextsorten basieren und andere nicht. Hierdurch ergäbe sich die Notwendigkeit, ein Unterscheidungskriterium einzuführen, um den Status eines Textexemplars – mediale Variante eines Printpendants vs. Instanz einer Hypertextsorte – zu bestimmen; Jakobs geht auf diesen Aspekt nicht ein. Wenn angenommen wird, dass *alle* Texte im WWW Textexemplare von Hypertextsorten sind, dabei aber verschiedene Ausprägungen aufweisen können, entfällt diese Notwendigkeit (vgl. Abschnitt 5.8.1).

„sich an den Rollen [orientiert], die potentielle Nutzer einnehmen können“ (ebd., S. 243).[8] Zur Charakterisierung der Strukturierungsmuster von Hypertextsorten führt Jakobs (2003, S. 243) zwei Beschreibungsebenen ein, die sich auf die (1) „Anordnung und Abfolge der Module eines Systems“ und auf die (2) „Anordnung und Abfolge von Inhalten eines Moduls“ beziehen. Der Terminus „Modul“ wird von Jakobs (2003, S. 245) synonym zu „Webseite“ verwendet (vgl. Fußnote 52, S. 84), d. h. es wird sowohl die Strukturierung der enthaltenen Knoten bezüglich des Gesamthypertextes, als auch die Binnenstrukturierung einzelner Knoten berücksichtigt. Zusätzlich werden für diese beiden Ebenen (a) globale, (b) mittlere und (c) lokale „Beschreibungstiefen“ angenommen (ebd., S. 244), die sich auf die Struktur der Ganzheit, die Strukturierung der Entitäten auf mittlerem Niveau und „Strukturen in Kontaktstellung“ beziehen.[9] Die in dem Modell von Sandig enthaltene Kategorie Formulierungsmuster muss Jakobs (2003, S. 244) zufolge nicht modifiziert werden; fokussiert werden sollten insbesondere die „hypertextspezifische Lexik und Formulierungsmuster für Phänomene wie Links und andere Navigationsmittel“. Die Kategorie der Textgestalt sollte um Beschreibungsmittel für multimediale Elemente erweitert werden (ebd., S. 245). Zusätzlich muss eine Kategorie zur Charakterisierung interaktiver Elemente eingeführt werden, da Interaktivität „[z]u den konstitutiven Merkmalen für Hypertext gehört“ (ebd.). Jakobs (2003, S. 245) unterscheidet Interaktivtiät „im Sinne des sprachlichen Interagierens von Individuen“ und „im Sinne des Reagierens der Computerumgebung auf menschliche Eingaben“, so dass Hypertextsorten in zweierlei Hinsicht „ein mehr oder weniger großes Interaktionspotential“ besitzen (ebd., S. 246; vgl. Fußnote 166, S. 249). Die Kategorie des Durchschnittsumfangs wird von Jakobs (2003, S. 246) aufgehoben, da sie als „im Falle von [...] Hypertextsorten wenig hilfreich“ bezeichnet wird.[10] Die praktische Anwendung dieses Beschreibungsapparats schätzt Jakobs (2003, S. 246) als kritisch und als „alles andere als einfach“ ein.

[8] Jakobs (2003, S. 243) exemplifiziert diese „rollenbasierte Strukturierung“ (in Kapitel 6 werden derartige Homepages als „zielgruppenspezifische Einstiegsseiten“ bezeichnet, vgl. Fußnote 15, S. 308) anhand eines Bildschirmabzugs der Einstiegsseite des Webauftritts eines Universitätsklinikums, die z. B. „Infos für Patienten“, „Infos für Ärzte“, „Infos für Besucher“, „Infos für Studierende“, aber auch „Infos über Pflege“ und „Infos über Einrichtungen“ zur Verfügung stellt, d. h. streng genommen liegt diesbezüglich eine heterogene Strukturierung nach Rollen und zusätzlich nach Themen vor.

[9] Die folgenden Beispiele werden von Jakobs (2003, S. 244) angegeben: (1 a) „mehrfachsequenziert vs. Matrixstruktur“; (2 a) „das Wichtigste zuerst, dann Details vs. Anfang – Mitte – Schluss“. (1 b) „eine thematische Subhierarchie als Teil mehrfachsequenzierter Hypertexte“; (2 b) „Abschnitte einzelner Seiten“. (1 c) „zwei direkt verlinkte Module“; (2 c) „Wortstellung des Satzes“. Bezüglich dieser Beispiele fallen verschiedene Aspekte auf: Zunächst kann nicht pauschal angenommen werden, dass die drei Beschreibungstiefen für *alle* Hypertexte gelten, z. B. setzt (1 b) einen komplex strukturierten Hypertext voraus, d. h. ein nur wenige Knoten umfassender Hypertext besitzt nicht notwendigerweise eine mittlere Strukturierungsebene. Weiterhin wird nicht deutlich, welche anderen Beispiele (neben dem bereits genannten) für (1 c) gelten können, weil die lokale Strukturierung in Form eines Hyperlinks zwischen zwei Knoten in gewisser Hinsicht die Mindestanforderung für diese Beschreibungsdimension darstellt. Ein praktisches Problem bezüglich (1 a) betrifft den Umstand, dass idealtypische Sequenziertheitsgrade nur selten beobachtet werden können. In der Regel liegen Mischtypen vor, wobei sich die Frage nach ihrer Notation und Gewichtung stellt. Die Beispiele für (2 a) und (2 b) zeigen, dass Jakobs die tatsächliche Makrostruktur eines HTML-Dokuments, die neben den eigentlichen Inhalten zusätzlich z. B. Navigationselemente, Suchboxen und Werbebanner etc. enthalten kann, außer Acht lässt.

[10] Diese Einschätzung ist kritisch zu hinterfragen, da diese Kategorie z. B. zur Charakterisierung konkreter Hypertextexemplare der Hypertextsorten *Lexikon* oder *Enzyklopädie* eingesetzt werden kann, bei denen – neben der inhaltlichen Qualität – möglichst viele Artikel vorliegen sollten, so dass eine Enzyklopädie mit z. B. lediglich 100 Einträgen als nicht vollständig ausgeprägtes Hypertextexemplar beschrieben werden könnte.

5.3 Zur Ausrichtung des Hypertextsortenmodells

Die nachfolgenden Ausführungen diskutieren die konzeptionellen Prämissen und die sich aus ihnen ergebenden Schwerpunkte des anschließend dargestellten Hypertextsortenmodells. Abschnitt 5.3.1 geht zunächst aus texttechnologischer Perspektive auf die *Hypertext Markup Language* ein. Aus dieser Betrachtung ergeben sich für das Modell verschiedene Konsequenzen und Anforderungen, die in Abschnitt 5.3.2 thematisiert werden. Daraufhin erläutert Abschnitt 5.3.3 die textlinguistische Ausrichtung des Modells, woraufhin Abschnitt 5.3.4 den Textbegriff des Modells vorstellt. Abschließend geht Abschnitt 5.3.5 auf die Problematik der Granularität der zu untersuchenden Einheiten ein.

5.3.1 Die Hypertext Markup Language aus Sicht der Texttechnologie

Abschnitt 5.2.1 hat verdeutlicht, dass DTDs unterschiedliche Grade von Strukturiertheit zugeschrieben werden können. Für die maschinelle Identifizierung von Hypertextsorten ist es notwendig, Webseiten, d. h. HTML-Dokumente, zu analysieren. Die DTDs der Auszeichnungssprachen HTML bzw. XHTML (Raggett et al., 1999, Pemberton, 2002) stellen ein extremes Beispiel für unterschiedliche Strukturiertheitsgrade dar, weil sie hinsichtlich der Makro- und Mikrostrukturierung von Textexemplaren nur sehr wenige explizite Vorgaben machen. Durch die in den verbreiteten Webbrowsern implementierte fehlertolerante Verarbeitung nicht wohlgeformter HTML-Dokumentinstanzen werden diese Vorgaben zusätzlich aufgeweicht. HTML spezifiziert nicht die Textstrukturmuster einer bestimmten Textsorte, sondern stellt vielmehr ein heterogenes Inventar von Bausteinen zur Verfügung, die aus Sicht des Produzenten ohne jegliche Einschränkungen flexibel kombiniert werden können. Ein zentrales Merkmal der HTML-DTDs betrifft die Unterscheidung von „inline"- und „block"-Elementen. Während sich erstere primär auf die Auszeichnung von Text beziehen (z. B. zur Realisierung von Kursiv- oder Schreibmaschinenschrift), operieren „block"-Elemente auf der makrostrukturellen Ebene und gestatten die freie Anordnung von Überschriften, Tabellen, Listen, Zitaten, Formularen und Absätzen, die ihrerseits „inline"- oder weitere „block"-Elemente enthalten können.[11] Diese äußerst flexiblen Kombinationsmöglichkeiten erlauben die Realisierung von Textexemplaren, die unterschiedlichen Hypertextknotensorten zugerechnet werden können (vgl. Kapitel 4), die wiederum *keine* expliziten Entsprechungen innerhalb des Markups der HTML-Dokumente besitzen. Der Aspekt der Strukturierung von Texten tritt demzufolge bei HTML zugunsten ihrer Präsentation in den Hintergrund (vgl. auch Fußnote 2, S. 262).

Ein weiterer Aspekt betrifft die von den Entwicklern von HTML vorgenommene Vermengung unterschiedlicher Ebenen der Textauszeichnung (vgl. Walker, 1999): Die *Hypertext Markup Language* definiert Auszeichnungselemente für strukturelles (z. B. `h1`, `h2`, `ul`, `ol`, `table` etc.), semantisch-logisches (z. B. `em` und `strong` zur Betonung wichtiger bzw. sehr wichtiger Textteile), präsentationsorientiertes (z. B. `i` und `b` für Kursiv- bzw. Fettdruck; gelegentlich auch als physikalisches Markup bezeichnet), referentielles (`a` und `link`) und funktionales Markup (`script`, `object` und `embed` zur Einbettung von Programmen). Es kann somit

[11] Die Deklaration des Elements `body` in der HTML-DTD umfasst *arbiträre* Sequenzen der Elemente `p`, `h1`, `h2`, `h3`, `h4`, `h5`, `h6`, `ul`, `dl`, `pre`, `dl`, `div`, `noscript`, `blockquote`, `form`, `hr`, `table`, `fieldset`, `address` und `script`.

festgehalten werden, dass es sich bei HTML nicht um eine traditionelle Markup-Sprache zur Annotierung von Textexemplaren *einer* spezifischen Textsorte, sondern vielmehr um eine heterogene, funktional ausgerichtete und in der Mehrzahl aller Dokumente primär für gestalterische Zwecke verwendete Sprache zur Auszeichnung *beliebiger* Textexemplare im *World Wide Web* handelt, die jedoch im Regelfall implizit einer spezifischen Hypertextknotensorte innerhalb einer übergreifenden Hypertextsorte zugehörig sind (vgl. Kapitel 4).

HTML 4.01 (Raggett et al., 1999) stellt mit 91 Elementen und 119 Attributen umfangreiche Möglichkeiten zur Auszeichnung von Texten zur Verfügung, jedoch schöpfen nur die wenigsten Produzenten dieses Inventar vollständig aus (vgl. Abschnitt A.4.2, S. 738 ff.). Stattdessen werden z. B. Überschriften sehr häufig durch eine explizit zentrierte Anordnung (etwa mittels des Elements `center`) und einen Wechsel in eine größere Schriftart realisiert (durch das Element `font`), obwohl HTML für diese textstrukturelle Komponente das Element `h1` bereithält.[12] Auch Listen werden sehr häufig nicht mit Hilfe der Elemente `ul` (*unordered list*) oder `ol` (*ordered list*) ausgezeichnet, sondern z. B. durch explizite Zeilen- oder Absatzwechsel, wobei jedem Listeneintrag eine Grafik (ein farbiger *bullet point*) oder die Aufzählungsnummer vorangestellt wird. Logische Struktureinheiten wie z. B. Überschriften und Listen werden nicht immer über entsprechende HTML-Elemente markiert. Außerdem werden viele Elemente nur sehr selten eingesetzt, da den Produzenten die Existenz dieser Elemente entweder nicht bewusst ist oder ihre Verwendung als überflüssig eingestuft wird.[13]

5.3.2 Sprach- und texttechnologische Anforderungen an das Modell

Aufgrund der im vorangegangenen Abschnitt aus texttechnologischer Perspektive dargestellten Spezifika von HTML ergeben sich mehrere Konsequenzen und Anforderungen für das Hypertextsortenmodell. Das *tag abuse syndrome* (vgl. Fußnote 12) macht es erforderlich, im Rahmen eines übergreifenden Hypertextsortenmodells ein maschinenlesbares Repräsentationsformat vorzusehen, das eine Abbildung textstrukturell und funktional identischer, jedoch mit unterschiedlichen HTML-Auszeichnungen realisierter makrostruktureller Bestandteile von Webdokumenten auf ein *einheitliches* und *überschaubares* Inventar von Makrostrukturkomponenten ermöglicht.[14] Diese Anforderung bezieht sich auf die Rekonstruktion der vom Produzenten intendierten Makrostruktureinheiten, da diese im Falle der Zweckentfremdung von Elementen den zugrunde liegenden HTML-Auszeichnungen nicht unmittelbar entnommen werden können. In diesem Fall ist es notwendig, eine Analyse der „visuellen Semantik" des Markups vorzunehmen: Es sind vor allem typografische Merkmale zu untersuchen, um

[12] Elemente werden von vielen Anwendern nicht zur Auszeichnung von Struktureinheiten, sondern aufgrund ihres typografischen Darstellungseffekts im Browser eingesetzt. Diese Vorgehensweise wird häufig als *tag abuse* bezeichnet. Barnard et al. (1996) gehen davon aus, dass es sich hierbei um die wohl üblichste Art der Dokumentauszeichnung im *World Wide Web* handelt.

[13] Hierzu gehören z. B. die Elemente `cite` (zur Markierung von Literaturverweisen), `dfn` (zur Auszeichnung von Definitionen), `code` (zur Markierung von Quelltexten), `q` (zur Strukturierung eingebetteter Zitate) sowie `abbr` und `acronym` (zur Auszeichnung von Abbreviaturen und Abkürzungen).

[14] Die Begriffe der Makrostruktur bzw. Makrostrukturkomponenten beziehen sich nicht auf die Ebene semantischer Text- oder Absatzpropositionen (van Dijk, 1980, vgl. auch Abschnitt 2.2.2), sondern auf die Ebene der Textoberfläche, die unterschiedliche Konstituenten wie z. B. Überschriften, Abschnitte von Fließtext, Listen, Abbildungen, Navigationshilfen und isolierte Textabschnitte umfasst. Ein generischer maschineller Zugriff auf diese Ebene ist in HTML nicht möglich.

die intendierte Funktion eines makrostrukturellen Bausteins bestimmen zu können. Ein solcher Rekonstruktionsprozess, der auch die Auszeichnungsebenen von HTML einbeziehen muss, (vgl. Abschnitt 5.3.1), sollte in einem System vollautomatisch erfolgen.

Die Repräsentation der Makrostruktureinheiten kann mittels einer Multiebenen-Annotation (vgl. Witt, 2004) unmittelbar in den untersuchten HTML-Dokumenten erfolgen. Sämtliche Inhalte und HTML-Auszeichnungen der zu analysierenden Webseite stellen die Primärdaten dar, die um zusätzliche Elemente angereichert werden können. Diese beziehen sich auf eine abstrakte Repräsentation der Makrostrukturbestandteile, die in den von den HTML-Elementen aufgespannten Baum integriert und durch einen spezifischen Namensraum identifiziert werden. Diese Vorgehensweise ist im Paradigma der Texttechnologie verankert, das auch auf allen weiteren Repräsentationsebenen des Modells verwendet wird, um eine formale Homogenität der beteiligten Datenstrukturen zu gewährleisten, so dass beliebige texttechnologische Werkzeuge und korrespondierende Analyseverfahren eingesetzt werden können.

Neben der Ermittlung und Annotation der Makrostrukturbestandteile kommt ein weiterer Aspekt hinzu: Die in Kapitel 14 vorgestellten prototypischen Komponenten setzen verschiedene sprachtechnologische Standardverfahren wie z. B. *Part-of-Speech*-Tagging ein. Damit diese Verfahren nur auf diejenigen makrostrukturellen Einheiten angewendet werden, die für ihre Funktionalität in Frage kommen (so dass z. B. eine Satzgrenzenbestimmung nur auf Abschnitte angewendet wird, die Fließtext enthalten), müssen zusätzliche Informationen über die Inhalte der Bestandteile ermittelt werden (in obigem Beispiel die Anzahl von Wörtern sowie die Anzahl von Interpunktionszeichen).

Eine übergreifende Anforderung des Hypertextsortenmodells betrifft die Modellierung und Repräsentation einer Typologie von Hypertextsorten, wofür Verfahren gegeben sein müssen, um die beteiligten Ebenen formal repräsentieren und Relationen abbilden zu können. Abschnitt 5.2.1 hat dargestellt, dass diese Anforderungen mit DTDs nicht realisiert werden können. Zusätzlich muss in Betracht gezogen werden, dass Vererbungsmechanismen zur Modellierung unterschiedlicher Spezifizierungsgrade von Hypertexttypen verwendet werden können. Auf diese Weise können z. B. die Hypertextsorten *persönliche Homepage eines Wissenschaftlers* und *private Homepage eines Studierenden* als subgenerische Varietäten des Hypertexttyps *Homepage einer Person* repräsentiert und die übergreifenden Merkmale des Superkonzepts an die Subklassen vererbt werden.

5.3.3 Textlinguistische Ausrichtung des Modells

Das Hypertextsortenmodell soll in der Lage sein, sowohl sprach- und texttechnologische Anwendungen als auch textlinguistische Analysen zu unterstützen, innerhalb derer es möglich sein sollte, auf beliebige linguistische Beschreibungsebenen[15] Bezug zu nehmen und quantitative Verfahren (z. B. Wortartenmarkierung) zu integrieren. Durch den Einsatz texttechnologischer Standardformalismen, die die Verwendung spezifischer Repräsentationsformate gestatten (z. B. die TEI-Dokumentgrammatiken), werden diese Anforderungen in technologischer Hinsicht durch die Einführung zusätzlicher Annotationsebenen erfüllt. Von wesentlicher Bedeutung ist hierbei, wie die Kapitel 3 und 4 gezeigt haben, die Ebene der Makro-

[15] Hiermit sind sowohl Beschreibungen als auch Annotationen gemeint, die sich z. B. auf syntaktische, semantische oder rhetorische Analysen beziehen (vgl. die Abschnitte 2.2.1 bis 2.2.5).

struktur einzelner HTML-Dokumente (vgl. auch Schütte, 2004a), die im Zuge der Modularisierung von Text- und Informationseinheiten in grundlegender Weise von traditionellen Printtexten differiert (vgl. Abschnitt 3.5.6).[16] Einen weiteren Schwerpunkt bildet die Repräsentation von Hyperlinkstrukturen, die einen relationalen Formalismus erfordert (vgl. Abschnitt 6.5). Da die praktische Durchführung von textlinguistischen Analysen in Bezug auf Hypertextsorten oftmals als sehr komplex eingestuft wird (vgl. z. B. Haas und Grams, 2000, und Jakobs, 2003), sollte darüber hinaus eine bestmögliche Unterstützung durch maschinelle Verfahren gewährleistet sein. Dies geschieht in dem vorliegenden Modell einerseits durch die Analyse-Schnittstellen der Korpusdatenbank (vgl. Kapitel 7), andererseits durch verschiedene Möglichkeiten der Visualisierung von Analyseergebnissen und grundlegenden Datenstrukturen, die die Relationen zwischen Entitäten grafisch verdeutlichen.

5.3.4 Der Textbegriff und die Charakterisierung von Hypertexten

Die Ausführungen in den Kapiteln 3 und 4 hinsichtlich der generellen Eigenschaften von Hypertexten im WWW verdeutlichen, dass eine gewisse Diskrepanz in Bezug auf die meist sehr abstrakten Annahmen besteht, auf denen viele Textdefinitionen basieren. Die Auffassung von Text „im Sinne einer komplex strukturierten, kommunikativ, konzeptuell und thematisch gegliederten Einheit, die vorrangig aus Sätzen besteht“ (Bittner, 2003, S. 113) kann nicht ohne Weiteres auf Hypertexte angewendet werden (vgl. auch Abschnitt 3.5.6). Gansel und Jürgens (2002, S. 31) betonen in diesem Zusammenhang, dass sich die linguistische Kategorie Text einer eindeutigen und exhaustiv für sämtliche Textexemplare zutreffenden Definition entzieht (vgl. Abschnitt 2.2.8). Auch Adamzik (2004, S. 39 f.) thematisiert diesen Aspekt und räumt der Textdefinition einen nur sekundären Stellenwert ein, da neuere textlinguistische Ansätze die Zusammenstellung unterschiedlicher Beschreibungsaspekte anstreben, die sich zur Charakterisierung von Texten eignen.[17] Aus diesem Grund verwendet Schütte (2004a, S. 140 ff.) in ihrer Studie „keine fest umrissene Textdefinition“, sondern einen „offenen Merkmalskatalog“. Vor dem Hintergrund der von Sandig (2000) vertretenen Sichtweise (vgl. die Abschnitte 2.2.9 und 2.3.6) geht Storrer (2004b, S. 23) auf diese Problematik im Kontext von Hypertexten ein, denn Dokumente und Dokumenttypen erweisen sich „als mehr oder weniger gute Vertreter des Textkonzepts, je nachdem, wie viele und welche Merkmale sie aufweisen. Die rein theoretische Frage, ob Hyperdokumente Texte sind oder nicht, verschiebt sich damit zur praxisrelevanten Frage, worin sich die verschiedenen Typen digitaler Dokumente vom prototypischen Textkonzept unterscheiden, wo Gemeinsamkeiten und Unterschiede zu gedruckten Dokumenten bestehen und durch welche texttechnologischen Verfahren man diesen Unterschieden Rechnung tragen kann.“

Es wurde mehrfach betont, dass die kommunikative Funktion einen zentralen Stellenwert bei der Charakterisierung von Textsorten besitzt und von aktuellen, primär an der Pragmatik ausgerichteten Textdefinitionen entsprechend reflektiert wird. Beispielsweise bezeichnet der

[16] Oftmals wird (z. B. von Bucher, 2004) lediglich zwischen einer „Mikroebene“ (Knoten) und einer „Makroebene“ (Verknüpfung von Knoten) unterschieden. In Anlehnung an van Dijk (1980) erscheint es adäquater, die Ebenen der Mikro-, Makro- und Superstruktur auch auf Hypertexte anzuwenden (vgl. Abschnitt 2.2.2).

[17] Kapitel 3 stellt diesbezüglich eine Vielzahl von Aspekten vor, die zur Charakterisierung von Hypertexten im *World Wide Web* eingesetzt werden können.

Terminus Text nach Brinker (2001, S. 17) „eine begrenzte Folge von sprachlichen Zeichen, die in sich kohärent ist und die als Ganzes eine erkennbare kommunikative Funktion signalisiert." Obwohl sich Hypertexte von Texten in vielerlei Hinsicht unterscheiden, besteht kein Grund zu der Annahme, dass Hypertexte *nicht* dem zentralen und Brinker zufolge obligatorischen Kriterium unterliegen, eine kommunikative Funktion besitzen zu müssen. Huber (2002, S. 51 f.) setzt diese Textdefinition als Grundlage seines Hypertextanalysemodells ein und hebt ihre Stärken bezüglich einer Anwendung auf das *World Wide Web* hervor: Zunächst wird Text als eine „begrenzte Folge von sprachlichen Zeichen" aufgefasst, weshalb das WWW in seiner Gesamtheit nicht in den Skopus dieser Definition fällt (vgl. auch Storrer, 2000b, sowie Jakobs, 2003, S. 236). Vielmehr können Hypertexte als Äquivalent des gedruckten Textes in Hypertextumgebungen gelten. Im Hinblick auf das Kriterium der „in sich [vorhandenen] Kohärenz" ist Huber (2002, S. 51) der Ansicht, dass Kohärenzstrukturen primär „zwischen Knoten *eines* Hypertextes" als in Hypertextnetzen wie dem WWW existieren, deren Übergänge als „tendenziell eher weniger kohärent" eingestuft werden (ebd.). Der entscheidende Vorteil dieser Textdefinition hinsichtlich ihrer Anwendung auf Hypertexte, die im *World Wide Web* publiziert werden, betrifft ihren sehr umfassenden Geltungsbereich: Dieser ermöglicht es, nahezu allen Hypertexten den Status „Text" zuzuweisen.

Im *World Wide Web* verfügbare Hypertexte werden üblicherweise als „Website" bezeichnet (vgl. auch Abschnitt 3.3.7).[18] Mit Bezug auf die Textdefinition von Brinker (2001, S. 17) kann eine Website somit als „funktional-thematisch bestimmte kommunikative Ganzheit" (Jakobs, 2003, S. 236) konzeptualisiert werden, die aus beliebig vielen Knoten in Form von HTML-Dokumenten besteht, die untereinander durch Hyperlinks verknüpft sind. Diese Knoten können wiederum als einzelne, das Kriterium der kohäsiven Geschlossenheit in der Regel erfüllende Teiltexte aufgefasst werden, die in spezifischen Relationen zum übergeordneten Kommunikat stehen. Die in Knoten enthaltenen Teiltexte besitzen somit ebenfalls Textstatus, deren Inhalte und Funktionen jedoch im Allgemeinen unterstützend auf den umfassenden Hypertext und dessen spezifische kommunikative Funktion abzielen. Die Teiltexte können neben Fließtext (im Sinne des „klassischen" Textes als monolinear sequenzierte Abfolge von Sätzen, vgl. Abschnitt 2.2) zahlreiche weitere Konstituenten besitzen. Hierzu gehören beispielsweise Listen in unterschiedlichen Ausprägungen, Textfragmente, einzelne Sätze, Phrasen und Wörter, Navigationshilfen, Tabellen, interaktive Formulare und multimediale Objekte aller Art (vgl. die Abschnitte 3.5.6 und 5.3.1).

Eine Website wird im Regelfall von einem spezifischen Thema dominiert. Schütte (2004a, S. 211) beschäftigt sich mit den Einstiegsseiten der Webpräsenzen russischer und deutscher Unternehmen. Bezüglich des Kommunikationsgegenstandes stellt sie für diese Untersuchungsdomäne (in Anlehnung an Brinker, 2001, S. 142) die prägnante Gleichung „Emittent = Thema" auf, d. h. „das Unternehmen [bildet] zugleich die gemeinsame Einordnungsinstanz, den einheitlichen thematischen Bezugspunkt, dem sich alle auf der Website anzutreffenden Teilthemen unterordnen." (vgl. auch Adamzik, 2004, S. 85). Obwohl diese Gleichung z. B. auch auf persönliche Homepages anwendbar ist, gilt sie jedoch nicht für beliebige Websites, die sich beispielsweise mit einem bestimmten Film-Genre oder spezifischen Aspekten

[18] Der Begriff „Website" ist mehrdeutig. Einerseits wird er als Synonym für einen WWW-basierten Hypertext verwendet, andererseits bezieht er sich auf die Ebene der Kommunikationsform (vgl. Abschnitt 4.2.4).

der Informationstechnologie beschäftigen. Dass nicht jede Website über eine Anbindung an einen Kommunikationsgegenstand verfügt, zeigen Suchmaschinen, die sich primär über ihre Funktion charakterisieren lassen. Auch Online-Zeitungen sind nicht durch ein einzelnes übergeordnetes Thema erfassbar. Im Falle universitärer Webangebote kann die Gleichung Emittent = Thema jedoch durchaus zur Beschreibung des übergreifenden Hypertexttyps herangezogen werden, da – mit Ausnahme einiger untergeordneter Hypertextsorten[19] – sämtliche Knoten eines universitären Hypertextes auf einen einheitlichen thematischen Bezugspunkt in Gestalt der anbietenden Hochschule bezogen werden können. Die genannten Beispiele können klar von literarischen Hypertexten (d. h. Hyperfiction) abgegrenzt werden, weshalb sie in Anlehnung an die in der Textlinguistik übliche Terminologie auch als Gebrauchshypertexte bezeichnet werden.[20] Bereits in Abschnitt 3.7 wurde argumentiert, dass fast alle Angebote im WWW zur Gruppe der Gebrauchshypertexte zählen und dass literarische Hypertexte in ihrer prototypischen Ausprägung als unsequenziert organisierte und zahlreiche voneinander differierende Lesepfade anbietende Konglomerate von Knoten eher den Status einer Randerscheinung besitzen.

Sowohl Texte als auch Hypertexte (insbesondere die im WWW angebotenen) sind, wie die Kapitel 3 und 4 gezeigt haben, sortenspezifisch, wobei im Falle von Hypertexten der Bezug zu einer Hypertextsorte in vielen Fällen nur impliziter Natur ist (vgl. Abschnitt 4.3). Die kommunikative Funktion und partiell auch das Thema eines Hypertextes im *World Wide Web* determiniert die zu verwendende Hypertextsorte, die wiederum in obligatorischer oder optionaler Weise zusätzliche Hypertextsorten umfassen kann. Bei universitären Webangeboten kann die anbietende Institution als Thema aufgefasst werden, so dass ein unmittelbarer Zusammenhang zwischen den innerhalb einer Universität existenten Organisationsstrukturen und den korrespondierenden Hypertexten besteht (vgl. die Abschnitte 6.3.4 und 6.5). Es ist also notwendig, den Kommunikationsgegenstand – d. h. die generische Struktur einer Hochschule – zu erfassen, um Korrespondenzen zu den jeweiligen Hypertexttypen bzw. Hypertextsorten und somit ihre Geltungsbereiche ermitteln zu können.[21]

[19] Hierzu gehören generellere Angebote wie z. B. Webseiten mit Informationen über die Region, in der eine Hochschule angesiedelt ist oder private Homepages von Studierenden. Diese nehmen zwar häufig Bezug auf die Universität, an der sie studieren, hierzu besteht jedoch keine Notwendigkeit.

[20] Brinker (2001, S. 20) verwendet den Begriff der Gebrauchstexte zur Bezeichnung nichtliterarischer Texte und greift dabei auf die Definition von Dimter (1981, S. 35) zurück, der unter diesem Terminus diejenigen Texte subsumiert, mit denen „kein besonderer ästhetisch-literarischer Anspruch“ verbunden ist. Brinker (2001, S. 20) weist darauf hin, dass nicht immer eine klare Grenze zwischen literarischen Texten und Gebrauchstexten gezogen werden kann, was insbesondere für die „literarischen Gebrauchsformen“ wie z. B. Briefe, Memoiren, Essays und Predigten gilt. Obwohl mit den meisten Hypertexten im WWW kein literarischer Anspruch verbunden ist, so ist gerade bei Angeboten, die von professionellen Webdesignern gestaltet wurden, ein ästhetischer Anspruch auszumachen. In Bezug auf gedruckte Texte ist dieser jedoch eher mit der Typografie und dem Textdesign (also mit dem Peritext) als mit dem eigentlichen Inhalt des Textes zu vergleichen.

[21] Storrer (2004b, S. 30) weist im Kontext der Diskussion von Kohärenzaspekten in Hypertexten darauf hin, dass Texte so gestaltet sein sollen, „dass die Rezipienten den Zusammenhang zwischen den Textkonstituenten erkennen und sich ein kohärentes mentales Modell des Textinhalts aufbauen können. Dies gilt insbesondere im Kontext des Wissenstransfers und ist weitgehend unabhängig von der Wahl des Mediums, gilt also auch für den Hypertext.“ Übertragen auf universitäre Webangebote ergibt sich hierdurch die Konsequenz, dass sie es dem Rezipienten ermöglichen sollten, sich ein möglichst kohärentes mentales Modell einer Hochschule aufbauen zu können. Dieser Umstand verdeutlicht ebenfalls den Stellenwert der generischen Strukturierung einer (deutschen) Universität bezüglich der korrespondierenden und naturgemäß generischen Hypertextsorte.

5.3.5 Die Granularität der zu analysierenden Entitäten

Analysen von Hypertexten und somit auch Charakterisierungen von Hypertextsorten betreffen unterschiedlich granulare Entitäten. Neben der Struktur, die durch die Verknüpfung von Knoten entsteht, bezieht sich die Frage nach der Granularität der zu analysierenden Entitäten vor allem auf die Größe und den Inhalt der in einem Hypertext enthaltenen Knoten (vgl. auch Jakobs, 2003, S. 236). Zwei Aspekte sind in diesem Zusammenhang von entscheidender Relevanz: Erstens kann ein Hypertext als kommunikativ bestimmte funktional-thematische Ganzheit weitere Hypertexte umfassen, die ebenfalls dieser Definition entsprechen und sich in aller Regel auf den übergreifenden Hypertext beziehen (vgl. die Abschnitte 4.5.2 bis 4.5.4). Zweitens existiert die Problematik, dass Instanzen von Hypertextsorten wie z. B. der persönlichen Homepage eines Wissenschaftlers nicht notwendigerweise lediglich einen Einzelknoten umfassen, sondern auf mehrere Knoten verteilt sind, so dass sich die Frage stellt, welchen Textsortenstatus in diesem Fall die Knoten besitzen.

Crowston und Williams (1997, S. 37) berichten im Rahmen der Darstellung ihrer Stichprobenanalyse (vgl. Abschnitt 4.4.1), dass zahlreiche HTML-Dokumente gefunden wurden, die Bestandteile längerer Dokumente sind (z. B. ein Index oder ein Kapitel). In vielen Fällen sind diese Dokumente, so Crowston und Williams, durchaus noch als Teile eines spezifischen Genres erkennbar, wobei jedoch häufig die Funktion eines Dokuments nicht eindeutig bestimmt werden kann. Im Hypertextsystem *World Wide Web* besitzen HTML-Dokumente den Status atomarer Knoten. Crowston und Williams (1997, S. 38) gehen davon aus, dass auch der einzelnen Webseite ein Genre zugewiesen werden sollte: Während in gedruckten Texten die Paginierung durch die physikalischen Abmessungen der bedruckten Seiten bestimmt wird und ihr somit im Regelfall keine Bedeutung zugesprochen werden kann, existiert eine derartige „natürliche“ Beschränkung im Hypertextsystem WWW nicht. Die Aufteilung in Seiten sollte also Crowston und Williams zufolge die tatsächliche Struktur des Kommunikats reflektieren, so dass dem Rezipienten der Inhalt und die Form eines Einzeldokuments deutlich werden. Bei dieser Einschätzung handelt es sich jedoch lediglich um eine Gestaltungs- bzw. Strukturierungsempfehlung, denn prinzipiell unterliegen die Autoren von Webseiten diesbezüglich keinen Restriktionen. Gerade anhand der in Kapitel 4 dargestellten Gebrauchshypertextsorten wird deutlich, dass die Aufteilung tatsächlich arbiträr ist: Auf der persönlichen Homepage eines Wissenschaftlers erwarten Rezipienten unter anderem eine Publikationsliste. Diese kann einerseits unmittelbar in denjenigen Knoten integriert werden, der als Einstiegsseite fungiert, andererseits jedoch auch in einen untergeordneten Knoten ausgelagert werden. Auf dieser Ebene können sowohl sehr heterogene und unterschiedliche Funktionen in sich vereinende, als auch funktional homogene Knoten existieren. Die Untersuchung von Hypertextsorten kann sich somit nicht ausschließlich auf die Ebene des Knotens beschränken, sondern muss zusätzlich den Status potenziell eingebetteter Entitäten einbeziehen.

5.4 Ebene 1: Hypertexttypen und Hypertextsorten

Das Hypertextsortenmodell bezieht sich auf drei Granularitätsebenen. Hierbei handelt es sich um die Ebenen des (i) Hypertexttyps bzw. der Hypertextsorte, (ii) des Hypertextknotentyps bzw. der Hypertextknotensorte (Abschnitt 5.5) und (iii) des Hypertextsortenmoduls (Ab-

schnitt 5.6).[22] Die Termini „Genre“, „Digital Genre“ oder „Web-Genre“ werden aus zwei Gründen nicht verwendet: Erstens existieren im *World Wide Web* vornehmlich Gebrauchshypertexte, so dass der (in der deutschsprachigen Forschung) literaturwissenschaftlich geprägte Begriff „Genre“ nicht dazu geeignet erscheint, die kommunikative Realität in adäquater Weise reflektieren zu können. Zweitens wird dieser Begriff (in der angelsächsischen Forschung) ohne präzise Definition verwendet (vgl. Abschnitt 2.3.7). Aus diesem Grund bietet es sich an, die zentralen Begrifflichkeiten an die etablierten Termini Texttyp und Textsorte anzulehnen (vgl. Abschnitt 2.3.2).[23] Abbildung 5.2 stellt die genannten Konzepte anhand des Webauftritts einer Universität als Beispiel vor.[24]

Die Begriffe Hypertexttyp bzw. Hypertextsorte bezeichnen den generellen Typ einer Website. Diese Termini beziehen sich auf den abstrakten Typ eines Hypertextes im Sinne einer funktional und gegebenenfalls thematisch markierten kommunikativen Ganzheit, die durch mindestens einen, in der Regel jedoch mehrere Hypertextknoten instanziiert wird. Die Gesamtheit der instanziierenden Knoten wird als Hypertextexemplar, Hypertextinstanz, Hypertext oder Website bezeichnet. Hypertexttypen legen – auf einer sehr generischen Ebene – den strukturellen und inhaltlichen Aufbau eines Hypertextes fest und umfassen sowohl obligatorische als auch optionale Konstituenten. Der Begriff der Konstituenten bezieht sich – in Anlehnung an Furuta (1989) – auf drei unterschiedliche Granularitätsstufen: Instanzen eines Hypertexttyps können Instanzen weiterer Hypertexttypen bzw. Hypertextsorten einbetten, die ebenfalls als inhaltlich-funktional bestimmte Ganzheiten fungieren. Weiterhin umfassen sie spezifische Hypertextknotentypen bzw. -sorten (Abschnitt 5.5) und Hypertextsortenmodule (Abschnitt 5.6). Hypertexttypen stellen Hypertextsorten[25] im eigentlichen Sinne dar, d. h. sie beinhalten mehrere Knoten, die wiederum individuelle und vom übergeordneten Hypertexttyp sowie den eingebetteten Hypertextsortenmodulen abhängige Typen bzw. Sorten besitzen. Die Termini Hypertexttyp und Hypertextsorte (sowie auch Hypertextknotentyp und Hypertextknotensorte) lehnen sich im Hinblick auf ihren Geltungsbereich

[22] Verschiedene Aspekte einer ersten Version dieses Modells wurden in Rehm (2002b, 2004b, 2004c) publiziert. Dieses Kapitel stellt eine revidierte, in Bezug auf zahlreiche Aspekte erweiterte und präzisierte Version dar.

[23] Von einem vollständigen terminologischen Konsens kann jedoch in der deutschsprachigen Textlinguistik (noch) nicht die Rede sein, doch bieten die genannten Begriffe eine hinreichend adäquate Möglichkeit zur Differenzierung der unterschiedlichen Bestandteile von Hypertextsorten.

[24] Es ist zu betonen, dass nahezu alle verwandten Arbeiten (vgl. insbesondere Abschnitt 4.4) sowie alle bislang vorliegenden Ansätze zur maschinellen Identifizierung von Web-Genres davon ausgehen, dass *ein einzelnes* HTML-Dokument die vollständig ausgeprägte Instanz *eines* Web-Genres darstellt, d. h. weder übergreifende Strukturen noch die Binnenstrukturen von Knoten werden berücksichtigt. In Abbildung 5.2 werden die Ebenen der Hypertextsorte und des Hypertextsortenmoduls visuell hervorgehoben. Die dritte Ebene der Hypertextknotensorte bezieht sich auf einzelne Knoten, die in der Abbildung klar als solche zu erkennen sind.

[25] Jakobs (2003, S. 250) berichtet von einem Versuch, den Verbreitungsgrad des Begriffs „Hypertextsorte“ durch eine Befragung von 90 Studierenden zu ermitteln. Die Teilnehmer sollten diejenigen Hypertextsorten aufzählen, die sie kennen und am häufigsten benutzen. Die Antworten umfassten primär technische Aspekte (Markup-Sprachen, Programmiersprachen, Datenformate und Programme) sowie „ein breites Sammelsurium von Phänomenen“ (z. B. Elemente von Hypertext, Adressen von Websites, das Internet, Kommunikationsdienste, Fachliteratur etc.). Jakobs (2003, S. 250) folgert hieraus, dass nur „die wenigsten etwas mit dem Begriff [Hypertextsorte] anfangen können, obwohl sie Fächer belegen (Germanistik, Kommunikationswissenschaft), die ein gewisses theoretisches und praktisches Interesse für Texte aller Art vermuten lassen.“ Es ist jedoch zu beachten, dass der Begriff „Hypertextsorte“ bislang nur in sehr wenigen Publikationen verwendet wird, was das von Jakobs ermittelte sehr breite Antwortspektrum erklärt.

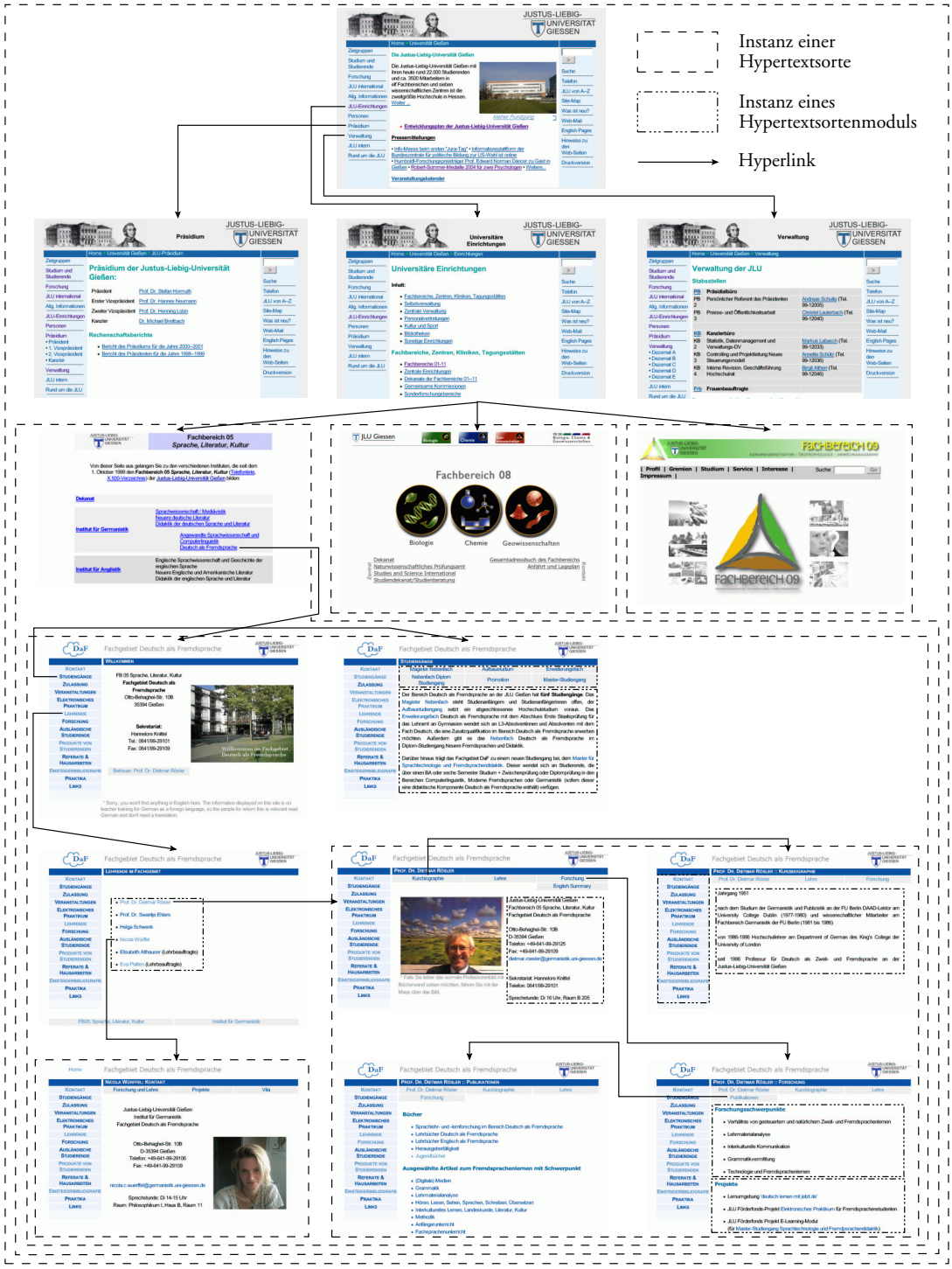


Abbildung 5.2: Einbettung der Instanzen von Hypertexttypen und Hypertextsortenmodulen

an Heinemann und Heinemann (2002, S. 142 ff.; vgl. auch Abschnitt 2.3.2) an. Der Begriff Hypertexttyp bezieht sich demnach auf Hypertextklassen mit einem sehr umfangreichen Geltungsbereich, wenigen distinktiven Merkmalen und einem entsprechend hohen Abstraktionsgrad, während sich der Terminus Hypertextsorte auf Hypertextklassen mit einem eher geringen Geltungsbereich, einer größeren Menge distinktiver Merkmale und einem niedrigen Abstraktionsgrad bezieht. Der Begriff Hypertextsorte ist, wie bereits in Fußnote 7 (S. 266) angesprochen, nicht hinsichtlich der Herkunft des zugrunde liegenden Musters markiert. Wenn z. B. eine ursprünglich aus dem Printbereich stammende Textsorte wie *wissenschaftlicher Artikel* als untereinander verknüpfte Gruppe von HTML-Dokumenten umgesetzt wird, dann handelt es sich hierbei – unter den in Abschnitt 4.3.2 dargestellten Prämissen – um eine Instanz der Hypertextsorte *wissenschaftlicher Artikel* (vgl. auch Abschnitt 3.3.6).

Zwischen Hypertexttypen und Hypertextknotentypen kann keine trennscharfe Grenze gezogen werden: In aller Regel werden die unterschiedlichen Konstituenten von Hypertexttypen zwar in einzelne Knoten aufgeteilt, doch entscheidet letzten Endes immer der Produzent, ob zwei Konstituenten, die zwar in der Mehrzahl der korrespondierenden Instanzen von Hypertexttypen als separate Knoten ausgeprägt sind, tatsächlich aufgeteilt oder innerhalb eines Einzelknotens aggregiert werden.[26] Hinsichtlich der Aggregierung oder Separierung einzelner Bestandteile können somit ebenfalls sortenspezifische Konventionen existieren, die – ebenso wie alle weiteren Bestandteile von Hypertexttypen – mittels empirischer Erhebungen zu ermitteln sind, die wiederum den Status Quo innerhalb einer Diskursgemeinschaft reflektieren (vgl. hierzu Abschnitt 4.3.2).[27] Um eine alternative Formulierung anzuführen: Hypertexttypen bzw. -sorten spezifizieren eine gewisse Menge obligatorischer und optionaler Hypertextsortenmodule, die in ihrer Gesamtheit wiederum in mindestens einem Knoten ausgeprägt werden. Wenn eine empirisch beobachtbare Konvention – z. B. im Hinblick auf die Anordnung und auch Positionierung zweier spezifischer Hypertextsortenmodule in einem Knoten – vorliegt, dann kann diesbezüglich – abhängig vom Abstraktionsgrad – von einem Hypertextknotentyp oder einer Hypertextknotensorte gesprochen werden. Die nachfolgenden Abschnitte gehen in detaillierter Form auf diesen Aspekt ein.

Von den modifizierten Beschreibungsebenen, die Jakobs (2003) für das Textmustermodell von Sandig (1997) vorschlägt, werden verschiedene Aspekte übernommen. Die Modifikationen betreffen im Wesentlichen die Einführung einer neuen Beschreibungskategorie zur Erfassung des Interaktivitätspotenzials einer Hypertextsorte und die Generalisierung der Sequenz- zu Strukturierungsmustern (vgl. Abschnitt 5.2.2). Das in dieser Arbeit eingeführte Hypertextsortenmodell sieht vor, dass Hypertextsortenmodule unterschiedliche Ausprägungen hinsichtlich der Typen Inhalt bzw. Thema, Interaktion, Kommunikation, Navigation,

[26] Abbildung 3.7 (S. 151) zeigt eine rudimentäre Typologie von Hypertexttypen im *World Wide Web*. Einige der Blattknoten dieser Typologie deuten die dargestellte Überlappung der Konzepte Hypertexttyp und Hypertextknotentyp an, so kann z. B. der Hypertexttyp *Homepage einer Person* (in der Abbildung verkürzend als „Private Sites“ bezeichnet) als ein einzelnes Dokument instanziiert werden, aber auch mehrere Dutzend HTML-Dokumente umfassen. Im letzten Fall fungiert der Hypertexttyp primär als organisierendes Schema, im ersten Fall liegt eine Verschmelzung von Hypertexttyp und Hypertextknotentyp vor.

[27] Ein wesentlicher Unterschied zwischen traditionellen Textsorten und Hypertextsorten betrifft den Umstand, dass Hypertextsorten nur mit äußerst detaillierten Kenntnissen der jeweiligen Untersuchungsgegenstandes (z. B. universitäre Webangebote) *intuitiv* erfassbar sind.

Metainformation, Dekoration und Textstrukturmuster besitzen.[28] Ein kleinformatiges Suchformular umfasst z. B. einen hohen Wert bezüglich der Interaktion, wohingegen das Merkmal Thema nicht ausgeprägt ist, so dass diesem Hypertextsortenmodul der primäre Typ Interaktion zugeschrieben werden kann (Abschnitt 5.6 geht genauer auf dieses Thema ein). In ähnlicher Weise können derartige Ausprägungen für Hypertextsorten in Bezug auf die Merkmale Inhalt bzw. Thema, Interaktion, Kommunikation und Dekoration erfasst werden.[29] Somit kann z. B. beschrieben werden, dass die Instanzen einer spezifischen Hypertextsorte im Hinblick auf das Webdesign in der Regel sehr dekorativ oder eher schlicht gestaltet werden, falls diesbezüglich Konventionen ermittelt werden können. Neben diesen Merkmalen wird für Hypertextsorten als übergreifendem Handlungstyp zusätzlich die zentrale Kategorie der kommunikativen Funktion, eine Beschreibungsdimension zur Erfassung von Strukturierungsmustern und ein Merkmal zur Repräsentation kontextueller Faktoren der Kommunikationssituation (Handlungsbeteiligte, Institution etc.) angenommen.[30] Jede Hypertextsorte besitzt also neben dem Inventar von Hypertextknotentypen[31] und Hypertextsortenmodulen eine Art standardisierte, durch die Fundierung auf empirischen Analysen jedoch die kommunikative Realität reflektierende Vorbelegung hinsichtlich der Merkmale kommunikative Funktion, Kontextfaktoren, Inhalt bzw. Thema, Interaktion, Strukturierung, Kommunikation und Dekoration. Eine derartige Vorbelegung gestattet es, Varietäten in Instanzen einer Hypertextsorte zu erfassen, die zusätzliche optionale Hypertextknotentypen sowie optionale Hypertextsortenmodule ausprägen. Die Vorbelegung fungiert somit als wesentlicher Bestandteil des prototypischen Kerns einer Hypertextsorte (vgl. Abschnitt 5.8.1).

Abbildung 5.2 umfasst zur Verdeutlichung der grundlegenden Konzepte einen sehr kleinen Ausschnitt einer Instanz der Hypertextsorte *Webauftritt einer Universität*. Die Instanz dieser übergreifenden Hypertextsorte umfasst zahlreiche weitere Instanzen, z. B. bietet ein von der Einstiegsseite erreichbarer Knoten eine Liste der unterschiedlichen Organisationseinheiten und Hyperlinks zu den Websites dieser Einheiten an. Diese Websites sind Instanzen der Hypertextsorte *Webauftritt eines universitären Fachbereichs* und obwohl sie doch grafisch in sehr unterschiedlicher Weise gestaltet sind, weil die Erstellung und Pflege im Verantwortungsbereich der jeweiligen Einheiten liegt (vgl. Abschnitt 6.3.4), umfassen sie nahezu identische Inventare von Konstituenten. Die übergeordnete Hypertextsorte sieht obligatorische Instanzen der Hypertextsorte *Webauftritt eines universitären Fachbereichs* vor, ebenso wie die Exemplare

[28] Diese sowie die im Folgenden genannten Typen sind nicht als statisches Inventar von Typen und Beschreibungsmerkmalen zu verstehen. Vielmehr bilden die Kriterien einen initialen Vorschlag, der wesentliche Beschreibungskategorien umfasst, jedoch im Hinblick auf ein spezifisches Forschungsinteresse entsprechend zu modifizieren bzw. zu spezifizieren ist.

[29] Eine globale Vorbelegung auf der Ebene der Charakterisierung einer Hypertextsorte hinsichtlich der Merkmale Navigation und Metainformation ist nicht möglich, da sich diese ausschließlich auf die Ebenen der Hypertextsortenmodule und Hypertextknotentypen beziehen.

[30] Es wurde mehrfach darauf hingewiesen, dass Strukturierungsmuster, wie sie z. B. von Storrer (2000b) dargestellt werden (vgl. Abschnitt 3.5.1), nur selten in Reinform vorliegen, weshalb diese Kategorie auch die Erfassung von Mischungen bezüglich der eingesetzten Sequenziertheitsgrade erlauben sollte.

[31] Diese prägen in ähnlicher Weise konkrete Werte mit Bezug auf die Merkmale Positionierung (von Hypertextsortenmodulen, die eine Hypertextknotensorte vorsieht), kommunikative Funktion (der kommunikativen Funktion der Hypertextsorte untergeordnet) und Dekoration (des einzelnen Knotens) aus. Weitere Merkmale werden durch die spezifischen Werte der in einem Knoten enthaltenen Instanzen von Hypertextsortenmodulen konstituiert. Dieser Aspekt wird in den nachfolgenden Abschnitten diskutiert.

dieser Kategorie Instanzen der Hypertextsorte *Webauftritt eines Instituts bzw. Seminars* (in der Abbildung nicht dargestellt) und diese wiederum Instanzen des *Webauftritts einer Professur bzw. Arbeitsgruppe* vorsehen. Für die Instanz der letztgenannten Hypertextsorte sind in Abbildung 5.2 zwei Knoten dargestellt, die Informationen über die von diesem Fachgebiet angebotenen Studiengänge sowie eine Liste der Lehrenden enthalten. Die Liste der Lehrenden verweist auf Instanzen der Hypertextsorte *persönliche Homepage eines Wissenschaftlers*, die wiederum Informationen zu aktuellen Projekten und Lehrveranstaltungen sowie Publikationslisten enthalten. Sämtliche der genannten untergeordneten Hypertextsorten beziehen sich auf die Instanz der Hypertextsorte *Webauftritt einer Universität*, die in funktional-thematischer Hinsicht das Rahmenmodell des gesamten universitären Webauftritts determiniert.

5.5 Ebene 2: Hypertextknotentypen und Hypertextknotensorten

In Abschnitt 5.4 wurden Hypertexttypen thematisiert, deren Instanzen von den Knoten gebildet werden, die einem Hypertext zugehörig sind. Die einzelnen Knoten, d. h. HTML-Dokumente, können wiederum auf Hypertextknotentypen bzw. Hypertextknotensorten basieren. Während Hypertexttypen das abstrakte, funktional-thematische Rahmenmodell einer Website und ihre obligatorischen und optionalen Komponenten (d. h. Hypertextsortenmodule) spezifizieren, entsprechen HTML-Dokumente und somit auch die Instanzen von Hypertextknotentypen aus Sicht der Rezipienten am ehesten dem Begriff der Textexemplare, weil in der Regel jeweils nur ein einzelner Knoten auf dem Bildschirm dargestellt wird. Da dem Produzenten die Aufteilung der Konstituenten eines Hypertexttyps freigestellt ist, ist diesbezüglich bei der Analyse konkreter Hypertexte eine größere Varianz zu erwarten als im Hinblick auf die tatsächliche Ausprägung der Hypertextsortenmodule eines Hypertexttyps (in einem Hypertextexemplar realisiert vs. nicht realisiert).[32] Die hier dargestellte Dimension des Hypertextknotentyps ist sowohl für die rein deskriptive Analyse als auch für maschinelle Methoden von Relevanz, da sie zwangsläufig auf die Ebene des einzelnen HTML-Dokuments anzuwenden sind.

Hypertextknotentypen bzw. -sorten werden durch einzelne HTML-Dokumente instanziiert, die auf der Textoberfläche Bausteine enthalten, die im Folgenden als Hypertextmodule bezeichnet werden. Diese Bausteine erfüllen die unterschiedlichsten Funktionen (Bereitstellung von Navigationselementen, Darstellung von Inhalt etc.), weshalb Instanzen von Hypertextknotentypen in funktionaler und thematischer Hinsicht hochgradig heterogener Natur sein können. Die Differenzierung zwischen Hypertextknotentypen und -sorten bietet sich

[32] Bei der in Abbildung 3.8 (S. 153) dargestellten Typologie von Hypertextknotentypen handelt es sich aus verschiedenen Gründen nur um eine rudimentäre Typologie. Zunächst geht sie von einer funktional-inhaltlichen Eindeutigkeit aus, die nur selten vorliegt. Weiterhin basiert die Typologie auf der Annahme, dass Hypertextknotentypen wie z. B. *persönliche Homepage*, *Handbuch* oder *Produktkatalog* existieren. Diese Typen können zwar als einzelne Knoten realisiert werden, doch manifestieren sie sich in der Regel in mehreren HTML-Dokumenten, weshalb sie vornehmlich der Ebene des Hypertexttyps bzw. der Hypertextsorte zuzurechnen sind. Andere Beispiele verdeutlichen die Notwendigkeit, die Ebene der Hypertextsortenmodule vorsehen zu müssen, z. B. kann der Hypertextknotentyp *Kontaktinformationen* (vgl. Abbildung 3.8) als einer von mehreren Bausteinen (d. h. als Hypertextsortenmodul) in die Einstiegsseite einer privaten Homepage integriert sein.

gerade hinsichtlich dieser Varianz an, denn neben der Ebene des Hypertexttyps existieren auch auf der Ebene des Knotens unterschiedliche Grade der Konventionalisierung. Hypertextknotentypen bzw. -sorten beschreiben, wie korrespondierende Knoten *in der Regel* von Hypertextproduzenten gestaltet und strukturiert werden, d. h. sie reflektieren empirische Erhebungen in Bezug auf bestimmte Untersuchungsdomänen. In Kapitel 4 werden verschiedene Studien vorgestellt, in denen die Einstiegsseiten der Webauftritte von Unternehmen analysiert werden (vgl. Abschnitt 4.6.2). Diese Klasse von Knoten unterliegt – ähnlich wie z. B. die Splash-Seite – bereits einem umfassenden Grad der Standardisierung, so dass diesbezüglich von einer Hypertextknotensorte gesprochen werden kann. Der generellere Terminus *Einstiegsseite* oder auch *Homepage* kann somit als Bezeichnung eines Hypertextknotentyps aufgefasst werden, dessen Existenz in diesem speziellen Fall für prinzipiell *alle* Hypertextexemplare *beliebiger* Hypertexttypen obligatorisch ist, sofern diese aus mehr als einem Knoten bestehen. Abbildung 5.2 macht deutlich, dass die angesprochene Varianz bezüglich der Instanzen von Hypertextknotentypen unmittelbar vom übergeordneten Hypertexttyp abhängig ist: Der Hypertexttyp *Webauftritt eines universitären Fachbereichs* besitzt eine *Einstiegsseite* als konventionalisierten Hypertextknotentyp. Da untergeordnete Hypertexttypen existieren (vgl. Abschnitt 5.4), bestehen die in der Abbildung dargestellten Einstiegsseiten vornehmlich aus Listen von Hyperlinks, die auf die Websites der jeweils untergeordneten Institute sowie der korrespondierenden Dekanate verweisen. In Bezug auf die übergeordnete Ganzheit kann dieser spezielle Hypertextknotentyp somit als Hypertextknotensorte *Einstiegsseite des Webauftritts eines universitären Fachbereichs* konzeptualisiert werden, die (unter anderem) durch ein Hypertextmodul vom Typ *Darstellung interner Hyperlinks* (umfasst sowohl die listenartige als auch die visuell aggregierte Präsentation) realisiert wird.

Neben Hypertextmodulen existiert die abstraktere Dimension der Hypertextsortenmodule, die in Abschnitt 5.6 eingeführt wird. In Bezug auf die Ebene des Knotens sind zwei wesentliche Eigenschaften von Hypertextsortenmodulen zu unterscheiden: Erstens kann ein einzelnes Hypertextsortenmodul als eigenständiger Knoten realisiert werden. Es fungiert in diesem Fall als Hypertextknotensorte. Zweitens kann ein Hypertextsortenmodul gemeinsam mit mindestens einem weiteren Hypertextsortenmodul in einen thematisch-funktional heterogenen Knoten eingebettet werden. Zur Verdeutlichung kann ein Beispiel aus der Ratgeberliteratur herangezogen werden: Reiss (2000, S. 130 ff.) unterscheidet verschiedene „secondary features“, die eine Website anbieten sollte, z. B. eine Sitemap, einen Index und eine Suchfunktion (vgl. Fußnote 8, S. 466). Insbesondere im Hinblick auf kommerziell ausgerichtete Websites führt Reiss ein „site tool“ auf, das er „About this site“ nennt:

> More and more, [sic] companies are combining their site map, webmaster e-mail address, disclaimer, and copyright information on a single page. Usually labeled “About this site”, this combination page seems well on its way to becoming an established web convention. (Reiss, 2000, S. 135)

Viele Websites von Unternehmen umfassen Reiss zufolge einen „About this site“ genannten Knoten, der von der Einstiegsseite aus mit einem in dieser Form benannten Hyperlinkanzeiger erreichbar ist. Bei den Bestandteilen dieser „combination page“ handelt es sich auf der Basis des Hypertextsortenmodells um Hypertextsortenmodule: Die „site map“, die „webmaster e-mail address“, der „disclaimer“ und „copyright information“ können von dem Pro-

duzenten eines Hypertextes potenziell in beliebiger Weise in einem oder mehreren Knoten realisiert werden (vgl. auch Abschnitt 4.5.2). Reiss beobachtet jedoch die Tendenz, dass die genannten Hypertextsortenmodule immer häufiger in einem Einzelknoten kombiniert werden und vermutet, dass es sich hierbei um eine in der Entwicklung befindliche Konvention handeln könnte. Bezüglich des Hypertextsortenmodells ist diese Konvention als eine Ausprägung auf der Ebene der Hypertextknotensorte aufzufassen. Auch für die Positionierung der einzelnen Konstituenten der Hypertextknotensorte „About this page" könnten sich demnach spezifische Konventionen entwickeln, z. B. wird zunächst die „site map" dargestellt, woraufhin der „disclaimer" präsentiert wird und im Schlussteil des Dokuments nennt der Produzent die „copyright information" sowie die „webmaster e-mail address" (Reiss, 2000, geht auf den Aspekt der Positionierung bzw. Binnensequenzierung der Bestandteile nicht ein).

Hypertextknotentypen bzw. -sorten besitzen konventionalisierte Vorbelegungen hinsichtlich der Hypertextsortenmodule, die in korrespondierenden Instanzen ausgeprägt werden. Weiterhin umfassen sie verschiedene grundlegende Merkmale, die sich auf die Positionierung, die kommunikative Funktion und die Dekoration beziehen. Das Merkmal der Positionierung betrifft die konventionalisierte Anordnung der Hypertextsortenmodule in einer Instanz, so dass z. B. das bei vielen Online-Zeitungen eingesetzte dreispaltige Layout erfasst werden kann (vgl. Indikator 38, S. 136).[33] Hypertextknotentypen können weiterhin eine spezifische kommunikative Funktion besitzen, die für eine entsprechende Instanz gilt. Diese bezieht sich in der Regel in unterstützender Weise auf die kommunikative Funktion der übergeordneten Hypertextsorte.[34] Das Merkmal Dekoration betrifft schließlich die Ebene des Webdesigns, das in vielen Fällen für alle Knoten eines Hypertextes einer konsistenten Gestaltung unterliegt. Da jedoch auch Abweichungen vorliegen können, ist die Einführung dieses Merkmals auf der Knotenebene notwendig. Weitere Merkmale werden durch ähnliche Vorbelegungen der enthaltenen Instanzen von Hypertextsortenmodulen in die übergeordneten Instanzen von Hypertextknotentypen bzw. -sorten propagiert.

Abschnitt 4.5.2 geht auf die Integration traditioneller Textsorten in Hypertexte ein. Bittner (2003, S. 127) führt in diesem Zusammenhang verschiedene Beispiele derartiger Einbettungen an, so werden etwa Textexemplare der Textsorten *Kochbuch*, *wissenschaftlicher Text* und *Fotogalerie* in 25 privaten Homepages ermittelt, weshalb Bittner dem Hypertexttyp *private Homepage* den Status eines „Containers" zuschreibt, der weitere Textsorten zu integrieren vermag. In der Regel werden die korrespondierenden Textexemplare derartiger Textsorten in einzelnen Knoten ausgeprägt, d. h. viele Hypertextknotentypen bzw. -sorten basieren unmittelbar auf traditionellen Textsorten, die sich in Bezug auf gedruckte Texte etabliert haben (vgl. Abschnitt 4.3.2). Häufig liegen die Texte, die als Grundlage der entsprechenden Knoten fungieren, bereits in digitaler Form vor, so dass der Knoten durch eine automatische

[33] Jakobs (2003, S. 237) führt die Beschreibungdimension der Strukturierungsmuster ein, die sich einerseits auf die Strukturierung (d. h. Sequenzierung) des gesamten Hypertextes und andererseits auf die „Anordnung und Abfolge von Inhalten eines Moduls" beziehen (vgl. Abschnitt 5.2.2). Die Ebene der globalen Strukturierung ist in dem hier vorgeschlagenen Modell ein zentrales Merkmal der Hypertextsorte, während die Strukturierung eines einzelnen Knotens durch die obligatorischen Hypertextsortenmodule einer Hypertextknotensorte und ihre jeweilige spezifische Positionierung ausgedrückt wird.

[34] Das textlinguistische Analysemodell für Hypertexte von Huber (2002) sieht ebenfalls mehrere Ebenen der Textfunktion vor, die sich auf den gesamten Hypertext, einen einzelnen Knoten, einen Absatz und einen Satz beziehen (vgl. Abschnitt 3.5.7 sowie Tabelle 3.1, S. 117).

Konvertierung des Dokuments in die *Hypertext Markup Language* erzeugt wird (vgl. Abschnitt 3.3.6). Dieser Aspekt bezieht sich in identischer Weise auf die Ebene der Hypertextsorte, falls das Exemplar einer traditionellen Textsorte als Hypertext realisiert wird, der mehrere Knoten umfasst. Wenn eine Person z. B. ein Kochbuch, das als LaTeX-Dokument vorliegt, mittels *LaTeX2HTML* in einen Hypertext konvertiert und diesen in ihrer *privaten Homepage* publiziert, dann ist die Hypertextsorte *Kochbuch*, die auf der gleichnamigen traditionellen Textsorte basiert, als untergeordnete oder auch eingebettete Hypertextsorte der *privaten Homepage* zu konzeptualisieren. Wenn davon ausgegangen wird, dass sich dabei jeweils ein *Rezept* in einem HTML-Dokument befindet, liegt in diesen Fällen die Hypertextknotensorte *Rezept* vor, da das Hypertextsortenmodul *Rezept* als Hypertextknotensorte fungiert.

5.6 Ebene 3: Hypertextsortenmodule

Nahezu alle Verfasser der in Abschnitt 4.4 dargestellten Analysen zur Ermittlung von Web-Genres auf der Grundlage zufällig zusammengestellter Stichproben geben an, dass die Zuweisung eines eindeutigen Hypertextsortenetiketts zu einem gegebenen HTML-Dokument aufgrund von Charakteristika wie z. B. der funktionalen oder strukturellen Heterogenität, nicht möglich war. Da Exemplare spezifischer Hypertextsorten weitere Exemplare anderer Hypertextsorten auf der *Knotenebene* einbetten können, kann oftmals keine eindeutige Bestimmung durchgeführt werden (vgl. ausführlich hierzu die Abschnitte 4.5.2 und 4.5.3). Somit muss die Annahme, dass es sich bei HTML-Dokumenten im Hinblick auf ihre Hypertextknotensorte um monolithische Entitäten handelt, zugunsten eines hochgradig modularen und sehr flexiblen Rahmenmodells zurückgewiesen werden.

Aus diesem Grund wird innerhalb des Hypertextsortenmodells angenommen, dass Hypertexttypen bzw. Hypertextsorten Hypertextsortenmodule umfassen, die insbesondere die thematischen makrostrukturellen Konstituenten eines Hypertexttyps spezifizieren.[35] In der in Abbildung 5.2 dargestellten Instanz der Hypertextsorte *persönliche Homepage eines Wissenschaftlers* sind verschiedene Beispiele hervorgehoben: Das Hypertextsortenmodul *Kontaktinformationen* ist unmittelbar in die Einstiegsseite dieser Instanz eingebettet, während ihr Produzent die Hypertextsortenmodule *Lebenslauf* (bzw. *biografische Informationen*) und *Forschungsschwerpunkte* in separate Knoten integriert hat. Dieses Beispiel verdeutlicht, dass sich Hypertextsortenmodule nicht nur auf die Einbettung innerhalb der Einstiegsseite beziehen, sondern alternativ auch als Hypertextknotensorten fungieren können, deren Instanzen über Hyperlinks in die Hypertextbasis eingebunden werden. Der generischere Begriff des Hyper-

[35] Es ist darauf hinzuweisen, dass sich der Begriff „Modul“, der in dem Hypertextsortenmodell in den Termini „Hypertextmodul“ und „Hypertextsortenmodul“ benutzt wird, nicht auf die Verwendungsweise von Storrer (1999a) und Jakobs (2003) als Synonym von „Knoten“, „Webseite“ oder „HTML-Dokument“ bezieht (vgl. Fußnote 52, S. 84). Hypertextsortenmodule entsprechen einer spezifischeren konzeptuellen Ebene als die von Kneece (1996, S. 195) angenommenen „digital modules“, die zur Bezeichnung der Konstituenten elektronischer Publikationen eingesetzt werden. Dabei beschränkt sich Kneece nicht auf Texte oder Hypertexte, sondern bezieht zusätzlich Datenbanken, Informationssysteme und auch das WWW ein. Ein Modul wird von Kneece (1996, S. 196) als „self-contained unit capable of being understood on its own“ definiert, das aus Text-, Grafik-, Video- oder Audioinformationen besteht. Insbesondere geht Kneece auf das „text module“ ein, das einem Textabschnitt oder einer Gruppe zusammengehöriger Abschnitte gleichbedeutend ist und zwangsläufig „unified, coherent, and complete“ ist (ebd.).

textmoduls bezieht sich auf einer strukturellen Ebene auf *alle* Konstituenten der Textoberfläche eines HTML-Dokuments (z. B. Überschriften, Listen, Absätze und Grafiken). Innerhalb einer maschinellen Verarbeitung von HTML-Dokumenten gestattet es diese Differenzierung, zunächst sämtliche Hypertextmodule zu ermitteln und diese anschließend auf ein Inventar von Hypertextsortenmodulen abzubilden. Dieser Prozess kann somit als eine Art Dekomponierung von Webseiten aufgefasst werden.

5.6.1 Verwandte Arbeiten

Horn (1989) richtet sich mit seiner „Information Mapping"-Methodologie insbesondere an die Verfasser von Gebrauchstexten und technischen Dokumentationen im Bereich der Wirtschaftskommunikation. Der Ansatz wird als eine modulare und strukturierte Technik zur Erstellung und Formatierung von Texten verstanden, um den Rezipienten einen schnellen Zugang zu den Kerninhalten zu gewährleisten.[36] Information Mapping ist primär auf den Arbeitsalltag in Unternehmen und Behörden ausgerichtet, weil, so Horn (1989, S. 99), nicht *jede* Person *jedes* einzelne Wort eines technischen oder administrativen Berichts lesen muss, weshalb keine „journalistic gimmicks" eingesetzt werden müssen, um den Rezipienten zur weiteren oder ausführlichen Lektüre zu motivieren. Stattdessen – und diesbezüglich existiert eine deutliche Parallele zu den Richtlinien der Webdesign- und Usability-Literatur – spricht sich Horn dafür aus, dem Leser einen schnellstmöglichen Zugriff auf die für ihn relevanten Textteile zu ermöglichen, indem „very clear chunks clearly designated by subheads" zur Verfügung gestellt werden (ebd.; vgl. Fußnote 28, S. 76). Eines der Kernprinzipien des Verfahrens betrifft die Ersetzung des Textabsatzes als „basic unit of meaning in functional written communication" durch den „information block", der in Kombination mit mindestens einem weiteren Block eine „information map" bildet (ebd., S. 80).[37] Der „information block", von dem Horn zufolge etwa 200 textsortenabhängige Typen existieren, ist also eine atomare Einheit, die z. B. aus mehreren Sätzen, einer Liste, einer Tabelle oder einer Grafik bestehen kann und zusätzlich über ein Etikett identifiziert werden muss.[38] Horn (1989, S. 92) zufolge existiert eine enge Verbindung zwischen spezifischen „document types" und den in ihnen enthaltenen „most frequent types of blocks". Das Verfahren zur Ermittlung dieser häufigsten Blocktypen wird nicht konkretisiert, Horn (1989, S. 109) gibt jedoch etwa 40 dieser Typen an, die sich auf die „discourse domain" der relativ stabilen Textinhalte beziehen („relatively stable subject matter", andere Diskursdomänen sind z. B. „experimental knowledge" und „part of subject matter under debate or consideration", ebd., S. 104).[39] Über eine Tabelle

[36] Der gesamte Band *Mapping Hypertext* (Horn, 1989) wurde vom Autor nach den Prinzipien des Information Mapping-Ansatzes verfasst und gestaltet. Folglich unterscheidet sich dieses Buch hinsichtlich seines Textsatzes und seiner Strukturierung deutlich von den traditionellen und etablierten typografischen Konventionen.

[37] Nach Meinung von Horn (1989, S. 90) ist das Konzept des Textabsatzes zu vage definiert und beinhaltet verschiedene Unzulänglichkeiten: Es sei z. B. unklar, wie ein Absatz eingeleitet und beendet werden soll, was wiederum zur Konsequenz habe, dass sowohl in Schule und Hochschule als auch in Weiterbildungskursen keine eindeutig definierten Regeln zur Verwendung von Absätzen vermittelt werden könnten.

[38] Horn (1989, S. 84) bezieht sich auf die kognitionspsychologische Arbeit von Miller (1956) und geht davon aus, dass sich maximal 7±2 Sätze in einem „information block" (nach Horn, 1989, S. 88 f., „the smallest meaningful chunk for most readers") und maximal 7±2 „information blocks" innerhalb einer „information map" befinden sollten, die wiederum einen Umfang von zwei Seiten Text nicht überschreiten sollte.

[39] Bei diesen Typen handelt es sich z. B. um „block diagram", „checklist", „definition" und „example".

kann daraufhin im Rahmen der Analyse eines bestimmten Dokumenttyps wie z. B. „user guides“, „equipment manuals“ oder „operations manuals“ ermittelt werden, mit welchen „block types“ spezifische „topic types“ realisiert werden. Horn (1989, S. 110) geht davon aus, dass Dokumente zwei Arten von Informationen enthalten: „basic information“ und „supplementary information“. Die erste Gruppe umfasst die sieben Informationstypen „structure“, „concept“, „procedure“, „process“, „classification“, „principle“ und „fact“. Diesen Typen können „key blocks“ zugeordnet werden, so dass z. B. spezifiziert werden kann, dass ein „concept“ als „definition“, „example“ oder „non-example“ realisiert werden sollte (ebd., S. 112). Neben diesen Grundlagen widmet sich Horn (1989) der Anwendung von Information Mapping auf Hypertexte und führt z. B. „Hypertrails“ als Mengen von Hyperlinks ein, die beliebige Informationsbestandteile verknüpfen, um einen Hypertext für spezifische Rezipientengruppen bestmöglich nutzbar zu machen (etwa mittels „classification hypertrails“, „chronological hypertrails“ oder „prerequisite hypertrails“). Ein „information block“ wird dabei auf exakt einen Knoten abgebildet. Eine „information map“ umfasst einen übergeordneten Knoten, der wiederum auf die enthaltenen „information blocks“ verweist.

Jakobs (2003, S. 236) geht auf den Umstand ein, dass sich viele Hypertexte aus „Bausteinen“ zusammensetzen, „die ihrerseits funktional-thematisch bestimmte kommunikative Ganzheiten repräsentieren.“ (vgl. Abschnitt 5.2.2). Die Website eines Unternehmens enthält Jakobs (2003, S. 237) zufolge (i) funktional systembezogene, (ii) funktional aufgabenbezogene, (iii) funktional interaktionsbezogene und (iv) thematisch bestimmte Bausteine. Die Bausteine der ersten Gruppe dienen der Navigation. Als Beispiele werden die „Homepage“, „Überblicksseiten“ und „Suchmasken“ genannt (ebd.). Die zweite Gruppe umfasst Bausteine, die „den Vollzug einer nicht-sprachlichen Handlung [erlauben]“ (ebd.), z. B. „Bestellformulare“ und „Download-Seiten“. Die „Interaktion mit anderen Personen“ wird durch die Bausteine der dritten Gruppe ermöglicht (etwa „E-Mail“, „Foren“ und „Gästebuch“, ebd.). Die Bausteine der vierten Gruppe „bringen spezifische Inhalte ein“ (z. B. „Geschäftsbericht“, „Stellenbörse“, „FAQ“, „Produktpräsentation“, „Mission Statement“, „Archiv“, „Pressespiegel“ oder „Spiele“).[40] Jakobs weist darauf hin, dass Bausteine wie z. B. der Geschäftsbericht „integrierten E-Print-Textsorten-Varianten“ entsprechen, während es sich bei anderen „um in Hypertext integrierte Hypertextsorten“ handelt (ebd.). In Bezug auf alle enthaltenen Bausteine fungiert die Website als „übergeordnete funktional-thematische Ganzheit“, „als einbettender, organisierender, kontextualisierender Rahmen“ (ebd.). Nach Jakobs (2003, S. 237) „bliebe zu diskutieren“, ob Bausteinen wie der Übersichtsseite, dem FAQ oder der Einstiegsseite „Textsortencharakter zuzuschreiben ist“. Zumindest bezüglich des FAQ tendiert die Verfasserin dazu, eher eine funktional-thematische Bestimmung wie z. B. bei einem Index oder Glossar vorzunehmen, das FAQ also als einen „Textbaustein“ aufzufassen, der von einem „übergeordneten Ganzen“ abhängig ist (vgl. Indikator 26, S. 132, sowie Abschnitt 4.5.3).

In der Studie von Schütte (2004a, vgl. Abschnitt 4.6.2) besitzt die Analyse der Makrostruktur einen zentralen Stellenwert. Der Begriff bezeichnet „die globale, sowohl inhaltlich-formal als auch funktional bedingte Textstruktur [...], die textsortenabhängig durch eine spezifische Abfolge und wechselseitige Determiniertheit der sie konstituierenden Teiltexte

[40] Die genannten Beispiele (z. B. „Homepage“, „Überblicksseite“, „Gästebuch“) verdeutlichen, dass sich Jakobs (2003, S. 236) mit dem Terminus „Baustein“ auf eine Ebene bezieht, die eher der von Storrer (1999a) vertretenen Auffassung von „Modul“ als der Ebene des Hypertextsortenmoduls entspricht.

(bzw. makrostrukturellen Bestandteile) markiert ist.“ (ebd., S. 213). Schütte geht davon aus, dass Makrostrukturen rekurrente Textbaumuster sind und dass die Exemplare einer spezifischen Textsorte einander ähnliche Makrostrukturen besitzen. Die Rezipienten sind in der Lage, die Makrostrukturen anhand von Gliederungssignalen, die sich an der Textoberfläche manifestieren, zumindest teilweise identifizieren zu können. Unter Gliederungssignalen versteht Schütte typografische Mittel, Kapitelüberschriften und metakommunikative Äußerungen des Produzenten, die zur Steuerung der Rezeption und zur Verdeutlichung der Textstruktur eingesetzt werden (ebd.). Gerade die Einstiegsseite einer Website enthält Schütte zufolge vielfältige Gliederungssignale, die nicht durch verbale, sondern durch „layoutbedingte, topographische“ Mittel hergestellt werden (ebd.). Makrostrukturelle Komponenten – Schütte spricht auch von „informationellen Modulen“ – sind durch eine „optische Isolierung und flächige Verteilung über den Bildschirm“ erkennbar (ebd.).

Die dem Hypertextsortenmodell zugrunde liegende Auffassung von Hypertextsortenmodulen als atomare Konstituenten von Hypertexttypen und Hypertextsorten, die in korrespondierenden Hypertextexemplaren unterschiedliche Ausprägungen erfahren können, folgt in vielerlei Hinsicht den Ansätzen von Horn (1989), Haas und Grams (2000), Jakobs (2003) und Schütte (2004a), den in Abschnitt 3.5.6 dargestellten Arbeiten zum Textdesign und zur Modularisierung von Text- und Informationseinheiten sowie den Ausführungen zur Einbettung von Hypertextsorten (Abschnitte 4.5.2 und 4.5.3). Dieser zentrale Aspekt des Modells trägt insbesondere dem Umstand Rechnung, dass HTML-Dokumente aus zahlreichen unterschiedlichen Bestandteilen bestehen, „deren Strukturen durch grafische Mittel visualisiert werden“ (Bucher, 1999, S. 12), wodurch nicht genuine Langtexte, sondern vielmehr „visuelle Texte“ (ebd.) entstehen, d. h. die ausschließliche Präsentation linear organisierter Texte wird „durch eine modulare Informationsaufbereitung abgelöst“ (Bucher, 1996, S. 35).

5.6.2 Die Ebene der Textoberfläche – Hypertextmodule

Von wesentlicher Bedeutung ist zunächst die Annahme, dass *sämtliche* Bestandteile der Textoberfläche eines HTML-Dokuments eindeutig identifizierbar sind und im Hinblick auf ihre typografisch-gestalterische Realisierung durch HTML-Elemente auf der Textoberfläche beschrieben werden können. Auf diese Weise kann eine Webseite z. B. in Listen (beispielsweise von Text, Textfragmenten oder Hyperlinks), Abschnitte von Fließtext, kleinformatige und großformatige Grafiken und isoliert positionierte Textfragmente dekomponiert werden. Diese Elemente können funktional charakterisiert und teilweise hierarchisiert werden, so dass die unterschiedlichen Hypertextmodule eines Webdokuments sichtbar werden. Diesbezüglich können etwa Überschriften verschiedener Stufen, ihnen untergeordnete Textabschnitte und mit typografischen Mitteln erzeugte Listen sowie unabhängige Navigationshilfen erkannt werden. Auf dieser Ebene existieren nun vollkommen generische Hypertextmodule, die in einem zweiten Schritt auf Hypertextsortenmodule abgebildet werden können.[41] Die Identifizierung von Hypertextmodulen ist, wie Kapitel 14 zeigen wird, maschinell realisierbar. Da in HTML unterschiedliche Dimensionen der Textauszeichnung möglich sind (vgl. Abschnitt 5.3.1), ist es notwendig, die sehr umfangreichen und für linguistische Beschreibun-

[41] Das Verhältnis zwischen Hypertextmodulen und Hypertextsortenmodulen entspricht konzeptionell in etwa den Beziehungen zwischen „information block types“ und „topic types“ bei Horn (1989).

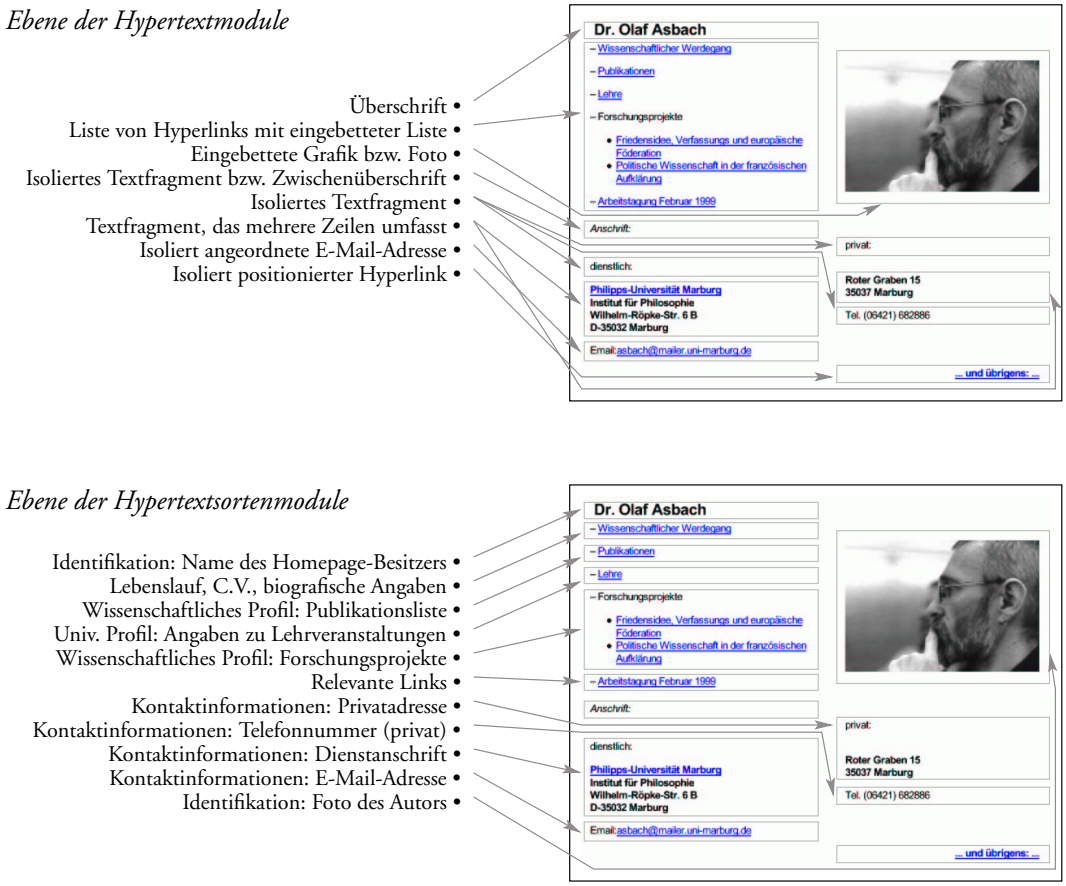


Abbildung 5.3: Die Ebenen der Hypertextmodule und Hypertextsortenmodule

gen oder maschinelle Verarbeitungsprozesse ungeeigneten Kombinationsmöglichkeiten von HTML zur (impliziten) Realisierung makrostruktureller Bestandteile durch ein übergreifendes und im Umfang begrenztes Inventar von Hypertextmodulen als generische Komponenten der Textoberfläche zu explizieren. In der oberen Hälfte von Abbildung 5.3 ist die Einstiegsseite einer Instanz der Hypertextsorte *persönliche Homepage eines Wissenschaftlers* dargestellt, in der die Hypertextmodule hervorgehoben wurden. Die Liste von Hyperlinks wurde z. B. nicht durch das HTML-Element `ul` realisiert, sondern als eine Tabelle (`table`), deren einzelne Zeilen mit einem Gedankenstrich beginnen. Bei diesem Konstrukt handelt es sich also nicht um isoliert positionierte Textfragmente, sondern – im übergreifenden Kontext betrachtet – um eine Navigationshilfe, über die der Rezipient auf die weiteren Bestandteile dieses Hypertextes zugreifen kann. Nach Fleming (1998, S. 64) sollten im Webdesign visuelle Hierarchien eingesetzt werden, um Beziehungen zwischen den Elementen eines HTML-Dokuments zu verdeutlichen, was unter anderem durch die relative Größe der Bestandteile und ihre Position realisiert werden kann. Die Überschrift in Abbildung 5.3 ist derjenige Bestandteil dieser Webseite, der in der größten Schrift dargestellt ist, d. h. alle weiteren Elemente dieses Dokuments beziehen sich in gewisser Hinsicht auf diese Überschrift.

5.6.3 Die Ebene der Makrostruktur – Hypertextsortenmodule

Hypertextsortenmodule sind, wie eingangs angesprochen, die grundlegenden makrostrukturellen und vornehmlich thematisch-funktional markierten Konstituenten der Textstrukturmuster von Hypertexttypen bzw. Hypertextsorten. Sie können in der Instanz eines Hypertexttyps entweder als separate physikalische HTML-Dokumente oder als makrostrukturelle Bestandteile funktional heterogener Knoten ausgeprägt sein. Hypertextsortenmodule können unterschiedliche multimediale Inhalte besitzen und z. B. aus Text[42] (einzelnen Wörtern, Phrasen oder Sätzen, Fließtext oder Textfragmenten), Grafiken, Video- und Audio-Dateien bestehen. Zusätzlich existieren Hypertextsortenmodule, die über charakteristische und eigenständige Text- bzw. Hypertextstrukturmuster verfügen (z. B. die *Publikationsliste*, die Angabe einer *Postadresse* oder der *wissenschaftliche Artikel*).[43] Insbesondere Einstiegsseiten sind nach Haas und Grams (1998b, S. 101) „good examples of the many different ways in which authors may divide essentially the same information into different configurations of actual pages." Dabei handelt es sich bezüglich der von der übergeordneten Hypertextsorte vorgesehenen Hypertextsortenmodule tatsächlich nur um unterschiedliche Konfigurationen, d. h. individuelle Ausprägungen, denen Crowston und Williams (1999, S. 2) in etwa den Status der Paginierung eines gedruckten Textes zuweisen. Oftmals besitzen diese Ausprägungen jedoch einen konventionalisierten, d. h. hochfrequenten Status – in diesem Fall liegen Hypertextknotensorten vor (vgl. das von Reiss, 2000, S. 135, genannte Beispiel; siehe S. 280).

Hypertextsortenmodule weisen unterschiedliche Eigenschaften auf. Es existieren z. B. Hypertextsortenmodule, die einen universalen Status besitzen und in den Instanzen *beliebiger* Hypertextsorten bzw. Hypertextknotensorten verwendet werden (z. B. die *Navigationshilfe* und der *Zugriffszähler*). Andere Hypertextsortenmodule sind restriktiver und werden typischerweise in einer begrenzten Menge von Hypertextsorten instanziiert (z. B. die *Publikationsliste* oder *Kontaktinformationen*). Darüber hinaus kann ein spezifisches Hypertextexemplar mit einem Hypertextsortenprofil verglichen werden, das seinerseits durch die empirische Analyse einer Stichprobe konstruiert wird. Ein derartiges Profil umfasst Angaben zu obligatorischen und optionalen Hypertextsortenmodulen. Erstere sind als hochfrequente, d. h. konventionalisierte Komponenten einer Hypertextsorte aufzufassen, während letztere ihre Peripherie und somit Varietäten der Hypertextsorte darstellen. Diesbezüglich kann durchaus der Fall eintreten, dass ein spezifisches Hypertextexemplar einer Hypertextsorte nicht zugehörig ist, wenn lediglich ein einzelnes obligatorisches Hypertextsortenmodul *nicht* ausgeprägt ist (z. B. *Identifikation* in Bezug auf die *persönliche Homepage eines Wissenschaftlers*).

[42] Dabei wird in der Regel die in Abschnitt 5.3.4 dargestellte Textdefinition von Brinker (2001, S. 17) erfüllt, so dass die meisten Hypertextsortenmodule im Kontext des umgebendes Hypertextes als Teiltexte und in isolierter Form – zumindest in vielen Fällen – als Texte aufgefasst werden können.

[43] Der Begriff des Hypertextstrukturmusters bezieht sich nicht auf eine Gruppe von zwei Knoten, sondern lediglich auf Ausgangsanker von Hyperlinkverweisen, die nach einer spezifischen Konvention benannt sind. Der wissenschaftliche Artikel ist ein Beispiel für ein Hypertextsortenmodul, das in der Regel als eigenständige Hypertextknotensorte fungiert, d. h. ein einzelner Knoten umfasst einen wissenschaftlichen Artikel als Hypertextsortenmodul und z. B. Navigationshilfen als zusätzliche Hypertextsortenmodule. Zugleich kann er jedoch auch als monosequenzierter Hypertext in mehreren Knoten umgesetzt werden, d. h. gerade diejenigen traditionellen Textsorten, die äußerst ausgeprägte Textstrukturmuster besitzen und über einen hohen Durchschnittsumfang verfügen (und somit durch die Aufteilung eines Textes in unterschiedliche Knoten einen potenziellen Mehrwert erzeugen), können zusätzlich als Hypertextsorten fungieren.

Abbildung 5.3 zeigt in der unteren Hälfte die Hypertextsortenmodule der bereits angesprochenen Instanz der Hypertextsorte *persönliche Homepage eines Wissenschaftlers*. Diese Hypertextsorte umfasst zunächst verschiedene Hypertextsortenmodule, die der *Identifikation* des Autors dienen (z. B. *Name* und *Foto*). Das *wissenschaftliche Profil* des Autors konstituiert sich durch die Angabe einer *Publikationsliste* und aktueller *Forschungsprojekte*. Der Vergleich der beiden Ebenen zeigt, dass nur eine partielle Korrespondenz zwischen Hypertextmodulen und Hypertextsortenmodulen existiert. Die Liste von Hyperlinks umfasst als einzelnes Hypertextmodul Verweise zu mehreren Knoten, die als Instanzen der korrespondierenden Hypertextknotentypen (z. B. *Lebenslauf* und *Publikationsliste*) fungieren, die ihrerseits wiederum auf Hypertextsortenmodulen basieren, da sie in vielen Fällen nur als einer von mehreren Bestandteilen eines Knotens ausgeprägt sind.[44] Andere Hypertextsortenmodule wie z. B. die (obligatorische) *Dienstanschrift* und die (optionale) *Privatadresse* sind unmittelbar in der Einstiegsseite realisiert, wodurch die eingangs erwähnte funktionale Heterogenität hervorgerufen wird. Kapitel 10 geht in detaillierter Form auf die Hypertextsorte *persönliche Homepage eines Wissenschaftlers* ein und stellt ihr Hypertextsortenprofil vor.

5.6.4 Die Charakteristika von Hypertextsortenmodulen

Hypertextsortenmodule stellen die grundlegenden Bausteine von Hypertextsorten dar. Im Folgenden werden ihre zentralen Charakteristika vorgestellt.

Status von Hypertextsortenmodulen Hypertextsortenmodule besitzen hinsichtlich eines Hypertexttyps oder einer Hypertextsorte einen individuellen Status: Spezifische Hypertextsortenmodule beziehen sich ausschließlich auf eine Hypertextsorte eines übergeordneten Hypertexttyps. Ein Beispiel hierfür ist die *Sprechstunde* (Teil des komplexen Hypertextsortenmoduls *Kontaktinformationen*) in Bezug auf die Hypertextsorte *persönliche Homepage eines Wissenschaftlers*. Generelle Hypertextsortenmodule können in den unterschiedlichen Hypertextsorten eines Hypertexttyps verwendet werden, wozu z. B. die Hypertextsortenmodule *Identifikation* und *Kontaktinformationen* hinsichtlich des abstrakten Hypertexttyps *Homepage einer Person* gehören. In gewisser Hinsicht stellen spezifische Hypertextsortenmodule somit die distinktiven Merkmale einer Hypertextsorte in Bezug auf den übergeordneten Hypertexttyp dar. Universale Hypertextsortenmodule können in beliebigen Hypertextsorten verwendet werden, z. B. der *Zugriffszähler* und der *Rückverweis zur Einstiegsseite*.

Verwendung von Hypertextsortenmodulen Die Hypertextsortenmodule einer Hypertextsorte können nur empirisch bestimmt werden. Für diesen Zweck sind Stichproben korrespondierender Exemplare zu analysieren. Die individuellen Frequenzen von Hypertextsortenmodulen innerhalb einer derartigen Analyse bestimmen ihre Verwendung. Falls es in mindestens der Hälfte der untersuchten Hypertexte instanziiert wird, ist von einem obligatorischen Hypertextsortenmodul die Rede, andernfalls handelt es sich um ein optionales

[44] Diese vereinfachende Darstellung dient lediglich der Veranschaulichung, denn das Hypertextmodul *Liste von Hyperlinks* entspricht dem Hypertextsortenmodul *primäre Navigationshilfe*. Die Problematik wird in Kapitel 11 genauer diskutiert, wobei auch gezeigt wird, dass unterschiedliche Typen der *Navigationshilfe* existieren.

Hypertextsortenmodul. Erstere bilden den Kern einer Hypertextsorte, letztere repräsentieren ein vielfältiges Spektrum von Varietäten. Zusätzlich existieren Hypertextsortenmodule, die in Instanzen von Hypertextsorten oder Hypertextknotentypen bzw. -sorten in der Regel *nicht* ausgeprägt werden. Als Beispiel kann das Hypertextsortenmodul *Sitemap* gelten, das in der Hypertextknotensorte *Einstiegsseite eines universitären Webauftritts* typischerweise nicht enthalten ist, sondern in einem separaten Knoten realisiert wird.

Separierung und Integrierung von Hypertextsortenmodulen Ein HTML-Dokument kann beliebige Hypertextsortenmodule der zugehörigen Hypertextsorte umfassen. Im Extremfall wird die Instanz einer Hypertextsorte in einem einzelnen Knoten ausgeprägt, was z. B. gelegentlich bei Exemplaren der *persönlichen Homepage eines Wissenschaftlers* zu beobachten ist. Typischerweise werden diejenigen Hypertextsortenmodule, die ein etabliertes Textstrukturmuster besitzen, in eigenständige Knoten separiert. In diesem Fall fungiert das (einzige in einem Knoten enthaltene) inhaltlich-thematische Hypertextsortenmodul als Hypertextknotensorte, die jedoch zusätzliche Komponenten, die sich auf den Peritext beziehen, beinhalten kann (z. B. eine *Navigationshilfe* oder einen *Zugriffszähler*). Einer Einstiegsseite untergeordnete Knoten umfassen im Regelfall ein oder zwei inhaltlich-thematisch markierte Hypertextsortenmodule, in Ausnahmefällen können auf dieser Ebene auch mehr Instanzen beobachtet werden. Hypertextsortenmodule können nicht nur als Hypertextknotensorten, sondern auch als Hypertextsorten fungieren: Die *Publikationsliste* kann z. B. in einen separaten Knoten ausgelagert werden, sie kann aber auch in Form eines Hypertextes aufbereitet werden, wobei die konventionalisierten Bestandteile dieses Hypertextsortenmoduls als Hypertextknotensorten fungieren – in diesem Beispiel handelte es sich etwa um mehrere thematisch oder nach dem Veröffentlichungstyp sortierte Publikationslisten, die sich jeweils in eigenständigen Knoten befinden. Es können allerdings nicht beliebige Hypertextsortenmodule als Hypertextknotensorten oder Hypertextsorten fungieren. Typischerweise gilt diese Eigenschaft für inhaltlich-thematisch markierte Hypertextsortenmodule, die ein etabliertes Textstrukturmuster besitzen (wie z. B. die angesprochene *Publikationsliste*, der *wissenschaftliche Artikel* oder das *Vorlesungsverzeichnis*). Andere Hypertextsortenmodule werden in der Regel in Kombination mit Hypertextsortenmodulen der primären Typen Inhalt/Thema, Navigation und Kommunikation instanziiert, z. B. der *Zugriffszähler* oder die *Navigationshilfe*.

Typen von Hypertextsortenmodulen Bislang wurden vornehmlich (i) inhaltlich-thematisch bestimmte Hypertextsortenmodule als generische Makrostrukturkomponenten spezifischer Hypertextsorten diskutiert, die einen der Schwerpunkte der in Teil III präsentierten Studien bilden. Es existieren weitere Kategorien von Hypertextsortenmodulen, die an die Überlegungen von Jakobs (2003, S. 237) anknüpfen. (ii) Interaktionale Hypertextsortenmodule umfassen insbesondere HTML-Formulare und eingebettete Programme, welche dem Rezipienten Interaktionen mit dem Hypertext ermöglichen, die über das sukzessive Aktivieren von Hyperlinks zur Navigation hinausgehen. (iii) Hypertextsortenmodule vom Typ Kommunikation gestatten es dem Rezipienten, mit Personen über das WWW in Kontakt zu treten, was z. B. über E-Mail-Hyperlinks oder Diskussionsforen geschehen kann. (iv) Hypertextsortenmodule der Kategorie Navigation erlauben es dem Rezipienten, sich in dem Hypertext zu bewegen. Diese Hypertextsortenmodule werden z. B. in Einstiegsseiten eingesetzt, so dass die

Navigation in die untergeordneten Knoten ermöglicht wird. (v) Der Typ Metainformation umfasst Hypertextsortenmodule, die Informationen über einen Knoten oder die gesamte Hypertextsorteninstanz enthalten (z. B. der *Zugriffszähler* oder das *Datum der letzten Änderung*). (vi) Hypertextsortenmodule des Typs Dekoration umfassen unter anderem multimediale Elemente, die ausschließlich dekorativen Charakter besitzen. (vii) Der Typ Textstrukturmuster schließlich stellt einen Sonderfall dar, der sich konzeptionell zwischen Hypertextsortenmodulen und Hypertextknotentypen bewegt. Hypertextsortenmodule dieses Typs spezifizieren generische Textstrukturmuster, die von den Produzenten einer Hypertextsorte üblicherweise zur Realisierung bestimmter Bestandteile von Knoten eingesetzt werden. Da sich dieser Typ ausschließlich auf die Ebene der Textstruktur bezieht, können sie in korrespondierenden Instanzen beliebige Inhalte umfassen, sie können jedoch ebenfalls Instanzen anderer, vornehmlich inhaltlich-thematisch markierter Hypertextsortenmodule beinhalten.[45]

Hypertextsortenmodule werden nicht genau einem dieser Typen zugeordnet. Die sieben Kategorien sind – in Anlehnung an Crowston und Kwasnik (2004, vgl. Abschnitt 5.2.2) – in sämtlichen Repräsentationen von Hypertextsortenmodulen ausgeprägt, so dass primäre und sekundäre Typen vorliegen, die die Kategorienzugehörigkeit markieren. Jeder Typ ist – sehr abstrakt formuliert – als eine Art Kontinuum aufzufassen und jedes Hypertextsortenmodul besitzt für jeden Typ eine Ausprägung, so dass in konzeptuell und funktional identischen Hypertextsortenmodulen unterschiedlicher Hypertextsorten sortenspezifische Varietäten ermöglicht werden. Die primär ausgeprägte Kategorie determiniert den primären Typ eines Hypertextsortenmoduls; die Typen können somit auch als grundlegende „facets“ im Sinne von Crowston und Kwasnik verstanden werden. Auf diese Weise bilden alle Hypertextsortenmodule, die in der Instanz einer Hypertextknotensorte vorliegen, kompositionell die korrespondierenden Ausprägungen des Knotens bezüglich der sieben Kategorien.

Ein Hypertextknotentyp umfasst in der Regel die Merkmale der Positionierung obligatorischer Hypertextsortenmodule, der kommunikativen Funktion und der Dekoration. Die spezifischen Ausprägungen der in einer Instanz existierenden Hypertextsortenmodule wirken sich kaskadierend auf ihren Knoten aus; sie bilden in ihrer Gesamtheit die Merkmalsausprägungen eines Knotens. Durch die zusätzliche Einbettung optionaler Hypertextsortenmodule werden vielfältige Modifikationen ermöglicht, die als Abweichungen von einem prototypischen Kern aufgefasst werden können (vgl. Abschnitt 5.8.1). Ein weiterer Aspekt betrifft die generelle Problematik der Typologisierung: Auf der Basis des Typeninventars können Typologien von Hypertextsortenmodulen konstruiert werden. Darüber hinaus können – dies wird bereits durch die Bezeichnungen der Ebenen angedeutet – Typologien von Hypertexttypen bzw. -sorten und Hypertextknotentypen bzw. -sorten erstellt werden.

Atomare und komplexe Hypertextsortenmodule Inhaltlich oder strukturell zusammengehörige Konstituenten einer Hypertextsorte können als komplexe Hypertextsortenmodule aufgefasst werden und somit Hierarchien auf der Ebene der Dokumentstruktur erlau-

[45] Jakobs (2003, S. 237) bezeichnet Typ (ii), der „den Vollzug einer nicht-sprachlichen Handlung“ erlaubt, als „funktional aufgabenbezogen“. Typ (iii) fokussiert Jakobs zufolge die „Interaktion mit anderen Personen“ und wird „funktional interaktionsbezogen“ genannt. Typ (iv) wird als „funktional systembezogen“ bezeichnet. Die Kategorie Textstrukturmuster bezieht sich ausschließlich auf die lokale Ausprägung eines Hypertextsortenmoduls, weshalb sie nicht auf die Ebene der Hypertextsorte propagiert wird.

ben. Diese Vorgehensweise der Einführung abstrakter Kategorien zur konzeptuellen Bündelung zweier oder mehrerer verwandter Konstituenten ist ein Grundprinzip der texttechnologischen Informationsmodellierung, das insbesondere in Dokumentgrammatiken angewendet wird (vgl. Maler und Andaloussi, 1996, und Lobin, 2001a). Beispielsweise kann das Hypertextsortenmodul *Zugriffszähler* als atomar angesehen werden, wohingegen *Angaben zu Lehrveranstaltungen* und *Funktionen* eines Wissenschaftlers innerhalb einer Universität als atomare Hypertextsortenmodule des komplexen Hypertextsortenmoduls *universitäres Profil* verstanden werden. Es besteht kein unmittelbarer Zusammenhang zwischen atomaren und komplexen Hypertextsortenmodulen und der Frage, ob ein spezifisches Hypertextsortenmodul in einen übergreifenden Knoten eingebettet oder in einen separaten Knoten integriert werden kann, oder ob es eingebettet werden *muss* (wie z. B. der *Zugriffszähler*). In der Regel lagern Produzenten diejenigen Hypertextsortenmodule in separate Knoten aus, die über ein etabliertes Textstrukturmuster verfügen (z. B. die *Publikationsliste*). Bei komplexen Hypertextsortenmodulen können oftmals alle untergeordneten Hypertextsortenmodule in eigenständige Knoten separiert werden, wobei die korrespondierenden Hypertextknotentypen auf den Strukturmustern der Hypertextsortenmodule basieren, aber zusätzlich peritextuelle Elemente (z. B. Navigationshilfen) enthalten können. Besitzt ein Hypertextsortenmodul hingegen kein eigenständiges Textstrukturmuster, so erfolgt in der Regel keine Auslagerung.

Komplexe Hypertextsortenmodule umfassen zwei oder mehr atomare Hypertextsortenmodule. Diese Beschreibungsebene wird aus zwei Gründen angenommen: Erstens dient sie zur Binnenstrukturierung von Hypertextsortenmodulen, die eine differenzierte und charakteristische Struktur besitzen (z. B. die einzelnen Bestandteile der *Kontaktinformationen*). Zweitens werden atomare Hypertextsortenmodule auf einer abstrakteren Ebene als komplexes Hypertextsortenmodul zusammengefasst, wenn die atomaren Komponenten verwandte Themen, Funktionen oder Strukturen besitzen (das komplexe Hypertextsortenmodul *wissenschaftliches Profil* in der *persönlichen Homepage eines Wissenschaftlers* besteht unter anderem aus den atomaren Bausteinen *Publikationsliste*, *Forschungsprojekte* und *Mitgliedschaft in Fachverbänden*). Komplexe Hypertextsortenmodule besitzen ebenfalls diese Charakteristika, die sich aus den Ausprägungen ihrer atomaren Komponenten ergeben. Es ist zu beachten, dass weitere Hierarchie- bzw. Strukturierungsebenen existieren, z. B. könnte für das atomare Hypertextsortenmodul *Publikationsliste* eine Binnenstrukturierung angenommen und im Rahmen einer Analyse untersucht werden. Die Bündelung von atomaren zu komplexen Hypertextsortenmodulen stellt daher nur eine rudimentäre und initiale Strukturierungs- und Hierarchisierungsebene dar, die auch auf weitere Ebenen ausgedehnt werden kann.

Merkmale von Hypertextsortenmodulen Die bislang angeführten Beispiele für Hypertextsortenmodule verdeutlichen, dass sich dieser Begriff auf ein extrem umfangreiches und heterogenes Inventar makrostruktureller Bausteine bezieht, die, um nur einige unterschiedlich granulare Beispiele zu nennen, von dem *Rückverweis zur Einstiegsseite* über den *Namen des Homepage-Besitzers*, die *Navigationshilfe* bis zum *Lebenslauf* reichen. Hypertextsortenmodule sind Komponenten, die vom Produzenten in einem Exemplar einer spezifischen Hypertextsorte entweder ausgeprägt werden oder nicht. Die Beispiele zeigen, dass es sich sowohl um konventionalisierte Hyperlinks (*Rückverweis zur Einstiegsseite*) als auch um deutlich umfangreichere und primär inhaltlich-thematische Bestandteile wie den *Lebenslauf* handeln kann.

Neben den Typenausprägungen besitzen Hypertextsortenmodule – in Bezug auf den Hypertext als übergeordnete funktional-thematische Ganzheit – den Status von Teiltexten. Es können prinzipiell alle in Kapitel 2 genannten Beschreibungsebenen für Texte bzw. Textsorten auch auf Hypertextsortenmodule angewendet werden. Sie besitzen z. B. in der Regel eine kommunikative Funktion, die sich etwa auf die übergeordnete kommunikative Funktion oder Navigationsaspekte beziehen kann und sie können auf etablierten Textsorten beruhen. Dieser Umstand wird bei der Etikettierung von Hypertextsortenmodulen deutlich: Falls sie auf traditionellen Textsorten oder Hypertextsorten basieren, wird das jeweilige Etikett übernommen (z. B. *Publikationsliste*, *Lebenslauf*, *Hotlist* und *Gästebuch*). Falls ein solcher Bezug nicht vorliegt, eine Komponente jedoch eindeutig von anderen differenziert und als Hypertextsortenmodul identifiziert werden kann, wird ein möglichst deskriptives Etikett gewählt.

Fazit Die in diesem Abschnitt in gebündelter Form vorgestellten Charakteristika von Hypertextsortenmodulen stellen nur einen Ausschnitt dar. Die in Teil III präsentierten Analysen beziehen sich vornehmlich auf diese Eigenschaften und stellen eine Vielzahl von Beispielen vor. Da Hypertextsortenmodule die grundlegenden Komponenten von Hypertextsorten darstellen, können zahlreiche weitere Eigenschaften in das Zentrum einer Untersuchung rücken (siehe hierzu auch das Modell von Huber, 2002, in Abschnitt 3.5.7, die Studie von Schütte, 2004a, in Abschnitt 4.6.2 und die Überlegungen von Jakobs, 2003).

5.7 Das Hypertextsortenmodell im Überblick

Abbildung 5.4 stellt den generischen Aufbau einer Hypertextsorte – und somit die zentralen Bestandteile des Hypertextsortenmodells – zusammenfassend in abstrakter Form dar.[46] Eine Hypertextsorte besteht aus mindestens einem obligatorischen Hypertextsortenmodul sowie mindestens einer Hypertextknotensorte (in diesem minimalen Fall fungiert das Hypertextsortenmodul als Hypertextknotensorte). Sie kann auch Hypertextsorten einbetten, die ihrerseits als funktional-thematisch markierte Ganzheiten fungieren und sich unterstützend auf die übergeordnete Ganzheit beziehen.[47] Zwingend notwendig ist die Einbettung einer Hypertextsorte auf einer generischen Ebene nicht, jedoch kann sie für eine spezifische Hypertextsorte (z. B. *Webauftritt einer Universität*) mit untergeordneten Hypertextsorten (etwa *Webauftritt einer Fakultät* oder *Webauftritt eines Fachbereichs*) obligatorisch sein. Eine Hypertextsorte muss mindestens ein obligatorisches Hypertextsortenmodul beinhalten, das den primären Typ Inhalt/Thema, Interaktion, Kommunikation oder Navigation besitzt. Die *minimale* Instanz einer beliebigen Hypertextsorte umfasst daher eine Instanz eines obligatorischen Hypertextsortenmoduls, das als Hypertextknotensorte fungieren kann, was typischerweise für Hypertextsortenmodule gilt, die ein etabliertes Textstrukturmuster besitzen. Zu-

[46] In Abbildung 5.4 werden die Begriffe „Hypertextsorte“ und „Hypertextknotensorte“ verwendet. Die Termini „Hypertexttyp“ und „Hypertextknotentyp“ beziehen sich auf abstraktere Klassen von Hypertexten bzw. Hypertextknoten, so wird die *Einstiegsseite des Webauftritts einer Institution* als Hypertextknotentyp verstanden, der spezifische Subkategorien wie z. B. *Einstiegsseite eines universitären Webauftritts* umfasst. Weiterhin ist anzumerken, dass sich dieses Modell auf Gebrauchshypertextsorten bezieht (vgl. Abschnitt 5.3.4).

[47] Eingebettete Hypertextsorten entsprechend ebenfalls dem Schema, das in Abbildung 5.4 dargestellt ist. Durch diese rekursive Verankerung können beliebig tiefe Einbettungen erfolgen.

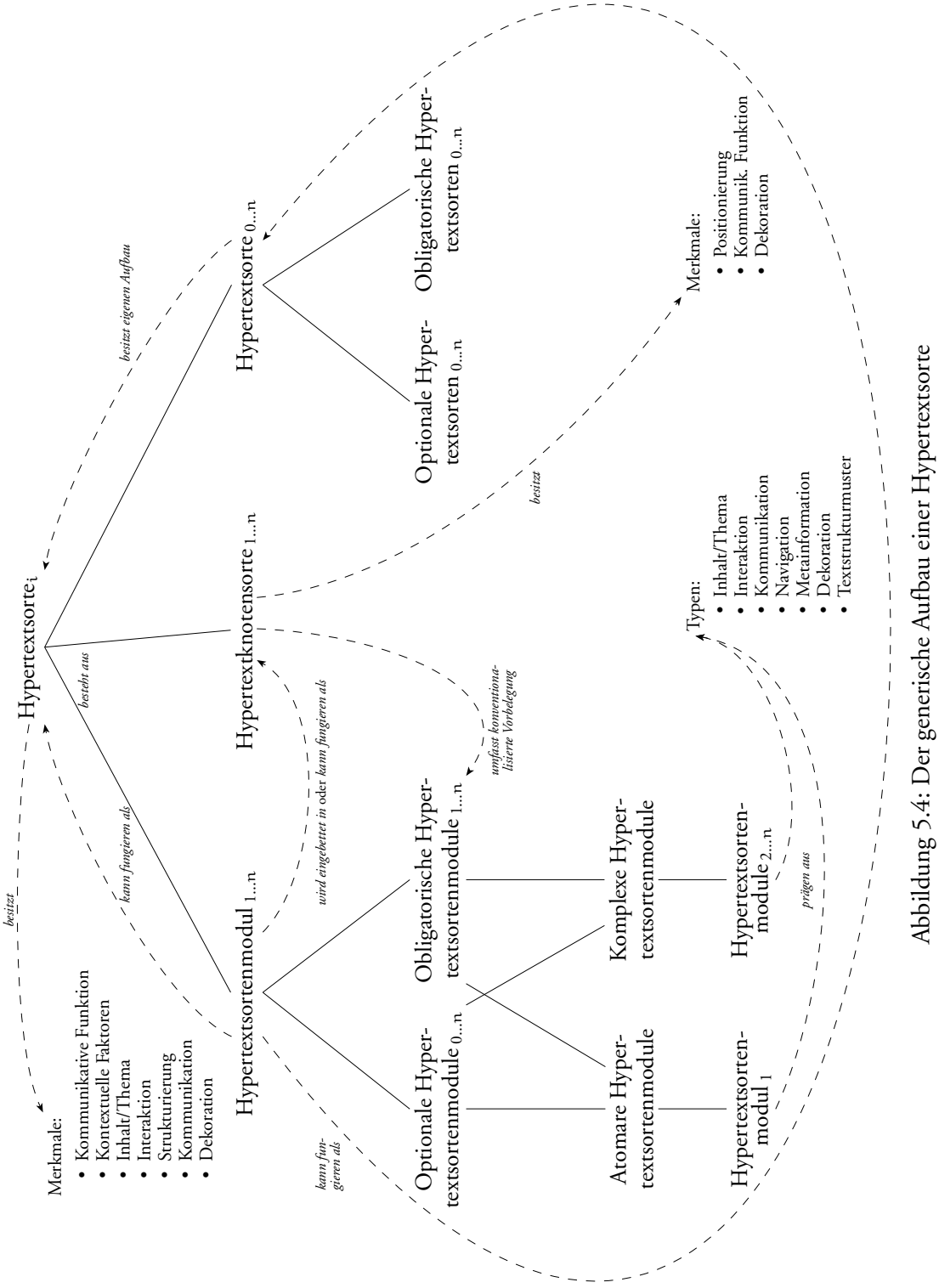


Abbildung 5.4: Der generische Aufbau einer Hypertextsorte

dem können derartige Hypertextsortenmodule als eigenständige Hypertextsorten fungieren (vgl. Fußnote 43, S. 287), was in Teil III anhand zahlreicher Beispiele erläutert wird.

Ein Inventar von Merkmalen wird zur Erfassung der grundlegenden Parameter einer Hypertextsorte eingesetzt. Die Ausprägungen dieser Merkmale – kommunikative Funktion, kontextuelle Faktoren, Inhalt bzw. Thema, Interaktion, Strukturierung, Kommunikation und Dekoration – sind von den empirischen Analysen abhängig, in denen Instanzen der korrespondierenden Hypertextsorte untersucht und dann induktiv generalisiert werden. Falls z. B. eine Markierung hinsichtlich der Dekoration in der Mehrzahl der Hypertextexemplare beobachtet werden kann, ist dieses Merkmal zu setzen. Im anderen Fall fungiert es als Leerstelle. Hypertextsortenmodule prägen unter anderem Typen aus (Inhalt/Thema, Interaktion, Kommunikation, Navigation, Metainformation, Dekoration, Textstrukturmuster), von denen eines den primären Typ des Hypertextsortenmoduls bestimmt. Hypertextknotensorten umfassen nur eine rudimentäre Menge von Merkmalen, da die Instanzen von Hypertextsortenmodulen, die in der Instanz einer Hypertextknotensorte enthalten sind, die untergeordneten Merkmalswerte auf die Ebene des Knotens und – bezogen auf die Gesamtheit aller Instanzen von Hypertextknotensorten – schließlich auf die Ebene der Hypertextsorteninstanz propagieren. Unterschiedliche Zusammenstellungen von Hypertextsortenmodulen resultieren somit in unterschiedlichen, Varietäten erfassenden Merkmalsausprägungen in den Instanzen der jeweiligen Hypertextknotensorte sowie der Hypertextsorte.

Der Aspekt der Typologisierung von Hypertextsorten wurde bislang nur am Rande angesprochen. Der Einsatz texttechnologischer Repräsentationsformate zur Modellierung der Textstrukturmuster traditioneller Textsorten sieht sich in diesem Zusammenhang mit dem Problem konfrontiert, dass sich Dokumentgrammatiken jeweils nur auf genau eine Textsorte beziehen. Die Abstraktion textsortenübergreifender Teilmuster ist nur eingeschränkt möglich (vgl. Abschnitt 5.2.1). Da die Auffassung von Hypertextsorten als Bündel individuell ausgeprägter Instanzen von Hypertextknotensorten und Hypertextsortenmodulen gerade im Hinblick auf diesen Aspekt sehr flexible Möglichkeiten der Modellierung erforderlich macht, ist der alleinige Einsatz DTD- oder XML Schema-basierter Dokumentgrammatiken ausgeschlossen. Hinzu kommt, dass konventionalisierte Hyperlinkstrukturen zwischen Knoten repräsentierbar sein müssen (vgl. Abschnitt 6.5). Im Kontext einer Typologie von Hypertextsorten ist es notwendig, Modularisierungs- und Vererbungsmechanismen einzuführen, um z. B. zu gewährleisten, dass subgenerische Varianten des Hypertexttyps *Homepage einer Person* adäquat repräsentiert werden können. Auf diese Weise können die Hypertextsorten *persönliche Homepage eines Wissenschaftlers* und *private Homepage eines Studierenden* als spezifische Ausprägungen des übergeordneten Typs modelliert werden, die dessen generische Eigenschaften erben. Zusätzlich ist es notwendig, neben der Typologie von Hypertexttypen und -sorten Typologien von Hypertextsortenmodulen und Hypertextknotensorten einzuführen, die als Bausteine der Typologie von Hypertexttypen fungieren und von den modellierten Kategorien flexibel referenziert werden können.

5.8 Zusätzliche Aspekte des Hypertextsortenmodells

Neben den in den Abschnitten 5.4 bis 5.6 dargestellten grundlegenden Konzepten des Hypertextsortenmodells sind zwei weiterführende Aspekte von Bedeutung. Dabei handelt es

sich um die Betrachtung von Hypertextsorten als Prototypen (Abschnitt 5.8.1) sowie um die Identifizierung und Sammlung von Hypertextsorten (Abschnitt 5.8.2).

5.8.1 Hypertextsorten als Prototypen

Das Hypertextsortenmodell fußt auf dem Prinzip, dass Hypertextsorten vornehmlich auf einem Inventar von Hypertextsortenmodulen basieren, die von Produzenten flexibel auf der Knotenebene strukturiert, im Rahmen einer empirischen Analyse rekonstruiert und zu einer Repräsentation der Hypertextsorte generalisiert werden können. Diese Repräsentation stellt daher in gewisser Weise den Prototypen einer Hypertextsorte dar. Die ermittelten hochfrequenten obligatorischen Hypertextsortenmodule bilden den Kern des Prototypen, wohingegen die niedrigfrequenten, aber ebenfalls in Exemplaren beobachteten Hypertextsortenmodule die äußere Peripherie des Prototyps konstituieren. Abschnitt 2.3.6 macht deutlich, dass Textsorten nicht notwendigerweise statische und restriktive Gebilde darstellen, sondern vielmehr eine große Variabilität besitzen, d. h. nicht nur „Textsorten sind prototypisch“ (Sandig, 1997, S. 29). Diese Eigenschaft ist auch Hypertextsorten zuzuweisen.

Charakterisierungen von Textsorten basieren in der Regel auf prototypischen bzw. idealtypischen Vertretern (Sandig, 2000, S. 103) und beziehen sich somit primär auf möglichst generische Verallgemeinerungen und nur sekundär auf Abweichungen von diesem prototypischen Kern (vgl. Abschnitt 2.3.6).[48] Sandig (2000, S. 103) zufolge sind Realisierungen von Textsorten „variabel je nach individueller Situation und individuell damit verfolgtem Zweck“, denn in Textsorten „sind bereits Grade der Variabilität vorgesehen.“ Diese Variabilität gilt umso mehr für Hypertextsorten (vgl. Kapitel 4 und Teil III); der flexible Mechanismus einzusetzender Hypertextsortenmodule, die als obligatorische oder optionale Konstituenten aufgefasst werden, trägt diesem Umstand Rechnung. Auf diese Weise können Hypertextexemplare als typische oder weniger typische Vertreter einer spezifischen Hypertextsorte konzeptualisiert werden. Dieser Grad der Zugehörigkeit kann ebenfalls auf der Ebene der Hypertextknotensorte ermittelt werden (vgl. Schütte, 2004a, S. 222 f., sowie Abschnitt 4.6.2).

Hypertexttypen bzw. -sorten besitzen als knotenübergreifende Rahmenkonzeptionen eindeutig kennzeichnende, prototypische Eigenschaften bezüglich verschiedener Beschreibungsebenen wie z. B. kommunikative Funktion, Thema und Interaktion. Spezifikationen dieser Eigenschaften stellen die typische Definition einer Hypertextsorte dar, die durch die Existenz optionaler Hypertextsortenmodule in einem Hypertextexemplar modifiziert werden kann. Da obligatorische Hypertextsortenmodule ein fester Bestandteil der Definition einer Hypertextsorte sind, besitzen optionale Hypertextsortenmodule somit die Möglichkeit, Varietäten auch inhaltlich, funktional und interaktional erfassen zu können. Ein optionales Hypertextsortenmodul *Suchformular*, das ebenfalls über eine definierende Menge von Merkmalen verfügt, könnte in einer Hypertextknoteninstanz verwendet werden und durch den definierten Status einer hohen Interaktion die Interaktionsmerkmale dieser Instanz und der übergreifenden Hypertextsorteninstanz inkrementieren.

[48] Bei der textlinguistischen Charakterisierung einer Textsorte ist bekannt, welche Textsorte vorliegt. Die vorliegende Arbeit wird mit der Problematik konfrontiert, dass zunächst Hypertextsorten, Hypertextknotensorten und Hypertextsortenmodule zu *identifizieren* sind (vgl. Abschnitt 4.4). Diese Vorgehensweise setzt die induktive und repetitive Analyse einer Stichprobe voraus, um Gemeinsamkeiten und Konventionen zu ermitteln, die die Annahme dieser Entitäten rechtfertigen. Diese Problematik wird in Teil III in detaillierter Form diskutiert.

5.8.2 Zur Sammlung und Identifizierung von Hypertextsorten

Crowston und Kwasnik (2004, S. 3) unterscheiden zwei Ansätze zur Erstellung einer Klassifikation von Web-Genres: Zunächst kann man *top-down* von existenten Genre-Bezeichnungen aus den Bereichen der Textlinguistik oder des Bibliothekswesens ausgehen (vgl. Dimter, 1981) und versuchen, die in einer Analyse ermittelten Web-Genres auf das vorhandene Inventar abzubilden. Eine derartige Vorgehensweise ist mit zwei Problemen verbunden (Crowston und Kwasnik, 2004, S. 4): Erstens konstituieren sich Genres in Diskursgemeinschaften. Ein Genre, das sich in einer Diskursgemeinschaft entwickelt hat, der die analysierende Person nicht angehört, wird also möglicherweise in nicht adäquater Weise auf ein Genre des vorhandenen Inventars abgebildet. Zweitens sei es problematisch, neuartige Web-Genres in ein derartiges Inventar einzugliedern, das von traditionellen Genres geprägt ist. Crowston und Kwasnik (2004, S. 4) kommen bezüglich dieser ersten Alternative zu dem Schluss, „that a traditional typology of genre or document forms will not be sufficient to describe the emerging and dynamic genres identifiable by users." Aus diesem Grund präferieren Crowston und Kwasnik die Alternative, eine Klassifikation von Genres *bottom-up*, d. h. aus Sicht der Benutzer aufzubauen, was sowohl den Prozess der Einordnung von Genres in ein Klassifikationsschema als auch ihre Etikettierung umfasst. So ist es Crowston und Kwasnik zufolge notwendig, eine Übereinstimmung mit den von Anwendern des WWW verinnerlichten Konzepten zu gewährleisten, wenn eine Genre-Klassifikation als zusätzliche Ressource in eine Suchmaschine integriert werden soll. Zur Repräsentation des Klassifikationsschemas schlagen Crowston und Kwasnik den Einsatz eines Inventars von Facetten vor (vgl. Abschnitt 5.2.2).

Crowston und Williams (1997, 2000), Shepherd und Watters (1999) sowie Haas und Grams (1998a, 1998b, 2000) werten Stichproben aus, die nach dem Zufallsprinzip aus den Datenbeständen von Suchmaschinen zusammengestellt wurden (vgl. Abschnitt 4.4). Bei der Datenerhebung wurden keinerlei Restriktionen vorgenommen, so dass die Resultate dieser Studien sehr heterogener Natur sind und nur eine partielle Deckungsgleichheit aufweisen, da die jeweiligen Verfasser auf unterschiedliche Granularitätsstufen der zu ermittelnden Entitäten abzielten. Ein Vergleich dieser nach Crowston und Kwasnik (2004) *top-down* durchgeführten Arbeiten mit der Studie von Brandl (2002), deren Inventar von Website-Typen aus Interviews mit Experten ermittelt wurde, zeigt, dass verschiedene Korrespondenzen existieren. Während die vollständig *bottom-up* (Crowston und Kwasnik, 2004) realisierte Fundierung einer Typologie von Hypertextsorten auf den von Rezipienten wahrgenommenen Kategorien zur Ermittlung der kommunikativen Realität zu präferieren ist, belegen diese Übereinstimmungen, dass den eingangs aufgeführten, induktiv und introspektiv von den jeweiligen Verfassern angefertigten Studien durchaus eine gewisse Aussagekraft und Plausibilität zuzuschreiben ist (vgl. Abschnitt 4.4.8). Zur Erstellung einer homogenen Typologie von Hypertextsorten ist es unumgänglich, die Untersuchungsdomäne einzuschränken, so dass einzelne Stichproben erhoben und hinsichtlich des spezifischen Forschungsinteresses analysiert werden können, um somit sukzessive Teilaspekte der Typologie zu spezifizieren. Die vorliegende Arbeit beschränkt sich auf die deutschsprachigen Dokumente der Webangebote deutscher Hochschulen. Diese Vorgehensweise basiert auf der Hypothese, dass diese Domäne einer relativ homogenen Diskursgemeinschaft zugehörig und hinreichend spezifisch ist, um bezüglich der Ergebnisse Homogenität gewährleisten zu können.

Im Hinblick auf die Identifizierung eigenständiger Hypertextsorten gehen die in Teil III dargestellten Stichprobenanalysen von drei unterschiedlichen Orientierungspunkten aus. Zunächst handelt es sich dabei um diejenigen Hypertextsorten, die bislang von der Forschung ermittelt und charakterisiert worden sind (vgl. Abschnitt 4.6). Hinsichtlich der Untersuchungsdomäne der universitären Webangebote kommt den traditionellen Textsorten, die Heinemann (2000b) für den Bereich Hochschule und Wissenschaft ermittelt hat (vgl. Abschnitt 6.2), eine besondere Bedeutung zu. Dies gilt ebenfalls für die von Schütte (2004a, S. 211) aufgestellte Gleichung „Emittent = Thema“, da ein universitäres Webangebot über zahlreiche Emittenten verfügt, die jeweils eigene Hypertexte im WWW pflegen (vgl. Abschnitt 5.3.4). Das in diesem Kapitel vorgestellte Hypertextsortenmodell besitzt einen komplexen Aufbau. Die Spezifizierung eines vollständigen und jedes einzelne Hypertextexemplar des Korpus (vgl. Kapitel 7) abdeckenden Hypertexttyps *Webauftritt einer Universität* ist nicht das Ziel dieser Arbeit. Vielmehr soll die Anwendung des Hypertextsortenmodells auf eine spezifische Untersuchungsdomäne exemplarisch demonstriert werden, um (unter anderem) eine initiale Sammlung und Typologisierung von Hypertextsorten durchzuführen.

5.9 Zusammenfassung

Der Geltungsbereich des in diesem Kapitel vorgestellten Hypertextsortenmodells bezieht sich auf beliebige Gebrauchshypertexte im WWW. Das Modell basiert auf drei Granularitätsebenen: Die Begriffe Hypertexttyp und Hypertextsorte bezeichnen den generellen Typ eines Hypertextes im Sinne einer funktional-thematisch markierten kommunikativen Ganzheit. Hypertextsorten und zugehörige Hypertextexemplare können über ihre kommunikative Funktion, kontextuelle Faktoren, ihren Inhalt bzw. ihr Thema, Möglichkeiten der Interaktion, ihre Strukturierung, Merkmale der Kommunikation sowie ihre Dekoration charakterisiert werden. Auf der zweiten Ebene existieren Hypertextknotentypen bzw. -sorten, die durch HTML-Dokumente instanziiert werden. Die Gesamtheit dieser Instanzen bildet die Instanz einer Hypertextsorte. Hypertextknotentypen können durch eine spezifische kommunikative Funktion, dekorative Elemente und Aspekte der Positionierung charakterisiert werden. Das Merkmal der Positionierung bezieht sich auf Instanzen von Hypertextsortenmodulen, die die dritte Granularitätsebene darstellen und als makrostrukturelle Bausteine fungieren. Neben dem inhaltlich-thematischen Typ existieren Hypertextsortenmodule der Typen Interaktion, Kommunikation, Navigation, Metainformation, Dekoration und Textstrukturmuster. Die zusätzliche Annahme einer Ebene von Hypertextsortenmodulen ist aus zwei Gründen notwendig: Erstens können die makrostrukturelleren Komponenten von einem Hypertextproduzenten prinzipiell beliebig in unterschiedlichen Knoten angeordnet werden, zweitens können in Hypertextexemplaren Konventionen hinsichtlich der spezifischen Verwendung und Anordnung auf der Knotenebene beobachtet werden. Darüber hinaus können Hypertextsortenmodule als Hypertextknotensorten und als Hypertextsorten fungieren, was in der Regel für Hypertextsortenmodule gilt, die ein etabliertes Textstrukturmuster besitzen. Instanzen von Hypertextsorten können Instanzen weiterer Hypertextsorten einbetten, die sich in diesem Fall vornehmlich in unterstützender Weise auf die kommunikative Funktion der übergeordneten Ganzheit beziehen.

Profile von Hypertextsorten auf der Grundlage dieses Modells sind mit empirischen Analysen zu erstellen, in denen Hypertextexemplare einer spezifischen Untersuchungsdomäne im Hinblick auf die beschriebenen Aspekte untersucht werden. Auf diese Weise können die einzelnen Konstituenten von Hypertextsorten ermittelt werden. Diejenigen Hypertextsortenmodule, die in der Mehrzahl der untersuchten Hypertexte einer spezifischen Hypertextsorte verwendet werden, können als obligatorisch betrachtet und von optionalen, d. h. niedrigfrequenten Hypertextsortenmodulen differenziert werden. Die obligatorischen Hypertextsortenmodule fungieren dabei als prototypischer Kern der Hypertextsorte. Auf diese Weise können abweichende Vertreter von Hypertextsorten als Varietäten konzeptualisiert werden. Die dabei erfassten Merkmalsausprägungen werden von den Hypertextsortenmodulen über die Ebene der Hypertextknotensorte bis zur Hypertextsorte propagiert und wirken sich auf dieser globalen Beschreibungsebene auf die Eigenschaften der übergreifenden Ganzheit aus.

5.10 Fazit – Von der Theorie zur Anwendung

Das in diesem Kapitel vorgestellte Hypertextsortenmodell wurde für zwei Anwendungen entwickelt. Einerseits stellt es drei grundlegende Ebenen der Analyse und assoziierte Kriterien zur Verfügung, die zur *textlinguistischen Charakterisierung von Hypertextsorten* eingesetzt werden können. Andererseits können derartige Charakterisierungen als *Grundlage sprach- und texttechnologischer Anwendungen* dienen – diese Arbeit fokussiert die Applikationen der maschinellen Identifizierung von Hypertextsorten sowie die Informationsextraktion. Eine Ontologie bildet hierbei die Schnittstelle zwischen den textlinguistischen Charakterisierungen und sprachtechnologischen Anwendungen, d. h. die Resultate der Untersuchungen von Hypertextexemplaren können zur Konstruktion einer Ontologie von Hypertextsorten eingesetzt werden, die im textlinguistischen Sinne als Typologie aufgefasst werden kann. Diese Ontologie kann mit verschiedenen Werkzeugen in einem für das *Semantic Web* entwickelten texttechnologischen Standardformat repräsentiert werden, so dass eine vollständige Unabhängigkeit von spezifischen Plattformen und Anwendungen gewährleistet ist.

Während dieses Kapitel die theoretischen Grundlagen des Hypertextsortenmodells vorgestellt hat, geht das nachfolgende Kapitel 6 zunächst auf die Untersuchungsdomäne der universitären Webangebote ein, woraufhin Kapitel 7 das dieser Arbeit zugrunde liegende Korpus sowie die Korpusdatenbank vorstellt. Teil III präsentiert verschiedene Analysen und Sammlungen von Hypertextsorten innerhalb der Untersuchungsdomäne, wobei unterschiedliche Aspekte des Hypertextsortenmodells exemplarisch aufgezeigt werden. Daraufhin erläutert Kapitel 13 die bereits angesprochene Ontologie von Hypertextsorten, die auf den Ergebnissen der Analysen aus Teil III beruht. Die Kapitel 8 bis 13 stellen daher in methodologischer und technologischer Hinsicht eine exemplarische Anwendung des Hypertextsortenmodells dar. Kapitel 14 diskutiert abschließend sprach- und texttechnologischen Anwendungen, die auf der Grundlage der Ontologie von Hypertextsorten implementiert werden können bzw. für ihre vollständige Nutzung in sprachtechnologischen Applikationen benötigt werden.

6

Die Untersuchungsdomäne: Universitäre Webangebote

6.1 Einleitung

In der vorliegenden Arbeit werden Hypertextsorten am Beispiel universitärer Webangebote untersucht, für die ein Korpus der deutschsprachigen HTML-Dokumente aller deutschen Universitäten aufgebaut wurde (vgl. Kapitel 7).[1] Analysen von Stichproben, die zufällig aus den Beständen von Suchmaschinen oder Katalogen zusammengestellt wurden, ohne dabei Beschränkungen hinsichtlich der Datensammlung vorzunehmen, können zwar die im WWW vorhandene Bandbreite von Hypertextsorten veranschaulichen, sie sind jedoch nicht geeignet, um homogene Ergebnisse zu erzielen, die im Kontext der maschinellen Identifizierung von Hypertextsorten oder hierauf aufbauenden Verarbeitungsprozessen eingesetzt werden können (vgl. Abschnitt 4.7). Die Wahl der eingangs genannten Untersuchungsdomäne erfolgte aus verschiedenen Beweggründen. Zunächst handelt es sich bei den Webauftritten von Hochschulen – von einer sehr abstrakten Perspektive betrachtet – um einen sehr homogenen Kommunikationsbereich, der zahlreiche gemeinsame Merkmale beinhaltet, die einer Generalisierung unterzogen werden können. Zugleich liegen sehr komplexe Inhalte, Funktionen

[1] Bereits Beghtol (2001) weist in ihrer sehr abstrakten Betrachtung digitaler Genres auf die Tatsache hin, dass der Untersuchungsbereich einzuschränken ist, denn „for genre analysis to be most useful, we need to identify the domain of interest and then assemble as complete a set of genre typologies as possible, based on as many characteristics of division as seem helpful. In this way, we can understand the structure of the domain and potentially use that understanding to increase the precision and success of information retrieval." Auch Schütte (2004a, S. 152) zufolge erscheint es angeraten, „zu Typologisierungszwecken zunächst domänenbezogen vorzugehen". Androutsopoulos und Kraft (2003, S. 4) erwarten innerhalb einer Untersuchgsdomäne wie z. B. universitären Webangeboten „ein Wechselspiel von Konvergenzen und Divergenzen, gemeinsamen Lösungen und individuellen Besonderheiten". Thelwall (2005, S. 610) beschäftigt sich mit dem Bereich „scientific Web intelligence" und hebt das Potenzial der gewählten Untersuchungsdomäne hervor: „[S]tudies restricted to the academic Web have more potential for a systematic approach than those aimed at the general Web."

und Strukturierungen vor, die Repräsentationsverfahren benötigen, welche wiederum nicht vom Untersuchungsgegenstand abhängig sein sollten (vgl. die Kapitel 5 und 13). Die Wahl der universitären Webangebote wurde ebenfalls durch die Tatsache beeinflusst, dass sie bezüglich der von ihnen umfassten Hypertextsorten bislang nicht untersucht wurden, und darüber hinaus meist manuell hergestellt werden, d. h. im Korpus existieren nur wenige Websites, die auf Datenbank-gestützten CMS-Systemen basieren, welche für die Untersuchung rekurrenter Strukturen weniger interessant sind, da die angebotenen Vorlagen oftmals unverändert übernommen werden. Abschnitt 6.2 geht zunächst auf die traditionellen Textsorten ein, die im Kommunikationsbereich Hochschule und Wissenschaft eingesetzt werden, woraufhin Abschnitt 6.3 die grundlegenden Merkmale universitärer Webangebote darstellt.

6.2 Textsorten in der Hochschule

Zur Kontrastierung der nachfolgenden Ausführungen ist es notwendig, Zusammenstellungen traditioneller Textsorten einzubeziehen. Als Ausgangspunkt dient der Beitrag „Textsorten des Bereichs Hochschule und Wissenschaft“ (Heinemann, 2000b), der den Kommunikationsbereich der Institution Hochschule als gesellschaftlich determinierten Rahmen beschreibt, in dem die Handelnden in charakteristischer Weise eine gewisse Menge von Zielen mittels typischer (sprachlicher) Handlungen verfolgen (ebd., S. 702). Zu der Institution gehören nicht nur Gebäude, Einrichtungen und Geräte als Voraussetzungen für das zweckgerichtete Handeln, sondern auch Individuen als Repräsentanten sozialer Gruppen, die über bestimmte Fähigkeiten verfügen, um die komplexen Ziele im Rahmen der Institution erreichen zu können. Die Institution wird durch ein hierarchisches, interpersonales Beziehungsgefüge gesteuert, so dass das Handeln nach Verbindlichkeiten und Konventionen abläuft, die als Handlungs- und Kommunikationsmuster aufgefasst werden können. Die institutionelle Kommunikation wird als „asymmetrisch“ („mit unterschiedlichem Rede- und Fragerecht“), „stärker verbindlich“ und „eher thematisch-fachlich als partner-orientiert“ charakterisiert (ebd.). Zu ihrem „Kern gehören [...] alle Handlungen und Sprachhandlungen/Texte, die auf das (verallgemeinernde, theoriebezogene) Eruieren, Erfassen und Beschreiben von Phänomenen der Welt und das Lösen von Problemen gerichtet sind.“ (ebd., S. 703). Es existieren aber auch Bereiche, in denen sprachliche Handlungen durchgeführt werden, die nicht der Wissenschaft zuzuordnen sind (z. B. administrative Aktivitäten, Baumaßnahmen und Alltagsgespräche). Heinemanns Darstellung der Textsorten des Bereichs Hochschule und Wissenschaft orientiert sich an drei charakteristischen Textsortenklassen, bei denen es sich um wissenschaftlich geprägte, wissenschaftspraktische und organisierende Textsorten handelt.

6.2.1 Forschung und Wissenschaft – Theoriebezogene Textsorten

In Bezug auf die theoriebezogenen Textsorten dient die „Orientierung auf Phänomene der *Wissenschaft*“ als „dominant inhaltliches Basiskriterium“ (Heinemann, 2000b, S. 704), als gemeinsame Einordnungsinstanz (vgl. Abschnitt 2.3.3). Als Ziele aller Formen wissenschaftlicher Tätigkeiten beschreibt Heinemann das Erkenennen von Zusammenhängen der objektiven Welt, das Beschreiben und Erläutern allgemeiner Merkmale und das Zuordnen von

Merkmalen der konkreten individuellen Erscheinungen, Gegenstände und Zustände zu komplexeren Ganzheiten oder Strukturen sowie das Prognostizieren von und Reflektieren über Anwendungsmöglichkeiten. Abstrakter formuliert: Das Ziel wissenschaftlicher Tätigkeiten ist die Gewinnung von Wissen. Dieses wiederum besitzt Heinemann zufolge ein „genuines »Transportmittel«, [...] die vielzitierte ›Wissenschaftssprache‹“ (ebd.), die die Reproduktion des Wissens zum Zwecke seiner Weitergabe ermöglicht. Daher wird die Wissenschaftssprache auch als „Grundlage für die Wirksamkeit wissenschaftlicher Forschungen“ bezeichnet. In den Fachsprachen besitzt die allgemeine Wissenschaftssprache Komplemente, die durch zusätzliche Fachjargonismen und Fachtextsorten charakterisiert werden.[2]

Die folgenden linguistischen und textlinguistischen Phänomene können nach Heinemann (2000b, S. 704 f.) insbesondere der allgemeinen Wissenschaftssprache zugeordnet werden: (i) „inhaltliche Bezogenheit auf wissenschaftliche Problemstellungen und -lösungen“; (ii) „Expertencharakter der Darstellung“; (iii) „primäre Orientiertheit auf Wissenschaftler als Partner“; (iv) „Para-Texte als Präsignale (Vorwort, Anmerkung, Bibliographie, Register)“; (v) „Themeneinleitungen und Abschluss-Sätze als textuelle Verweis-Signale“; (vi) „eindeutige Strukturiertheit/Gegliedertheit der Texte (Haupt-, Zwischen-, Subtitel)“; (vii) „Fach-Code-Gemeinschaft“; (viii) „hohe Frequenz vor allem substantivischer Termini“; (ix) „häufige Zitationen als Autoritätsargumente“; (x) „Verlagerung der wichtigsten Informationen in den nominalen Bereich (Informationsverdichtung [...]) und Desemantisierung präsenter Verben“; (xi) „relativ hohe Frequenz von Passiv- und Infinitkonstruktionen“; (xii) zusätzlich gelten „drei fundamentale strategische Maximen“, so dass „die Fakten ›objektiv‹ für sich selbst sprechen: das Ich-Verbot, Erzähl-Verbot, Metaphern-Verbot“.

Durch diese Merkmale sind verschiedene theoriebezogene Textsorten gekennzeichnet: „Als Prototypen solcher wissenschaftlicher (Schrift-)Textsorten gelten *Monographien*, *Abhandlungen/Aufsätze*, *Forschungsberichte*, *Rezensionen*, *Dissertationen*, *Abstracts*, [...] *Magister-/Diplomarbeiten*, da diesen Textsorten wissenschaftliches Sprachhandeln als selbstverständlich unterlegt wird.“ (ebd., S. 705). Diese Textsorten (Sandig, 1997, zählt einige weitere auf) werden fast ausschließlich in einer Hochschule produziert, ihre Terminologie „geht über das Terminologieverständnis allgemeinwissenschaftlicher [...] Texte hinaus, sie sind also auf die konkrete Lösung eines Einzelproblems [...] direkt ausgerichtet“ (ebd.). Diese Textsorten können weiter unterteilt werden, so gelten z. B. *Monographien* oder *Dissertationen* als Primärtextsorten, wohingegen etwa *Rezensionen* oder *Abstracts* als Sekundärtextsorten fungieren (basierend auf Göpferich, 1995). Zu dieser Klasse gehören auch *wissenschaftliche Artikel in Enzyklopädien, Lexika und Wörterbüchern*, da sie auf bestehenden wissenschaftlichen Arbeiten beruhen und Wissen in komprimierter Form vermitteln.[3]

In den Grenzbereich zwischen wissenskonstituierenden/theoriebezogenen und wissenstransmittierenden/didaktischen Textsorten fällt nach Heinemann das *wissenschaftliche Gutachten*. Für dessen Konstitution sind „wissenschaftssprachliche Verfahren wie *argumentieren*,

[2] Sandig (1997, S. 31) bezeichnet die „allgemeine Wissenschaftssprache“ als „fächerübergreifende Wissenschaftssprache bzw. Wissenschaftsstil“ und nennt zusätzlich „verschiedene spezifischere Wissenschaftsstile“ innerhalb einzelner Domänen, was dem von Heinemann gewählten Terminus entspricht.

[3] Eine differenziertere Typologie geht von Partnerbeziehungen als dominantem Basiskriterium aus: Hiernach können Textsorten der fachinternen (zwischen Experten), der interfachlichen (geringer Grad an Textsorten- und Fachkenntnis beim Partner) und der fachexternen Kommunikation (zwischen Experten und Laien) unterschieden werden (Heinemann, 2000b, S. 705).

beweisen, *begründen*, *definieren*, *widerlegen*, *erörtern* usw. notwendig, so dass auch mit Hilfe dieser Textsorte (neues) Wissen vermittelt wird, wenn auch vor dem Hintergrund vorhandenen Textmaterials." (Heinemann, 2000b, S. 706). Für die Untersuchung von Hypertextsorten ist eine zweite Textklasse relevant, denn „in diesen Grenzbereich" sind auch „Textsorten in und mit den »neuen Medien« einzuordnen" (ebd.); deren Nutzung

> ist einer Vielzahl von Rezipienten (noch) nicht geläufig, zudem sei der technische Aufwand für viele abschreckend; und die Texte seien nur selektiv, als Teiltexte oder in komprimierten Formen, nutzbar – so zumindest lauten die Argumente der Skeptiker. Andererseits kann auf die Aktualität und schnelle Verfügbarkeit neuester Informationen verwiesen werden. Als Informationsquellen sind die externen Wissensspeicher für die wissenschaftliche Tätigkeit unverzichtbar geworden; deshalb sind sie auch den konstitutiven Texten bzw. Textsorten des Bereichs Hochschule und Wissenschaft zuzuordnen. (Heinemann, 2000b, S. 706)

Auf die „Argumente der Skeptiker" soll an dieser Stelle nicht näher eingegangen werden, da sie in einer Zeit, in der z. B. mehr als die Hälfte aller Bundesbürger über 14 Jahren das Internet nutzen, nicht mehr greifen. Dass sich Heinemann auf das Internet bzw. die Internet-Dienste bezieht, kann nur indirekt erschlossen werden („schnelle Verfügbarkeit", „externe Wissensspeicher"), da die Extension des nicht mehr zeitgemäß erscheinenden Terminus „neue Medien" nicht eindeutig ist. Heinemann betont zwar, dass es „noch an umfassenden Untersuchungen fehlt" (ebd.), jedoch bleibt unklar, weshalb sie eine Pauschalzuweisung dieser Textklasse zu den „konstitutiven Texten bzw. Textsorten des Bereichs Hochschule und Wissenschaft" vornimmt, ohne zunächst eine genauere Bestimmung der Medien des Gegenstandsbereichs (CD ROM, SMS, WWW, E-Mail etc.) und anschließend eine Spezifizierung der betrachteten Merkmale oder Informations- bzw. Kommunikationsangebote durchzuführen (z. B. *von einer Universität angebotene Webdokumente*, *interaktive E-Learning-Module*, *wissenschaftliche Informationssysteme* oder *digitale Bibliotheken*).[4]

6.2.2 Wissenschaftsdidaktik – Wissenstransmittierende Textsorten

Der Wissenschaftsdidaktik sind unterschiedliche Textsorten zuzuordnen, die der Vermittlung von Wissenschaftswissen dienen: *Vorlesungs-* und *Seminarkonzeptionen*, *Vorlesungs-/Seminarmanuskripte*, *Rahmenpläne für Vorlesungen und Seminare*, *Literaturlisten*, *Terminilisten*, *Handouts/Tischvorlagen* (mit Tabellen und Grafiken), *Manuskripte für Publikationen* (primär wissenschaftsdidaktischen Inhalts) etc. „Besondere Hervorhebung verdienen in diesem Zusammenhang verschiedene Formen von *Lehrmaterialien* [...], vor allem natürlich *wissenschaftliche Artikel in Hochschul-Lehrbüchern* bzw. die komplexe Kommunikationsform des *Hochschul-Lehrbuchs* selbst, das zielgruppenorientiert geordnete und im wesentlichen abgesicherte wissenschaftliche Einsichten [...] vermittelt." (Heinemann, 2000b, S. 706).[5]

[4] Das Beispiel *interaktive E-Learning-Module* macht weiterhin deutlich, dass bei der Diskussion digitaler Dokumente eine differenzierte Betrachtungsweise notwendig ist, in diesem Fall ist z. B. auch der Bereich wissenstransmittierender Textsorten betroffen (vgl. auch Abschnitt 2.1).

[5] Viele Texte, die insbesondere in Hauptseminaren verwendet werden, stammen gerade nicht aus „Hochschul-Lehrbüchern", sondern aus Monografien, Sammelbänden oder auch Konferenzbänden. Aus diesem Grund liegt zumindest bei diesen Lehrmaterialien ebenfalls ein Grenzbereich zu den theoriebezogenen Textsorten vor.

Hinzu kommen diejenigen Textsorten, die von Studierenden zum Zweck der Wissensaufnahme und Wissensverarbeitung verwendet werden: *Vorlesungsnachschriften*, *Exzerpte*, *Konspekte*, *Stichwortzettel* und *Protokolle* (z. B. von Lehrveranstaltungen, Experimenten, Analysen etc.).[6] Bei diesen Beispielen handelt es sich nach Heinemann um „komprimierend-zweckbestimmte Formen der Wissensaneignung und -aufbereitung", die „einerseits inhaltlich-sachliche Präzision [voraussetzen] (*Vorlesungsmitschriften*) [...], andererseits aber verfügen sie über ein umfangreiches Spektrum individueller Gestaltung (z. B. im Hinblick auf Abkürzungen, Hervorhebungen, den Grad der Vollständigkeit, die Reihenfolge der Fakten)." (ebd.). Eine weitere Besonderheit betrifft hierbei ihr „Eingebundensein in größere Textkomplexe", denn anders „als die Sekundär-Textsorten sind die hier genannten Teil-Textsorten als wissenordnende und wissenaufbewahrende Textklassen in diesem Bereich unverzichtbar." (ebd.).[7]

Die dritte Gruppe wissenschaftsdidaktischer Textsorten umfasst studentische Schrifttexte mit „Als-ob-Status" (Heinemann, 2000b, S. 706), d. h. *Referate*, *schriftliche Hausarbeiten*, *Magister-* und *Diplomarbeiten* etc. Bei diesen Textsorten wird „die Fiktion aufrechterhalten [...], dass es sich in diesen und ähnlichen Fällen um ein Kommunizieren zwischen Experten/Spezialisten eines Fachgebiets handelt" (ebd., S. 707).

Die letzte Gruppe von Textsorten hat die Funktion, Kenntnisse, Fähigkeiten und Fertigkeiten der Studierenden zu überprüfen, um bei Erfolg entsprechende Leistungsnachweise auszuhändigen: *Klausuren*, *Hausarbeiten*, *Seminararbeiten*, *Handouts* (für studentische Vorträge) und *Referate* (in Schriftform). „Diese Formen der Leistungskontrolle sind wissenschaftstheoretischen Anforderungen verpflichtet; sie müssen aber auch [...] nach Form, Umfang und teilweise Inhalt verwaltungstechnischen Vorgaben entsprechen" (ebd.).[8]

6.2.3 Textsorten der Wissenschaftsverwaltung

Als dritten und letzten Bestandteil des Kommunikationsbereichs Hochschule identifiziert Heinemann die Wissenschaftsverwaltung und betont ihre Bedeutung für die Universität:

> [Ohne] straffe Organisation aller innerinstitutionellen [...] Beziehungen [...] wäre das Funktionieren von Wissenschaft [...] nicht denkbar. In diesem Sinne bilden die Hochschul-Verwaltungen gleichsam die organisatorische Mitte der Institution, das, was Uni-

[6] Texte der nicht ausschließlich auf Lehrveranstaltungen bezogenen Textsorten *Konspekt*, *Stichwortzettel*, *Protokoll* etc. werden im Rahmen von Forschungsaktivitäten auch vom wissenschaftlichen Personal (z. B. Hochschullehrer, Projektmitarbeiter und wissenschaftliche Mitarbeiter) angefertigt, weshalb diese Textsorten zusätzlich dem Bereich theoriebezogener Textsorten zugerechnet werden müssen. Heinemann sieht offenbar die Publikation eines Textes als konstitutive Eigenschaft für Textsorten eben dieses Bereiches an.

[7] Die Frage, weshalb die genannten Textsorten als „Teil-Textsorten" und nicht als Sekundärtextsorten bezeichnet werden, wird von Heinemann nicht thematisiert. Nach Göpferich (1995, S. 132 f.) sind Sekundärtextsorten „Texte zuzurechnen, die durch Selektion, Komprimierung, Kommentierung und/oder Evaluation der Informationen aus Primärtextsorten hervorgehen." Weiterhin wird eine zusätzliche Unterscheidung getroffen: „Der Terminus der Sekundärtextes darf nicht mit demjenigen des Auxiliartextes [...] verwechselt werden. Ersterer betont die Entstehungsbedingungen des Textes, letzterer die Funktion. Ein Auxiliartext kann im Gegensatz zu einem Sekundärtext niemals unabhängig vom Rest des Globaltextes erworben werden" (ebd., S. 132), und er ist weiterhin auch zwingend mit dem Medium seines Globaltextes verbunden. Demnach können die genannten Textsorten also durchaus als Sekundärtextsorten bezeichnet werden.

[8] Die Liste der Textsorten wurde hier ebenfalls unmittelbar von Heinemann übernommen. Differenzierungskriterien für die Textsorten *Hausarbeit*, *Seminararbeit* und *Referat* (in Schriftform) werden nicht angegeben.

> versitäten und Hochschulen trotz zahlreicher auseinanderdriftenden Tendenzen – im Innersten [...] zusammenhält, indem Ämter/Dezernate und Behörden der Wissenschaftsverwaltung [...] weitgehend normierte Handlungsabläufe des Bereichs koordinieren und kontrollieren sowie nicht zuletzt Wissen (Ergebnisse von wissenschaftlichen Tätigkeiten [...] und von Verwaltungsprozessen) speichern. (Heinemann, 2000b, S. 707)

Universitäten besitzen zahlreiche Organisationseinheiten, in deren Varietät sich die Vielzahl der verbundenen Aufgaben widerspiegelt: Die *Universitätsverwaltung* wird hierarchisch strukturiert in verschiedene Organe (*Pressestelle*, *Personalrat*, *Amt für Auslandsbeziehungen* etc.), in Dezernate (z. B. *Haushalts- und Personaldezernat*, *Dezernat für Akademische Verwaltung*, *Dezernat für Betriebstechnik*, *Haushaltsplanung*) sowie Sachgebiete (häufig innerhalb einzelner Dezernate bzw. Referate angeordnet, z. B. *Akademische Angelegenheiten*, *Studentenangelegenheiten* und *Akademisches Auslandsamt*, letzteres kann weiter untergliedert sein in *EU-Programme*, *Universitätskontakte*, *Gästehäuser* etc.). Auf der dezentralen Ebene befinden sich mit den Organisationseinheiten der Fakultäten, Fachbereiche und Institute weitere Abteilungen, in denen Verwaltungsaufgaben durchgeführt werden.[9]

Die zentralen und dezentralen Organisationseinheiten führen eine Vielzahl von Verwaltungshandlungen mit unterschiedlichen Aufgaben und Zwecken durch. Heinemann teilt die *Textsorten der Verwaltungskommunikation*[10] in drei Klassen ein. (1) „Juristische bzw. politische und damit verwaltungsexterne Bezugs- und Rahmentexte für alle Verwaltungsprozesse“, die normierte Rahmenvorgaben für das Handeln in der Verwaltung darstellen und die aus diesem Grund in zugehörigen Texten referenziert werden: (i) *Gesetzestexte*, z. B. *Hochschulrahmengesetz*, *Hochschulerneuerungsgesetz*, *Gesetze über die Hochschulen im Freistaat Sachsen*; (ii) *Verordnungen*, z. B. *über die Vergabe von Studienplätzen*, *über Art und Umfang der dienstlichen Aufgaben*, *über Nebentätigkeiten an staatlichen Hochschulen*; (iii) *Erlasse*, z. B. *über den Aufenthalt ausländischer Studienbewerber*, *über die Grundsätze für den Hochschulzugang*; (iv) *Satzungen*, z. B. von *An-Instituten*. (2) „Verwaltungsinterne *Dienstanweisungen* und *Geschäftsordnungen*“ fungieren ebenfalls als normierende Rahmentexte und regeln das Handeln der Verwaltungsangestellten. (3) Die „verwaltungsinternen Textsorten“ sind diejenigen Textklassen, die innerhalb der Institution verfasst werden: (i) „Textsorten des internen Verwaltungsverkehrs“, z. B. *Berichte*, *Protokolle*, *Verwaltungsgliederungsgepläne*, *Mitteilungen*, *Erklärungen*, *Abrechnungen*, interne *Anfragen* und *Anträge*; (ii) „Textsorten, die sich an Angehörige der Institution (aber nicht an die Agenten der Verwaltung selbst) wenden“, z. B. *Bekanntmachungen*, *Bescheinigungen*, *Rechnungen*, *Bescheide* (mit zahlreichen Subtypen), *Urkunden*, *Zertifikate*; (iii) „Textsorten, die – von Außenstehenden – an die Verwaltung gerichtet werden“, z. B. *Anträge*, *Anfragen*, *Eingaben*, *Widersprüche*, *Erklärungen* (eventuell mit *Anlagen*). Zu dieser Klasse gehören nach Heinemann auch die von Studierenden, Mitarbeitern und Nichtuniversitätsangehörigen ausgefüllten *Formulare*. Abschließend nennt Heinemann die *Verwaltungshandbücher*, die interne Angelegenheiten und Prozessabläufe festlegen und regulieren, nimmt jedoch keine Zuordnung zu einem der drei Bereiche vor.

[9] Die Darstellung von Heinemann ist sehr abstrakt und geht nicht auf unterschiedliche Benennungen (z. B. „Institut“ vs. „Seminar“) oder Gemeinsamkeiten der einzelnen Hochschulgesetze der Länder bezüglich des Aufbaus einer Universität ein. Dieser Themenkomplex ist für die vorliegende Arbeit von besonderem Interesse und wird in Abschnitt 13.4 (S. 586 ff.) genauer thematisiert.

[10] Die Bezeichnungen der Bereiche und Textsorten werden zitiert nach Heinemann (2000b, S. 708).

6.2.4 Fazit

Abbildung 6.1 stellt – ausgehend von Heinemann (2000b) – eine zusammenfassende (und keineswegs vollständige) Typologie der Textsorten in der Hochschule dar, die um einige bei Heinemann fehlende Textsorten ergänzt wurde.[11] Der gesamte Bereich der digitalen Medien wird dabei nicht berücksichtigt. Ebenfalls nicht dargestellt werden die einzelnen Textsorten der drei Subkategorien der Verwaltungstextsorten.

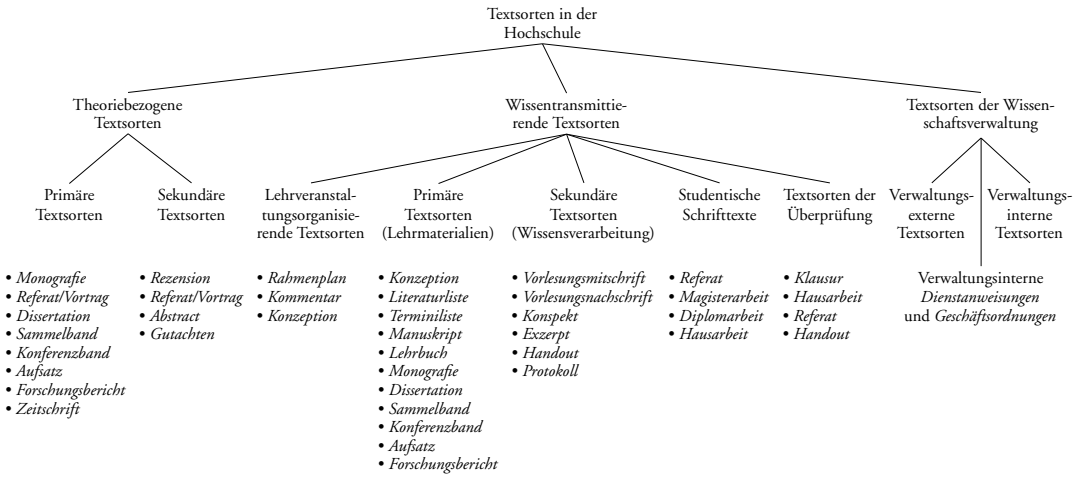


Abbildung 6.1: Textsorten im Kommunikationsbereich Hochschule und Wissenschaft (nach Heinemann, 2000b)

Die theoriebezogenen Textsorten werden wie bei Heinemann in primäre und sekundäre Textsorten aufgeteilt. Das *wissenschaftliche Gutachten* befindet sich nicht im Grenzbereich zu den wissentransmittierenden Textsorten, sondern wird eindeutig den sekundären Textsorten zugeordnet (vgl. die Definition von Göpferich, 1995, in Fußnote 7, S. 303). Mehrfachzuordnungen werden zugelassen, so ist es z. B. einerseits möglich, dass ein *Referat/Vortrag* auf einem vorhandenen und möglicherweise bereits publizierten (vgl. Fußnote 6) Artikel beruht und somit den sekundären Textsorten zuzuordnen ist. Andererseits kann ein Vortrag auf einem unpublizierten Entwurf basieren, in diesem Fall handelt es sich um eine primäre Textsorte. Im Bereich der wissentransmittierenden Textsorten werden ebenfalls primäre und sekundäre Textsorten angenommen: Textexemplare der primären Textsorten werden beinahe ausschließlich von Lehrenden (*Veranstaltungskonzeption*, *Literaturliste* etc.) bzw. wissenschaftlich Tätigen (*Hochschullehrbuch*, *Monografie*, *Sammelband* etc.) erstellt und fungieren als Lehrmaterialien. Textexemplare der sekundären Textsorten *Vorlesungsmitschrift*, *Konspekt*, *Exzerpt* etc. werden hingegen primär (vgl. jedoch Fußnote 6) von Studierenden zum Zweck der Wissensaufnahme und Wissensverarbeitung angefertigt. Eine neue Subkategorie, die auch den Verwaltungstextsorten zugeordnet werden kann, bündelt lehrveranstaltungsorganisierende Textsorten und umfasst unter anderem *Lehrveranstaltungskommentare*.

[11] Bei Heinemann fehlen z. B. *Lehrveranstaltungskommentar*, *Konferenzband*, *Sammelband*, *Handbuch*, *Overhead-Foliensatz*, *digital projizierter Foliensatz* und *Poster*.

6.3 Webangebote von Universitäten und Hochschulen

Die Webangebote von Hochschulen und Universitären sind bezüglich der Hypertextsorten, die sie beinhalten, hochgradig heterogener Natur.[12] Die im vorangegangenen Abschnitt wiedergegebene Überblicksdarstellung von Heinemann (2000b) ist auf den Umgang mit Texten innerhalb des regulären Hochschulbetriebs ausgerichtet und kann daher zwangsläufig in Bezug auf das WWW nur eine sehr eingeschränkte Gültigkeit besitzen. Die in universitären Websites verwendeten Hypertextsorten können ebenfalls nicht durch eine pure Reduktion der von Heinemann präsentierten Klassen auf diejenigen Textsorten, die für die Rezipienten eines Webangebots nützlich sein könnten, in Kombination mit den in Kapitel 4 dargestellten Hypertextsorten charakterisiert werden. Bevor diese Problematik im Folgenden detailliert betrachtet wird, müssen zunächst verschiedene Kernaspekte herausgearbeitet werden, die für nahezu alle Webangebote von Universitäten gelten.[13]

Wie heterogen die Websites von Universitäten tatsächlich sind, verdeutlicht Siegel (1999b, S. 9) anhand einer Querschnittsdarstellung, die jedoch auch nur einen Teil des Spektrums reflektiert: Physiker wollen mit anderen Physikern in Kontakt treten, um Forschungsergebnisse und Daten auszutauschen. Theaterwissenschaftler möchten das WWW einsetzen, um virtuelle Spielorte einzurichten, wohingegen Universitätsverwaltungen digitale Formulare für die verschiedensten Zwecke anbieten möchten. Über den Campus können zahlreiche Webcams verstreut sein, die einer weltweiten Öffentlichkeit in Echtzeit Impressionen zur Verfügung stellen. Lehrveranstaltungen können als Präsenzlehre durchgeführt werden, und parallel wird eine Videoaufnahme des Dozenten angefertigt, die über das WWW verfolgt werden kann und zu einem späteren Zeitpunkt für Prüfungsvorbereitungen in einem Archiv zugänglich sein wird. Bibliotheken realisieren komfortable Schnittstellen, mit deren Hilfe Kataloge online recherchiert und Bücher vorgemerkt werden können. Authentifizierungsmechanismen erlauben den Angehörigen einer Universität den Zugriff auf digitale Zeitschriftenartikel. Die Mensa bietet den Speiseplan und die Öffnungszeiten ebenfalls online an. Auch Studierende können mit dem WWW vielfältige Interessen verfolgen. Falls Lehrveranstaltungen aus Krankheitsgründen nicht besucht werden können, sollte das gesamte Material online zur Verfügung stehen, Diskussionsforen sollen Gruppenarbeiten ermöglichen. In einem Praktikum entwickelte Roboter sollen über eine WWW-Schnittstelle interessierten Personen vorgeführt werden können. Zugleich plädieren die im Rechenzentrum beschäftigten Techniker und In-

[12] Im Gegensatz zu unterschiedlichen Typen kommerzieller Websites existieren zu universitären Webangeboten kaum Vorarbeiten. Der einzige Bereich, der sehr umfassend untersucht wurde, sind die Webauftritte von Universitätsbibliotheken, insbesondere im Kontext der veränderten Rahmenbedingungen, die sich für digitale Bibliotheken ergeben (vgl. z. B. Endres und Fellner, 2000, Kengeri et al., 1999, und McGillis und Toms, 2001, sowie die dort genannten Literaturverweise). Aus diesem Grund werden die Websites von Bibliotheken und von ihnen angebotene *digital libraries* im Folgenden nur am Rande diskutiert.

[13] Der Begriff des universitären Webangebots bezeichnet *alle* HTML-Dokumente, die sich auf Webservern innerhalb eines Hochschulnetzwerks befinden (z. B. `.uni-bonn.de` oder `.uni-hamburg.de`). Der Terminus umfasst somit unter anderem den Webauftritt der Universitätsleitung, die privaten Homepages von Studierenden, das Informationsangebot einer Mensa und die Websites einzelner Fachbereiche und Institute. Die Webauftritte aller größeren deutschen Hochschulen und Universitäten werden in technischer Hinsicht über das deutsche Forschungsnetz (DFN, vgl. `http://www.dfn.de`) betrieben, dessen Infrastruktur das Gigabit-Wissenschaftsnetz (WiN) darstellt. Das DFN wird vom 1984 gegründeten DFN-Verein (Verein zur Förderung eines Deutschen Forschungsnetzes e. V.) betrieben und dabei vom Bundesministerium für Bildung und Forschung unterstützt.

genieure, die den oder die Webserver betreuen, dafür, alle HTML-Dokumente barrierefrei zu gestalten, so dass z. B. sehbehinderte oder blinde Personen, die einen *screen reader* verwenden, um sich Webseiten vorlesen zu lassen, nicht aus der Gruppe der Rezipienten ausgeschlossen werden. Gleichzeitig liegt es der Verwaltung am Herzen, dass *alle* von einer Universität angebotenen Webseiten ein konsistentes und unverwechselbares Erscheinungsbild aufweisen und dass sie optimale Navigations- und Suchfunktionen umfassen.

6.3.1 Rezipienten- und Zielgruppen

Die Webangebote von Universitäten und Hochschulen werden von unterschiedlichen Personengruppen rezipiert und in gleicher Weise richten sich spezifische Angebote an bestimmte Zielgruppen. Zunächst kann zwischen internen und externen Benutzern differenziert werden. Zu den internen Benutzern gehören alle Angehörigen, d. h. das gesamte wissenschaftliche sowie technisch-administrative Personal und alle Studierenden. Middleton et al. (1999, S. 220) bezeichnen die internen Benutzer als „existing customers“ in einem „captive market“. Die Institution sollte dafür Sorge tragen, die jeweiligen Bedarfe bestmöglich zu unterstützen, indem der Arbeitsalltag durch einen effizienten Zugang zu Informationen und hilfreichen Dienstleistungen erleichtert wird (z. B. Telefonverzeichnisse, Formulare und Dokumente aller Art, Datenbanken, Bibliothekskataloge, Zeit- und Raumbelegungspläne, Skripte, Lehrmaterialien, Diskussionsforen etc.; vgl. auch Hinman, 2002). Die sehr heterogene Gruppe der externen Benutzer bildet Middleton et al. (1999, S. 220) zufolge einen „target market“, für den unter anderem Informationen über Studiengänge, Lehrveranstaltungen und Zulassungsvoraussetzungen, die Stadt, in der sich die Universität befindet, FAQ-Dokumente, Informationen zur Anreise sowie Übernachtungsmöglichkeiten, Pressemitteilungen, Kontaktinformationen, Stellenangebote, Formulare und Unterhaltungsangebote zur Verfügung gestellt werden sollten. Die Reputation einer Universität kann insbesondere durch nützliche generische oder einem bestimmten Fachgebiet zugehörige Ressourcen erhöht werden. Middleton et al. (1999, S. 221) unterscheiden acht Gruppen externer Benutzer[14] (vgl. auch Gullikson et al., 1999, S. 293): (i) Hierzu gehört zunächst die Gruppe von Studieninteressenten, die sich ihrerseits zusammensetzt aus Personen, die soeben ihre Schulausbildung abgeschlossen haben, das Studienfach und/oder die Universität wechseln wollen oder ein Auslandssemester an der Hochschule absolvieren möchten. (ii) Die Gruppe der zukünftigen Mitarbeiter kann unterteilt werden in diejenigen Personen, die sich vor der Entscheidung für eine Bewerbung um eine vakante Position, vor einem Vorstellungsgespräch oder unmittelbar vor dem Antritt einer Stelle über die Universität und ihre internen Strukturen informieren wollen. (iii) Externe Wissenschaftler sind insbesondere an Publikationen, Forschungsergebnissen, laufenden Projekten und eventuell einer Kontaktaufnahme mit an der Hochschule tätigen Wissenschaftlern interessiert. (iv) Repräsentanten aus der Privatwirtschaft interessieren sich für ähnliche Aspekte, jedoch vor dem Hintergrund des Wissens- und Technologietransfers, um z. B. wissenschaftliche Expertise in Form einer Studie für in der Planung befindliche Produkte oder Dienstleistungen einzuholen. (v) Journalisten benötigen einen effektiven Zugriff auf Pressemitteilungen und müssen in der Lage sein, Experten für tagesaktuelle und kontroverse

[14] Hinzu kommen diejenigen Benutzer, die ohne die spezielle Intention, ein universitäres Webangebot aufzusuchen, auf ein solches gelangen, z. B. über eine Suchmaschine oder einen Katalog.

Themen lokalisieren zu können. (vi) Der Gesetzgeber sowie Drittmittelgeber greifen auf universitäre Websites zu, um sich über aktuelle Entwicklungen, Forschungsschwerpunkte und finanziell von ihnen unterstützte Förderprogramme und Projekte zu informieren. Die beiden abschließend von Middleton et al. genannten Gruppen gelten für deutsche Hochschulen bislang nur eingeschränkt, hierzu gehören (vii) Alumni, die sich regelmäßig über Neuigkeiten informieren möchten sowie (viii) Gönner und Stifter, die an der Geschichte einer Hochschule interessiert sind und Kontaktinformationen benötigen, um – wie möglicherweise auch einige Alumni – die Institution oder einen Förderverein finanziell zu unterstützen. Es wird deutlich, dass diese acht Gruppen sehr unterschiedliche Informationsbedarfe besitzen, die ein universitäres Webangebot reflektieren sollte.[15] Weiterhin gehören individuelle Rezipienten nicht notwendigerweise ausschließlich einer dieser Kategorien an, so kann z. B. ein Studierender freiberuflich für die ortsansässige Tageszeitung tätig sein und somit als interner und externer Benutzer eingestuft werden. Außerdem ist zu beachten, dass Rezipienten im Laufe ihres Lebens unterschiedlichen Zielgruppen angehören (z. B. vom Schüler zum Studierenden zum wissenschaftlichen Mitarbeiter, vgl. auch Kraus, 2000, S. 4 f.).

6.3.2 Funktionen

Die mit kommerziellen Websites verbundenen Ziele sind gut dokumentiert. Während die Nutzung des WWW aus der Sicht eines Unternehmens bereits intuitiv erfassbar ist, liegen bei universitären Webangeboten aufgrund der Existenz mehrerer Zielgruppen zahlreiche potenzielle Funktionen vor, von denen einige bereits im vorangegangenen Abschnitt angerissen wurden. Middleton et al. (1999, S. 220) liefern drei Gründe, weshalb eine Hochschule eine Website betreiben sollte: Zunächst dient sie als Mittel der Kommunikation zwischen Personen und Gruppen in den Bereichen Forschung, Lehre und Verwaltung. Weiterhin realisiert sie den Zugriff auf nützliche Informationsangebote, Datenbanken und Verzeichnisse besser und effizienter als jedes andere Medium. Abschließend dient die Website der Repräsentation einer Universität im WWW, sie fungiert als eine Art permanente, flexible, tagesaktuelle und weltweit erreichbare Werbemaßnahme. Aus diesen Funktionen können Charakteristika der Informationen abgeleitet werden, die den Benutzern zur Verfügung gestellt werden sollten. Hierzu gehören Materialien, die im weitesten Sinne der Werbung dienen und sich an Studieninteressenten und zukünftige Mitarbeiter richten. Außerdem sind Informationen, Dienstleistungen und Ressourcen bereitzustellen, die für die Rezipienten einen Mehrwert besitzen und sie zum wiederholten Besuch einladen, wodurch eine Hochschule als innovative Institution positioniert und die interne sowie externe Kommunikation – auch zum Zwecke der Lehre – unterstützt werden kann. Die Bedarfe der Benutzer (maximal relevante, aktuelle und zugreifbare Inhalte) und der anbietenden Institution (Werbung, Repräsentation, Innovation, bestmögliche Abdeckung der Nutzerbedarfe) überlappen sich, wie Middleton et al. (1999, S. 225) anmerken, nur partiell, weshalb die Hochschule dafür Sorge tragen sollte, die Bedürfnisse der Anwender vollständig zu erfüllen, um die eigenen Ziele erreichen zu können. In diesem Zusammenhang sind die Gründe von Interesse, die die 24 von Gullikson

[15] Einige Hochschulen bieten auf ihren Einstiegsseiten Hyperlinks zu Überblicksdokumenten mit Informationen an, die für spezifische Zielgruppen zusammengestellt wurden. Derartige zielgruppenspezifische Navigationshilfen werden in Kapitel 11 diskutiert.

et al. (1999, S. 296) befragten Angehörigen einer kanadischen Hochschule für die Benutzung ihrer Website genannt haben: „To look for information about: Courses“ (66,7% aller Teilnehmer), „Library“ (45,8%), „Services“ (37,5%), „Professors“ (29,1%), „Other“ (25%, bestehend aus „job postings“, „off-campus housing“, „graduate program information“, „exam schedule“, „degree requirements“), „Societies“ und „Events“ (jeweils 12,5%) sowie „Regulations“ (8,3%). Poock und Lefond (2001, S. 16) weisen zudem darauf hin, dass Universitäten das WWW als Mittel zur Anwerbung von Studierenden nicht unterschätzen sollten.

6.3.3 Autoren und Produzenten

Der Kreis der Produzenten universitärer Webangebote ist ebenso heterogen wie die Gruppe ihrer Rezipienten und er unterliegt, wie bereits in Abschnitt 4.8 angedeutet wurde, einem Spannungsverhältnis zwischen Privatheit und Kommerzialität. Der Faktor der Privatheit bezieht sich auf die traditionellen Herstellungsbedingungen universitärer Webangebote, die beinahe ausnahmslos im Rahmen von Aufgaben der universitären Selbstverwaltung von Angehörigen einer Hochschule erstellt werden (Mitarbeitern von Rechenzentren, Pressestellen, Forschungsprojekten, Technikern, Hilfskräften etc.). Obwohl viele universitäre Webauftritte unter optischen und technischen Gesichtspunkten äußerst professionell gestaltet sind, so handelt es sich doch dabei in den meisten Fällen um eine Tätigkeit, die mit den eigentlichen Aufgabenfeldern und Kernkompetenzen der jeweiligen Produzenten nur sekundär zu tun hat und von ihnen nur teilweise oder gar nicht abgedeckt wird: Wenn beispielsweise Mitarbeiter eines Rechenzentrums vollständig für eine Website verantwortlich sind, so ist meist die technische Infrastruktur hervorragend ausgebaut, aber hinsichtlich der z. B. für die Kohärenz des Angebots wichtigen Faktoren Web- und Textdesign und im Bereich der Benutzerfreundlichkeit dürften Defizite zu verzeichnen sein.[16] Falls diese Aufgaben hingegen vollständig von Mitarbeitern einer Pressestelle oder eines Dezernats für Öffentlichkeitsarbeit durchgeführt werden, so bieten diese typischerweise professionell gestaltete Texte an, wobei jedoch technische Aspekte des Angebots und seine Benutzerfreundlichkeit in den Hintergrund treten. Seit der Anfertigung des Korpus haben sich die Webauftritte der deutschen Universitäten in dieser Hinsicht drastisch verbessert, die ehemals isolierte Anfertigung des Einstiegsbereichs durch eine einzelne Abteilung wird nur noch an wenigen Hochschulen praktiziert. Neben der Privatheit bezieht sich der Faktor der Kommerzialität auf die Tatsache, dass vielen deutschen Hochschulen von den jeweiligen Landesregierungen immer weniger Mittel zugeteilt werden, weshalb von den Präsidien und Rektoraten professionelle und konkurrenzfähige Websites

[16] Diesbezüglich ist die Anmerkung einer Versuchsperson von Relevanz, die in dem von Gullikson et al. (1999, S. 300) durchgeführten Usability-Test der Website einer kanadischen Universität die Aufgabe hatte, die Information zu finden, wie eine E-Mail-Adresse beantragt wird: „Some computer person wrote this page, it's very technical and uses a lot of jargon, words that if all you wanted to know was how can I get email and didn't know anything about Internet Services or usernames or anything like that.“ Gullikson et al. merken zu dieser Einschätzung an: „The above participant was not an anomaly. Even when the participant was familiar with the process, he/she still had difficulties navigating the menu structure.“ Cunliffe (2000, S. 296 f) schlägt ein informelles Modell zur Entwicklung benutzerfreundlicher Websites vor und charakterisiert – basierend auf verschiedenen Studien – die Gruppe der Autoren nichtkommerzieller Webangebote: Diese besitzen häufig keine Webdesign-, Web-Usability- oder HCI-Expertise, arbeiten in zu kleinen Teams und haben neben der Entwicklung einer Website noch weitere Aufgaben zu erledigen. Weiterhin stehen ihnen nur begrenzte oder gar keine finanziellen Mittel zur Verfügung und sie führen nur sehr selten Usability-Tests durch.

gefordert werden, um es z. B. Studieninteressenten und anderen Wissenschaftlern zu erleichtern, sich über bestimmte Studiengänge und die Strukturen einer Hochschule zu informieren. Nicht mehr zeitgemäße Websites könnten daher dem Ruf einer Universität schaden, was mittelbar auch finanzielle Auswirkungen haben kann.

Die heterogene Produzentengruppe wirkt sich auf die Qualität universitärer Webangebote aus. Neben Websites, die zwar eingerichtet, aber nicht aktualisiert oder wieder entfernt werden (z. B. eine Ankündigung für einen internen Workshop), existieren komplexe Anwendungen, über das WWW zugreifbare Technologiestudien, visuell sehr ansprechende Angebote, die jedoch Lücken bezüglich ihrer Inhalte und optisch nur rudimentär aufbereitete, aber qualitativ hochwertige themenspezifische Ressourcen aufweisen. Eine Beachtung der gesamten Bandbreite zu berücksichtigender Aspekte (Benutzerfreundlichkeit, Text- und Webdesign, Umsetzung von Hypertext- und HCI-Richtlinien etc.) ist abhängig von der Expertise der Produzenten, die, wie Poock und Lefond (2001, S. 17) anmerken, im universitären Bereich oftmals Autodidakten sind: „Unfortunately, those who develop these Web pages tend to be self-taught employees [. . .] or students who are given little supervision".[17] Somit haben die Produzenten dieser Websites nur selten klare Zielvorstellungen über das zu entwickelnde Angebot, denn „efforts to date have been built largely on enthusiasm and 'best guesses' as to what should be done." (Middleton et al., 1999, S. 219). Und da keine Qualitätskontrolle existiert und sich das *Peer Reviewing* von Webangeboten innerhalb einer Universität[18] (noch) nicht etablieren konnte, ist damit zu rechnen, dass sich die angesprochene Bandbreite somit auch in Zukunft nicht signifikant in Richtung qualitativ durchgehend hochwertiger Angebote ändern wird. Storrer (1999a, S. 40) weist auf eine zusätzliche Problematik hin:

> Inhalte und Struktur einer Site können umso besser geplant werden, je stärker ihre Verwaltung institutionalisiert ist und von einer begrenzten Gruppe von Personen kontrolliert wird. Die Site einer Online-Zeitung mit einer gut organisierten Online-Redaktion kann deshalb wesentlich einheitlicher gestaltet werden als [. . .] die Site einer großen Universität, in der es angesichts der großen Zahl von Institutionen, Gruppen und Einzelpersonen schwer ist, ein einheitliches Design durchzusetzen. (Storrer, 1999a, S. 40)

Nicht nur die inhaltliche Qualität der Angebote leidet unter den genannten Faktoren, sondern auch das für die Universitätsleitung so wesentliche Ziel der Präsentation einer einheitlichen Marke, die von einem visuell konsistenten und somit die globale Kohärenz positiv beeinflussenden Webdesign geprägt ist.

[17] Siegel (1999a, S. 34–39) stellt ein Projekt dar, in dem von einer Universität und einer Agentur ein Forschungsportal entwickelt wurde, wobei der immense konzeptionelle und technische Aufwand, der bei einer spezialisierten Website auch im universitären Kontext entstehen kann, sehr deutlich wird. In Deutschland ist damit zu rechnen, dass nur dann externe Agenturen mit der Entwicklung einer universitären Website beauftragt werden, wenn ausreichende Drittmittel vorliegen oder die Agentur von sich aus ein Referenzprojekt erstellen möchte. Meines Erachtens handelt es sich dabei – zumindest im deutschen Hochschul- und Wissenschaftsbereich – um Ausnahmeerscheinungen. Besonderes Augenmerk legen die Hochschulleitungen jedoch auf die Gestaltung der Einstiegsseiten, die mittlerweile häufig von externen Unternehmen angefertigt werden. Für die im Korpus verfügbaren Websites gilt dies nur in sehr eingeschränkter Weise (vgl. Kapitel 11, S. 461 ff.).

[18] Dieses könnte z. B. durch (anonyme) Annotationen oder Rankings erfolgen, die nur von den Rechnern von Universitätsangehörigen aus durchgeführt oder eingesehen werden können (z. B. auf der Grundlage des von Greenhill und Venkatesh, 1999, vorgeschlagenen Systems).

6.3.4 Strukturierung und Informationsarchitektur

Verschiedene Aspekte ziehen sich wie ein roter Faden durch die gesamte Literatur zu den Themen Webdesign und Web-Usability. Hierzu zählt der Ratschlag, die Struktur einer Website unter keinen Umständen auf Grundlage der internen Strukturierung oder der Organisationshierarchie derjenigen Institution zu entwerfen, die die Website anbietet (vgl. Abschnitt 6.2.3). Stattdessen sollte sich ihre Architektur an den Bedürfnissen der Benutzer und den von ihnen verfolgten Aufgaben und Prozessen orientieren (vgl. z. B. Nielsen, 1999, S. 15). Rosenfeld und Morville (1998, S. 25) weisen ausdrücklich auf diesen Umstand hin:

> The fact is that labeling and organization systems are intensely affected by their creators' perspectives. We see this at the corporate level with web sites organized according to internal divisions or org[anization] charts. In these web sites, we see groupings such as *marketing*, *sales*, *customer support*, *human resources*, and *information systems*. [...] To design usable organization systems, we need to escape from our own mental models of content labeling and organization. (Rosenfeld und Morville, 1998, S. 25)

Es wird argumentiert, dass Websites gerade *nicht* nach den Organisationseinheiten strukturiert sein sollten, weil diese nicht notwendigerweise mit den Bedürfnissen der Nutzer korrelieren. Eben dies trifft jedoch für universitäre Webangebote nur in den seltensten Fällen zu.[19] Zunächst ist die Gruppe der Rezipienten sehr heterogen, weshalb es extrem aufwändig ist, die korrespondierenden Bedarfe zu antizipieren (oder gar empirisch zu erheben) und diese in einer Terminologie, die bezüglich ihrer Verständlichkeit sämtliche Gruppen bestmöglich abdeckt (von Schülern bis zu Hochschullehrern), in einem attraktiven und verständlichen Webdesign anzuordnen. Die Gliederung in Organisationseinheiten, dies haben die durchgeführten Korpusanalysen ergeben (vgl. auch Kapitel 11, S. 461 ff.), stellt ein zentrales Grundgerüst beinahe jedes universitären Webauftritts dar, weil sie den Produzenten oftmals intuitiv nachvollziehbar erscheint und den Aufbau der Universität selbst reflektiert (vgl. auch Gullikson et al., 1999, S. 295 f., sowie Kraus, 2000, S. 12).

Zu den Aufgaben der universitären Selbstverwaltung gehört auch die Entwicklung und Pflege von Webangeboten. Auf den Ebenen einzelner Professuren, Institute, Seminare, Fachbereiche, Fakultäten und in den zentralen Einrichtungen (Verwaltung, Rechenzentrum etc.) sind somit jeweils Personen tätig, deren Aufgabenbereich unter anderem die initiale Erstellung sowie die Betreuung des Webangebots umfasst. Jede Organisationseinheit ist für die Pflege des eigenen Angebots verantwortlich und sollte im besten Fall dafür Sorge tragen, dass die Informationen vollständig, korrekt und aktuell sind und sich gleichzeitig an den möglicherweise vorhandenen globalen Richtlinien hinsichtlich Webdesign und Gestaltung

[19] Middleton et al. (1999, S. 222) sind ebenfalls der Ansicht, dass es nicht ausreichend ist, Hyperlinks zu den einzelnen Fakultäten oder Fachbereichen anzubieten, denn „these links may or may not contain information which the user wants, but there is no way of telling until the link has been explored. This approach also encourages a 'many isolated parts' model of development, in which individual departments take varying approaches to style, navigation and content, often duplicating effort and ultimately failing to address the users' needs which are not necessarily specific to a department." Boardman (2005, S. 24) berichtet zu den Hyperlinkanzeigern der Einstiegsseite eines universitären Webauftritts: „Many of the hyperlinks offered under each of the headings can be seen to correspond with actual physical locations within the university campus – [...] a default institutional page [...] functions very much like the reception area of a large building."

orientieren. Umgekehrt lassen die Prozesse der Selbstverwaltung nur wenig Spielraum, *alle* beteiligten Websites eines universitären Webauftritts einer zentralen Instanz zu unterstellen, es sei denn, für diese Aufgaben werden Mitarbeiter eingestellt, die eine redaktionelle und gestalterische Kontrolle ausüben, indem ein *Content Management System* eingeführt wird, so dass die Eingabe von Inhalten auch weiterhin durch die dezentralen Einheiten selbst erfolgen kann.[20] Zusammenfassend kann festgehalten werden, dass sich die Struktur einer Universität als einzelne, hierarchisch gestufte, zentrale und dezentrale Organisationseinheiten unmittelbar in der Architektur ihres Webauftitts widerspiegelt, wobei die Webangebote einzelner Einheiten in der Regel den Status einer in sich abgeschlossenen Website innerhalb des jeweils übergreifenden funktionalen Ganzen besitzen (vgl. Gillenson et al., 2000). Während der Autor einer derartigen Website im Regelfall ein Angehöriger der jeweiligen Einheit ist und im Auftrag des korrespondierenden Leitungsgremiums handelt, fungiert die Organisationseinheit als abstrakter Produzent und stellt sich selbst dar, d. h. auch in den einzelnen Websites eines universitären Webangebots gilt die von Schütte (2004a, S. 211) für Unternehmenshomepages aufgestellte Gleichung „Emittent = Thema".

6.3.5 Ratschläge zur Gestaltung

Die vorangegangenen Abschnitte haben bereits Teilaspekte der Gestaltung universitärer Webangebote thematisiert. Nach Ansicht von Gullikson et al. (1999, S. 293) kann eine nicht intuitiv nachvollziehbare Informationsarchitektur oder ein wenig lesefreundliches Webdesign zu verschiedensten Problemen führen. So könnten sich z. B. Studieninteressenten, deren Informationsbedarf von dem Webangebot einer Universität beim ersten Besuch nicht abgedeckt wird, stattdessen auf den Websites anderer Hochschulen informieren.[21] Auch die Effektivität des Angebots im Hinblick auf die internen Benutzer kann durch Mängel in puncto Web- und Navigationsdesign beeinträchtigt werden. Im Übrigen projiziert ein mangelhaftes Webdesign ein schlechtes Image in Bezug auf potenzielle Mitarbeiter und mögliche Sponsoren.

Middleton et al. (1999) geben Ratschläge zum Aufbau eines benutzerfreundlichen universitären Webangebots: Die Bedarfe der verschiedenen Benutzergruppen sollten bestmöglich unterstützt werden (das Paradigma des *User-Centred Design* ist *der* essenzielle Bestandteil aller aktuellen Webdesign- und Usability-Ratgeber, vgl. Abels et al., 1997, 1998). Alle Informationen, von denen sich die Universitätsverwaltung oder -leitung wünscht, dass die internen Benutzer von ihnen Kenntnis besitzen, sollten online zugänglich sein. Von wesentlicher Bedeutung ist nach Auffassung von Middleton et al. die saliente Präsentation aktueller Neuigkeiten in der Einstiegsseite des Webangebots, die somit aus einem statischen (den Navigationshilfen und weiteren peritextuellen Komponenten) und einem dynamischen Teil (den fortwährend aktualisierten Neuigkeiten) bestehen sollte. Es ist eine zentrale redaktionelle Kontrollinstanz einzurichten, die dafür Sorge trägt, dass die Dokumente der obersten

[20] Middleton et al. (1999, S. 225) sprechen sich für die Einführung von „editorial and administrative policies" aus, die für das gesamte Webangebot einer Universität gelten sollen, „to ensure control is exercised across the university, without stifling creativity."

[21] Dieser Umstand kann verschiedene Ursachen haben, z. B. die nicht gegebene Verständlichkeit der angebotenen Informationen, Probleme bezüglich der Auffindung relevanter Informationen, proprietäre WWW-Technologien, die der von einem Interessenten verwendete Browser nicht darstellen kann oder aufgrund mangelnden Kontrastes zwischen Text- und Hintergrundfarbe nur schlecht lesbare Texte.

Hierarchieebenen konsistent gestaltet und strukturiert sind. Weiterhin sollte das Webangebot eine separate Einstiegsseite für die internen Benutzer umfassen.[22]

Die internen Benutzer benötigen im Arbeitsalltag keine Materialien, die primär werbende Funktion besitzen. Weiterhin kann in diesem Bereich Middleton et al. zufolge weniger Aufwand hinsichtlich der grafischen Aufbereitung betrieben werden, da die Funktionalität der Website im Vordergrund steht. Sie sollte digitale Formulare beinhalten, um bürokratische Vorgänge zu erleichtern (z. B. Reiseanzeigen und -abrechnungen, Einstellungsformulare etc.). Da derartige Informationen für externe Benutzer keine Relevanz besitzen, sollte ein Zugangsschutz eingerichtet werden. Die Informationen für externe Benutzer sollten es gewährleisten, dass sie sich möglichst effizient ein Bild von der Universität machen können, um sie – im Falle von Lesern mit wirtschaftlichen Interessen – zu einer Zusammenarbeit mit der Institution anzuregen oder – bei Studieninteressenten – für ein Studium an dieser Universität zu begeistern. Besonderes Augenmerk legen Middleton et al. auf eine flache Informationsarchitektur, da die Konzeptualisierung, die Webdesigner einer tiefen Strukturierung zugrunde legen, nicht notwendigerweise derjenigen der Benutzer entspricht. Die Leser sollten so wenig Entscheidungen wie möglich treffen müssen, um zu einer spezifischen Information zu gelangen. Falls keine flache Hierarchie realisiert werden kann, sollten sich die einzelnen Bereiche der Website auf funktionale und nicht auf inhaltliche Kategorien stützen. Die Etiketten einzelner Rubriken und die auf sie verweisenden Hyperlinkanzeiger sollten so benannt sein, dass sie intuitiv interpretiert werden können.

6.3.6 Zur Benutzerfreundlichkeit

Dass die realen Gegebenheiten im WWW die im vorangegangenen Abschnitt dargestellten Gestaltungsratschläge nur partiell reflektieren, zeigen verschiedene Studien, in denen die Benutzerfreundlichkeit universitärer Webangebote analysiert wurde. Gullikson et al. (1999) untersuchen den Einfluss der Informationsarchitektur (mit Bezug auf Rosenfeld und Morville, 1998) und des Designs der Website der Dalhousie University (Halifax, Nova Scotia, Kanada, `http://www.dal.ca`) auf das Navigationsverhalten von 24 Probanden (Studierende aus dem Grund- und Hauptstudium sowie wissenschaftliche Mitarbeiter und Professoren).[23] Im

[22] Während Middleton et al. (1999, S. 221) empfehlen, unterschiedliche Einstiegsseiten für interne und externe Benutzer zu pflegen, sprechen sie sich *gegen* Überblicksdokumente aus, die Informationen enthalten, die für bestimmte Zielgruppen relevant sein könnten und auch dazu dienen, die oftmals sehr umfangreichen universitären Angebote auf die für einen bestimmten Kreis von Rezipienten wesentlichen Kerninhalte zu reduzieren (vgl. Fußnote 15): „[I]t is wrong for designers to narrowly categorise users – as is often the case“ (Middleton et al., 1999, S. 222). Diese Begründung erscheint zumindest dann nicht plausibel, wenn ausreichende Ressourcen zur Erstellung und Pflege der zielgruppenspezifischen Überblicksseiten bereit gestellt werden können. Poock und Lefond (2001, S. 19) sind der Ansicht, dass derartige Dokumente eine hilfreiche Zusatzfunktion besitzen können: „home pages that have links grouped by “prospective students”, “current students” etc., tend to greatly enhance the architecture as compared to groupings by function category [. . .].“ Außerdem kann auf diese Weise eine Terminologie verwendet werden, mit der die Mitglieder der Zielgruppe vertraut sind.

[23] Gullikson et al. (1999) führen einen klassischen Usability-Test durch, in dem die Probanden einzeln bei der Durchführung vorgegebener Aufgaben beobachtet werden (in diesem Fall z. B. „Can a student be expelled from Dalhousie for committing plagiarism?“, „Where does one go to activate an email account?“ oder „Who is the chair of Dalhousie's theatre programme?“). Ein Vorfragebogen dient zur Ermittlung demografischer Daten, in einem Nachfragebogen und Interview schildern die Versuchspersonen ihren Eindruck des Webangebots.

November 1998 enthielt die Einstiegsseite dieser Universität fünf Menüpunkte („About Dalhousie“, „Academics“, „Departments“, „Campus Life“, „Library Services“), die von knappen Erläuterungen flankiert waren und auf weitere Zugriffs- bzw. Navigationsseiten verwiesen. Von den sechs gestellten Aufgaben konnten die Probanden im Schnitt 3,7 beantworten, wofür durchschnittlich 88,3 Sekunden benötigt wurden. Bei wenig intuitiv erscheinenden Auswahloptionen verbrachten sie bis zu zwei Minuten auf den Menüseiten der zweiten Ebene, was auf Navigationsschwierigkeiten aufgrund einer unklaren Benennung der Hyperlinkanzeiger hindeutet.[24] Es können weitere Probleme auf wenig intuitive Linkanzeiger zurückgeführt werden, die den Zielknoten nicht präzise umschreiben, so berichten die Teilnehmer, dass bereits auf der Einstiegsseite unklar sei, worin sich die Menüpunkte „Academics“ und „Departments“ unterscheiden. Ein zweiter Kritikpunkt betrifft die Strukturierung und Kategorisierung des Angebots, die trotz einer kohärenten visuellen Gestaltung der oberen Ebenen zu Verwirrungen und dem „lost in hyperspace“-Effekt geführt hat (vgl. Abschnitt 3.4.2). Zu Beginn einer Aufgabe gingen die Teilnehmer systematisch vor und verfolgten einen spezifischen Pfad, indem sie von Menü zu Menü navigierten. Sobald jedoch der initiale Pfad fehlschlug, gingen sie zunehmend unlogischer vor und wählten beliebige Menüpunkte aus, d. h. die angebotenen Navigationshilfen boten bei der Suche keine Unterstützung. Gullikson et al. (1999, S. 300) bezeichnen die Leistungen der Teilnehmer bei der Auffindung spezifischer Informationen auf der Grundlage einfacher und im akademischen Bereich typischer Fragen insgesamt als „poor“. Die Probanden selbst gaben der Website eine Bewertung von 55%, was – basierend auf den Standards dieser Universität – der Note D entspricht. Die abschließende Beobachtung reflektiert den Status Quo der Webangebote vieler deutscher Hochschulen:

> Not unlike many other academic web sites, this site contains the pages from a group of loosely organised units – a set of academic silos – united under a common umbrella. While this describes the organisational culture of many universities, the result is more likely associated with the need to exert autonomy and academic freedom, and establish corporate identity within an individual unit or faculty, rather than servicing the needs of the organisation's client base. Participants on the other hand see the site as *the university* – as an integrated whole – a tightly coupled grouping of informational units organised to suit their particular needs. (Gullikson et al., 1999, S. 300)

Zu ähnlichen Ergebnissen gelangen Poock und Lefond (2001): 55 Schüler von High-Schools in North Carolina und Michigan wurden gebeten, Websites von Hochschulen zu beurteilen und Aufgaben durchzuführen. Den Inhalt einer universitären Website bezeichnen 97% aller Probanden als wichtigen oder sehr wichtigen Bestandteil. Die Architektur wird von 95% als wichtig oder sehr wichtig bewertet – die effektivsten Websites seien diejenigen, die ein nach Zielgruppen differenziertes Angebot enthielten (Poock und Lefond, 2001, S. 18; vgl. Fußnote 15, S. 308) und „visually intuitive“ seien. Die für andere Zielgruppen (Alumni, Mitarbeiter etc.) intendierten Informationen wurden als überflüssig betrachtet. Nach Einschätzung von Poock und Lefond waren die meisten Probanden sehr versiert im Umgang mit

[24] Sowohl die erfolgreichen als auch die weniger erfolgreichen Probanden schätzen die Architektur der Website als nicht intuitiv ein und sind mit dem Angebot unzufrieden (Gullikson et al., 1999, S. 298). Diese Einschätzungen sind nicht abhängig von der Webkompetenz der Probanden, Gullikson et al. geben jedoch an, dass die Probanden, die häufig auf diese Website zugreifen, sie als insgesamt hochwertiger einschätzen.

dem WWW, was sich in der häufig geäußerten Meinung widerspiegelt, dass eine unprofessionelle Website ein Indikator für eine qualitativ minderwertige Institution sei.

6.3.7 Universitäre Webangebote aus Sicht der Marktforschung

Kamenz et al. (1998) nehmen eine Analyse universitärer Webangebote auf der Grundlage von Bewertungskriterien vor, die aus der Marktforschung stammen.[25] Es geht den Verfassern unter anderem um die „Erarbeitung aller internettypischen und hochschulbezogenen Merkmale von Internetseiten“, die „vollständige Darstellung und Bewertung aller deutschsprachigen Internet-Angebote der in Deutschland registrierten Hochschulen“ und ein „Ranking aller Auftritte“ (ebd., S. 7).[26] Zur Charakterisierung und Bewertung wurde ein Katalog von 83 Kriterien erstellt, die den vier Gruppen „Inhalte“ (40 von 100 möglichen Punkten), „Layout“, „Handling“, und „Interaktivität“ (jeweils 20 Punkte) zugeordnet wurden (vgl. Kamenz et al., 1998, S. 316–318).[27] Insgesamt wird ein durchschnittliches Ergebnis von 46 Punkten ermittelt, was vor dem Hintergrund, dass im Schnitt lediglich vier Punkte auf die Gruppe „Interaktivität“ entfallen, als „nicht zufriedenstellend“ eingestuft wird (ebd., S. 9).[28] Die „enormen Defizite bei den Internet-Präsentationen deutscher Hochschulen“ (ebd., S. 11) machen sich Kamenz et al. zufolge insbesondere in den Bereichen verfügbares Sprachangebot, Vorstellung der Mitarbeiter durch Fotos, Vorlesungsverzeichnis, Suchfunktion, Auslandsstudien, akademisches Auslandsamt, BAföG-Zuweisungen, Stipendiengeber und Telefonverzeichnis bemerkbar, es liegt also eine sehr große Streuung in Bezug auf die Ausprägung der vier Kategorien vor.[29] Aus diesem Grund fordern Kamenz et al. (1998, S. 13) „Customizing“, worunter „Kundenorientierung und Kreativität“ verstanden wird. Hierdurch seien die „aufgezeigten Mängel [...] leicht und mit überschaubaren Mitteln zu beheben.“[30]

[25] Im Jahr 2003 wurde eine Folgestudie publiziert, die jedoch bis zur Fertigstellung der vorliegenden Arbeit nicht bezogen werden konnte.

[26] Kamenz et al. (1998) geben an, dass die Webangebote von 257 Universitäten und Hochschulen untersucht wurden. Es werden jedoch keine Angaben zum Umfang oder zur Anzahl der jeweils analysierten HTML-Dokumente gemacht. Da sich 39 Kriterien auf inhaltliche Aspekte wie z. B. „Öffnungszeiten von Hochschuleinrichtungen“ oder „grafisches Organigramm“ beziehen, könnte daher der Umstand existieren, dass das Vorhandensein eines Kriteriums in einem spezifischen Webauftritt von Kamenz et al. nicht bestätigt werden konnte, weil das entsprechende HTML-Dokument innerhalb des zentralen Webangebots nicht referenziert wird.

[27] Die Kriterien sind in Tabelle 6.1 dargestellt und wurden Kamenz et al. (1998, S. 9 f., sowie S. 316–318) entnommen; es werden nicht für alle Kriterien Frequenzen angegeben. Bezüglich dieses Katalogs sind einerseits verschiedene Zuordnungen zu den vier Gruppen sowie die Geltungsbereiche einiger Kriterien unklar, so wird z. B. nicht deutlich, weshalb „Besucherzähler“ der Gruppe „Handling“ und „Unterhaltung, Spiele“ dem Bereich „Inhalt“ (anstatt „Interaktivität“) zugeordnet wurden. Nicht thematisiert wird, auf welche Weise die Ausprägungen von Kriterien wie z. B. „Links – sonstige“ (auch: „sonstige hochschulbezogene Links“), „Foto der Mitarbeiter, Dozenten“ und „Links zu anderen thematischen Seiten“ bestimmt werden, d. h. auf welcher Datengrundlage ein derartiges Kriterium als „existent“ vs. „nicht existent“ eingestuft wird, schließlich können viele der Kriterien prinzipiell von beliebigen Dokumenten eines universitären Webauftritts instanziiert werden.

[28] Die mit 76 Punkten beste Bewertung entfällt auf den Webauftritt der FernUniversität Hagen.

[29] Teil III der vorliegenden Arbeit wird zeigen, dass die von Kamenz et al. als defizitär empfundenen Aspekte in Bezug auf die analysierten Stichproben nur eine eingeschränkte Gültigkeit besitzen.

[30] Interessanterweise gehen Kamenz et al. (1998) nicht auf den typischen Erstellungs- und Pflegeprozess der Webauftritte von Hochschulen ein, der – zumindest im Jahr 1998 – ausschließlich auf den Rahmen der akademischen Selbstverwaltung beschränkt gewesen sein dürfte, wodurch sich zwangsläufig die verschiedensten Restriktionen ergeben, die ein „leichtes Beheben“ der „aufgezeigten Mängel“ verhindern.

Layout	1. Nutzung bewegter Elemente (37%), 2. Nutzung von Musikeinspielungen (4,6%), 3. Nutzung von Videofilmen (3,1%), 4. Nutzung von Fotos, 5. Nutzung von Grafiken, 6. Nutzung von Texten, 7. Nutzung der eigenen CI, 8. CI der Webseiten („Side identity") und der Navigation, 9. Textliche Gestaltung, 10. Ästhetische Gestaltung
Handling	1. Eigene Domain (95%), 2. Suchfunktion (interne), Index (externe) (52%), 3. Besucherzähler (15%), 4. Hilfefunktion (12%), 5. Link zur eigenen Homepage (12%), 6. Anzahl der Seiten mit der Information „not found", „under construction" (7,7%), 7. Größe der Homepage (in KB), 8. Nutzung der Frametechnik, 9. Vor- und Rücksprung-Icons, 10. Navigation über Icons, 11. Navigation über Text, 12. Sicherheitshinweise bei risikobehafteten Anwendungsmöglichkeiten, 13. Browser-Kompatibilität (Netscape, Microsoft), 14. Vergrößerung von Fotos
Inhalte	1. Angabe der Adresse (100%), 2. Übersicht über Lehrstühle, Fachbereiche (99%), 3. Informationen über die Hochschule („Wir über uns", Geschichte) (94%), 4. Links – sonstige (z. B. regionale Seiten) (73%), 5. Präsentation in englischer Sprache (71%), 6. Telefonverzeichnis und/oder e-mail-Verzeichnis (Mitarbeiter, Dozenten) (62%), 7. Wegbeschreibung (60%), 8. Informationen über das Akademische Auslandsamt (59%), 9. Informationen über Präsidium, Direktorium, Rektor, Kanzler (58%), 10. Informationen über Forschungsschwerpunkte, -berichte (56%), 11. Übersicht über Weiterbildung, Aufbaustudien (54%), 12. Informationen über das BAföG, Stipendien, Förderprogramme (49%), 13. Fotos der Hochschule (49%), 14. Campus-, Gebäudeplan (47%), 15. Links zu Partnerhochschulen (46%), 16. Informationen für die Presse, Pressemitteilungen (46%), 17. Informationen über die Transferstelle (40%), 18. Foto der Mitarbeiter, Dozenten (36%), 19. Vorlesungsverzeichnis (33%), 20. Datenbanken, Statistiken, Rechner (32%), 21. Speiseplan der Mensa (27%), 22. Organigramm (8,1%), 23. Verwendung von Downloads (5%), 24. Einsatz einer WebCamera (3,1%), 25. Aktualität (jünger als 1 Monat), 26. Angabe des Update-Datums, 27. Links zu Partnern (z. B. Institute, Unternehmen), 28. Links zu anderen thematischen Seiten, 29. Verwendung von Plug-Ins, 30. Anzahl der wählbaren Fremdsprachen (ohne Englisch), 31. Kennzeichnung neuer Seiten, 32. Sonstiges Dienstleistungsangebot (z. B. Schulungen), 33. Unterhaltung, Spiele, 34. Öffnungszeiten von Hochschuleinrichtungen, 35. Aktuelle Informationen (z. B. Veranstaltungsänderungen), 36. Informationen über Hochschuleinrichtungen, 37. Informationen – sonstige, 38. Informationen der Presse, Pressestimmen, 39. Informationen über die Hochschulzeitung
Interaktivität	1. Jobangebote (46%), 2. Intranet, Extranet (17%), 3. Bestellmöglichkeit von Prospekten (8,5%), 4. Frage- und Antwortsammlungen, sog. „FAQ" (7,7%), 5. Eintragungsmöglichkeit in ein Gästebuch (7,3%), 6. Newsgroups, Foren, Mailinglisten (6,9%), 7. Zielgruppenbefragung (Marktforschung) (4,2%), 8. Telefon-Hotline (1,5%), 9. Kommunikationsmöglichkeit in einem „Chatroom" (1,2%), 10. Setzen eines Cookies zur Kundenidentifikation (0,4%), 11. Push-Technik (0,4%), 12. Webphone (0%), 13. Internet-Aktionen (z. B. Ratespiele, Wettbewerbe), 14. Hochschulaktionen außerhalb des Internet (z. B. Sonderveranstaltungen, Workshops), 15. Kontaktmöglichkeit über e-mail – allgemeines Eingabefenster, 16. Kontaktmöglichkeit über e-mail – vorbereitetes Eingabeformular, 17. Online-Katalog-Abfrage der Bibliotheken (Bestellung, Verlängerung Reservierung), 18. Interaktive Online-Anträge (z. B. Einschreibung), 19. Shopping-Angebote über sonstige Dienstleistungen

Tabelle 6.1: Der von Kamenz et al. (1998) eingesetzte Kriterienkatalog

6.3.8 Traditionelle Textsorten in universitären Webangeboten

Bei einem Vergleich der Textsorten des Kommunikationsbereichs Hochschule und Wissenschaft (vgl. Abschnitt 6.2) und der wesentlichen Merkmale universitärer Webangebote in Verbindung mit den in Kapitel 4 vorgestellten Hypertextsorten fallen verschiedene Aspekte auf: Heinemann beschränkt ihre Darstellung auf theoriebezogene und wissenstransmittierende Textsorten sowie Textsorten der Hochschulverwaltung. Eine der zentralen Funktionen des Webangebots einer Hochschule betrifft jedoch die Werbung im weitesten Sinne (vgl. Raisman, 2003). Diese Funktion wird im Bereich der Printmedien, die von der Institution selbst veröffentlicht werden, z. B. von Hochglanzprospekten, Forschungsberichten (von Organisationseinheiten wie z. B. einem Zentrum, einer Professur oder der gesamten Universität), Faltblättern und Flyern sowie Studiengangsbroschüren und Studienführern, aber auch Lebensläufen und Publikationslisten (eines Wissenschaftlers) erfüllt. Zu diesem Bereich gehören auch Texte, die Veranstaltungen der gesamten Hochschule oder einzelner Organisationseinheiten ankündigen (öffentliche Ringvorlesungen und Kolloquien, Konzerte und Aufführungen, Tagungen, Kongresse etc.), z. B. Plakate und Poster, Presseanzeigen, Handzettel und Zeitungs- sowie Zeitschriftenartikel.

Bezüglich der Darstellung Heinemanns stellt sich die Frage, welche dieser Textsorten in universitären Webangeboten zu finden sind. Hierzu gehören zunächst fast alle schriftlich realisierten theoriebezogenen Textsorten, so werden z. B. Dissertationen in eigens eingerichteten digitalen Bibliotheken veröffentlicht. Viele wissenschaftliche Zeitschriften sind ebenfalls im WWW verfügbar, befinden sich jedoch meist auf den Webservern der Verlagshäuser. Einzelne in Zeitschriften oder Sammelbänden erschienene oder noch erscheinende Artikel befinden sich in der Regel – entweder als Entwurfs- oder endgültige Version – auf den

persönlichen Homepages von Wissenschaftlern oder den übergreifenden Publikationslisten von Arbeitsgruppen oder Instituten. Konferenzbeiträge oder vollständige Proceedings-Bände werden ebenfalls häufig in dieser Form veröffentlicht. Die einzige theoriebezogene Textsorte, die nicht im WWW angeboten wird, ist das Gutachten (z. B. einer Qualifikationsarbeit oder Anmerkungen anonymer Gutachter zu einem bei einer Zeitschrift oder einer Konferenz eingereichten Beitrag). Auch im Hinblick auf die wissenstransmittierenden Textsorten finden verschiedene Textklassen im WWW keine Verwendung. Hierzu gehören zunächst bezüglich der lehrveranstaltungsorganisierenden Textsorten übergreifende Rahmenpläne und Konzeptionen. Lehrveranstaltungskommentare werden hingegen durchaus auf den persönlichen Homepages von Dozenten, übergreifenden Angeboten eines Instituts oder Fachbereichs oder den zentralen Webangeboten oder E-Learning-Plattformen der Universität angeboten. Die primären Textsorten (Lehrmaterialien) sind ebenfalls in vielen Fällen online verfügbar, gelegentlich existiert eine Zugangsbeschränkung, um z. B. die PDF-Datei eines Zeitschriftenbeitrags oder sogar einer Monografie lediglich einer Gruppe von Studierenden, denen Kennung und Passwort mitgeteilt wurde, oder innerhalb des Netzwerks einer Universität zum Download anzubieten. Textexemplare der sekundären Textsorten, die der Wissensverarbeitung dienen, werden vornehmlich von Studierenden erstellt. Ob sie im WWW veröffentlicht werden, kann von verschiedenen Faktoren abhängig sein, eine Vorlesungsmitschrift kann z. B. unmittelbar auf dem Laptop angefertigt und binnen Minuten per drahtlosem Netzwerk auf der privaten Homepage oder im *Content Management System* der Fachschaft publiziert werden. Dies kann auch für Konspekte oder Exzerpte gelten, die von Diplomanden oder Doktoranden im Laufe der Anfertigung einer Qualifikationsarbeit erstellt werden und z. B. im WWW in gesammelter Form einen Überblick über ein spezifisches Themengebiet geben können. Handouts sind aufgrund ihres fragmentarischen Charakters ohne den zugehörigen Vortrag oder das Referat kaum verständlich, ihr Nutzen für eine weltweite Öffentlichkeit ist somit eingeschränkt, was ihre eher geringe Verbreitung im WWW erklärt. Einige Lehrende publizieren digitale Handouts auf lehrveranstaltungsbegleitenden Seiten, so dass sie zur Nachbereitung einer Sitzung und Vorbereitung einer Prüfung eingesetzt werden können. Protokolle enthalten häufig sensitive Daten, weshalb sie üblicherweise lediglich einem eingeschränkten Kreis von Rezipienten zugänglich gemacht werden. Falls sie im WWW publiziert werden, erfolgt der Zugang meist passwortgeschützt. Die Publikation studentischer Schrifttexte erfolgt im Falle von Diplom- oder Magisterarbeiten in aller Regel auf Initiative der Verfasser, die ihre Arbeit auf der privaten Homepage zur Verfügung stellen. Dies kann in gleicher Weise für Hausarbeiten, schriftliche Versionen von Referaten oder Foliensätze gelten, die darüber hinaus auch gelegentlich in das Webangebot einer Lehrveranstaltung integriert werden. Die Textsorten der Überprüfung wurden bereits teilweise angesprochen. Während von Prüflingen angefertigte Klausuren naturgemäß nicht im WWW verbreitet werden, bieten einige Dozenten oder auch Fachschaften Altklausuren oder typische Fragestellungen an. Abschließend werden zahlreiche Textexemplare online zur Verfügung gestellt, die auf Textsorten der Wissenschaftsverwaltung basieren. Textexemplare der Gruppe juristischer und politischer Textsorten, die normierende Rahmenvorgaben enthalten, sind insbesondere Teil der Webangebote des Gesetzgebers, d. h. des Bundes oder der Länder. Oftmals werden PDF-Versionen von Gesetzestexten oder Erlassen auf universitären Webservern gespiegelt. Zu dieser ersten Gruppe gehören auch Satzungen universitärer Einheiten, die in der Regel Bestandteil des

zentralen Webangebots einer Hochschule sind. Die zweite Gruppe von Textsorten umfasst Dienstanweisungen und Geschäftsordnungen und regelt das Handeln der Verwaltungsangestellten innerhalb der Institution. Derartige Texte werden nur sehr selten im WWW veröffentlicht, weil sie ausschließlich für die in diesem Bereich tätigen Personen bestimmt sind. Die dritte Gruppe beinhaltet verwaltungsinterne Textsorten, deren Textexemplare innerhalb der Institution verfasst werden. Von dem internen Verwaltungsverkehr (Berichte, Protokolle, Gliederungspläne, Mitteilungen, Anfragen, Anträge etc.) werden lediglich diejenigen Texte im WWW publiziert, die für alle Angehörigen der Hochschule von Interesse sein könnten – alle anderen Texte können, von der Öffentlichkeit abgeschirmt, im Intranet der Verwaltung veröffentlicht werden. Bestimmte Textsorten, die sich mit einer Informationsfunktion an alle Angehörigen der Institution richten (z. B. Rundschreiben oder Mitteilungen der Universitätsleitung), werden ebenfalls häufig in das universitäre Webangebot integriert. Textsorten mit Deklarations- oder Obligationsfunktion wie der Vertrag, die Bescheinigung, die Rechnung, die Urkunde oder das Zertifikat sind hiervon ausgenommen. Die dritte Untergruppe der verwaltungsinternen Textsorten umfasst Heinemann zufolge diejenigen Texte, die von Außenstehenden an die Verwaltung gerichtet werden, z. B. Anträge, Anfragen, Eingaben, Widersprüche, Erklärungen oder Formulare. Texte dieser Klassen können zwar in Einzelfällen auch auf elektronischem Wege an die Verwaltung geschickt werden, dies dürfte jedoch primär per E-Mail und nicht per WWW geschehen. Formulare werden zwar von vielen Universitätsverwaltungen im WWW zum Download und anschließendem Ausdruck angeboten, die Rücksendung erfolgt jedoch in aller Regel auf postalischem Wege.

6.4 Zusammenfassung

Dieses Kapitel geht auf grundlegende Eigenschaften universitärer Webangebote ein. Die Websites von Hochschulen sind auf verschiedenen Ebenen von einer sehr ausgeprägten Heterogenität gekennzeichnet: Als Rezipientengruppen existieren die Angehörigen der Institution (Studierende sowie das wissenschaftliche und administrativ-technische Personal) und externe Benutzer, z. B. Wissenschaftler anderer Universitäten und Institute und potenzielle Mitarbeiter und Studieninteressenten. Diese unterschiedlichen Gruppen besitzen individuelle Informationsbedarfe, die von den Webangeboten der Hochschulen nur partiell reflektiert werden. Ebenso heterogener Natur ist die Gruppe der Produzenten. Diese umfasst studentische Hilfskräfte, die die Webauftritte von Professuren oder Instituten betreuen, Mitarbeiter von Rechenzentren, die die technische Infrastruktur verwalten und z. B. Dokumentationen für Netzwerke und ihre Dienste erstellen (vgl. z. B. El-Bayoumi, 1999, und Hopkins, 2000) sowie die Pressestelle bzw. Abteilung für Öffentlichkeitsarbeit, die häufig für einen Großteil des zentralen Angebots verantwortlich ist (Einstiegsseite, Webangebot der Hochschulleitung etc.). Der Kreis der Produzenten zeichnet sich vor allem dadurch aus, dass die Tätigkeit der Erstellung und Pflege von WWW-Angeboten meist der universitären Selbstverwaltung zugehörig ist, d. h. es handelt sich um Aufgaben, die nicht von den eigentlichen Kernkompetenzen eines Arbeitspsychologen, Betriebswirtschaftlers, Molekularbiologen, Fremdsprachendidaktikers oder Lehramtsstudierenden abgedeckt werden. Die mangelnde Expertise wird zwar in vielen Fällen durch Enthusiasmus und Engagement und das Adaptieren der Kon-

ventionen guter Webangebote kompensiert (vgl. Abschnitt 4.3.2), aber dennoch existieren in den Webangeboten gerade größerer Universitäten zahlreiche Inseln, die keine einheitliche Gestaltung besitzen und nach unterschiedlichen Prinzipien strukturiert sind. Diese Inseln korrespondieren – wiederum im Einklang mit den Grundsätzen der universitären Selbstverwaltung – mit den hierarchisch angeordneten Organisationseinheiten einer Hochschule: Die Website einer Universitätsbibliothek wird in der Regel von den Mitarbeitern ihrer Abteilung für Informationstechnologie betreut, während das Angebot eines soziologischen Instituts von einem wissenschaftlichen Mitarbeiter oder einer Hilfskraft administriert wird. Die Angehörigen des Rechenzentrums pflegen die eigene Website, ebenso wie die Mitarbeiter einer Professur für Kunstgeschichte für ihren Webauftritt verantwortlich sind. Diese Darstellung bezieht sich auf denjenigen Stand, der von dem Korpus reflektiert wird, das für die vorliegende Arbeit aufgebaut und bezüglich der hier thematisierten Aspekte explorativ untersucht wurde. Da uneinheitlich gestaltete Websites innerhalb eines universitären Webangebots nur bedingt ein professionelles und innovatives Image vermitteln und somit eine seiner Kernfunktionen nicht erfüllen, ist mittlerweile der Trend zu beobachten, dass der *gesamte* Webauftritt von der Universitätsleitung als wichtiges Marketing-Instrument wahrgenommen wird. So werden – z. B. mit Hilfe eines zentral administrierten CMS-Systems (vgl. Krause, 2003) – einheitlich gestaltete Schablonen angeboten, die von den Angehörigen einer Organisationseinheit über ein WWW-Interface mit Inhalten befüllt und anschließend gepflegt werden können.

6.5 Fazit – Zum Bedarf einer relationalen Repräsentation

Die Organisationsstruktur einer Universität wird in nahezu allen im Korpus enthaltenen universitären Webauftritten als Blaupause der jeweiligen Informationsarchitektur verwendet und ist somit maßgeblich an der Konstituierung von Hypertextsorten in dieser Domäne beteiligt: Eine Universität besitzt z. B. verschiedene Fachbereiche, die ihrerseits mehrere Institute umfassen, in denen einzelne Arbeitsgruppen tätig sind. Auf jeder dieser Ebenen werden korrespondierende Webauftritte eingesetzt, die unter anderem auf die Websites der spezifischeren Organisationseinheiten verweisen. Das Webangebot einer Organisationseinheit wird also von ihrem Status innerhalb der Institution determiniert. Daher stellt sich zwangsläufig die Frage, welche *generischen* Strukturen innerhalb der deutschen Hochschulen vorliegen – gemeint sind *nicht* ihre Webauftritte, sondern die Institutionen selbst – und inwiefern diese Strukturen zu einer einheitlichen Informationsarchitektur des Hypertexttyps *Webauftritt einer Universität bzw. Hochschule* generalisierbar sind.

Das Hypertextsortenmodell (Kapitel 5) ist, wie auch Abbildung 1.3 (S. 11) verdeutlicht, domänenunabhängig, d. h. die universitären Webangebote stellen ein Fallbeispiel zur Anwendung des Modells dar. Von besonderer Bedeutung sind in diesem Modell relationale Strukturen, die es ermöglichen, die angesprochene Organisationsstruktur einer Hochschule ebenso zu reflektieren wie die Einbettung von Hypertextknotensorten (vgl. Abschnitt 4.5.2). Das Modell geht dabei von zwei wesentlichen Prämissen aus: Einerseits soll es sprachtechnologische Anwendungen bestmöglich unterstützen, andererseits soll es Analysen aller linguistischen Beschreibungsebenen ermöglichen. Schlobinski (2000a, S. 824) weist darauf hin, dass semantische Netze zur Analyse und Charakterisierung von Hypertexten eingesetzt werden können. In jedem Fall ist jedoch, so Schlobinski, ein relationaler Ansatz zu wählen, der

eindeutige Definitionen der Elemente und Relationen erlaubt, „was trivial erscheint, in der Praxis aber keineswegs so trivial ist“ (ebd.). Darüber hinaus sollte dieser Formalismus auch unterschiedliche Typen von Hyperlinks repräsentieren können. Hinzu kommen thematische Relationen und, wie dieses Kapitel gezeigt hat, institutionelle Relationen, die sich auf die Organisationsstruktur einer Hochschule beziehen (z. B. Universität *has-part* Fachbereich), die wiederum ein wesentlicher Bestandteil des abstrakten Domänenmodells ist. Kapitel 13 geht detailliert auf diese Problematik ein und stellt die *Web Ontology Language* als übergeordneten Formalismus zur Repräsentation von Hypertextsorten vor, der die Potenziale semantischer Netze und die maschinelle Anwendbarkeit in sich vereint.

7

Sammlung, Zugriff und Analyse von Webdokumenten mittels einer Korpusdatenbank

7.1 Einleitung

Die in Abschnitt 4.4 diskutierten Arbeiten zeigen, dass eine systematische Untersuchung von Hypertextsorten nur mit Hilfe von Restriktionen bezüglich des Untersuchungsgegenstandes vorgenommen werden kann – eine präzise Identifizierung und Generalisierung von Hypertextsorten basierend auf Stichproben *beliebigen* Inhalts ist geradezu unmöglich. Die vorliegende Arbeit bezieht sich auf die Webseiten der deutschen Hochschulen (vgl. Kapitel 6).

Da sowohl das Layout als auch die Inhalte von HTML-Dokumenten häufig aktualisiert werden (Koehler, 2002), war es notwendig, einen Schnappschuss (vgl. Roberts, 1998, S. 81, Hawking et al., 1999a, Walker, 1999, Storrer, 1999b, S. 4, Schütte, 2004a, S. 143–148) der Untersuchungsdomäne anzufertigen: Es wurde ein Korpus erzeugt (vgl. Thompson, 2000), das die deutschsprachigen Dokumente der deutschen Hochschulen umfasst (vgl. Grefenstette und Nioche, 2000). Zu diesem Zweck wurden die Webserver sämtlicher Universitäten mit Hilfe eines *Crawlers* und eines automatischen Sprachenidentifizierers rekursiv traversiert, um *alle* verfügbaren deutschsprachigen HTML-Dokumente zu sammeln und auf einem lokalen Server zu spiegeln.[1] Im Folgenden wird die für diese Arbeit implementierte Korpusdatenbank vorgestellt, die aus drei wesentlichen Bausteinen besteht: (i) Ein *Webserver* stellt die im

[1] Jucker (2004, S. 18) weist im Kontext der sprachwissenschaftlichen Untersuchung von Webseiten auf „nicht unerhebliche Probleme der *Datensammlung*“ hin, die durch das in diesem Kapitel vorgestellte System zumindest in Teilen gelöst werden (vgl. auch Amitay, 2000a). Unter anderem stellt Jucker die Frage, ob ein „adäquates Korpus“ beispielsweiser einer Online-Zeitung auch die Zieldokumente externer Hyperlinks umfassen sollte (ebd., S. 19). In der vorliegenden Arbeit werden Dokumente auf externen Websites nicht berücksichtigt, da sie für das verfolgte Ziel nicht relevant sind. Die Frage nach dem „adäquaten Korpus“ kann somit nur vor dem Hintergrund eines spezifischen Erkenntnisinteresses beantwortet werden.

Korpus enthaltenen Dokumente zur Verfügung. (ii) Eine *relationale Datenbank* umfasst zahlreiche Metadaten über diese Dokumente und dient der Speicherung verschiedener Analyse- und Administrationsinformationen. (iii) *Zugriffsfunktionen* erlauben sowohl die manuelle als auch die automatische Exploration und Analyse des Datenbestandes. Der direkte Zugriff erfolgt über ein Web-basiertes Front-End, der indirekte Zugriff gestattet autarken Analysemodulen einen netzwerkbasierten Zugang zum Korpus.

Abschnitt 7.2 geht zunächst auf die Datensammlung ein, bespricht das Werkzeug zur Sprachenidentifizierung, diskutiert den Aufbau der relationalen Korpusdatenbank und die Aufbereitung sowie den Import der Metadaten.[2] Anschließend werden die direkten und indirekten Zugriffsmöglichkeiten auf die Korpusdatenbank thematisiert (Abschnitte 7.3 und 7.4). Da die Korpusdatenbank eine sehr spezifische Funktionalität besitzt, zu der meines Wissens keine anderen Arbeiten existieren, werden abschließend Berührungspunkte zu verwandten Arbeiten aus unterschiedlichen Themengebieten diskutiert (Abschnitt 7.5).

7.2 Datensammlung und Datenhaltung

Dieser Abschnitt thematisiert die Sammlung der im Korpus enthaltenen Dokumente sowie die Aufbereitung der Daten für den Import in eine relationale Datenbank (Rehm, 2001). Im Einzelnen geht es zunächst um die Frage, wie die Webserver der Universitäten traversiert wurden, um rekursiv – basierend auf den in HTML-Dokumenten enthaltenen Hyperlinks – alle Dokumente verschiedener Dateiformate zu erfassen und in den Korpusdatenbank-Server zu integrieren (Abschnitt 7.2.1). An dem Prozess der Datensammlung ist ein automatischer Sprachenidentifizierer beteiligt, um nicht deutschsprachige Dokumente von der Sammlung auszuschließen (Abschnitt 7.2.2). Abschnitt 7.2.3 diskutiert das Design der Korpusdatenbank, die die zentrale Komponente für den direkten und indirekten Dokumentzugriff darstellt. Abschnitt 7.2.4 thematisiert die konzeptionelle Verbindung zwischen den gesammelten Daten und der Korpusdatenbank: Ein als *Finite State Transducer* fungierendes *Perl*-Skript bereitet die Rohdaten so auf, dass sie in die relationale Datenbank importiert werden können. Abschnitt 7.2.5 geht auf den Umgang mit Duplikaten ein, die bei der Erstellung von Korpora, die aus Webseiten bestehen, eines der zentralen Probleme darstellen, woraufhin Abschnitt 7.2.6 den Inhalt und den Umfang des Korpusbestandes vorstellt.

Abbildung 7.1 zeigt den Ablauf des Dokumentsammlungszyklus, dessen Basis eine zufällig angeordnete Liste aller deutschen Hochschulen ist.[3] Diese Liste beinhaltet weitere Einrichtungen (z. B. Musik- und Kunsthochschulen), die ebenfalls in das Korpus aufgenommen wurden, da hierfür – wie sich nach der initialen Datensammlung gezeigt hat – ausreichend viel Festplattenplatz vorhanden war. Die auch als *Crawl* bezeichnete Datensammlung beginnt mit der URL der Einstiegsseite des Webauftritts einer Einrichtung (z. B. `http://www.uni-giessen.de`). Daraufhin traversiert das System alle verfügbaren HTML-Dokumente und alle erreichbaren Webserver der entsprechenden Domäne (`.uni-giessen.de`), um die Dateien, die den Restriktionen entsprechen, in das Korpus zu übertragen.

[2] Anhang A (S. 721 ff.) stellt eine statistische Charakterisierung der Untersuchungsdomäne vor, die zahlreiche Merkmale der im Korpus enthaltenen HTML- und XML-Dokumente diskutiert.

[3] Abbildung 7.2 erläutert die Werkzeuge, die in Abbildung 7.1 in abgerundeten Boxen dargestellt sind.

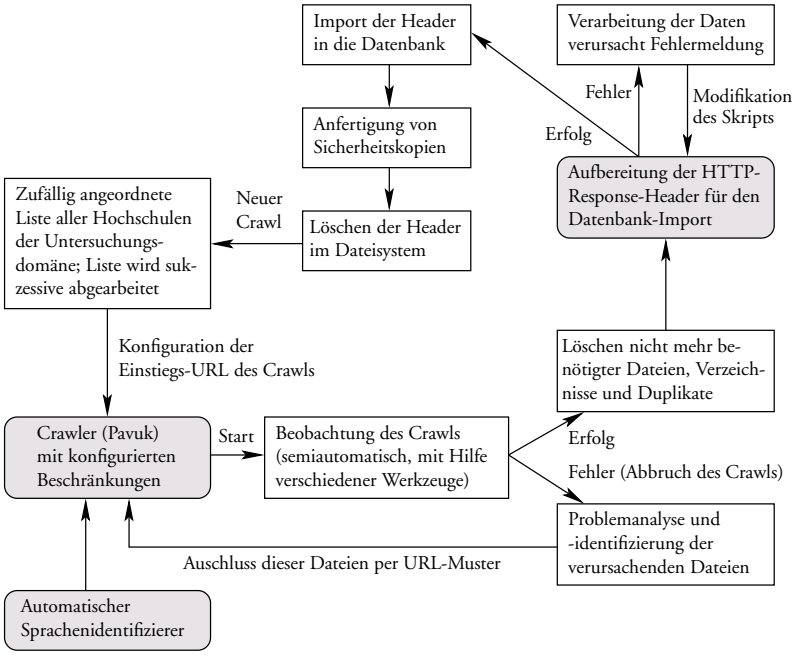


Abbildung 7.1: Schematischer Ablauf des Dokumentsammlungszyklus

7.2.1 Automatische Traversierung des World Wide Web

Crawler, häufig auch als *Robots*, *Spider* oder *Worms* bezeichnet, ermöglichen es, die Dokumentbestände entfernter Webserver automatisch auf einen lokalen Rechner zu transferieren (vgl. etwa Turau, 1998c, Brin und Page, 1998, Heydon und Najork, 1999, Raghavan und Garcia-Molina, 2001, Chakrabarti, 2003, Hemenway und Calishain, 2003). Der *Crawler* fungiert dabei aus Sicht des *Hypertext Transfer Protocol*s als Client, der bestimmte, möglicherweise sogar jeden gefundenen Hyperlink verfolgt.[4] Da ein Browser von einer Person bedient wird, die entscheidet, welche Webseiten angesteuert werden, stellt sich die Frage nach der Navigationsautomatisierung: Für einfache Aufgaben werden oftmals Werkzeuge wie z. B. *wget* benutzt, das auf der Kommandozeile mit einer URL als Argument gestartet wird, woraufhin die angegebene Datei auf das lokale System übertragen wird.[5] Gruppen von Dokumenten können rekursiv heruntergeladen werden, wobei die Rekursionstiefe oder erlaubte und nicht erlaubte Domains spezifiziert werden müssen, damit der *Crawl* terminiert.

Vor Beginn der Datensammlung wurde der Inhalt des aufzubauenden Korpus spezifiziert als die deutschsprachigen Webseiten aller Webserver aller deutschen Hochschulen. Neben

[4] Dieser Vorgang entspricht dem Beschreiten eines gerichteten Graphen (vgl. Gudivada et al., 1997). Der im Rahmen dieser Arbeit eingesetzte *Crawler Pavuk* benutzt in der Grundeinstellung ein *depth-first*-Verfahren.

[5] Das Werkzeug *wget* ist Teil der GNU-Programmsammlung, die von der *Free Software Foundation* (FSF) gepflegt wird (vgl. `http://www.gnu.org/software/wget/wget.html`). Burke (2002) zeigt, wie einfache *Crawler* mit Hilfe des *Perl*-Moduls *LWP* implementiert werden können. Hemenway und Calishain (2003) geben zahlreiche Hinweise zur Realisierung unterschiedlicher Anwendungen von *Spidern* und *Crawlern*. Ein Beispiel für einen integrierten *Crawler* ist der *Gatherer* des verteilten Systems *Harvest* (`http://harvest.sourceforge.net`), das unter anderem in dem Projekt GERHARD eingesetzt wird (Wätjen et al., 1998, vgl. Abschnitt 13.3).

Dateiformat	Erläuterung	Medientyp
HTML-Dokumente	*Hypertext Markup Language* (Raggett et al., 1999)	`text/html`
CSS-Dateien	*Cascading Style Sheets* (Bos et al., 1998)	`text/css`
XML-Dateien	*Extensible Markup Language* (Bray et al., 2004b)	`text/xml`
SGML-Dateien	*Standard Generalized Markup Language* (ISO 8879, 1986)	`text/sgml`
ASCII-Dateien	ASCII-Textdateien ohne weitere Spezifizierung des Dokumenttyps (*American Standard Code for Information Interchange*)	`text/plain`
Usenet-Artikel	RFC 850- bzw. RFC 1036-konforme Dateien	`message/news`
E-Mails	RFC 822- bzw. RFC 2822-konforme Dateien	`message/rfc822`

Tabelle 7.1: Die im Korpus enthaltenen Medientypen

HTML-Dokumenten sollten auch Dateien verschiedener anderer Medientypen in das Korpus aufgenommen werden (vgl. Tabelle 7.1 sowie Abschnitt 7.2.4). Für die Datensammlung ergibt sich hierdurch die Restriktion, dass der *Crawler* ausschließlich Dateien der spezifizierten Medientypen in das Korpus aufnehmen soll. Eine weitere Beschränkung betrifft die Dateigröße, die 512 000 Bytes (500 Kilobyte) nicht überschreiten soll, um zu verhindern, dass sehr große Dokumente unnötig viel Platz im Korpus einnehmen.[6] Eine weitere Restriktion stellt die Aufgabe dar, mit Hilfe eines automatischen Sprachenidentifizierers zu erkennen, ob ein Dokument deutschsprachig ist oder nicht (vgl. Abschnitt 7.2.2). Weiterhin sollten nur diejenigen HTML-Dokumente in das Korpus aufgenommen werden, die auf Webservern hinterlegt sind, die Dokumente auf dem HTTP-Standardport 80 anbieten. Webserver, die auf anderen Ports arbeiten, sind meist – so haben vorab durchgeführte Tests ergeben – experimentelle, WAP- oder Proxy-Server und daher für das Vorhaben nicht relevant.

Aufgrund der Komplexität dieser Beschränkungen war es nicht möglich, *wget* als *Crawler* zu benutzen, da die Restriktionen nicht robust realisiert werden konnten. Stattdessen wurde das Werkzeug *Pavuk*[7] eingesetzt, das über eine ähnliche Funktionalität verfügt, zusätzlich jedoch bei der Traversierung einer sehr großen Zahl von Webseiten zuverlässiger arbeitet und fortgeschrittene Konfigurationsmöglichkeiten bietet. Wenn alle Beschränkungen erfüllt sind und *Pavuk* ein HTML-Dokument auf dem lokalen Rechner ablegt, findet eine minimale inhaltliche Modifikation statt: Damit ein Betrachten der Dokumente mit einem Webbrowser möglich ist, werden alle eingebetteten URLs so modifiziert, dass sie auf die – sofern ebenfalls im Korpus vorhanden – lokalen Dokumente bzw. auf die entfernten Dateien verweisen. Bei der Anzeige der im Korpus enthaltenen HTML-Dokumente werden somit keine *broken image*-Icons angezeigt, da mittels des HTML-Elements `img` eingebettete Grafiken aufgrund der modifizierten URL noch immer eine gültige Referenz besitzen.

[6] Derart umfangreiche Dateien kommen in der Realität nur sehr selten vor: Lediglich neun HTML-Dateien innerhalb der Domäne `.uni-giessen.de` sind größer als dieser Schwellwert (vgl. Abschnitt A.4.1, S. 736 ff.).

[7] *Pavuk* unterliegt der *GNU General Public License* (GPL) und wurde von Stefan Ondrejicka implementiert (vgl. `http://www.idata.sk/~ondrej/pavuk/`). Während des Einsatzes dieser Software für den Korpusaufbau wurden zahlreiche Fehler entdeckt und dem Entwickler mitgeteilt. Stefan Ondrejicka hat freundlicherweise verschiedene umfangreiche Funktionen implementiert, die für die Datensammlung benötigt wurden. Seit dem Sommer 2003 wird *Pavuk* über *SourceForge* entwickelt und vertrieben (`http://pavuk.sourceforge.net`).

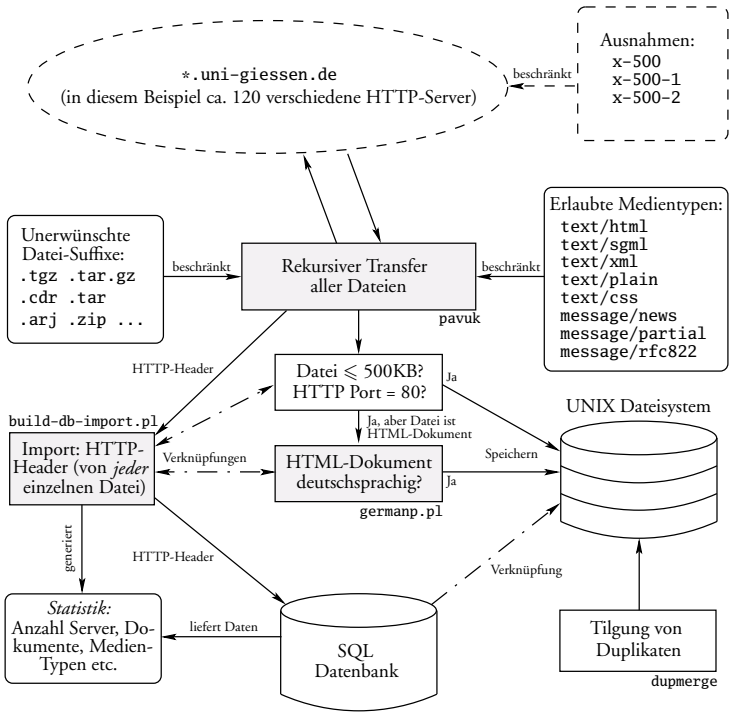


Abbildung 7.2: Schematischer Ablauf der Sammlung der Dateien einer Universität mit konfigurierten Beschränkungen am Beispiel `*.uni-giessen.de`

Abbildung 7.2 verdeutlicht am Beispiel der Webserver der Universität Gießen den Ablauf der Datensammlung sowie die Beschränkungen, die hinsichtlich der in das Korpus aufzunehmenden Daten spezifiziert wurden: *Pavuk* ermöglicht das automatisierte und rekursive Übertragen von Dateien mittels HTTP (RFC 2616).[8] Neben der Startadresse – im Beispiel handelt es sich um `http://www.uni-giessen.de` – wird *Pavuk* mit verschiedenen Restriktionen konfiguriert. Eine wichtige Einschränkung, die bislang noch nicht angesprochen wurde, bezieht sich auf die Angabe eines Musters, das spezifiziert, welche Hyperlinks verfolgt werden sollen (im Beispiel handelt es sich um URLs innerhalb von `.uni-giessen.de`).[9] Drei Server werden explizit ausgeschlossen, da sie ausschließlich die Daten des X.500-Verzeichnisses zur Pflege von Kontaktinformationen der Hochschulangehörigen enthalten. Die in Tabelle 7.1 dargestellten Medientypen können ebenfalls mit Hilfe der Konfigurationsmöglichkeiten von *Pavuk* spezifiziert werden. Aufgrund der Tatsache, dass von sehr vielen Webservern Dokumente mit fehlerhaften Angaben bezüglich des `Content-Type` (d. h. des Medientyps einer Datei) ausgeliefert werden, wird zusätzlich eine umfangreiche Liste nicht erwünschter Datei-

[8] Während der Datensammlung wurde der *Robot Exclusion Standard* eingehalten (`http://www.robotstxt.org`, vgl. auch Gourley et al., 2002, S. 230 ff.). Dieser Quasi-Standard sieht vor, dass *Crawler* vor der Traversierung eines Servers überprüfen sollten, ob im Wurzelverzeichnis eine Datei namens `robots.txt` existiert, in der der Administrator spezifizieren kann, welche Teile des Webangebots ein Roboter besuchen darf, um z. B. Datenbank-Anwendungen auszuschließen, die extrem viele Seiten dynamisch generieren.

[9] Vereinzelt werden weitere Bereiche, die häufig als *Crawler-* bzw. *Spider*-Fallen bezeichnet werden, manuell ausgeschlossen. Der nachfolgende Abschnitt geht genauer auf dieses Thema ein.

suffixe spezifiziert.[10] Diese Beschränkungen gelten ausschließlich für die Aufnahme einer Datei in das Korpus. Für *alle* referenzierten Dateien werden diejenigen Metadaten, die in Form des HTTP-Response Headers (RFC 2616) vom Webserver geliefert werden, lokal gespeichert, so dass diese Informationen in die relationale Datenbank importiert werden können.

Das Speichern der Dokumente im Dateisystem bietet den Vorteil, dass Analysewerkzeuge (z. B. *sed*-, *awk*- oder *Perl*-Skripte) ohne den Performanz-intensiven Umweg über die Datenbank eingesetzt werden können. Prinzipiell existieren zwei Möglichkeiten, die Dokumente in einer Datenbank abzulegen: Einerseits könnten sie in einem relationalen Datenbanksystem wie *MySQL* als Daten vom Typ `BLOB` (*Binary Large Object*) gespeichert werden, andererseits könnte eine native XML-Datenbank Verwendung finden, wobei HTML- und SGML-Dateien in wohlgeformte XML-Dateien überführt werden müssten. Dateien der verbleibenden Medientypen müssten in XML-Instanzen gekapselt werden, was einen zusätzlichen Verwaltungsaufwand erforderte. Bei einer Datenmenge von mehr als 4 000 000 Dateien scheiden beide Optionen aus Performanzgründen aus. Ein zweiter Vorteil betrifft die ständige Verfügbarkeit der Daten durch einen in der Datenbanktabelle `http_header` enthaltenen Verweis sowie die Tatsache, dass der auf dem Korpusdatenbank-Server ebenfalls installierte Webserver die im Korpus enthaltenen Dokumente per HTTP ausliefern kann. Dieser Aspekt ist für die direkten und indirekten Zugriffsmöglichkeiten auf das Korpus von zentraler Bedeutung.

Verwandte Arbeiten

Crawler werden vornehmlich von Suchmaschinen zur Dokumentsammlung eingesetzt. Brin und Page (1998) stellen die Architektur von *Google* vor: Verteilte und simultan operierende *Crawler* werden von einem URL-Server mit Adresslisten bedient und senden Dokumente an den Store-Server, der Indexierungs- und Komprimierungsfunktionen durchführt (vgl. auch Hirai et al., 2000, Arasu et al., 2001). Obgleich für den produktiven Betrieb einer Suchmaschine zweifelsfrei sehr viel umfangreichere Einschränkungen hinsichtlich Performanz und Funktionalität gelten, sind die von Brin und Page (1998) dargestellten *Crawler* durchaus mit den in diesem Kapitel beschriebenen Konzepten vergleichbar. Aus diesem Grund werden lediglich die Kernmerkmale dieses Systems vorgestellt, die angesichts des aktuellen Umfangs dieser Suchmaschine[11] nur noch von historischem Interesse sind, jedoch die einzigen veröffentlichten Daten über die Architektur des Systems darstellen: Die *Crawler* und der URL-Server wurden in *Python* implementiert. Üblicherweise arbeiteten drei *Crawler* parallel, von denen jeder simultan ca. 300 HTTP-Verbindungen bediente. Zu Spitzenzeiten konnte das System pro Sekunde etwa 100 Dokumente herunterladen, die in ca. 600 Kilobyte Daten resultierten. Ein die Performanz beeinträchtigender Flaschenhals waren diesbezüglich DNS-Anfragen, weshalb jeder *Crawler* mit einem eigenen DNS-Cache ausgestattet wurde.[12]

[10] Viele Webserver versenden Archive mit dem Suffix `.zip` als Dateien des Medientyps `text/plain`, obwohl `application/zip` der korrekte Typ wäre. Um die Integration derartiger Dateien in das Korpus zu vermeiden, wurde der explizite Ausschluss basierend auf dem Dateisuffix realisiert.

[11] Im August 2005 indexiert *Google* (`http://www.google.com`) etwa acht Milliarden HTML-Dokumente. Der Datenbestand wird in mehreren Server-Farmen verwaltet, die sich auf mehr als 15 000 *Linux*-Rechner verteilen (Barroso et al., 2003). Die *Google*-eigenen *Crawler* werden als *Googlebots* bezeichnet.

[12] Das *Domain Name System* bildet die symbolischen Namen von Rechnern, die mit dem Internet verbunden sind, auf ein numerisches Adressschema ab (vgl. RFC 1034 und RFC 1035).

Heydon und Najork (1999) ersetzen in ihrem *Crawler Mercator* die synchronisierte UNIX-Funktion `gethostbyname()` durch eine unsynchronisierte Version, so dass ein verzögerter DNS-Aufruf das System nicht blockieren kann. Darüber hinaus gehen sie auf Probleme ein, die bei umfangreichen *Crawls* sehr häufig auftreten: Hierzu gehören URL-Aliase, die den multiplen Download einer Datei verursachen, falls sie nicht detektiert werden (vgl. Abschnitt 7.2.5). Das schwerwiegendste Problem sind *Crawler*-Fallen, die bewirken können, dass ein *Crawl* nicht terminieren kann. Einige dieser Fallen sind nicht beabsichtigt, z. B. ein falsch gesetzter symbolischer Link im Dateisystem des Webservers (vgl. Gourley et al., 2002, S. 221), der immer wieder neue und gültige URLs erzeugt. Zu den beabsichtigten Fallen gehören CGI-Skripte, die unendlich viele erfolgreich adressierbare URLs generieren. Für diese sehr strittige Vorgehensweise existieren zwei Motive: Kommerzielle Anbieter verfolgen das Ziel, im Index einer Suchmaschine möglichst hoch gelistet zu werden. Die entsprechenden Algorithmen auf Seiten der Suchmaschine hängen unter anderem von der Anzahl eingehender Verweise auf ein Dokument ab. Wenn extrem viele Dokumente, die von der *Crawler*-Falle automatisch generiert werden, auf ein spezielles Dokument verweisen, wird dessen Rang in Suchmaschinen drastisch erhöht, falls der Suchmaschinen-*Crawler* der Falle nicht entgehen kann. Die zweite Motivation betrifft die gewollte Irritation von *Crawlern*, deren Aufgabe es ist, für Zwecke des Direktmarketings möglichst viele E-Mail-Adressen zu sammeln.

Ein Algorithmus zur Vermeidung von *Crawler*-Fallen wurde bislang nicht veröffentlicht, so dass die problematischen Adressen, wie dies auch beim Korpusaufbau geschieht, manuell in eine schwarze Liste aufgenommen werden müssen. Es ist jedoch anzunehmen, dass die Betreiber der großen Suchmaschinen entsprechende Heuristiken einsetzen, da es kaum eine Möglichkeit gibt, *Crawler*-Fallen vollständig manuell zu erfassen, wenngleich verschiedene Indizien auf ihre Existenz hinweisen können (überproportional viele Dokumente in einem bestimmten Ast eines Dokumentbaums, rekurrente Verzeichnisnamen in einer URL oder besonders lange URLs). Derartige Indizien stellt Chakrabarti (2003) dar, die im Rahmen einer generischen *Crawler*-Architektur von einem „URL approval guard“ detektiert werden, so dass einige Typen von Fallen als automatisch identifizierbar gelten können. Das Kapitel „Crawling the Web“ in Chakrabarti (2003) stellt die umfassendste Quelle für die Implementierung von *Crawlern* dar, die für den Aufbau von Kollektionen in Millionengröße gedacht sind. Chakrabarti bespricht alle problematischen Aspekte, z. B. die Vermeidung von Duplikaten, die Blockgröße des Dateisystems, DNS-Caching und die Normalisierung von URLs. Hemenway und Calishain (2003) stellen Anwendungsmöglichkeiten von *Crawlern* und *Screen-Scrapern* dar. Als *Screen-Scraper* oder auch *Wrapper* werden Werkzeuge bezeichnet, die in der Lage sind, vorgegebene Informationen in HTML-Dokumenten zu extrahieren (vgl. Abschnitt 14.6.3). Eine Anwendung, die mit Hilfe eines *Crawlers* und eines *Scrapers* realisiert werden kann, ist z. B. das automatische Absetzen einer Suchabfrage an die Website eines Buchhändlers, die Identifizierung von Treffern und die Sammlung von Benutzerkommentaren.[13]

[13] Dieses Beispiel verdeutlicht, dass *Crawler* nicht in der Lage sind, das *hidden* bzw. *deep Web* zu traversieren: Es umfasst diejenigen HTML-Dokumente, die nach dem Ausfüllen eines Formulars dynamisch aus den Beständen von Datenbanken generiert werden (vgl. Bergman, 2000). Raghavan und Garcia-Molina (2001) stellen einen *Crawler* vor, der HTML-Formulare einer rudimentären Inhaltsanalyse unterzieht, woraufhin in einer Datenbank vorgehaltene Suchanfragen für das automatische Ausfüllen und Abschicken des Formulars eingesetzt werden, um auf diese Weise auch die im *hidden Web* enthaltenen Dokumente erreichen zu können.

7.2.2 Maschinelle Sprachenidentifizierung

Da das Korpus ausschließlich deutschsprachige Dokumente enthalten soll, war die Implementierung eines Sprachenidentifizierers notwendig.[14] Im Folgenden wird das für die Datensammlung implementierte Werkzeug `germanp.pl` vorgestellt. Eine Evaluation zeigt, dass dessen Lexikon-basiertes Verfahren bezüglich der Präzision dem Status Quo der Language-Identification-Algorithmen entspricht. Abschließend werden verwandte Ansätze zur Erzeugung von Korpora mit Hilfe automatischer Sprachenidentifizierer diskutiert.

Im Bereich der *Automatic Language Identification*[15] werden einerseits statistische Methoden (vgl. Dunning, 1994) und andererseits Lexikon- bzw. Wortlisten-basierte Ansätze eingesetzt (z. B. Giguet, 1995). Langer (2002) diskutiert die jeweiligen Vor- und Nachteile: Statistische Methoden basieren meist auf n-Gramm-Techniken, d. h. der Algorithmus wird z. B. mit Byte-Trigrammen trainiert, die aus manuell kategorisierten Texten extrahiert werden können, was die Trainingsphase vereinfacht (vgl. Cavnar und Trenkle, 1994). Sie erreichen bei der ausschließlichen Verarbeitung regulärer Dokumente – Langer spricht von „Standardtexten“ – Präzisionswerte von annähernd 100%, weshalb die automatische Sprachenidentifizierung als „weitgehend gelöstes Problem“ gilt (Langer, 2002, S. 99).[16] Neben statistischen Verfahren werden auch häufig Algorithmen eingesetzt, die auf unterschiedlich umfangreichen Lexika basieren. Für einige Anwendungen ist es ausreichend, eine Liste hochfrequenter Wortformen (Funktionswörter etc.) als Lexikon einzusetzen; häufig werden auch mittel- und niedrigfrequente Wörter zur Steigerung der Präzision aufgenommen. Die Zusammenstellung der Wortliste stellt einen sehr aufwändigen Prozess dar, der bestensfalls halbautomatisch durchgeführt werden kann, da Wortlisten spezieller Sprachen *ohne* Internationalismen oder Eigennamen kaum erhältlich sind. Aus diesem Grund müssen derartige Wortlisten manuell korrigiert werden, um alle Wörter, die nicht sprachspezifisch sind, zu entfernen. Da solche Algorithmen auf dem Abgleich der in einem Dokument gefundenen Wörter mit Lexikoneinträgen basieren, bereiten diesen Verfahren insbesondere sehr kurze Texte Probleme, in denen eventuell von einem eher fachsprachlichen – und daher nicht im Lexikon enthaltenen – Vokabular Gebrauch gemacht wird, wodurch die Fehlerrate auf mehr als 10% steigen kann. Fachsprachliche Ausdrücke werden von statistischen Verfahren eher kompensiert, da sie auf Wortfragmenten operieren, doch wird die Präzision auch dieser Methoden durch sehr kurze Dokumente, die zwischen einem und etwa 20 Wörtern umfassen, deutlich beeinträchtigt.[17] Ein Vorteil Lexikon-basierter Verfahren ist hingegen, dass die zur Erkennung benutzten Sprachmodelle manuell manipulierbar bleiben, da Modifikationen der Wortlis-

[14] Auch Thelwall (2005, S. 617) hebt „the multilingual nature of university Web sites“ hervor.

[15] Muthusamy und Spitz (1998) stellen Ansätze für geschriebene und gesprochene Sprache dar. Die Verarbeitung gesprochener Sprache wird im Folgenden, ebenso wie die Problematik der Zeichensatzkodierung (vgl. Beesley, 1988, Constable und Simons, 2000, Langer, 2002), nicht weiter betrachtet. Gertjan van Noord pflegt eine Liste verfügbarer Sprachenidentifizierer (`http://odur.let.rug.nl/~vannoord/TextCat/competitors.html`).

[16] Eine derart hohe Präzision ist nur dann erreichbar, „wenn von der Unterscheidung extrem eng verwandter Sprachen abgesehen wird“ (Langer, 2002, S. 102), z. B. Norwegisch und Dänisch oder Serbisch und Kroatisch.

[17] Langer (2002, S. 100) merkt – kontrastierend zu der hier gewählten Darstellung – an, dass „[beide Ansätze] eine wichtige Gemeinsamkeit [haben]: die Klassifizierung beruht letztlich auf dem Abgleich von Byte-Sequenzen im Dokument mit Byte-Sequenzen in einem vor der Laufzeit erstellten Lexikon von spezifischen Sequenzen des entsprechenden Typs. Bei N-Gramm-Algorithmen sind dies Sequenzen mit fixer Länge, bei wortbasierten Ansätzen von Leer- und Satzzeichen begrenzte Einheiten variabler Länge.“

ten jederzeit vorgenommen werden können. Darüber hinaus existieren weitere Dokumenteigenschaften, die einen negativen Einfluss auf die automatische Sprachenerkennung haben können: Im Bereich der Lexik sind dies etwa multilinguale Dokumente (vgl. Giguet, 1996), Eigennamen (z. B. bei Dokumenten, die lediglich Namenslisten darstellen), Wiederholungen irreführender Sequenzen (etwa Abkürzungen, die in anderen Sprachen als Lexem existieren) sowie der Bereich der Morphologie, der es insbesondere für die Lexikon-basierte Erkennung hochflektierender Sprachen notwendig macht, entweder ein Vollformlexikon zu pflegen oder die flektierten Formen mit einem Morphologiemodul generieren zu lassen.

Der Sprachenidentifizierer germanp.pl

Aufgrund der ausschließlichen Erkennung deutschsprachiger Dokumente ist es beim Aufbau des Korpus nicht notwendig, möglichst viele verschiedene Sprachen zu identifizieren. Stattdessen ist die Minimalklassifizierung ausreichend, ob die Sprache, in der ein Dokument verfasst wurde, *deutsch* oder *unbekannt* ist. Es wurden verschiedene Experimente mit im Quelltext verfügbaren Tools durchgeführt.[18] Die Tests haben gezeigt, dass diese Systeme für die im Kontext der automatischen Sprachenidentifizierung sehr spezielle Klassifikationsaufgabe entweder keine befriedigenden Kategorisierungsergebnisse liefern oder schlicht zu langsam ablaufen. Gerade die Performanz spielt bei der Datensammlung eine wesentliche Rolle, da für *jedes* HTML-Dokument eine Überprüfung der Sprache notwendig ist.

Prinzipiell können – neben der Analyse des in einem Dokument enthaltenen Textes – weitere Informationen zur Identifikation der Sprache eingesetzt werden. Hierzu gehören etwa die in der URL eines Dokuments enthaltene *top-level*-Domäne (`.de`, `.fr`, `.es` etc.) oder eine explizite Angabe der Sprache innerhalb eines `meta`-Tags (vgl. Kelly et al., 2003). In der Realität sind derartige Informationen kaum nutzbar: Die *top-level*-Domäne kann lediglich einen sehr groben Hinweis auf den Standort des Servers geben; explizite Sprachangaben in Form entsprechender `meta`-Tags waren bei vorab durchgeführten Tests kaum verfügbar (vgl. Abschnitt A.4.4, S. 759 ff.), weshalb der hier beschriebene Ansatz zur Identifizierung der Sprache ausschließlich auf der Analyse des in einem Dokument enthaltenen Textes basiert.

Da mit den frei verfügbaren Implementierungen keine befriedigenden Resultate erzielt werden konnten, wurde das Werkzeug `germanp.pl` implementiert, das mit einem Lexikon-basierten Sprachenidentifizierungsverfahren arbeitet (Rehm, 2001).[19] Vor der Verarbeitung wird zunächst durch den Terminal-basierten Browser *lynx*, der von einem kapselnden, von *Pavuk* aufgerufenen Shell-Skript gestartet wird, eine Tilgung des in einem HTML-Dokument enthaltenen Markups durchgeführt (vgl. Abbildung 7.3).[20] Es werden alle HTML-Elemente entfernt und Entitäten aufgelöst (z. B. `&ouml;` → ö in der Latin 1-Kodierung), indem *lynx* mit Hilfe der Optionen `-force_html` und `-dump` veranlasst wird, das HTML-Dokument zu interpretieren und den Text an die Standardausgabe zu schicken, wo er von `germanp.pl` eingelesen

[18] Beispielsweise mit dem n-Gramm basierten *TextCat*, implementiert von Gertjan van Nord, das im Quelltext unter `http://odur.let.rug.nl/~vannoord/TextCat/` verfügbar ist und auf Cavnar und Trenkle (1994) basiert.

[19] Zur Benennung des Tools siehe den Abschnitt „The -P convention“ in Raymond (1996), online verfügbar unter `http://www.catb.org/jargon/`: „Turning a word into a question by appending the syllable ‘P’; from the LISP convention of appending the letter ‘P’ to denote a predicate (a boolean-valued function). The question should expect a yes/no answer, though it needn't.“

[20] Der Browser *lynx* ist Teil praktisch jeder *Linux*-Distribution und erhältlich unter `http://lynx.isc.org`.

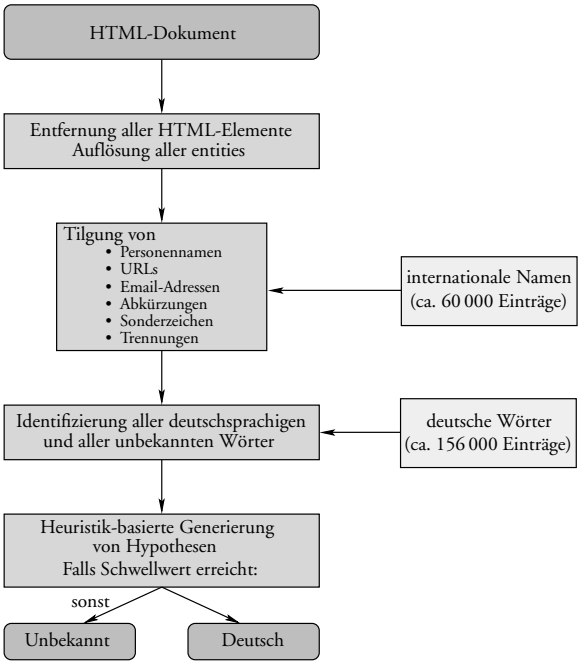


Abbildung 7.3: Die Architektur des automatischen Sprachenidentifizierers `germanp.pl`

wird. Zunächst findet eine rudimentäre Tokenisierung (vgl. Grefenstette und Tapanainen, 1994) statt, um Wörter von Nicht-Wörtern zu separieren, wobei unter anderem alle Großbuchstaben in Kleinbuchstaben konvertiert werden, denn in sehr vielen HTML-Dokumenten findet sich durchgängige Groß- oder Kleinschreibung. Daraufhin werden durch `germanp.pl` mit Hilfe eines Lexikons und regulärer Ausdrücke diejenigen Token entfernt, die vollkommen unabhängig von der benutzten Sprache zu betrachten sind und sich negativ auf die Präzision auswirken könnten (z. B. URLs, E-Mail-Adressen, Abkürzungen, verschiedene Sonderzeichen und Eigennamen).[21] In der folgenden Verarbeitungsstufe werden für alle deutschsprachigen und alle unbekannten Token Frequenzlisten erstellt, wobei nur 10 000 zufällig ausgewählte Token betrachtet werden, damit auch umfangreiche Texte effizient verarbeitet werden. Das Lexikon umfasst etwa 156 000 Einträge und wurde halbautomatisch auf seine Korrektheit überprüft: Einerseits wurde die Wortliste mit Hilfe anderssprachiger Listen abgeglichen, d. h. Einträge, die sowohl in einer anderen als auch in der deutschsprachigen Liste enthalten waren, wurden entfernt. Andererseits wurden große Teile der Liste manuell überprüft.[22] Sobald die Frequenzen der deutschsprachigen und unbekannten Wörter vorlie-

[21] Aus einem Projekt zur Eigennamenerkennung standen Namenslisten zur Verfügung, die hierfür eingesetzt wurden (vgl. Kober et al., 1998). Die Datenquellen zum Aufbau von Eigennamenlexika basieren unter anderem auf einer Telefonbuch-CD ROM sowie auf der extrem „reine“ Daten enthaltenden „Normdaten-CD ROM für Personennamen und Schlagwörter“ der Deutschen Bibliothek (PND, 1996).

[22] Die Wortliste basiert unter anderem auf dem Synonymlexikon des Systems LEU/2 (vgl. Emde, 1991), den von Heinz Knutzen zusammengestellten Wortlisten (z. B. erhältlich unter `ftp://ftp.th-darmstadt.de/pub/dicts/ispell/dictionaries/hk-deutsch.tar.gz`) für die frei verfügbare Rechtschreibhilfe *ispell* sowie auf ca. 60 000 Wörtern, die in Dokumenten enthalten waren, die in einer Vorstudie akquiriert wurden.

gen, werden die Häufigkeiten bezüglich der Anzahl der *Wörter* und in Bezug auf die *Zeichen*[23] eines Wortes in Beziehung zur Länge des Textes gesetzt und der prozentuale Anteil als *deutsch* erkannter Token ermittelt. Dieser beträgt für sehr viele der verarbeiteten Dokumente zwischen 85 und 100%. Die Festlegung des Schwellwertes gestaltete sich schwierig, da verschiedene Grenzfälle – etwa sehr viele fachsprachliche Termini, englischsprachige Navigationshilfen etc. – den Anteil erkannter Wörter teilweise erheblich reduziert haben. Aus diesem Grund arbeitete `germanp.pl` während der Datensammlung mit einem Schwellwert von 40% (in der Literatur wird verschiedentlich für vergleichbare Verfahren ein Wert von 50% vorgeschlagen). Dieser ist klein genug, um Grenzfälle korrekt zu klassifizieren, und er ist groß genug, um Dokumente unbekannter Sprachen nicht in das Korpus aufzunehmen.

Evaluation von germanp.pl

Zur Evaluation wurden aus einer Liste der 156 926 Webdokumente, die am 29./30. Dezember 2000 innerhalb von `*.uni-giessen.de` angeboten wurden, 150 URLs zufällig ausgewählt. Von diesen 150 Dokumenten werden 144 korrekt klassifiziert, 97 als *deutsch* und 47 als *unbekannt*. Ein Dokument konnte nicht klassifiziert werden, da es nicht vom Medientyp `text/html` war, wodurch sich ein Recall von $R = 99,3$ und eine Precision von $P = 96,64$ ergibt.[24] Bei den fünf falsch klassifizierten Dokumenten lag in vier Fällen Bilingualität im weitesten Sinne vor, d. h. die Dokumente enthielten sowohl deutschsprachige als auch ähnlich viele anderssprachige Token. Abbildung 7.4 zeigt zwei dieser Dokumente, bei denen eine objektive Zuordnung zu einer einzigen Sprache auch von einem menschlichen Klassifizierer nur schwer durchzuführen ist. Derartige multilinguale Dokumente mit ausgewogenen Verteilungen der beteiligten Sprachen kann der Erkenner – bedingt durch den Lexikon-basierten Ansatz – nicht korrekt klassifizieren. Das fünfte Dokument umfasst sieben Token, die die Beschreibung eines Fotos darstellen und fast vollständig als fachsprachlich bezeichnet werden können, so dass auch in diesem Fall die Erkennung fehlschlägt, da sie nicht in der Wortliste enthalten sind.[25] Es zeigt sich, dass auch dieser Ansatz für extrem kurze – weniger als etwa 10–12 Wörter umfassende – fachsprachliche Dokumente ungeeignet ist (vgl. die Annahmen von Langer, 2002, zu Standardtexten, die mindestens 20 Wörter enthalten sollten).

Bei der Datensammlung kann ein Problem entstehen: Falls ein HTML-Dokument, dessen Sprache als *unbekannt* klassifziert wurde, Hyperlinks auf Dokumente enthält, die als *deutsch* klassifiziert werden, existiert keine korpusinterne Verknüpfung der deutschsprachigen Dokumente: Sie werden zwar in das Korpus integriert, doch fehlt die Webseite, die die Verknüpfungen beinhaltet. Für den Datenbank-getriebenen Zugriff ist dieses Problem irrelevant. Der direkte Zugriff über die Web-Oberfläche ist jedoch nicht unmittelbar möglich. Ein *PHP*- oder *Perl*-Skript könnte zur Überprüfung der Verknüpfungen eingesetzt werden.

[23] Damit ein positiver Treffer wie „Abgeordnetenentschädigungsgesetzes“ stärker gewichtet wird als z. B. „absolut“. Hierdurch wird die Trennschärfe des Algorithmus drastisch gesteigert.

[24] Vom Webserver bezüglich des HTTP-Response-Header-Feldes `Content-Type` fehlerhaft markierte Dokumente werden bei der Korpuserstellung *nicht* an das Werkzeug zur Identifizierung der Sprache weitergereicht.

[25] Die Token lauten: „Landwirtschafts GmbH Aschara Rapspreßkuchen in der Biodieselanlage“. Die Abkürzung „GmbH“ wird von `germanp.pl` getilgt, so dass 11,01% als *deutsch* erkannter Token berechnet werden, da sich nur „der“ im Lexikon befindet; „in“ ist aufgrund von Überschneidungen nicht in der Liste enthalten; „Aschara“ ist nicht Teil der Eigennamenliste (vgl. Mikheev et al., 1999).

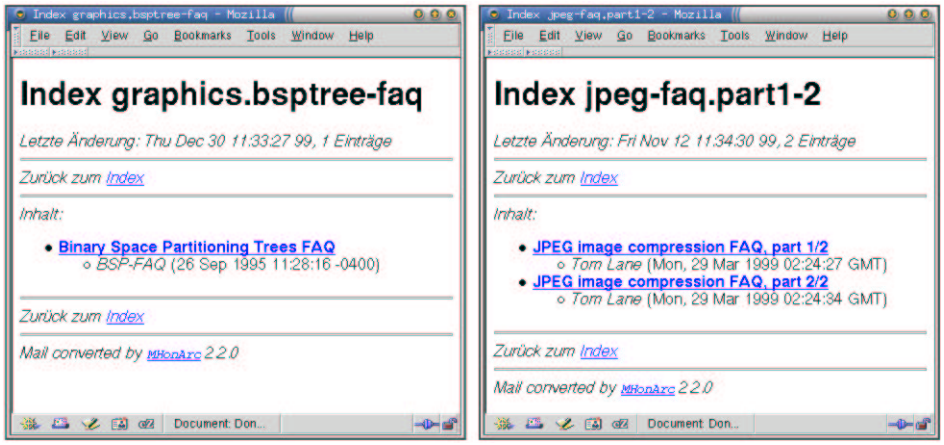


Abbildung 7.4: Zwei von `germanp.pl` fehlerhaft klassifizierte Dokumente

Verwandte Arbeiten

Automatische Sprachenidentifizierer werden in Suchmaschinen, als Bestandteile von Vorverarbeitungsketten in natürlichsprachlichen Systemen und für den Korpusaufbau eingesetzt. Der von Cowie et al. (1998) präsentierte *Crawler* baut sprachspezifische Kollektionen auf. Hierzu wird er mit einer oder mehreren URLs sowie der gewünschten Sprache konfiguriert. Der *Crawler* tilgt das HTML-Markup und leitet den Text an einen Sprachenidentifizierer weiter. Falls eine Übereinstimmung mit der gewünschten Sprache vorliegt, wird der Text gespeichert. Hyperlinks werden gesammelt und sukzessive von dem *Crawler* besucht. Der Sprachenidentifizierer unterscheidet 34 Sprachen und arbeitet mit einem statistischen n-Gramm-Verfahren. Es wurden monolinguale Korpora von bis zu 50 Megabyte Größe für das Türkische, Arabische und Russische erzeugt. Ein weiterer Aspekt betrifft die Zusammenstellung von Korpora für Minderheitensprachen (vgl. Constable und Simons, 2000). Jones und Ghani (2000) stellen Verfahren vor, die – ausgehend von einem in Tagalog, der auf den Philippinen verbreitetsten Sprache, geschriebenen Dokument – automatisch Suchanfragen konstruieren, so dass weitere in Tagalog verfasste Webseiten gefunden werden. Der Algorithmus arbeitet auf der Wortebene, wobei Unigramme aus dem Ausgangsdokument gewonnen und in den nachfolgenden Anfragen an verschiedene Suchmaschinen weiter benutzt werden. Die von Jones und Ghani (2000) dargestellten Verfahren sind insbesondere in Verbindung mit der automatischen Inhaltsanreicherung von Belang, da linguistisch annotierte Korpora für Minderheitensprachen kaum erhältlich sind.

Der zweite Schwerpunkt betrifft die Integration von Sprachenidentifizierern in Suchmaschinen sowie IR- und IE-Anwendungen (vgl. auch Abschnitt A.6). Hierdurch wird dem Benutzer die Möglichkeit gegeben, eine Suchanfrage auf Webseiten zu beschränken, die in einer bestimmten Sprache verfasst wurden oder um diejenigen Dokumente zu indexieren, die in bestimmten Sprachen verfasst worden sind (vgl. Benczúr et al., 2003). Langer (2001) stellt den Sprachenidentifizierer *Elexir* der Suchmaschine *AllTheWeb* vor, der 45 Sprachen unterscheidet und mit einem hybriden Verfahren arbeitet: Für 41 Sprachen werden Lexika benutzt, zur Identifizierung der vier verbleibenden Sprachen werden n-Gramm-Modelle

eingesetzt, da in diesen Sprachen (Japanisch, Chinesisch, Koreanisch und Thai) Wortgrenzen nicht markiert sind; es wird ein Recall von $R = 96$ und eine Präzision von $P = 100$ angegeben. Capstick et al. (2000) setzen ebenfalls eine Komponente zur automatischen Sprachenidentifizierung ein. Das System gestattet es Anfragen zu stellen, woraufhin relevante Dokumente in die Muttersprache des Anwenders übersetzt werden. Zunächst werden HTML-Dokumente mit Hilfe eines *Crawlers* gesammelt, woraufhin eine Identifizierung der Sprache stattfindet. Der Algorithmus basiert auf dem von Cavnar und Trenkle (1994) vorgeschlagenen Verfahren, wobei n-Gramme zwischen einem und fünf Zeichen Länge benutzt werden. Die 300 häufigsten n-Gramme aus einem 40 Sprachen umfassenden Trainingskorpus bilden die Basis der Sprachmodelle. Zu verarbeitende Dokumente werden auf ihre häufigsten n-Gramme reduziert und mit den vorhandenen Modellen verglichen. In einer Evaluation bestätigen Capstick et al. die Ergebnisse anderer Autoren: Der Algorithmus operiert bei Dokumenten mit 21 oder mehr Wörtern mit einer Präzision von beinahe konstant $P = 100$, die jedoch bei Dokumenten mit weniger Wörtern drastisch abnimmt.

7.2.3 Aufbau der Korpusdatenbank

Die im Korpus enthaltenen Dokumente befinden sich im *Linux*-Dateisystem. Metainformationen werden in einer relationalen Datenbank gepflegt, was eine große Flexibilität bezüglich des Datenzugriffs gewährleistet (vgl. die Abschnitte 7.3 und 7.4).[26] Abschnitt 7.2.4 diskutiert die Aufbereitung der von *Pavuk* erzeugten HTTP-Response-Header-Dateien für den Import in die Datenbanktabellen, die nachfolgend erläutert werden.

HTTP-Response-Header

In der Tabelle `http_header` befinden sich Informationen über *alle* im Zuge eines *Crawls* besuchten Dateien. Die Struktur der Tabelle wurde auf Basis der Definition des HTTP-Response-Headers modelliert, da sie mit derartigen Daten gefüllt wird (vgl. Abschnitt 7.2.4). Es werden nur diejenigen Informationen abgelegt, die für die vorliegende Arbeit relevant sind. Einige Felder des HTTP-Response-Headers (z. B. `Warning` und `Upgrade`) wurden gekürzt übernommen. Die Feldinhalte, die korrespondierenden Datentypen sowie ein Beispieldatensatz sind in Tabelle 7.2 dargestellt. Die beispielhaft aufgeführte Datei ist nicht im Korpus verfügbar, andernfalls enthielte das Feld *Korpus-Datei* eine lokale Pfadangabe, die ihren Speicherort im Korpus angäbe (Tabelle B.1, S. 783, enthält die tatsächlichen Feldnamen).

Webserver-Daten

Die Tabelle `server_info` nimmt Daten auf, die sich auf einen Webserver beziehen. Es wird sein Name, der benutzte HTTP-Port, der Webserver-Typ (d. h. die eingesetzte Webserver-Software), die HTTP-Version (diese Felder sind ebenfalls Bestandteile des HTTP-Response-Headers, vgl. Abschnitt 7.2.4) sowie die jeweilige Universität gespeichert. Tabelle 7.2 stellt die Felder sowie Beispieldaten dar (vgl. Tabelle B.2, S. 784).

[26] Eingesetzt wird die unter der GNU GPL verfügbare relationale Datenbank *MySQL* in der Version 3.22.32 (vgl. `http://www.mysql.com` sowie DuBois, 1999).

http_header

Feld	Datentyp	Beispiel
ID	int(10)	72926
Partieller URI	text	/~g91063/ps/textannotation.ps.gz
Korpus-Datei	text	
Verweis zum Server	int(10)	8
HTTP-Status-Code	smallint(5)	200
Content-Length	int(10)	156432
Content-Type	varchar(30)	postscript
Content-Encoding	tinytext	x-gzip
Content-Language	tinytext	
Content-Location	tinytext	
Location	tinytext	
Date	datetime	2001-01-16 15:48:19
Expires	datetime	
Last-Modified	datetime	1999-07-28 20:41:38
WWW-Authenticate	*boolean*	0
Cache-Control	*boolean*	0
Content-MD5	*boolean*	0
Pragma	*boolean*	0
Set-Cookie	*boolean*	0

server_info

ID	int(10)	8
Server-Name	varchar(255)	www.uni-giessen.de
Port	mediumint(5)	80
Server-Typ	varchar(255)	Apache/1.3.14 (Unix) Frontpage/3.0.4.2 PHP/4.0.4 mod_ssl/2.7.1 OpenSSL/0.9.6
HTTP-Version	varchar(5)	1.1
Universität	char(3)	GI

Tabelle 7.2: Die Tabellen http_header und server_info

Generierung und Verwaltung von Stichproben

Mittels der Web-Oberfläche der Korpusdatenbank können Stichproben generiert werden, indem in einem HTML-Formular Parameter gesetzt werden, woraufhin die Datenbank eine Stichprobe nach einem Zufallsverfahren berechnet. An der Speicherung und Analyse von Stichproben sind mehrere Datenbanktabellen beteiligt. Die Tabelle sample enthält Metadaten über Dokumente, die Teil einer oder mehrerer Stichproben sind (vgl. Tabelle B.8, S. 785). Neben einer fortlaufenden Nummer wird die ID der Stichprobe abgelegt, um Tabellenzuordnungen zu ermöglichen. Um effizient auf die zentralen Informationen eines Dokuments zugreifen zu können, wird ebenfalls dessen ID (bezogen auf *ID* in http_header), die ID des Webservers (bezogen auf *ID* in server_info), die URL sowie die Universität notiert.

Die Tabelle meta_sample enthält Informationen, die bei der Generierung von Stichproben automatisch angelegt werden (vgl. Tabelle B.5, S. 784). Neben der ID besitzt jede Stichprobe einen Titel, einen Zeitstempel, die Anzahl der enthaltenen Dokumente, die Anzahl der in der Datenbank verfügbaren Dokumente, die auf die SQL-Query zutreffen, die die Stichprobe zufällig generiert hat, sowie eine Liste der in der Korpusdatenbank verfügbaren

Universitäten. Weiterhin skizziert eine Beschreibung den Verwendungszweck der Stichprobe. Die SQL-Query, die eine Stichprobe erzeugt hat, wird zu Protokollzwecken ebenfalls notiert. Abschließend folgt eine Liste von Benutzern, die Schreibzugriff auf diese Daten haben.

Analyseschemata werden in der Korpusdatenbank als *Templates* bezeichnet. Eine Stichprobe kann zu Analysezwecken auf eines oder mehrere Templates abgebildet werden, so dass die Web-Oberfläche ein zugehöriges HTML-Formular für die manuelle Auswertung präsentiert. Diese Zuordnung kann mittels der Web-Oberfläche vorgenommen werden und wird in der Tabelle `sample_template` gespeichert (vgl. Tabelle B.7, S. 785).

Analyse von Dokumenteigenschaften – Verwaltung von Templates

In einer Analyse erhobene Dokumenteigenschaften werden in Template-bezogenen Tabellen abgelegt. Die Tabelle `meta_template` enthält – ähnlich wie `meta_sample` – allgemeine Angaben zu einzelnen Analyseschemata (vgl. Tabelle B.6, S. 785). Auch die Ergebnisse einer Analyse werden in der Datenbank gespeichert. Hierfür existieren vier Tabellen (`template1` bis `template4`; vgl. die Tabellen B.9, B.10, B.11 und B.12). An dieser Stelle wird lediglich `template1` erläutert: Die Web-Oberfläche ist auf verschiedene Protokollinformationen angewiesen, weshalb die Felder *ID*, *Template-ID*, *Sample-ID*, *Dokument-ID*, *Benutzer* und *Datum* in jeder Template-Tabelle vorhanden sein müssen. Die analysierbaren Eigenschaften in `template1` beziehen sich auf die Hypertextsorte (`hts`), einen Kommentar (`comment`) und die Eigenständigkeit eines Dokuments (`standalone`). Für jedes Template wurden in der Web-Oberfläche entsprechende Funktionen in den zugehörigen *PHP*-Skripten implementiert.

Universitätsbezogene Metadaten

In der Tabelle `universities` werden universitätsbezogene Metadaten gespeichert, die für Präsentations- und Auswahlmöglichkeiten innerhalb der Web-Oberfläche und zur automatischen Generierung von Statistiken benötigt werden (vgl. Tabelle B.3, S. 784).[27] Neben der korpusinternen Abkürzung der Institution wird der generelle Universitätstyp, die Adresse der Einstiegsseite, der offizielle Name sowie der oder die Standorte vermerkt. Das Feld *Plattenplatz* enthält die Angabe des Festplattenspeichers, der für die Ablage der Daten im Korpus eingesetzt wird.[28] Das Feld *Vollständiger Umfang* gibt den tatsächlichen Umfang *aller* von *Pavuk* besuchten Dateien (deren Dateigrößen in den HTTP-Response-Headern gespeichert werden) der jeweiligen Domäne an. Die Informationen, die in dieser Tabelle enthalten sind, wurden automatisch aus den von `build-db-import.pl` (vgl. Abschnitt 7.2.4) generierten Statistiken gewonnen und daraufhin von einem *Perl*-Skript in die Datenbank importiert.

Benutzerdaten

Die Tabelle `user` enthält Informationen über einzelne Benutzer und wird für die Authentifizierung gegenüber der Web-Oberfläche eingesetzt. Die Benutzerinformationen werden auch

[27] Die Tabellen in Anhang E auf der beiliegenden CD ROM wurden von einem *Perl*-Skript generiert, das per Datenbank-API unter anderem die in der Tabelle `universities` enthaltenen Datensätze ausgelesen hat.

[28] Diese Angabe bezieht sich auf den Wert, den das Kommando `du` (*disk usage*) für die jeweilige Verzeichnishierarchie meldet. Der tatsächliche Wert ist aufgrund der ebenfalls enthaltenen dateisystembezogenen Verwaltungsinformationen niedriger anzusetzen.

bei der Analyse von Stichproben verwendet: Der Benutzer, der eine Stichprobe anlegt kann anderen Benutzern die Bearbeitung einer Stichprobe gestatten (vgl. Tabelle B.4, S. 784).

7.2.4 Aufbereitung der HTTP-Response-Header für den Import in die Korpusdatenbank

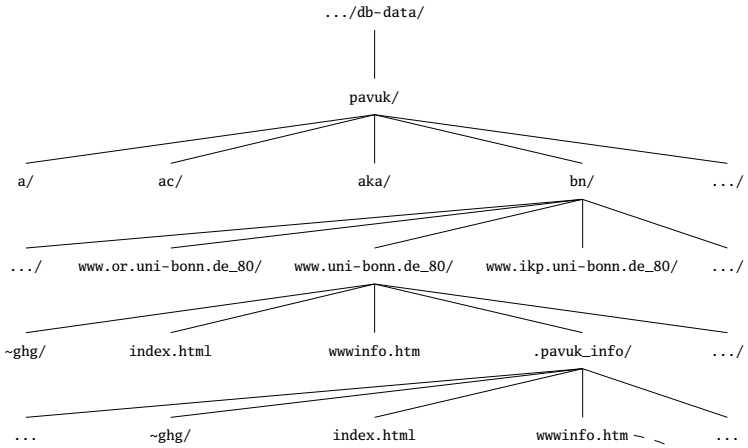
Der *Crawler Pavuk* sammelt Dokumente, die in das Korpus integriert werden sollen. Für jede mittels HTTP besuchte Datei (dies gilt auch für diejenigen Dateien, die *nicht* in das Korpus integriert werden) legt *Pavuk* eine Art Metadatei an, die den zugehörigen HTTP-Response-Header (RFC 2616) enthält und sich in Unterverzeichnissen der gespiegelten Verzeichnisstruktur befindet. Die HTTP-Response-Header, die durch die Sammlung der Daten einer Universität entstehen, werden von dem *Perl*-Skript `build-db-import.pl` verarbeitet. Dieser *Finite State Transducer* stellt sicher, dass die jeweiligen Daten den korrespondierenden Datentypen der relationalen Datenbank entsprechen und überführt die enthaltenen Informationen in ein Format, das mit Hilfe des Werkzeugs `mysqlimport` in die *MySQL*-basierte Korpusdatenbank importiert werden kann (vgl. Abbildung 7.2). Im Folgenden wird die Funktionsweise von `build-db-import.pl` erläutert, wobei insbesondere Probleme thematisiert werden, die durch HTTP-Response-Header hervorgerufen werden, die keine Konformität zum HTTP-Standard (RFC 2616) aufweisen. Gerade derartige fehlerhafte Header haben es immer wieder notwendig gemacht, umfangreiche Änderungen und Erweiterungen an dem *Perl*-Skript vorzunehmen, um vorliegende Daten erfolgreich verarbeiten zu können.

Ablage der Daten durch Pavuk

Pavuk erstellt bei der Datensammlung auf dem Korpusdatenbank-Server eine Verzeichnishierarchie zur Aufnahme der Korpusdaten, die diejenigen Dateien umfassen, die den in Abschnitt 7.2 dargestellten Beschränkungen genügen. Auf der obersten Ebene dieser Hierarchie befindet sich – für jede im Korpus enthaltene Universität – ein zentrales Verzeichnis, in dem wiederum Verzeichnisse für jeden zum Zeitpunkt der Datensammlung per HTTP erreichbaren Webserver enthalten sind. Der Name eines Webservers entspricht dabei dem Verzeichnisnamen, wobei der HTTP-Port als Suffix angehängt wird, um Eindeutigkeit zu gewährleisten. Die Verzeichnisstruktur reflektiert unmittelbar die ursprünglichen Strukturen und Inhalte der gespiegelten Webserver. Für alle besuchten Dateien werden in zusätzlichen Dateien, die *Pavuk* jeweils in dem Verzeichnis `.pavuk_info` hinterlegt, die HTTP-Response-Header einer Anfrage gespeichert (vgl. Abbildung 7.5). Ein Response-Header entsteht, wenn von einem HTTP-Client (z. B. einem Browser oder einem Werkzeug wie *Pavuk*) ein HTTP-Request erfolgt, der üblicherweise in der Anforderung einer Datei des Webservers besteht. Vor ihrer Auslieferung schickt der Webserver den HTTP-Response-Header an den Client, um z. B. über den Status der Auslieferung und den Medientyp zu informieren.

Die Struktur des HTTP-Response-Headers

Das *Hypertext Transfer Protocol* spezifiziert für den HTTP-Header 47 Felder. Ein Beispiel befindet sich in Abbildung 7.5 (die Zeile `Original-URL: ...` wird von *Pavuk* hinzugefügt und



Inhalt des HTTP-Response-Headers des Dokuments mit der URL `http://www.uni-bonn.de/wwwinfo.htm`

Abbildung 7.5: Die Verzeichnisstruktur der im Korpus enthaltenen Daten und ein Beispiel für einen HTTP-Response-Header

ist kein Bestandteil des eigentlichen Headers). Der HTTP-Response-Header umfasst potenziell zahlreiche Felder, die in RFC 2616 ausführlich diskutiert werden, weshalb an dieser Stelle nur eine knappe Erläuterung erfolgt. In der ersten Zeile des eigentlichen Headers wird der Status der Anfrage spezifiziert. Dieser besteht aus der Angabe des Protokolls (`HTTP`) und seiner Versionsnummer (`1.1`), einem Status-Code (`200`) und seiner natürlichsprachlichen Beschreibung („Reason Phrase“). Es existieren fünf Gruppen von Status-Codes:

- 1xx: *Informational* – Anfrage empfangen, Verarbeitung erfolgt
- 2xx: *Success* – Die Aktion wurde erfolgreich empfangen, verarbeitet und akzeptiert
- 3xx: *Redirection* – Zur Verarbeitung der Anfrage sind zusätzliche Aktionen notwendig
- 4xx: *Client Error* – Fehlerhafte Anfragesyntax oder Verarbeitung nicht möglich
- 5xx: *Server Error* – Der Server kann eine offenbar gültige Anfrage nicht verarbeiten

Der übliche Status-Code einer Anfrage lautet 200, d. h. der Request wurde erfolgreich verarbeitet. Weitere Status-Codes sind etwa 403 („Forbidden“), 404 („Not found“) oder 500 („Internal Server Error“).[29] In der zweiten Zeile folgt ein Zeitstempel, der eine Datumsangabe des Webservers darstellt und laut RFC 2616 in einem von drei unterschiedlichen Formaten

[29] Die sehr häufig auftretenden Status-Codes 200 und 404 sind in der Korpusdatenbank mit 14 754 739 bzw. 891 585 Vorkommen vertreten (vgl. Tabelle A.6, Abschnitt A.3.8, S. 734).

vorliegen muss (entweder nach dem Standard für elektronische Post, RFC 0822, RFC 1123, RFC 1349 und RFC 2181, dem Standard für Usenet-Nachrichten, RFC 0850, RFC 1036 oder dem Format, das die C-Funktion `asctime()` liefert). Die folgende Angabe identifiziert die auf dem Webserver eingesetzte Software, im Beispiel handelt es sich um ein Produkt der Firma IBM, das auf dem Server *Apache* aufsetzt. In der vierten Zeile ist das Datum der letzten Änderung der angefragten Datei vermerkt; für diesen Zeitstempel gelten ebenfalls die oben angesprochenen Formatrestriktionen. Daraufhin folgt die Angabe eines Entity Tags, das für Caching-Zwecke eingesetzt werden kann. Das Feld `Accept-Ranges` kann in Verbindung mit weiterführenden Requests benutzt werden, um nur Teile einer angefragten Datei zu transferieren (im Beispiel könnte dies byteweise erfolgen). Das Feld `Content-Length` gibt die Länge der angefragten Datei in Bytes an. Das letzte Feld – die Angabe des Medientyps – spielt, in Verbindung mit dem Feld `Content-Encoding`, für den Browser eine entscheidende Rolle, da diese Angabe bestimmt, ob und wie eine bestimmte Datei dekodiert und dargestellt werden kann. Im Beispiel lautet der Wert `text/html`, es handelt sich also um ein Dokument, das mit Hilfe von HTML ausgezeichnet wurde. Die benötigte Zuordnung von Dateisuffixen (z. B. `.html`, `.htm` oder `.shtml`) auf Medientypen wird in Form einer Konfigurationsdatei durchgeführt und findet ausschließlich auf der Seite des Webservers statt. Der Browser erkennt den Dateityp daher nicht an der Endung, sondern anhand der `Content-Type`-Angabe und kann bei Bedarf entsprechende Hilfsprogramme starten, um z. B. ein Dokument im *Portable Document Format* (`application/pdf`) anzeigen zu können.

Funktionsweise von build-db-import.pl

Nach der Traversierung einer Domain wie z. B. `.uni-bonn.de` liegt eine gespiegelte Verzeichnisstruktur vor (vgl. Abbildung 7.5). Der *Finite State Transducer* `build-db-import.pl` wird nun mit verschiedenen Informationen konfiguriert (Einstiegspunkt in die Verzeichnishierarchie, Datenbank-interne Abkürzung der Universität, Offset in die Tabelle `server_info`), um den Dateibaum zu traversieren. Das Skript verarbeitet die HTTP-Response-Header und erzeugt drei Dateien, von denen zwei für den Datenbankimport in die Tabellen `http_header` sowie `server_info` eingesetzt werden; die dritte Datei enthält statistische Informationen. Obwohl eine Verarbeitung der in RFC 2616 spezifizierten Header prinzipiell mit einfachen Mitteln realisierbar ist, sind in dem *Perl*-Skript zahlreiche Ausnahmeregeln enthalten, die eine robuste Verarbeitung gewährleisten. So werden Plausibilitätsüberprüfungen durchgeführt, um im Falle eines Fehlers darauf hinweisen zu können, dass die Integrität der für den Import bestimmten Daten gefährdet sein könnte. Die Implementierung dieser Tests wurde notwendig, nachdem sich herausgestellt hat, dass zahlreiche Webserver fehlerhaft formatierte bzw. nicht RFC 2616-konforme Header-Zeilen versenden. Aufgrund dessen wurden Heuristiken implementiert, die Header-Zeilen trotz enthaltener Fehler interpretieren können.

Das Skript basiert auf einer Traversierung der Verzeichnishierarchie, die im Basisverzeichnis einer Universität beginnt (z. B. `bn/` in Abbildung 7.5). Von dort werden rekursiv alle Dateien verarbeitet, die sich in den `.pavuk_info`-Verzeichnissen befinden. Daraufhin wird eine Datei zeilenweise eingelesen und verarbeitet. Da die in Header-Feldern enthaltenen Informationen in Datenbanktabellen integriert werden, die wiederum pro Feld einen bestimmten Datentyp aufweisen, garantieren spezielle Funktionen, dass eine Kompatibilität des extrahier-

ten Wertes mit dem jeweiligen Datentyp vorliegt (vgl. Abschnitt 7.2.3). Weitere Funktionen erledigen die robuste Verarbeitung von Datums- und Zeitausdrücken, die an mehreren Positionen von Bedeutung sind. Da Informationen über Webserver in einer separaten Tabelle abgelegt werden, enthält das Skript die Funktion, einen Webserver, der noch nicht in der assoziierten Hash-Tabelle enthalten ist, dort aufzunehmen, so dass er nach einem Durchlauf des Skripts in den für die Tabelle `server_info` bestimmten Daten enthalten ist.

Das Skript umfasst etwa 1 500 Zeilen *Perl*-Code. Es werden die Module `Time::localtime` (zur Darstellung von Zeitstempeln), `File::Find` (zur Traversierung von Verzeichnissen) und `URI` (zum Parsing von *Uniform Resource Identifier*s und *Uniform Resource Locator*s) eingesetzt. Oftmals wurde es für die 100 in der Korpusdatenbank enthaltenen Universitäten mehrfach eingesetzt, um fehlerhafte Header-Zeilen zunächst identifizieren zu können und nach einer Anpassung durch die Software korrigieren zu lassen (vgl. Abbildung 7.1). Hiervon waren ca. 40 Modifikationen betroffen, z. B. die Adaptierung an Datumsformate, die nicht den drei vorgegebenen Schemata entsprechen, oder an fehlerhafte Jahresangaben (`:0` statt `2000`, `:1` statt `2001` oder `19100` statt `2000`), die auf Software zurückzuführen ist, die nicht Jahr-2000-sicher ist (z. B. die Webserver *CERN* 3.0 und *NCSA* 1.3). Es treten auch Werte auf, die laut RFC 2616 nicht existieren dürfen, so schickt z. B. der *Internet Information Server* (*IIS*) 3.0 den Wert `now` für `Expires`, obwohl eine Datumsangabe nach RFC 1123 vorgesehen ist.

Nach der Transformation der HTTP-Response-Header, dem Import in die Datenbank und der Erstellung von Sicherheitskopien werden diese Dateien gelöscht, da die `.pavuk_info`-Verzeichnisse nicht länger benötigt werden. Vor der Generierung der Import-Dateien werden auch Verzeichnisse gelöscht, die zum Zeitpunkt der Datensammlung zwar per HTTP erreichbar waren, deren eigentlicher Zweck aber nicht darin besteht, die enthaltenen Dateien im WWW zu veröffentlichen. Hierzu zählen insbesondere Verzeichnisse namens `_vti_bin/`, `_vti_cnf/` oder `_vti_pvt/`, die Verwaltungsinformationen von *Microsoft Frontpage* beinhalten.[30] Ein weiterer Aspekt zur Reduktion des benötigten Plattenplatzes betrifft die Tilgung von Duplikaten, die im folgenden Abschnitt thematisiert wird. Insgesamt wurden aus den vorliegenden HTTP-Response-Headern von `build-db-import.pl` 16 196 511 zu importierende Datensätze für die Tabelle `http_info` und 14 968 Einträge für `server_info` generiert.

7.2.5 Behandlung von Duplikaten

Sobald große Mengen von HTML-Dokumenten gesammelt werden, tauchen zwangsläufig identische Kopien verschiedener Dokumente auf, z. B. Teile der Dokumentation der Sprache *Java* oder *Self HTML*, eine Anleitung zum Umgang mit der *Hypertext Markup Language*.[31] Da Duplikate unnötige Datenträgerkapazitäten einnehmen, können Verfahren eingesetzt werden, um sie während oder nach der Datensammlung zu erkennen und zu tilgen.

Beim Aufbau des Korpus wurden Duplikate nach der vollständigen Sammlung der Daten einer Universität aus dem Datenbestand entfernt. Die einfachste Methode ist die Traversierung der Verzeichnishierarchie, um Dateien mit identischen Namen und identischen Größen Byte für Byte auf identischen Inhalt zu untersuchen. Hierzu wird das UNIX-Werkzeug

[30] Hiervon waren beispielsweise innerhalb von `*.uni-giessen.de` mehr als 123 000 Dateien betroffen.

[31] Siehe auch Kilgarriff (2001): „[The Web] contains duplicates, near duplicates, documents pointing to duplicates that may not be there, and documents that claim to be duplicates but are not.“

`dupmerge`[32] eingesetzt, das von der Standardeingabe eine Liste von Dateien erwartet, wie sie etwa von `find` produziert wird, um daraufhin identische Dateien zu lokalisieren. Sobald eine identische Kopie einer Datei gefunden wird, legt `dupmerge` einen harten Link[33] an, der das ursprüngliche Duplikat ersetzt und auf das verbleibende Original verweist, so dass das Duplikat weiterhin adressierbar ist. Der eigentliche Vorteil dieses Werkzeugs ist seine Effizienz, da es selbst umfangreiche Datenbestände von mehr als 100 000 Dateien in kurzer Zeit verarbeiten kann; andere frei verfügbare Lösungen arbeiten in der Regel deutlich langsamer. Insgesamt hat `dupmerge` auf diese Weise Duplikate im Umfang von ca. zwei Gigabyte gelöscht.

Verwandte Arbeiten

Die Erkennung von Duplikaten wird bei der Erstellung von Korpora aus Webseiten und insbesondere bei der Aufbereitung von Kollektionen durch Suchmaschinen angewendet (um die Präsentation ähnlicher Dokumente zu vermeiden, d. h. das Ranking zu verbessern): Cooper et al. (2002) stellen ein Verfahren vor, das auf der Berechnung von Frequenzen basiert, deren Grundlage Phrasen darstellen, die von einem Textmining-System aus einer Kollektion extrahiert werden. Mit der Zuweisung einer Art „Fingerabdruck“, der durch die Addition einfacher Hash-Funktionen – angewendet auf die Wörter eines Dokuments – entsteht, werden grobe Einschätzungen bezüglich der Ähnlichkeit ermöglicht. Durch den tatsächlichen Vergleich der jeweiligen Phrasen zweier Dokumente kann deren exakter Unterschied bestimmt werden, um aktuelle Revisionen eines Textes oder auch Plagiate zu identifizieren, d. h. dieses Verfahren kann nicht nur exakte Kopien erkennen, sondern auch beinahe identische oder auch sehr ähnliche Dokumente voneinander differenzieren. Cho et al. (2000) beschreiben ein Verfahren, das während der Datensammlung arbeitet und diesen Prozess um 40% reduziert (vgl. auch Chakrabarti, 2003, S. 29). Hierbei steht insbesondere die Erkennung sehr umfangreicher Kollektionen (z. B. das *Linux Documentation Project*, LDP, oder die „Java 1.0.2 API Documentation“) im Vordergrund, da ihre Erkennung während des *Crawls* diesen Prozess erheblich beschleunigt: Die LDP-Seiten besitzen einen Umfang von ca. 25 Megabyte und sind laut Cho et al. auf etwa 180 Webservern weltweit verfügbar, d. h. der *Crawler* vermeidet den Download von insgesamt 4,5 Gigabyte Daten.

Heydon und Najork (1999) gehen unter anderem auf *Crawler*-Fallen ein, zu denen URL-Aliase und Session-IDs gehören und die ebenfalls zur Vermeidung von Duplikaten relevant sind: Session-IDs werden vom Webserver dynamisch generiert und in einer URL hinterlegt, um Navigationspfade protokollieren zu können. Auf diese Weise entstehen für ein Dokument potenziell unendlich viele URLs, die *Mercator* anhand von Fingerabdrücken unterscheiden kann. Diese werden sowohl für den Inhalt eines Dokuments als auch für dessen URL mit einem MD5-ähnlichen Verfahren erzeugt (vgl. Abschnitt 7.3.1). Derartige Fingerabdrücke können auch benutzt werden, um die vier unterschiedlichen Formen von URL-Aliasen verarbeiten zu können. Zwei URLs stehen in einer Alias-Beziehung, wenn sie auf das gleiche

[32] Implementiert von Phil Karn und unter der GNU GPL verfügbar: `http://www.ka9q.net/code/dupmerge/`.

[33] Im Gegensatz zu einem symbolischen Link, der prinzipiell quer über alle Verzeichnisstrukturen zeigen darf. Harte Links werden nicht über einen Pfadnamen realisiert, sondern über eine identische inode-Angabe (unter UNIX die Angabe des unmittelbaren Adressbereiches auf einer Festplatte), weshalb harte Links nur innerhalb einer einzigen Partition angelegt werden können. Diese Einschränkung trifft auch auf `dupmerge` zu.

Dokument verweisen: (i) Mit Hilfe des *Domain Name System*s (DNS, RFC 1034, RFC 1035) können *mehrere* symbolische Namen für *eine* IP-Nummer angegeben werden. Heydon und Najork (1999) nennen das Beispiel der Rechnernamen `coke.com` und `cocacola.com`, die auf eine identische IP-Nummer zeigen, weshalb jedes Dokument dieses Webservers mindestens drei URLs besitzt. *Mercator* benutzt von derartigen URLs jeweils nur die kanonische Angabe (*canonical name* bzw. CNAME), die ebenfalls im DNS eingetragen werden kann oder die kleinste IP-Nummer. (ii) HTTP arbeitet in der Grundeinstellung auf Port 80, so dass bei Webservern, die auf diesem Port operieren, die Angabe der Portnummer nicht notwendig ist, jedoch trotzdem vorliegen kann, so dass ein Dokument zwei unterschiedliche URLs besitzt. *Mercator* fügt, ebenso wie *Pavuk*, den Default-Wert 80 hinzu. (iii) Mehrere URLs können auf das gleiche Dokument verweisen, indem z. B. auf Seiten des Webserver-Dateisystems symbolische Links eingesetzt werden, so dass eine Datei namens `index.html` auch als `home.html` adressierbar ist; in speziellen Fällen können hierdurch Zyklen verursacht werden. (iv) Weiterhin können Dokumente auf mehrere Webserver verteilt sein. Diese Problematik bezieht sich auf gespiegelte Sammlungen populärer Angebote wie das *Linux Documentation Project*. Heydon und Najork (1999) geben an, dass bei einem Test-*Crawl*, der acht Tage dauerte, insgesamt 8,5% aller heruntergeladenen Dokumente Duplikate waren.

7.2.6 Inhalt und Umfang des Korpus

Für die Untersuchungsdomäne (vgl. Kapitel 6) wurde ein Korpus angefertigt, das die deutschsprachigen Dokumente deutscher Universitäten und Hochschulen umfasst, um gewährleisten zu können, dass im Verlauf der Anfertigung dieser Arbeit eine statische, sich nicht verändernde Menge von Dokumenten als „sprachlicher Schnappschuss“ zur Verfügung steht.[34]

Das Korpus umfasst die zwischen Januar 2001 und September 2002 gesammelten HTML-Dokumente von insgesamt 100 Hochschulen (vgl. Anhang E auf der beiliegenden CD ROM). Die allgemeinen und technischen Universitäten sind vollständig enthalten, zusätzlich wurden die Webauftritte von 26 Musik- und Kunst-, Wirtschafts- und sonstigen Hochschulen in das Korpus integriert. Die Korpusdatenbank verzeichnet 14 968 Webserver und die Metadaten von insgesamt 16 196 511 per HTTP erreichbaren Dateien innerhalb der 100 universitären Webauftritte – bei 8 465 105 dieser Dateien handelt es sich um HTML-Dokumente. Das Korpus selbst umfasst eine Größe von ca. 40 Gigabyte und beinhaltet 4 294 417 Dateien. Neben 3 956 692 HTML-Dokumenten enthält die Kollektion 270 400 ASCII-Dateien, 35 651 CSS-Dateien, 25 871 XML-Dokumentinstanzen und 956 SGML-Instanzen. Anhang A beinhaltet eine detaillierte statistische Charakterisierung des Korpus.

[34] Diese Vorgehensweise lehnt sich an Roberts (1998, S. 85) an, der eine Analyse des gesamten *World Wide Web* vorschlägt, die jedoch mit zahlreichen Problemen verbunden ist (vgl. Kapitel 4): „One interesting direction to explore is a study of the entire WWW. A study of this undertaking would involve creating and maintaining an electronic corpus (or several inter-related corpora) and developing tools to efficiently analyze these data. The corpus would be a valuable archival source for future diachronic studies [...]. The data of the WWW are not permanent, a corpus is one way of ensuring a rich source of “historically valuable” data.“

7.3 Die Web-Oberfläche der Korpusdatenbank

Die Korpusdatenbank besteht aus drei Bausteinen. In der Datenbank werden Metadaten abgelegt, die von den besuchten Webservern für jede angefragte Datei in Form von HTTP-Response-Headern geliefert werden. Die Dokumente selbst werden im UNIX-Dateisystem gespeichert; in der Datenbanktabelle `http_header` befindet sich eine voll spezifizierte Pfadangabe für jede im Korpus enthaltene Datei (vgl. die Tabellen 7.2 und B.1). Der Datenbankserver ist mit dem Webserver *Apache* ausgestattet, wobei sich die im Korpus verfügbaren Dokumente in Verzeichnisstrukturen befinden, auf die der Webserver Zugriff hat.[35]

Als dritter zentraler Baustein der Korpusdatenbank existiert eine Web-Oberfläche für den manuellen Korpuszugriff. Abbildung 7.6 zeigt die indirekten (oben) und direkten (unten) Zugriffsmöglichkeiten auf die Korpusdatenbank sowie die vereinfachte Tabellenstrukturierung. Die in der Skript-Sprache *PHP* implementierte Oberfläche erlaubt dem Benutzer zahlreiche Zugriffsmöglichkeiten. Hierzu gehört z. B. die Suche nach Dokumenten oder Webservern, die Betrachtung und Visualisierung von Dokumenten, die Generierung, interaktive Analyse und Darstellung von Stichproben, die Korpusexploration und eine Benutzerverwaltung. Unter `http://hypnotic.germanistik.uni-giessen.de` ist die Web-Oberfläche für freigeschaltete IP-Nummern verfügbar. Es handelt sich um ein LAMP-System (vgl. z. B. Pollem, 1999, Behme, 2000b), da die Komponenten *Linux* (Betriebssystem), *Apache* (Webserver), *MySQL* (Datenbank) und *PHP* (Middleware) eingesetzt werden:

1. *Linux* – Die im Korpus verfügbaren Dokumente befinden sich im Dateisystem des *Linux*-basierten Korpusdatenbank-Servers; eingesetzt wird *Red Hat Linux* 6.2.
2. *Apache* – Der auf dieser Maschine installierte Webserver *Apache* hat lesenden Zugriff auf die Verzeichnishierarchien, in denen sich die Korpusdateien befinden, d. h. es besteht von Seiten des Benutzers Zugriff auf die Dokumente mittels HTTP.
3. *MySQL* – Metadaten über die Dokumente befinden sich in der *MySQL*-Datenbank, wodurch eine gezielte Zugriffssteuerung und Datenselektion ermöglicht wird.
4. *PHP* – Die Web-Oberfläche wurde in der Skript-Sprache *PHP*[36] realisiert, d. h. der Benutzer interagiert mit dem System über *PHP*-Skripte, die die Zugriffe auf die relationale Datenbank kapseln. Zur Betrachtung eines im Korpus enthaltenen Dokuments ist lediglich ein Redirect auf die lokale – auf den Webserver bezogene – URL notwendig. Da in den Dokumenten enthaltene Verweise auf Grafiken auf die entfernten Server umgeschrieben wurden, wird das Dokument vollständig dargestellt, falls die entfernten Dateien noch immer per HTTP verfügbar sind.

[35] Eingesetzt wird *Apache* in der Version 1.3.17. Dieser Webserver steht unter der von der *Open Source Initiative* (OSI, `http://www.opensource.org`) zertifizierten *Apache Software License* (`http://www.apache.org`).

[36] *PHP* (*PHP: Hypertext Preprocessor*, vgl. `http://www.php.net`) ist eine ursprünglich von Rasmus Lerdorf entwickelte Programmiersprache, die in HTML-Dokumente eingebettet werden kann und unter einer Open-Source-Lizenz verfügbar ist. Zur Laufzeit, d. h. unmittelbar nach der Anfrage eines derartig ausgezeichneten Dokuments bzw. Skripts von einem Browser, wird der enthaltene *PHP*-Code von einem speziellen Webserver-Modul ausgeführt, so dass dynamisch generierte Webseiten realisierbar sind. In der Web-Oberfläche des Korpusdatenbank-Servers wird *PHP* in der Version 4.0.4 Patchlevel 1 verwendet.

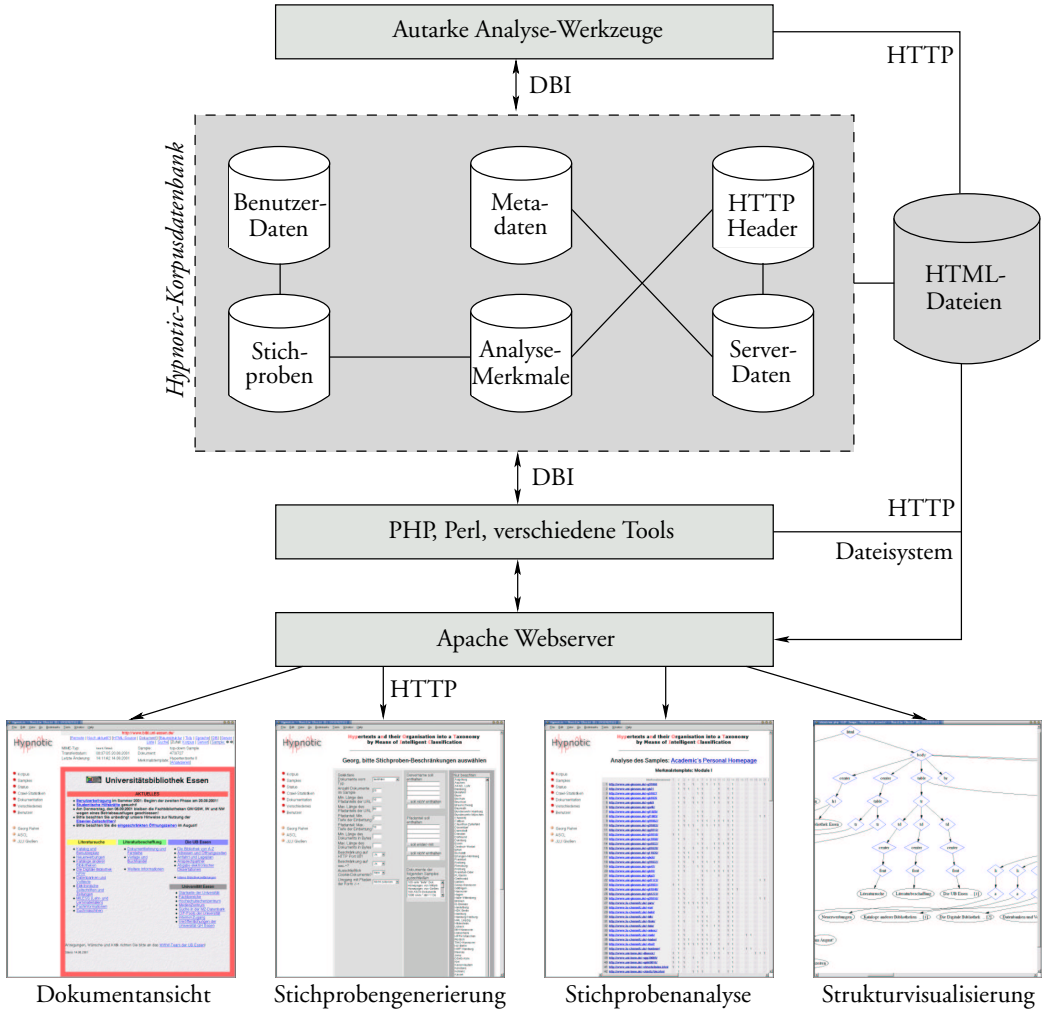


Abbildung 7.6: Zugriffsmöglichkeiten auf die Korpusdatenbank

Die in der Einstiegsseite der Web-Oberfläche enthaltene Navigationshilfe gestattet den Zugriff auf die wichtigsten Funktionen (Zugriff auf das Korpus; Generierung, Verwaltung und Analyse von Stichproben; generelle Statusinformationen; Zugriff auf statistische Informationen der einzelnen *Crawls*; Benutzerverwaltung). Nachfolgend werden die zentralen Funktionen und implementationsbezogene Aspekte thematisiert.

Die Web-Oberfläche besteht aus etwa 40 *PHP*-Skripten, die ca. 10 000 Zeilen Code umfassen. Von mehreren Skripten benötigte Funktionen (z. B. die Erzeugung eines Datenbank-Handles oder die Routine zur Benutzerauthentifizierung) werden in der Library `hypnotic_db.inc` gebündelt. Der Zugriff auf Informationen, die in der Korpusdatenbank enthalten sind, wird in *PHP* – ähnlich wie in anderen Programmiersprachen, z. B. *Perl* oder *Java* – mit Hilfe eines abstrakten Datenbank-APIs realisiert, das Funktionen zur Interaktion mit der *MySQL*-Datenbank zur Verfügung stellt. Der Ablauf eines Skripts umfasst den Aufbau der Datenbankverbindung, das Absetzen einer statischen oder dynamisch generierten SQL-

Anfrage, die sukzessive Verarbeitung des Anfrageresultats und die Ausgabe der Datenbankinhalte in Form einer HTML-Datei. Das per definitionem zustandslose Protokoll HTTP wird dabei mittels verschiedener HTTP-Variablen, die in der Browser-Umgebung gesetzt werden, um eine weitere Applikationslogik-Schicht ergänzt, so dass Zustände explizit von Skript zu Skript weitergereicht bzw. aus der Umgebung ausgelesen werden können.

7.3.1 Benutzer-Authentifizierung

Der Benutzer muss sich der Web-Oberfläche gegenüber authentifizieren. Die Verwaltungsinformationen, z. B. derBenutzername und das Passwort, werden – in Anlehnung an DuBois (1999, S. 378 ff.) – in der Datenbanktabelle `user` gepflegt (vgl. Tabelle B.4) und von der Funktion `user_authen` eingesetzt, die beim Aufruf eines *PHP*-Skripts aktiviert wird. Falls der Benuzter nicht bereits angemeldet ist, wird per Übermittlung des HTTP-Headers `WWW-Authenticate` zur Eingabe einer Kennung und eines Passworts aufgefordert.[37] Eine erfolgreiche Anmeldung wird in globalen Variablen, die zu jedem aufgerufenen *PHP*-Skript weitergereicht werden, vermerkt, so dass keine erneute Authentifizierung notwendig ist.

Die Pflege der Benutzerkonten erfolgt mit Hilfe zweier *PHP*-Skripte. Die Liste aller Benutzer ist in der primären Navigationshilfe der Oberfläche verfügbar, von der aus jeder Benutzer die Möglichkeit hat, die eigenen Daten zu ändern, was eine erneute Anmeldung voraussetzt. Benutzer mit dem Status `admin` können neue Benutzer anlegen. Für die Speicherung der Passwörter wird das von *PHP* unterstützte unidirektionale Verschlüsserungsverfahren MD5 eingesetzt. Neben der Brute-Force-Suche mit einem umfangreichen Lexikon besteht keine Möglichkeit, aus einem solchen Passwort den Klartext zu extrahieren. MD5 wurde zur Erzeugung von *fingerprints* für Textdokumente und digitale Signaturen entwickelt (die Abkürzung MD steht für *Message Digest*). Der Algorithmus nimmt eine Textdatei entgegen und produziert eine 128 Bit lange Ausgabe (vgl. RFC 1321, sowie Garfinkel und Spafford, 1996, S. 172 ff.). Der eigentliche Zugriff auf die in der Tabelle `user` abgelegten Passwörter ist – dies gilt für alle in der Datenbank gespeicherten Informationen – ebenfalls zugangsgeschützt: Einerseits dürfen nur Rechner mit im System eingetragenen IP-Nummern zugreifen, andererseits erfolgt der Zugriff auf den *MySQL*-Server selbst ebenfalls ausschließlich passwortgeschützt.

7.3.2 Möglichkeiten des Dokumentzugriffs

Der Benutzer wird beim Zugriff auf einzelne Dokumente oder Gruppen von Dokumenten unterstützt: Zum einen kann ein Dokument unmittelbar per Eingabe seiner Dokument-ID angefordert werden (`docbyid.php`). Zum anderen besteht die Möglichkeit der Anzeige eines zufällig aus dem gesamten Korpus ausgewählten Dokuments (`randomdocument.php`). Die weiteren Zugriffsfunktionen benötigen die Angabe zusätzlicher Informationen: Mit Hilfe des Skripts `allservers-form.php` erhält der Benutzer die Möglichkeit, eine HTTP-Version (`1.0` oder `1.1`), den HTTP-Port (`80`, `8080` etc.) und eine Hochschule zu wählen. Weiterhin gestattet es dieses HTML-Formular, in den Namen der in der Korpusdatenbank enthaltenen Webserver zu suchen. Abbildung 7.7 zeigt die erste Hälfte eines Beispiels: Zunächst

[37] Empfängt ein Browser diesen HTTP-Response-Header, wird üblicherweise ein Fenster eingeblendet, in das die Authentifizierungsinformationen einzutragen sind.

wird die Suche auf die Webserver der Universität Osnabrück eingeschränkt, wobei alle Server aufgelistet werden sollen, in deren Namen die Zeichenkette `cl-ki` enthalten ist. Der mittlere Screenshot zeigt das Ergebnis der Suchanfrage. Durch Aktivierung des Namens `www.cl-ki.uni-osnabrueck.de` gelangt der Benutzer zu der von dem Skript `showserver.php` generierten Seite, die in dem unteren Bildschirmabzug dargestellt ist. Die enthaltenen Hyperlinks erzeugen Dokumentlisten, die mit den jeweiligen Medientypen korrespondieren. In der unten dargestellten Navigationshilfe bieten sich drei weitere Zugriffsmöglichkeiten: Neben der Darstellung einer übergreifenden Liste (`showlist.php`) kann ein Dokument des Servers zufällig ausgewählt (`showdocument.php`) oder die Pfad- und Dateinamen durchsucht werden (`searchuri.php`). Abbildung 7.8 zeigt die zweite Hälfte des Beispiels, in dem zunächst nach HTML-Dokumenten gesucht wird, deren Pfad- bzw. Dateinamen die Zeichenkette `papier` enthalten. Der mittlere Screenshot zeigt die Suchergebnisse. Der untere Bildschirmabzug stellt die Dokumentansicht des ersten Treffers dar.[38]

7.3.3 Die Dokumentansicht

Die von dem Skript `showdocument.php` implementierte Dokumentansicht ist zweigeteilt: In der unteren Hälfte wird das Dokument – eingebettet in einen breiten, rot eingefärbten Rahmen – dargestellt, die obere Hälfte (die Kopfzeile) zeigt Metadaten des dargestellten Dokuments und gestattet den Zugriff auf Funktionen (vgl. den unteren Screenshot in Abbildung 7.8): Der Link „Remote“ öffnet ein neues Browser-Fenster und stellt zum Zwecke des direkten Inhalts- und Strukturvergleichs die entfernte Version des im Korpus enthaltenen Dokuments dar. Die Funktion „Noch aktuell?“ vergleicht das Änderungsdatum sowie die Dokumentgröße des im Korpus enthaltenen Dokuments mit den korrespondieren Informationen der entfernten Version (`showdocchange.php`). Auf diese Weise kann geprüft werden, ob das im Korpus enthaltene Dokument noch immer der Originalversion entspricht. Über „HTML-Source“ wird von dem Skript `showdocsource.php` das *Perl*-Skript `code2html` mit der dargestellten HTML-Datei als Eingabe aktiviert.[39] Das Skript generiert eine farblich hervorgehobene Version des HTML-Quelltextes, um den häufig sehr komplexen HTML-Code eines Dokuments möglichst transparent darzustellen (vgl. den oberen Screenshot in Abbildung 7.9). Die Funktion „Dokument“ schaltet zurück zur eigentlichen Dokumentansicht. Auch die folgenden drei in Funktionen benutzen externe Werkzeuge: Das Skript `showtree.php` generiert eine Visualisierung der Baumstruktur, die von den in einem Dokument enthaltenen HTML-Elementen aufgespannt wird (vgl. Abschnitt 14.4.2, S. 655 ff.). Die Funktion „Tidy“ aktiviert das Werkzeug *Tidy* mit dem betrachteten HTML-Dokument als Eingabe (vgl. Abschnitt 14.4, S. 652). *Tidy* enthält einen robusten und fehlerkorrigierenden

[38] Der mittlere Screenshot zeigt Dateigrößen von `0` Bytes und der untere Screenshot enthält keine Information über das Datum der letzten Änderung, was auf die fehlenden Felder `Content-Length` und `Last-Modified` innerhalb der Response-Header der Dokumente zurückzuführen ist. Sie werden bei HTML-Dokumenten, die mittels Server-seitig interpretiertem HTML generiert wurden, nicht gesetzt. Server-parsed HTML (häufig an dem Suffix `.shtml` zu erkennen) besteht aus einer *Apache*-eigenen Makro-Erweiterung für HTML, die es gestattet, mit sehr rudimentären Möglichkeiten dynamische Webseiten zu realisieren.

[39] Dieses Skript erzeugt HTML-Versionen von Quelltexten verschiedener Programmier- und Auszeichnungssprachen und färbt unterschiedliche syntaktische Elemente ein (*syntax highlighting*). In der Web-Oberfläche wird Version 0.9.1 von `code2html` eingesetzt (vgl. `http://www.palfrader.org/code2html/`).

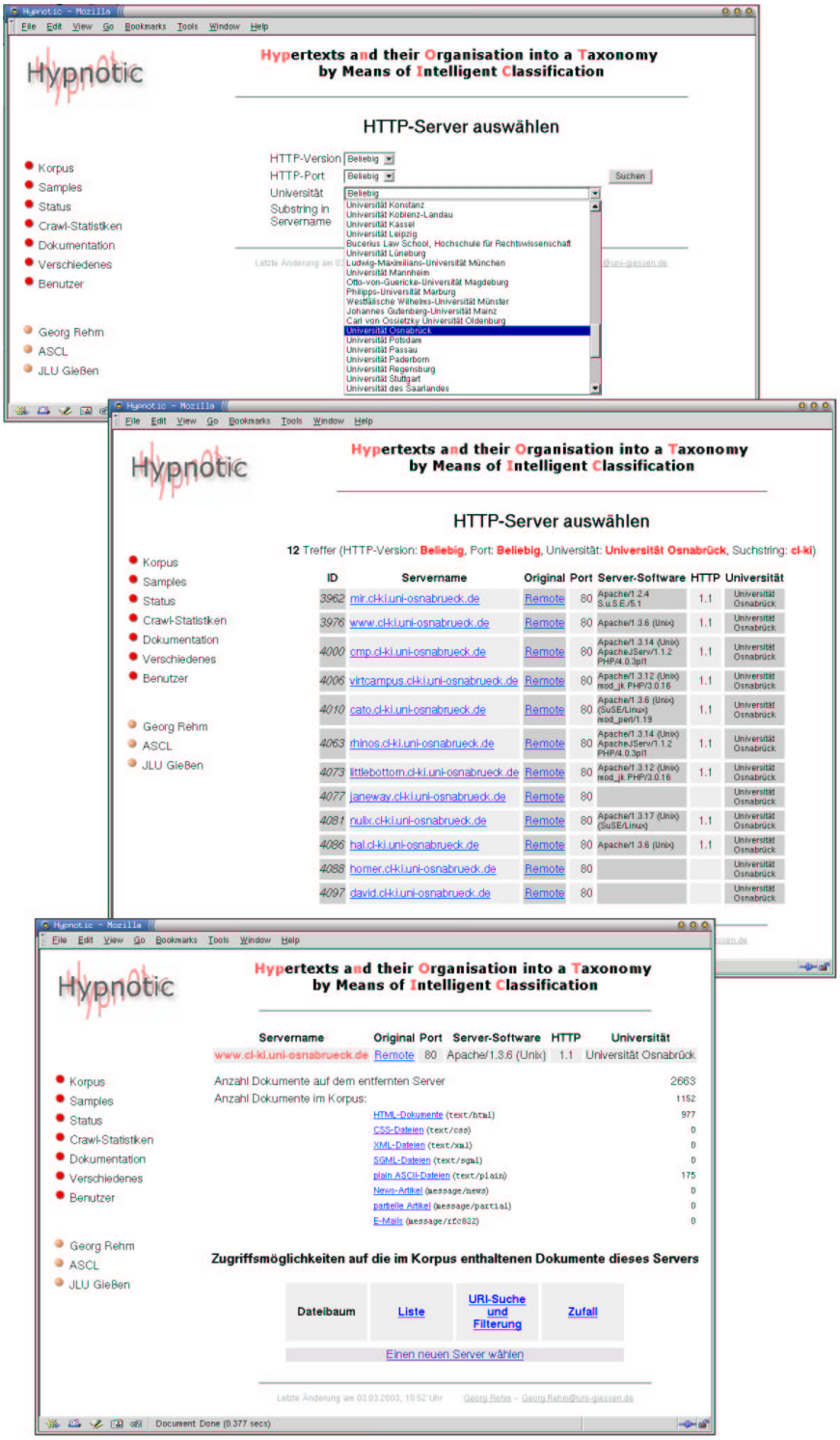


Abbildung 7.7: Navigation zu einem spezifischen Webserver in der Web-Oberfläche

Abbildung 7.8: Navigation zu einem spezifischen Dokument in der Web-Oberfläche

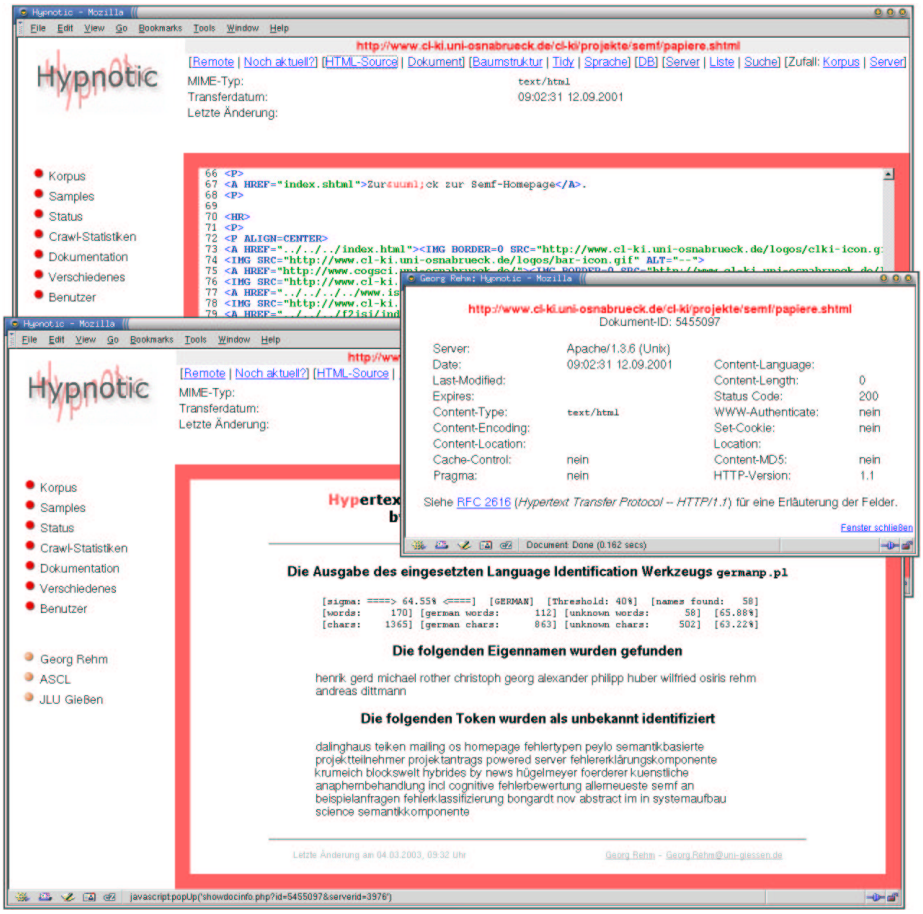


Abbildung 7.9: Verschiedene in der Dokumentansicht verfügbare Funktionen

HTML-Parser und stellt mit der Kommandozeilenoption `-errors` eine Liste der enthaltenen Syntaxfehler bezüglich der korrespondierenden HTML-DTD auf, die heuristisch ermittelt wird. Der Link „Sprache“ aktiviert den automatischen Sprachenidentifizierer `germanp.pl` (vgl. Abschnitt 7.2.2) mit der dargestellten HTML-Datei als Eingabe, um einen schnellen Zugriff auf den Anteil der erkannten Token zu gewährleisten (vgl. den unteren Screenshot in Abbildung 7.9). Mit Hilfe der Funktion „DB“ (für *Datenbank*) kann der Benutzer für ein Dokument dessen Datensatz der Korpusdatenbank (bezogen aus `http_header` und `server_info`) anzeigen lassen (vgl. den mittleren Screenshot in Abbildung 7.9). Über die Hyperlinks „Server“, „Liste“ und „Suche“ gelangt der Benutzer zu den Funktionen, die in Abschnitt 7.3.2 erläutert wurden. Mittels der Zufallsfunktionen kann ein zufällig ausgewähltes Dokument des derzeit untersuchten Webservers oder des gesamten Korpus angezeigt werden.

Die Darstellung eines Dokuments erfolgt durch einen HTTP-Redirect (RFC 2616). Das Dokument wird von `document.php` angezeigt, das als Parameter eine Dokument-ID benötigt und dessen Ausgabe sich auf den rot markierten Frame auswirkt. Das Skript extrahiert die Pfadangabe (`http_header.file`) des Dokuments aus der Datenbank und konstruiert hieraus

mit der Basisadresse des Webservers eine zugreifbare URL; dies setzt voraus, dass sich alle im Korpus verfügbaren Dokumente in Verzeichnisstrukturen befinden, auf die der Webserver lesenden Zugriff hat. Die generierte URL wird mit Hilfe des HTTP-Headers `Location` an den Browser geschickt, der das Dokument daraufhin automatisch vom Webserver anfordert.

7.3.4 Die Generierung von Stichproben

Ein zentrales Merkmal der Web-Oberfläche ist die Möglichkeit, zufällige Stichproben zu generieren. Eine solche Stichprobe entspricht der Extraktion einer Teilmenge des Korpus, die auf unterschiedlichen Kriterien basieren kann. Nach der Generierung werden die entsprechenden Informationen in Datenbanktabellen hinterlegt (vgl. Abschnitt 7.2.3). Dieser Abschnitt thematisiert zunächst die unterschiedlichen Möglichkeiten der Realisierung von Restriktionen sowie die Implementierung der Stichprobengenerierung, woraufhin der Umgang mit Stichproben und das Verfahren der manuellen Analyse dargestellt werden.

Das Skript `generate_sample.php` bietet die Möglichkeit, eine Stichprobe beliebigen Umfangs nach verschiedenen Kriterien zufällig generieren zu lassen. Nach der Generierung wird der Benutzer aufgefordert, der Stichprobe einen Titel zu geben und eine Beschreibung (Auswahlkriterien, Einsatzgebiet etc.) einzutragen, wobei auch weiteren Benutzern der Zugriff gestattet werden kann. Abbildung 7.10 stellt diesen Prozess anhand eines Beispiels dar: Über das HTML-Formular können verschiedene Kriterien spezifiziert werden, die die Zusammenstellung beeinflussen. Diese wird durch eine dynamisch erzeugte SQL-Query realisiert, für die ein gerüstartiges Beispiel in Listing 7.1 (S. 351) dargestellt ist. Die Query selektiert aus der Tabelle `http_header` (Zeile 8) eine Liste von Dokumenten basierend auf Bedingungen, die durch `WHERE`-Klauseln realisiert werden. Zeile 2 zeigt die Aufnahme eines neuen Feldes in die Resultatstabelle, das für jedes Dokument einen Zufallswert enthält. Zeile 13 ordnet *alle* gefundenen Dokumente in einer zufälligen Reihenfolge an, woraufhin die Anzahl der Dokumente, die das Ergebnis der Query umfassen soll, auf einen zuvor festgelegten Wert limitiert wird. Auf diese Weise kann das Resultat einer SQL-Anfrage, die z. B. mehrere tausend Treffer zurückliefert, auf 50 zufällig ausgewählte Treffer beschränkt werden. Die in dem HTML-Formular spezifizierbaren Beschränkungen werden in `generate_sample.php` auf schablonenartige Fragmente von `WHERE`-Klauseln abgebildet, so dass die in das Formular eingetragenen Werte benutzt werden, um die SQL-Query zusammenzufügen. Die folgende Auflistung stellt die einzelnen Einschränkungsmöglichkeiten der Stichprobengenerierung in der Reihenfolge dar, in der sie in dem HTML-Formular vorhanden sind:

1. *Selektiere Dokumente vom Typ* – Dieses Feld spezifiziert den Medientyp der Dokumente. Ein Menü gibt die möglichen Werte vor (`text/html`, `text/css`, `text/xml`, `text/sgml`, `text/plain`, `message/news` und `message/rfc822`).
2. *Anzahl Dokumente* – Dieser vom Benutzer frei wählbare Wert limitiert die Größe der Stichprobe (vgl. Zeile 13 in Listing 7.1).
3. *Minimale* und *maximale Länge des Pfadanteils der URL* – Eine URL wie beispielsweise `http://www.uni-giessen.de/germanistik/ascl/` besteht aus der Angabe der Methode bzw. des Protokolls (`http`), dem Namen eines Webservers (`www.uni-giessen.de`) sowie einem Pfadanteil, der ein Dokument oder eine Datei auf dem Webserver adressiert

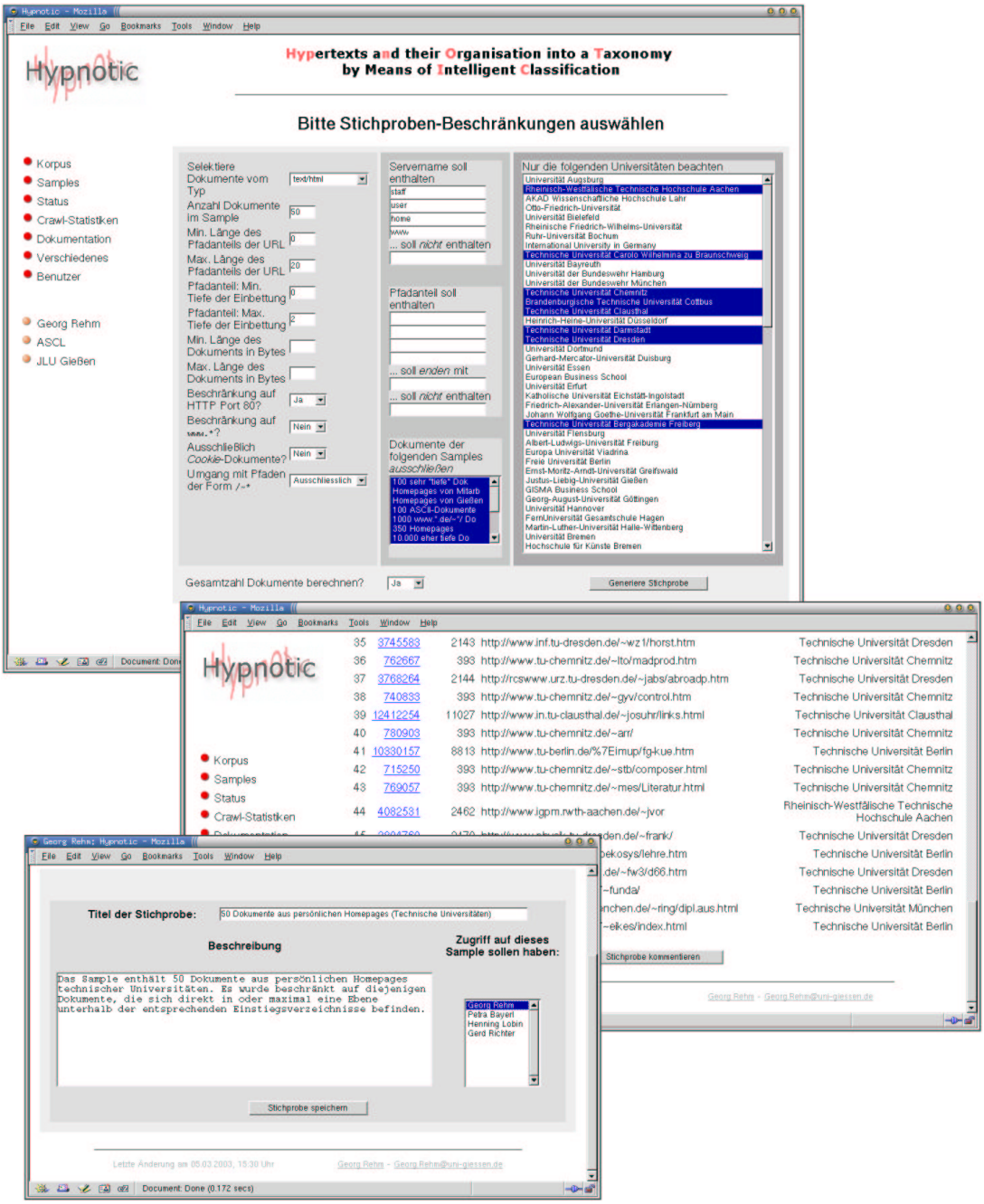


Abbildung 7.10: Die Generierung einer zufällig zusammengestellten Stichprobe

```
SELECT http_header.id,
       http_header.id*0+rand() as rand_col,
       server_info.servername,
       http_header.uri,
       http_header.server_info,
       server_info.city,
       server_info.port
       FROM http_header,server_info
            WHERE http_header.server_info = server_info.id and
                  http_header.file != "" and
                  http_header.cont_type = "html" and
                  ...
            ORDER BY rand_col LIMIT 100
```

Listing 7.1: Gerüst einer SQL-Query zur Generierung einer zufälligen Stichprobe

(`/germanistik/ascl/`).[40] Mit Hilfe der expliziten Spezifizierung der minimalen und maximalen Länge des Pfadanteils einer URLs kann eine Stichprobe auf Dokumente mit sehr kurzen oder sehr langen Pfadanteilen eingeschränkt werden. Die eintragbaren Werte *min* und *max* werden bei Benutzung wie folgt in die SQL-Query übernommen: `...` and `length(http_header.uri)` >= *min* and `length(http_header.uri)` <= *max*.

4. *Pfadanteil: Minimale* sowie *maximale Tiefe der Einbettung* – RFC 2396 spezifiziert das Zeichen `/` als Trennelement der Segmente des Pfadanteils (vgl. Fußnote 40); das Zeichen `?` trennt die Anfrage- von der Pfadkomponente ab. Diese Segmente beziehen sich in den meisten Fällen auf Verzeichnisse, die per HTTP zugreifbar sind und adressierbare Dateien enthalten. Mittels der Einschränkung bezüglich dieser Einbettungs- bzw. Verzeichnistiefe können Dokumente in eine Stichprobe aufgenommen werden, die sich nahe der Wurzel oder sehr tief innerhalb des Dokumentbaumes befinden (vgl. Abbildung 7.5). Die Werte *min* und *max* werden wie folgt in die SQL-Query eingetragen: `...` and `http_header.uri regexp "^(/[^/]*){`*min*`,`*max*`}$"`. Der reguläre Ausdruck wird mittels `^` und `$` am Zeilenanfang und -ende verankert. Innerhalb einer durch runde Klammern markierten Gruppe ist der Segmenttrenner `/` das erste Zeichen, woraufhin beliebige Zeichen mit Ausnahme des `/` folgen dürfen. Die Angaben in den geschweiften Klammern geben die minimale bzw. maximale Anzahl der möglichen Sequenzen von Treffern für diesen Ausdruck an.
5. *Minimale* und *maximale Länge des Dokuments in Bytes* – Diese Werte ermöglichen Stichproben, die ausschließlich sehr lange oder sehr kurze Dokumente enthalten. Der Wert bezieht sich auf die Länge eines Dokuments in Bytes, d. h. das HTML-Markup ist in den entsprechenden Angaben, die sich auf das Feld `Content-Length` des HTTP-Response-Headers beziehen, enthalten (vgl. Fußnote 38 auf S. 345).
6. *Beschränkung auf HTTP-Port 80* – Dieser binäre Schalter erlaubt die Aufnahme der Dokumente von Webservern, die nicht auf dem HTTP-Standard-Port arbeiten. Ist dieser Schalter aktiviert, wird die `WHERE`-Klausel `... and server_info.port = 80` in die

[40] RFC 2396 (vgl. RFC 1738) spricht statt des Protokolls von einem „Scheme“, die Server-Angabe wird als „Authority Component“ bezeichnet, und der „Path Component“ kann eine „Query Component“ folgen. Das W3C verwaltet die offizielle Liste gültiger Protokolle: `http://www.w3.org/Addressing/schemes.html`.

SQL-Query aufgenommen. Da sich nur HTML-Dokumente im Korpus befinden, die ursprünglich auf Port 80 angeboten wurden, ist diese Einschränkung nur sinnvoll, wenn Medientypen wie `text/xml` oder `text/plain` untersucht werden sollen.

7. *Beschränkung auf* `www.*` – Mit diesem binären Schalter werden lediglich diejenigen Webserver einbezogen, deren Namen mit `www` beginnen. Diese mittlerweile kanonische Benennung wurde erstmalig 1993 für den Webserver des NCSA (*National Center for Supercomputing Applications*, `http://www.ncsa.uiuc.edu`) benutzt und wurde zu einem de-facto-Standard, da am NCSA zu dieser Zeit die am häufigsten eingesetzte Webserver-Software entwickelt wurde (vgl. Abschnitt A.3.3). Über diesen Schalter kann der Benutzer eher unwichtige bzw. spezialisierte Webserver, die meist abweichende Namen besitzen, ausschließen. Realisiert wird dieser Schalter über einen Mustervergleich, der in der SQL-Syntax von *MySQL* mittels `like` und dem Metazeichen `%` realisiert wird: `... and server_info.servername like "www.%"`.
8. *Ausschließlich Cookie-Dokumente* – Dieser binäre Schalter erlaubt die Aufnahme von Dokumenten in eine Stichprobe, die im Webbrowser einen Cookie setzen, was durch das Feld `Set-Cookie` im HTTP-Response-Header realisiert wird. Dieses Feld ist nicht Teil der Kernspezifikation von HTTP, sondern wird in RFC 2965 beschrieben. Cookies werden häufig benutzt, um in dem zustandslosen Protokoll HTTP Funktionen wie Session-Management zu realisieren (vgl. Abschnitt A.3.7). Die Existenz des Feldes `Set-Cookie` im HTTP-Reponse-Header wird durch einen Wert in `http_header.set_cookie` vermerkt, der durch eine `WHERE`-Klausel abgefragt wird.
9. *Umgang mit Pfaden der Form* `/~*` – Derartige Pfade kennzeichnen oftmals HTML-Dokumente in Benutzerverzeichnissen, bei denen es sich meist um persönliche Homepages handelt. Das in dem HTML-Formular vorhandene Menü bietet die Werte *zulassen*, *nicht zulassen* und *ausschließlich* an. Auf diese Weise kann der Benutzer Pfadkomponenten, die mit `/~` (oder den Ersatzdarstellungen `/%7e` bzw. `/%7E`) beginnen, bei der Generierung einer zufälligen Stichprobe berücksichtigen, um z. B. eine Liste persönlicher Homepages aus verschiedenen Universitäten zusammenzustellen. Die `WHERE`-Klauseln werden über einen Mustervergleich realisiert.
10. *Servername soll enthalten* bzw. *nicht enthalten* – Über diese fünf Textfelder können Muster angegeben werden, die – *oder*-verknüpft – auf Namen von Webservern zutreffen sollen (z. B. `user`, `student`, `wohnheim`) bzw. nicht zutreffen sollen (beispielsweise `staff`), um eine Stichprobe auf bestimmte Typen von Webservern einzuschränken. Wie auch bei einigen weiteren Beschränkungsmöglichkeiten erforderen diese Funktionen eine manuelle Inspektion des Korpus, um geeignete Webserver zu ermitteln. Diese Beschränkungen werden ebenfalls per Mustervergleich realisiert.
11. *Pfadanteil soll enthalten* bzw. *nicht enthalten* bzw. *soll enden mit* – Diese Beschränkungen entsprechen den unter 10. beschriebenen Funktionen, wirken sich jedoch auf die Pfadkomponente aus. Die Funktionen erlauben es, positive und negative Muster in Pfadkomponenten anzugeben. Zusätzlich kann eine Zeichenkette spezifiziert werden, die an den abschließenden Zeichen einer URL verankert wird, um z. B. nur Dokumente mit bestimmten Dateisuffixen in eine Stichprobe aufnehmen zu können.

12. *Dokumente der folgenden Samples ausschließen* – Das Menü dieser Funktion umfasst alle verfügbaren Stichproben, von denen eine oder mehrere selektiert werden können. Die in diesen Stichproben enthaltenen Dokumente werden in diesem Fall bei der Generierung der aktuellen Stichprobe nicht berücksichtigt, um bereits analysierte Dokumente ausschließen zu können. Die Funktion wird z. B. benötigt, um ausschließlich neue Dokumente eines bestimmten Webservers, der bereits partiell untersucht wurde, in eine Stichprobe zu übernehmen. Für jede selektierte Stichprobe werden in dem *PHP*-Skript innerhalb von Schleifen alle beteiligten Dokument-IDs gesammelt und dann mittels einer `WHERE`-Klausel (`... and http_header.id != ID ...`) ausgeschlossen.
13. *Nur die folgenden Universitäten beachten* – Dieses Menü umfasst die Bezeichnungen der 100 in der Korpusdatenbank enthaltenen Universitäten. Diejenigen Hochschulen, deren Dokumente in die Stichprobe aufgenommen werden sollen, können selektiert werden. Die Hochschul-IDs werden innerhalb der SQL-Query mit einer Sequenz von *oder*-verknüpften `WHERE`-Klauseln der Form `... and ( server_info.city = "Bezeichnung1" or server_info.city = "Bezeichnung2" or ... ) and ...` spezifiziert.

Das *PHP*-Skript `generate_sample.php` umfasst etwa 800 Zeilen Code. Der Parameter `action`, der in der URL übergeben wird, spezifiziert eine der Funktionen „HTML-Formular darstellen“, „Sample auflisten“, „Sample kommentieren“ und „Sample speichern“. Derzeit werden die Datensätze *aller* Dokumente in nur einer Tabelle gespeichert (`http_header`), die mehr als 16 000 000 Einträge enthält. Daher ist die Berechnung von Stichproben, die nicht auf bestimmte Universitäten eingeschränkt werden, sehr aufwändig und kann mehr als 15 Minuten in Anspruch nehmen, weshalb ältere Browser nicht zur Interaktion mit der Web-Oberfläche eingesetzt werden können, da die während dieser Zeit bestehende HTTP-Verbindung oftmals nach drei Minuten unterbrochen wird.

Neben der interaktiven Generierung ist es auch möglich, mittels eines *Perl*-Skripts Stichproben automatisch in die Datenbank eintragen zu lassen. Die in Kapitel 11 analysierte Stichprobe sollte z. B. die Einstiegsseiten der ersten 35 Universitäten sowie diejenigen Webseiten umfassen, auf die die Einstiegsseiten verweisen. Da diese Kriterien mit Hilfe der Web-Oberfläche nicht bzw. nur durch aufwändige Modifikationen an `generate_sample.php` spezifizierbar wären, wurde ein Skript implementiert, das die Einstiegsseiten mittels des Moduls `HTML::Parser` bezüglich der enthaltenen Hyperlinks analysiert. Für jedes referenzierte Dokument wird überprüft, ob es im Korpus enthalten ist. Ist dies der Fall, wird die Dokument-ID bestimmt und mit weiteren Informationen, die für die Tabelle `sample` benötigt werden, gesammelt. Anschließend werden die extrahierten Daten von dem Skript in einem Format ausgegeben, das von `mysqlimport` in die Tabelle `sample` eingetragen werden kann.

Mit ähnlichen Mitteln wurde eine Stichprobe von 100 Instanzen der Hypertextsorte *persönliche Homepage eines Wissenschaftlers* erzeugt (vgl. Kapitel 10). Für diese Stichprobe mussten zunächst diejenigen Webserver identifiziert werden, die viele persönliche Homepages anbieten, von denen möglichst viele die persönlichen Webangebote von Universitätsangehörigen sein sollten. Zu diesem Zweck wurde mit Hilfe des in Listing 7.2 dargestellten Shell-Skripts eine Liste von Dokumenten erzeugt, bei denen die Pfadkomponenten der korrespondierenden URLs mit einer Tilde (~) beginnen. Die SQL-Query wird dabei von einem Shell-Skript gekapselt, d. h. die Ausgabe der Datenbank kann unmittelbar von nachgeschal-

```
#!/bin/bash
echo "select server_info.servername,http_header.uri from
                        http_header,server_info where
                                        http_header.server_info = server_info.id and
                                        http_header.uri regexp \"^/~[^/]+/$\";" | \
mysql -h hypnotic hypnotic_db | awk '{print $1}' | sort | uniq -c | sort -n
```

Listing 7.2: Generierung einer nach Häufigkeiten sortierten Liste von Webservern, die persönliche Homepages anbieten, mit Hilfe eines Shell-Skripts

teten UNIX-Werkzeugen weiter verarbeitet werden: Mit Hilfe von *awk* wird zunächst das jeweils erste Feld jeder Zeile (der Name des Webservers) extrahiert, woraufhin diese Liste sortiert wird. Anschließend werden identische Zeilen gezählt, woraufhin erneut numerisch sortiert wird, um die entstandene Liste von Webservern, die persönliche Seiten anbieten, nach den Häufigkeiten der verfügbaren Homepages zu sortieren. Von denjenigen Webservern mit den meisten persönlichen Homepages (z. B. `staff-www.uni-marburg.de`, `www.uni-giessen.de`, `www.tu-chemnitz.de` und `www.uni-bonn.de`) wurden daraufhin auf der Basis verschiedener Kriterien manuell 100 Einstiegsseiten gesammelt und mit dem *Perl*-Skript `build-sample-from-ids.pl` in die Datenbank-Tabelle `sample` eingetragen.

7.3.5 Die Analyse von Stichproben – Einsatz von Templates

Die Web-Oberfläche unterstützt die manuelle Stichprobenanalyse. Für diesen Zweck wird eine Stichprobe auf ein Template abgebildet, das diejenigen Informationen spezifiziert, die im Rahmen der Analyse erhoben werden sollen. Abschnitt 7.2.3 geht auf die beteiligten Datenbanktabellen ein, weshalb sie an dieser Stelle nur kurz erläutert werden: Die Tabelle `sample` enthält in einer fortlaufenden Liste verschiedene Informationen (Dokument-ID, Server-ID, originäre URL und Universität) über alle Dokumente, die sich in Stichproben befinden. Die Verbindung eines Datensatzes zu einer Stichprobe wird über ein zusätzliches Tabellenfeld realisiert. In der Tabelle `meta_sample` werden Metainformationen über Stichproben abgelegt, z. B. ihr Titel, das Generierungsdatum und die Anzahl der auf die erzeugende SQL-Query zutreffenden Dokumente. Die Tabelle `meta_template` enthält vergleichbare Informationen über die verfügbaren Templates, z. B. einen Titel und eine Beschreibung. Die Tabellen `template1` bis `template4` spezifizieren die Informationen, die im Rahmen einer Stichprobenanalyse erhoben werden. Die Tabelle `sample_template` verknüpft Stichproben und Templates, indem eine Sample-ID auf eine Template-ID abgebildet wird.

Die Abbildungen 7.11 und 7.12 zeigen mit Hilfe der Web-Oberfläche durchgeführte Schritte der Stichprobenauswertung. In Abbildung 7.11 werden die verfügbaren Stichproben der Datenbank aufgelistet. Das Skript `list_samples.php` greift auf die Tabellen `meta_sample` und `meta_template` zu, um die Liste dynamisch zu generieren. Weitere Funktionen, die von dieser Seite aus aufgerufen werden können, betreffen das Löschen einer Stichprobe sowie aller assoziierten Analyseergebnisse, die Verknüpfung einer Stichprobe mit einem Template, die Generierung einer neuen Stichprobe sowie das Anzeigen der verfügbaren Metadaten. Der mittlere Bildschirmabzug stellt den – ebenfalls von `list_samples.php` erzeugten – Inhalt einer Stichprobe dar, der die Dokument-ID, die ursprüngliche URL und die jeweilige

Abbildung 7.11: Auflistung der verfügbaren Stichproben (oben), Auswahl einer Stichprobe (mitte) und Darstellung eines Dokuments (unten; vgl. Abbildung 7.12)

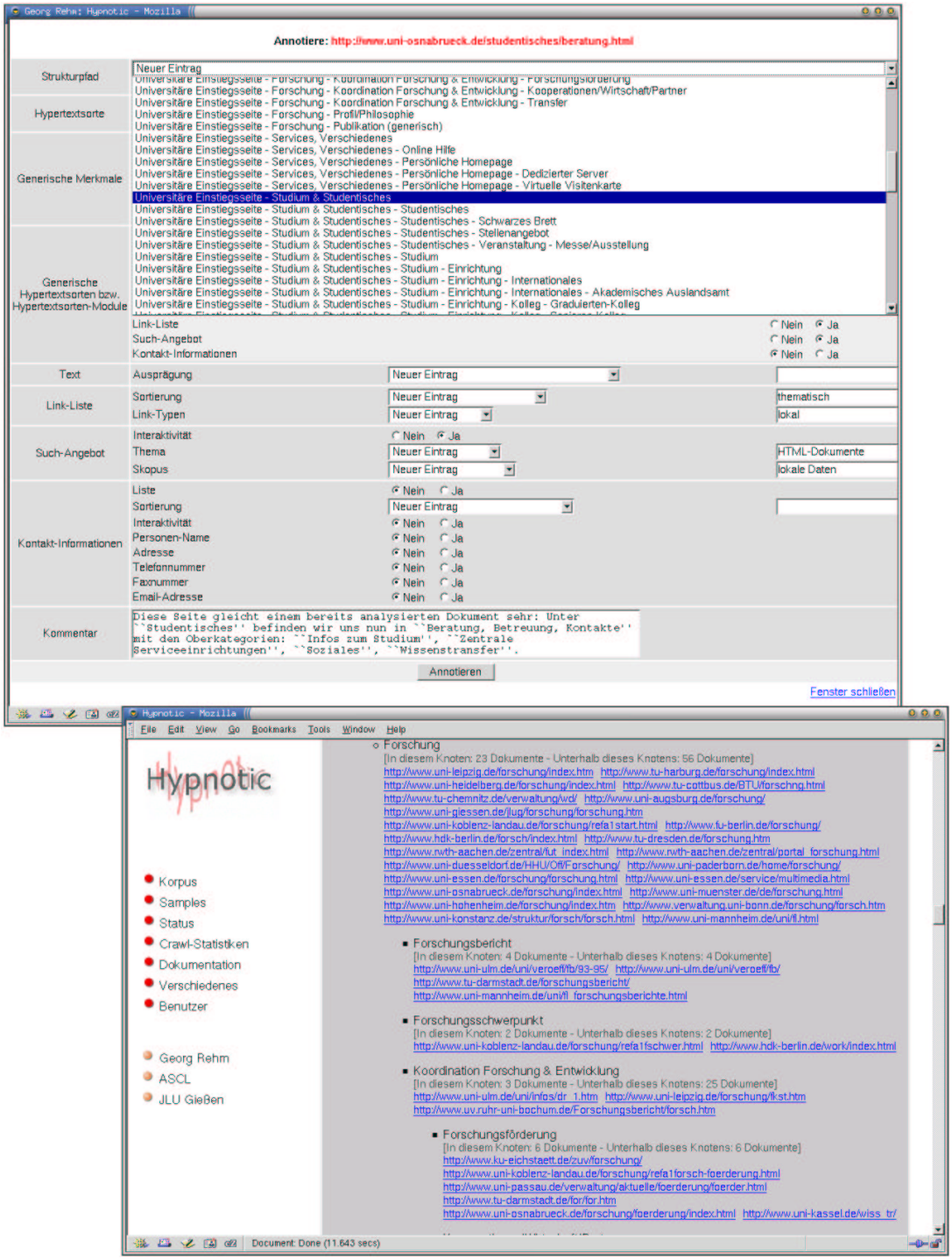


Abbildung 7.12: Darstellung des Analyse-Formulars für das in Abbildung 7.11 unten dargestellte Dokument sowie Ergebnisse der Stichprobenanalyse

Universität umfasst. Zusätzlich ist in der rechten Spalte entweder ein grüner oder ein roter Punkt dargestellt, der den Analysestatus eines Dokuments kennzeichnet. Von dieser Seite aus kann ein Dokument gezielt angesteuert oder die Analyseergebnisse angezeigt werden. Der unterere Bildschirmabzug zeigt die Dokumentansicht, die im Kontext der Stichprobenanalyse mit Erweiterungen versehen wird: In der Kopfzeile befinden sich oben rechts Navigationselemente, die das Aufrufen des nächsten oder vorherigen Dokuments der Stichprobe bzw. eine Auflistung des Inhalts einer Stichprobe erlauben. Weiterhin werden in der Kopfzeile die Namen der Stichprobe und des zugeordneten Templates sowie die Nummer des aktuellen Dokuments dargestellt. Durch den Hyperlink „Analysieren“ gelangt der Benutzer zu einem HTML-Formular, das in einem Pop-up-Fenster dargestellt wird und die interaktive Analyse des betrachteten Dokuments auf der Basis des zugeordneten Templates ermöglicht. In Abbildung 7.12 (oben) ist ein derartiges Formular mit den analysierten Merkmalen für das Dokument aus Abbildung 7.11 (unten) dargestellt. Realisiert wird die Maske von dem *PHP*-Skript `annotate.php`, das für jedes Template ein spezielles Formular implementiert. Die Implementation erfolgt mit Hilfe interaktiver Elemente, die von HTML spezifiziert werden (Pull-Down-Menüs, Textfelder zur freien Eingabe etc.). Die eingetragenen Werte werden mittels des Knopfes „Annotieren“ in der korrespondierenden Template-Tabelle gespeichert. Der untere Screenshot aus Abbildung 7.12 stellt schließlich ein Analyseresultat dar. Im Beispiel handelt es sich um eine hierarchische Visualisierung der oberen Ebenen der Domänenontologie mit den zugehörigen Dokumenten der Stichprobe.

Die Funktionen der Web-Oberfläche zur Analyse von Stichproben auf der Grundlage eines zugehörigen Templates werden von `list_samples.php` und `annotate.php` realisiert. Diese Skripte sind so modularisiert, dass der Aufwand zur Implementierung der benötigten Funktionen zum Umgang mit einem neuen Template so gering wie möglich ist. In `list_samples.php` muss die *PHP*-Funktion zur Anzeige der Analyse-Resultate erweitert werden, und in `annotate.php` ist das einem Template zugeordnete HTML-Formular sowie der Zugriff auf die korrespondiere Template-Tabelle der Datenbank zu implementieren. Falls die Web-Oberfläche von einem umfangreicheren Benutzerkreis genutzt werden soll, wäre es prinzipiell möglich, diese derzeit manuell durchzuführenden Erweiterungen zu automatisieren. Zu diesem Zweck wäre die Implementierung eines dynamischen „Template-Baukastens“ notwendig, der die freie Spezifizierung von Analyseschemata, zugehörigen HTML-Formular-Objekten und vorgefertigten Möglichkeiten der Darstellung von Analyse-Resultaten erlaubt.

7.4 Indirekter Korpuszugriff mittels API und HTTP

Die im Korpus enthaltenen Dokumente werden in einem Bereich des Dateisystems des Korpusdatenbank-Servers aufbewahrt, der auch von dem ebenfalls auf dem Datenbankserver installierten Webserver *Apache* zugreifbar ist. Da die Datenbank netzwerkfähig ist, kann sie von entfernten Rechnern angesprochen werden. Somit kann ein entfernter Client, d. h. ein autarkes Analyse-System, die Anfrage nach einer bestimmten Dokument-ID, die z. B. aus dem Inhalt einer Stichprobe ermittelt wurde, an die Datenbank richten, um daraufhin das korrespondierende Dokument per HTTP anzufordern (vgl. Abbildung 7.6, S. 343).

Listing 7.3 zeigt einen Teil der Implementierung dieser indirekten Zugriffsstrategie, die als *Perl*-Modul realisiert wurde. Der Dokumentzugriff findet häufig auf der Grundlage einer

```
1  package Hypnotic::DBLib;
2
3  use DBI;
4  use LWP::Simple;
5
6  sub getDbHandle {
7      my $db_host = "hypnotic.germanistik.uni-giessen.de";
8      my $d = "DBI:mysql:hypnotic_db:$db_host";
9      my $dbh = DBI->connect($d, "user", "passwort", { RaiseError => 1 });
10     return $dbh;
11 }
12
13 sub getDirectUrlFromId($) {
14     my ($docid) = @_;
15     my $doc_server = "hypnotic.germanistik.uni-giessen.de";
16     my $dbh = Hypnotic::DBLib::getDbHandle();
17     my $query = "select file from http_header where id = $docid;";
18     my $sth = $dbh->prepare($query);
19     $sth->execute();
20     my $path = $sth->fetchrow_array();
21     $sth->finish();
22     $dbh->disconnect();
23     $path =~ s/^\/usr\/local\/db-data(2)?\//korpus-db(2)?\//;
24     return "http://$doc_server/" . $path;
25 }
26
27 sub getDocumentFromUrl($) {
28     my ($url) = @_;
29     my $doc = get($url);
30     return $doc;
31 }
```

Listing 7.3: Implementierung des indirekten Korpuszugriffs (gekürzt)

Dokument-ID statt. Die Funktion `getDirectUrlFromId()` nimmt eine solche ID als Argument und generiert ein Datenbank-Handle (Zeile 16), um auf die Datenbank zugreifen zu können. Daraufhin wird eine SQL-Query an die Datenbank gerichtet, die den korrespondierenden Wert `http_header.file` ermittelt. Da es sich um eine vollständig qualifizierte Pfadangabe handelt, muss der Präfix bis zur Wurzel des in der *Apache*-Konfigurationsdatei spezifizierten Dokumentbaums entfernt werden, so dass eine per HTTP zugreifbare URL entsteht, die als Rückgabewert übergeben wird (Zeile 24). In einem zweiten Schritt wird die Funktion `getDocumentFromUrl()` benutzt, um das Dokument per HTTP auf das Analysesystem zu übertragen. Hierzu wird das Modul `LWP::Simple` eingesetzt, das die Funktion `get()` bereitstellt (Zeile 29).[41] Diese Funktion nimmt eine URL als Argument, richtet eine HTTP-Anfrage an einen Webserver und liefert das Dokument im Form einer Skalarvariable zurück, woraufhin es mit dem Modul `Hypnotic:HTML2XHTML` in eine wohlgeformte XML-Datei konvertiert wird (vgl. Abschnitt 14.4, S. 652 ff.).

[41] Das im CPAN erhältliche Paket *LWP* (bzw. *libwww-perl*, vgl. `http://www.linpro.no/lwp/`) ist die „World Wide Web library for Perl“ und enthält Module zum Zugriff und zur Verarbeitung von HTML-Dokumenten (vgl. Burke, 2002). Neben `http` werden die Methoden `https`, `ftp`, `news`, `gopher`, `file` und `mailto` unterstützt.

7.5 Verwandte Arbeiten

Es existieren mehrere Fragestellungen, Untersuchungsansätze und Systeme, die für die Konzeptionierung der Korpusdatenbank relevant sind bzw. ursprünglich ausschlaggebend waren. Den nachfolgend aufgeführten Projekten, Werkzeugen, Initiativen und Produkten ist gemein, dass es auf einer abstrakten Ebene um die Repräsentierung, die Speicherung, den Zugriff und teilweise auch um die Analyse von HTML-Dokumenten geht. Da unterschiedliche Forschungsperspektiven vorliegen, werden sie im Folgenden thematisch gruppiert dargestellt.

7.5.1 Einsatz von Suchmaschinen

Die ersten Untersuchungen von Web-Genres arbeiten zur Erstellung von Stichproben mit herkömmlichen Suchmaschinen. Beispielsweise setzen Crowston und Williams (1997) und Haas und Grams (1998a, 1998b, 2000) die mittlerweile nicht mehr verfügbare „Surprise"-Funktion von *AltaVista* ein (`http://www.altavista.com`), um 100 bzw. 75 von der Suchmaschine zufällig aus ihrem Index ausgewählte Dokumente zu sammeln (vgl. Abschnitt 4.4).[42]

Google bietet seit dem Frühjahr 2002 ein API (*Application Programming Interface*) an, mit dessen Hilfe Anwendungen auf den Datenbestand und eine Teilmenge der Funktionalität dieser Suchmaschine zugreifen können (vgl. `http://www.google.com/apis/` und Calishain und Dornfest, 2003). Das API basiert auf dem Web-Services-Paradigma (vgl. Wolff, 2004), benutzt also die Standards WSDL (*Web Services Description Language*, Christensen et al., 2001) und SOAP (*Simple Object Access Protocol*, Box et al., 2000, Mitra et al., 2002). Aus diesem Grund ist das API nicht auf spezifische Programmiersprachen oder Entwicklungsumgebungen beschränkt. Eine Anwendung muss sich gegenüber *Google* anhand eines Schlüssels identifizieren, der in jeder Anfrage zu hinterlegen ist, weshalb jeder Benutzer dieses Dienstes gezwungen ist, sich zunächst zu registrieren. Es können maximal 1 000 Anfragen pro Tag abgesetzt werden, von denen jeweils nur die ersten zehn Resultate übertragen werden.

7.5.2 Computer- und korpuslinguistische Ansätze

Es existieren zahlreiche Initiativen, die über das WWW liguistische oder computerlinguistische Ressourcen anbieten. Romary et al. (1999) stellen z. B. TEI-annotierte Korpora für das Französische zur Verfügung. Diese Daten beschränken sich jedoch ausschließlich auf traditionelle Quellen – die Betrachtungsweise, das gesamte *World Wide Web* selbst als ein Korpus aufzufassen, wird von Kilgarriff (2001) propagiert, denn das Web „presents the most provocative questions about the nature of language.":

> As corpus linguists, we are in the fortunate position of having a particular perspective and channel of attack for examining the web – perhaps the most extraordinary phenomenon of our time – which also just happens to provides [sic] solutions to many of our practical problems and an endless stream of new data. [...] The corpus of the new millenium is the web. (Kilgarriff, 2001)

[42] Bar-Ilan (2001) diskutiert die Nutzung von *AltaVista* für infometrische Analysen (vgl. auch Park und Thelwall, 2003). Hierzu zählen z. B. Untersuchungen zur Größe des WWW, zur Rate der Veränderungen von HTML-Dokumenten oder zur Analyse der von Hyperlinks aufgespannten Graphstruktur (vgl. Abschnitt A.6).

Eine derartige Perspektive (vgl. Rehm, 2004a) wird mittlerweile von zahlreichen Arbeiten vertreten (vgl. Abschnitt 7.2.2): Fairon (2000) stellt das System „GlossaNet“ vor, das periodisch verschiedene Websites besucht und Gruppen von HTML-Dokumenten herunterlädt, um aus diesen das Markup zu tilgen, so dass der enthaltene Text verschiedenen Tagging- und Parsing-Modulen zur Verfügung gestellt werden kann. Die Ergebnisse (z. B. Konkordanzen) werden dem Benutzer per E-Mail zugeschickt. Kehoe und Renouf (2002) und Morley et al. (2003) präsentieren die Metasuchmaschine WebCorp (`http://www.webcorp.org.uk`), die für eine Phrase Konkordanzen und Kollokationen anhand der Suchergebnisse von *Google*, *AltaVista* und *Northern Light* erzeugt.[43] Resnik und Smith (2002) stellen das System STRAND vor, das unter anderem das „Internet Archive“ (`http://www.archive.org`) verwendet, um Webseiten zu finden, die in mehrsprachigen Versionen existieren. Da Autoren häufig identische Dokumentstrukturen verwenden, wenn Inhalte in mehreren Sprachen auf jeweils eigenständigen Webseiten publiziert werden, können Lernverfahren, die sowohl die Dokumentstruktur als auch den Inhalt berücksichtigen, eingesetzt werden, um bilinguale Korpora für Sprachen aufzubauen, für die derartige Ressourcen bislang nicht erhältlich sind, z. B. Arabisch und Englisch (vgl. auch Li et al., 2001). Grefenstette (1999) benutzt die Suchmaschine *AltaVista*, um Übersetzungskandidaten für Komposita im Rahmen eines maschinellen Übersetzungssystems zu identifizieren. Die Bestandteile einer Phrase oder eines Kompositums werden für die Sprachpaare Deutsch und Englisch sowie Spanisch und Englisch für alle Lesarten einzeln übersetzt, woraufhin alle Kombinationen gebildet und deren individuelle Frequenzen anhand der Suchmaschine ermittelt werden. Grefenstette zeigt, dass auf diese Weise in ca. 88% aller Fälle die beste Übersetzung gefunden werden kann. Mit derartigen Frequenzangaben arbeitet auch Volk (2000, 2001, 2002), um einen Algorithmus zur syntaktischen Verankerung von Präpositionalphrasen anhand der Menge gruppierter Kookkurrenzangaben zu verbessern. Heyer et al. (2001) beschreiben das Projekt „Deutscher Wortschatz“ (`http://www.wortschatz.uni-leipzig.de`), in dem unter anderem HTML-Dokumente als Quellen zur Erstellung monolingualer Korpora eingesetzt werden. Filterkomponenten werden benutzt, um z. B. Eigennamen zu erkennen oder Kollokationen aufzubauen, die auch visualisiert werden können. Walker (1999) stellt das Projekt Snapshot vor, mit dessen Hilfe ein TEI-annotiertes Korpus von HTML-Dokumenten entstehen soll. Ein *Crawler* traversiert verschiedene Webserver und gibt Dokumente an einen Parser weiter, der sowohl Hyperlinks identifiziert und dem *Crawler* zurückmeldet, als auch robuste Strategien zur fehlertoleranten Verarbeitung und Konvertierung von HTML-Dokumenten beinhaltet. Auf diese Weise können TEI-annotierte Versionen der traversierten Webdokumente produziert werden.

7.5.3 Digitale Bibliotheken und Archivierung des World Wide Web

Digitale Bibliotheken (Endres und Fellner, 2000) zielen nicht nur auf die effizientere Recherche von Buch- und Zeitschriftenbeständen ab. Ein Aspekt betrifft auch die Archivierung von HTML-Dokumenten und vollständigen Websites (Sanz et al., 1998). Die prominenteste und fortgeschrittenste Initiative ist das 1996 ins Leben gerufene „Internet Archive“ (`http://www.archive.org`), das mit effektiven *Crawlern* (Burner, 1997) und aufwändigen

[43] Derartige Funktionen werden mittlerweile auch für das Sprachenlernen und -lehren auf der Grundlage von Korpora diskutiert (vgl. Foucou und Kübler, 2000, und Fletcher, 2001).

Archivierungsverfahren versucht, „a digital library of Internet sites and other cultural artifacts in digital form“ zu erstellen.[44] Mittels der „Wayback Machine“ kann auf den im August 2005 mehr als 40 Milliarden Webseiten umfassenden Datenbestand zugegriffen werden. Für eine URL können alle im Archiv enthaltenen Versionen aufgelistet, angezeigt und navigiert werden. Neben dem Einsatz zentraler Server für die „monumentale Aufgabe“ (Burner, 1997) der Archivierung des WWW wird auch der Einsatz von Peer-to-Peer-Methoden diskutiert.

Es existieren verschiedene Projektverbünde, die sich der dauerhaften Archivierung von WWW-Inhalten widmen (vgl. z. B. Campos und Silva, 2001). Die meisten dieser Ansätze wurden von Nationalbibliotheken initiiert. Hierzu zählen z. B. Projekte der Österreichischen Nationalbibliothek (AOLA, *Austrian On-Line Archive*, Rauber et al., 2002) oder der Bibliothèque nationale de France (Abiteboul et al., 2002). Vergleichbare Projekte wurden in Skandinaiven (*Nordic Web Archive*, `http://nwa.nb.no`), Australien (*Preserving and Accessing Networked Documentary Resources of Australia*, `http://pandora.nla.gov.au`), den USA (`http://www.digitalpreservation.gov`) und Großbritannien (*Digital Preservation Coalition*, `http://www.dpconline.org`) ins Leben gerufen. Eine derartige Initiative existiert zwar in Deutschland bislang nicht, mit dem vom Bundesministerium für Bildung und Forschung geförderten Projekt „Langzeitverfügbarkeit digitaler Dokumente“ (`http://www.dl-forum.de/Foren/Langzeitverfuegbarkeit/`) wird jedoch in Kooperation mit Projekten anderer Länder bis Mai 2006 „eine Initialzündung für eine kooperative Lösung angestrebt“.

7.5.4 Text Retrieval und Information Retrieval

Zahlreiche Forschungsprojekte, die sich dem *Text Retrieval* oder dem *Information Retrieval* widmen, arbeiten mit Kollektionen, die auf umfangreichen *Crawls* spezifischer Teile des WWW bestehen. Exemplarisch werden an dieser Stelle der „Web Track“ der TREC-Reihe sowie das WebBase-Projekt dargestellt. Die jährlich stattfindenden TREC-Konferenzen beschäftigen sich mit der Anwendung von IR-Verfahren auf sehr große Textkollektionen. Forschungsprototypen und Produktionssysteme partizipieren in einzelnen Tracks, die spezielle IR-Szenarien reflektieren („Question Answering Track“, „Query Track“ etc.). Bei der TREC-7-Konferenz im Jahr 1998 wurden den Teilnehmern erstmals Aufgaben in einem „Very Large Collection Track“ gestellt (Hawking et al., 1999b). Hawking et al. (1999a) stellen die hierfür eingesetzte VLC2-Kollektion vor: Es handelt sich um ein etwa 100 Gigabyte umfassendes Korpus, das vom *Crawler* des Internet Archive im Frühjahr 1997 gesammelt und auf Datenbändern zur Verfügung gestellt wurde. Die Kollektion umfasst 18,5 Mio. Dokumente von 116 102 unterschiedlichen Webservern. Die einzelnen Webseiten der Kollektion werden inhaltlich nicht modifiziert, aber jeweils in eine SGML-Instanz eingebettet, die dem üblichen TREC-Format entspricht (vgl. auch Hawking et al., 2000). Das WebBase-Projekt, das an der Stanford University durchgeführt wurde und ein Folgeprojekt von *Google* (Brin und Page, 1998) ist, verfolgt die Anfertigung einer möglichst effizienten technologischen Infrastruktur für ein „Web Repository“, das für Experimente mit IR-Algorithmen und Textklassifikationsverfahren eingesetzt werden kann (Arasu et al., 2001). Hirai et al. (2000) stellen diese verteilte

[44] Der vom „Internet Archive“ eingesetzte *Crawler* „Heritrix“ steht seit Anfang 2005 als Open-Source-Software zur Verfügung (vgl. `http://crawler.archive.org`).

Infrastruktur vor, die auf einem *Crawler*, einem Storage-Manager, einer Query-Engine, einem Multicast- und einem Indexing-Modul beruht. Es existieren drei Zugriffsmöglichkeiten auf die in einem Korpus verfügbaren Dokumente: Eine Webseite kann über ihre URL oder eine ID auf einen Client übertragen werden, mit Datenbankanfragen können Dokumente nach spezifischen Suchkritierien extrahiert werden, und der Datenstrom-basierte Zugriff gestattet schließlich die Übertragung einer Teilmenge des Gesamtbestandes. Ein Ziel dabei ist, die Daten auch anderen Institutionen zur Verfügung zu stellen, denn dies „would make it unnecessary for other sites to crawl and store the web themselves“ (Hirai et al., 2000).

7.5.5 Kommerzielle Werkzeuge

Haas und Grams (1998a,b, 2000) benutzen *AltaVista* zur Erstellung einer 75 Einträge umfassenden Liste von Webseiten, die daraufhin – gemeinsam mit den HTML-Dokumenten der ersten Verknüpfungsebene – mit dem kommerziellen Werkzeug *Web Whacker* (vgl. `http://www.bluesquirrel.com/products/whacker/`) traversiert und auf einen lokalen Rechner kopiert wurden. Mittlerweile sind derartige kommerzielle Werkzeuge in zahlreichen Ausprägungen verfügbar, die meisten richten sich an Heimanwender mit Betriebssystemen der *Microsoft Windows*-Familie (vgl. LeFurgy, 2001, sowie die von Braun, 2005, zusammengestellte Liste von Werkzeugen). Einige dieser Systeme (beispielsweise *Teleport Exec* in der „Very Large eXploration“-Version, `http://www.tenmax.com/teleport/execvlx/`) bieten sehr umfangreiche Funktionen, z. B. die Traversierung von 40 Millionen URLs oder die Verarbeitung von CSS-Dateien und *JavaScript*-Code zur Auffindung möglichst vieler Dateien.

7.5.6 Fazit

Eine der zentralen Thesen der vorliegenden Arbeit lautet, dass eine präzise Sammlung von Hypertextsorten ohne Einschränkungen bezüglich der Untersuchungsdomäne nicht durchführbar ist. Die von Crowston und Williams sowie Haas und Grams realisierte zufällige Auswahl von Webseiten anhand einer Suchmaschine (vgl. Abschnitt 7.5.1) stellt keinen geeigneten Weg für ein derartiges Vorhaben dar. Das *Google*-API könnte prinzipiell zur Datensammlung eingesetzt werden, es weist jedoch verschiedene Nachteile auf. Die Limitierung auf 1 000 Anfragen pro Tag stellt diesbezüglich einen wesentlichen Flaschenhals dar. Weiterhin ist nicht garantiert, dass der *GoogleBot* tatsächlich *alle* HTML-Dokumente eines bestimmten Webservers oder einer bestimmten Domäne indexiert hat oder dass eine spezifische Version einer bereits analysierten Webseite permanent im *Google*-Cache verfügbar ist.

Die in Abschnitt 7.5.2 thematisierten Ansätze, die nur einen kleinen Teil der WWW-bezogenen korpus- und computerlinguistischen Arbeiten darstellen, weisen zwei Charakteristika auf: Zum einen besitzen HTML-Dokumente meist den Status einer Ressource, die in sprachverarbeitenden Szenarien gewinnbringend eingesetzt werden kann (wie z. B. bei Grefenstette oder Volk). Zum anderen steht praktisch immer der eigentliche textuelle Inhalt einer Webseite im Vordergrund, d. h. Markup-Informationen spielen nur eine untergeordnete Rolle. Die HTML-Annotation wird in der Regel unwiderruflich aus den Dokumenten entfernt (z. B. bei Fairon, 2000). Die Auffassung von Webseiten als eigenständigen, genuin linguistischen Untersuchungsgegenstand in dezidiert korpus- oder computerlinguistischen Kontexten wird meines Wissens bislang nicht berücksichtigt.

Die Liste von Projekten, die sich im Kontext digitaler Bibliotheken mit der Archivierung des WWW beschäftigen sowie die in Abschnitt 7.5.2 dargestellten Bemühungen deuten an, dass mit einer Konvergenz korpuslinguistischer und bibliotheksbezogener Ansätze zu rechnen ist. Die möglichst vollständige Gewinnung, effiziente Speicherung und umfassende Analyse großer Teilbereiche des WWW – bzw. im Falle des Internet Archive des gesamten Dokumentbestandes – nimmt unter beiden Gesichtspunkten eine zentrale Position ein. Das Internet Archive bietet Wissenschaftlern mittlerweile einen Zugang an, der auf dem Telnet-ähnlichen SSH-Protokoll basiert. Auf diese Weise erhalten registrierte Benutzer einen rein textbasierten Zugriff auf die gesammelten Bestände. Die Voraussetzung hierfür sind jedoch UNIX- und Programmierkenntnisse, da nur sehr wenige, äußerst rudimentäre Werkzeuge zur Analyse der Daten zur Verfügung gestellt werden. Prinzipiell wäre das Internet Archive zwar eine Alternative zu dem in diesem Kapitel vorgestellten Korpus, jedoch existierten die genannten Zugangsmöglichkeiten bei Aufnahme der vorliegenden Arbeit noch nicht, zudem hätten entsprechende Zugriffs- und Analysemöglichkeiten implementiert werden müssen.

Die exemplarisch dargestellten Projekte aus den verwandten Bereichen TR und IR (Abschnitt 7.5.4) zeigen, dass bei derartigen Fragestellungen andere Aspekte im Vordergrund stehen als bei der computer- oder korpuslinguistischen Analyse von HTML-Dokumenten. Die VLC2-Daten werden von den am Web Track teilnehmenden Systemen vollautomatisch untersucht, d. h. ein Zugriff auf die Dokumente selbst muss, falls er benötigt wird, von Seiten der Teilnehmer implementiert werden. Die Architektur des WebBase-Systems ist sehr komplex, effizient und bietet prinzipiell die Möglichkeit, Daten auch an Dritte zu distribuieren. Einige für die vorliegende Arbeit zentralen Funktionen, die z. B. die Analyse der Daten mit einer grafischen Oberfläche betreffen, sind jedoch nicht gegeben.

Kommerzielle Werkzeuge (Abschnitt 7.5.5) wurden beim Aufbau der Korpusdatenbank für die Phase der Datensammlung aus mehreren Gründen nicht in Betracht gezogen, denn die Leistungsfähigkeit der preisgünstigen Produkte ist zu gering, um mehrere Millionen Webseiten zu traversieren. Diejenigen Tools hingegen, die hierzu in der Lage sind, besitzen einen entsprechenden Preis (z. B. mehrere tausend US-Dollar im Fall von *Teleport Exec VLX*).

7.6 Zusammenfassung

Dieses Kapitel stellt die Korpusdatenbank vor, die zur Sammlung, Betrachtung und Analyse deutschsprachiger HTML-Dokumente und verschiedener zusätzlicher Dateiformate entwickelt wurde. Zur Datensammlung wird ein *Crawler* eingesetzt, der – konfiguriert mit einer Startadresse – vollautomatisch sämtliche Webseiten einer spezifizierten Domäne rekursiv besucht und in das Korpus integriert, sofern nicht verschiedene Aufnahmebeschränkungen verletzt werden. Zur Identifizierung der Sprache von Dokumenten wurde ein automatischer Sprachenidentifizierer entwickelt, der mit einer Präzision von 96,64% arbeitet. Die Dokumente selbst werden vom *Crawler* im Dateisystem eines *Linux*-Rechners gespeichert, wohingegen diverse Metadaten, d. h. Informationen *über* die Dokumente, in einer relationalen Datenbank abgelegt werden. Diese hybride Datenhaltung ermöglicht flexible Zugriffsfunktionen, die indirekt und automatisch von speziellen Programmen oder direkt durch die manuelle Benutzung eines Browsers verwendet werden können. Für diesen Zweck wurde ein Front-End implementiert, das umfangreiche Navigations- und Analysefunktionen anbietet

und modular aufgebaut ist. Einige dieser Funktionen setzen nicht arbiträres HTML-Markup, sondern wohlgeformte XML-Strukturen voraus, weshalb ein *Perl*-Modul implementiert wurde, das eine automatische Transformation beliebiger HTML-Dokumente nach XHTML Transitional erlaubt (vgl. Abschnitt 14.4, S. 652 ff.). Die indirekt oder direkt durchgeführte Übertragung von im Korpus verfügbaren Dokumenten kann mit Hilfe des *Hypertext Transfer Protocol*s erfolgen, da auf dem Datenbank-Server ebenfalls ein Webserver installiert ist, der wiederum lesenden Zugriff auf den gesamten Korpusbestand besitzt.

7.7 Fazit

Es existieren verschiedene korpusbezogene Projekte und Prototypen mit einer Ausrichtung auf das WWW, in denen oftmals Verfahren zur automatischen Sprachenidentifizierung eingesetzt werden. Keines dieser Systeme bietet jedoch die Ausrichtung, den Funktionsumfang oder die Flexibilität, die die in diesem Kapitel vorgestellte Korpusdatenbank aufweist: Das System umfasst die drei Bausteine der Metadaten (relationale Datenbank), Datenhaltung (Dateisystem) und des Datenzugriffs (direkt mittels Web-Oberfläche oder indirekt mittels autark arbeitender Analyse-Werkzeuge). Diese Bausteine können, falls sehr umfangreiche Korpora aufgebaut werden sollen, zum Zwecke der Lastverteilung auf separaten Maschinen betrieben werden. Da die Kommunikation der Komponenten ausschließlich netzwerkbasiert arbeitet und beide Zugriffsmöglichkeiten das Client-Server-Prinzip nutzen, können mehrere Personen simultan mit dem System arbeiten. Die Korpusdatenbank basiert meines Wissens als einziges System auf dem intuitiven Paradigma der lokalen Spiegelung von HTML-Dokumenten, die anschließend – eingebettet in die durch das Front-End implementierte Navigations- und Analyseoberfläche – mit einem Browser betrachtet und untersucht werden können. Neuartig ist auch der Ansatz, Stichproben von Webseiten mit Hilfe von SQL-Querys bezüglich eines Bestandes von Metadaten zufällig berechnen zu lassen. Weiterhin sind durch Modifikationen der *PHP*-Skripte des Front-Ends vielfältige Erweiterungen und Anpassungen an weitere Forschungsschwerpunkte denkbar. Ein Nachteil des Systems betrifft die Tatsache, dass in der derzeit vorliegenden Version der Aufbau eines dynamischen Monitorkorpus nicht möglich ist, d. h. es ist nicht vorgesehen, im Korpus enthaltene HTML-Dokumente in regelmäßigen Intervallen erneut von den originären Webservern in das Korpus zu übertragen, um z. B. Aktualisierungen der Dokumente untersuchen zu können. Für die vorliegende Arbeit ist dieser Aspekt nicht relevant, jedoch sollten derartige Mechanismen für eine etwaige zweite Version des Systems in Betracht gezogen werden. Die Quelltexte des Systems werden nach Fertigstellung der Arbeit im WWW als Open-Source-Software veröffentlicht. Hierzu gehören die Tabellendefinitionen und Lexika sowie die zahlreichen Shell-, *PHP*- und *Perl*-Skripte, die für die Datensammlung, Sprachenidentifizierung, Verwaltung und Bedienung der Korpusdatenbank notwendig sind. Das *Perl*-Skript `build-db-import.pl` implementiert ein robustes Verfahren, die vom *Crawler Pavuk* gespeicherten HTTP-Response-Header zu verarbeiten. Aufgrund dieser spezifizischen Ausrichtung soll dieses Skript als zusätzliches Werkzeug in *Pavuk* einfließen, da es auch in anderen Kontexten eingesetzt werden kann (z. B. zur Konvertierung der Header-Informationen in ein XML-basiertes Format).

Teil III

Analysen und Sammlungen von Hypertextsorten

Überblick

Die nachfolgenden fünf Kapitel stellen Analysen vor, in denen verschiedene Aspekte von Hypertextsorten, Hypertextknotensorten und Hypertextsortenmodulen untersucht werden. Persönliche und private Homepages können als diejenigen Hypertextsorten gelten, die bislang am umfangreichsten erforscht wurden, weshalb sich die ersten drei Analysen dieser Thematik widmen. Kapitel 8 präsentiert eine vergleichende Untersuchung, in der die Frequenzen maschinell detektierbarer Merkmale für konzeptionelle Mündlichkeit in privaten Homepages von Studierenden und in persönlichen Homepages von Wissenschaftlern ermittelt werden. In Kapitel 9 werden 15 private Homepages von Studierenden einer Inhalts- und Makrostrukturanalyse unterzogen, um die beteiligten Hypertextsortenmodule und ihre Frequenzen zu erheben. Gegenstand von Kapitel 10 sind 100 persönliche Homepages von Wissenschaftlern, die ebenfalls hinsichtlich ihrer Inhalte und Makrostrukturkomponenten untersucht werden. Ebenso wie die zweite Analyse mündet auch diese Untersuchung in einem Hypertextsortenprofil; darüber hinaus werden metadiskursive Äußerungen der jeweiligen Produzenten thematisiert, die das in Kapitel 4 vorgestellte zyklische Modell der Entwicklung von Hypertextsorten bestätigen. Kapitel 11 umfasst zwei Phasen: Zunächst werden die – als institutionelle Homepages konzeptualisierbaren – Einstiegsseiten der Webauftritte von 35 Universitäten mit dem Ziel untersucht, sie mit kommerziell ausgerichteten Homepages zu vergleichen und ein Profil dieser Hypertextknotensorte zu erstellen. Die zweite Phase fokussiert diejenigen Dokumente, die durch Hyperlinks von den 35 Einstiegsseiten erreichbar sind und basiert auf der Hypothese, dass die Sammlung und Identifizierung der korrespondierenden Hypertextknotensorten und Hypertextsorten geeignete Werkzeuge zur Konstruktion der „oberen“ Ebenen einer Typologie von Hypertextsorten sind. Kapitel 12 komplementiert diesen Ansatz, indem für 750 zufällig ausgewählte und „tief“ eingebettete Dokumente die Hypertextknotensorten und übergeordneten Hypertextsorten ermittelt werden, um somit die „unteren“ Ebenen der Typologie bestimmen zu können. Die untersuchten Dokumente stammen aus der in Kapitel 7 vorgestellten Korpusdatenbank und wurden mittels der dort dargestellten Zugriffsmöglichkeiten analysiert. Es ist zu betonen, dass sich diese fünf Kapitel an dem in Kapitel 5 eingeführten Hypertextsortenmodell orientieren, weshalb – dem Modell entsprechend – eine abstraktere Perspektive eingenommen wird, die sich unter anderem auf die Identifizierung und Sammlung von Hypertextsorten, Hypertextknotensorten und Hypertextsortenmodulen bezieht. Für die Untersuchungsdomäne ist also zunächst ein Inventar von Hypertextsorten und der beteiligten Konstituenten zu erarbeiten.

8

Analyse 1: Quantitative Auswertung persönlicher Homepages

8.1 Einleitung

Da die *persönliche Homepage* als die von der linguistischen Literatur bislang am umfangreichsten untersuchte Hypertextsorte gelten kann (vgl. Abschnitt 4.6.3), beziehen sich die ersten drei Analysen auf diese Thematik. Die erste Studie betrifft eine vornehmlich quantitative Auswertung, um grundlegende Unterschiede zwischen privaten Homepages von Studierenden und persönlichen Homepages der Mitarbeiter von Universitäten zu ermitteln.[1]

Abschnitt 8.2 geht zunächst auf die Ziele der Untersuchung ein, woraufhin Abschnitt 8.3 in den Bereich der linguistischen Betrachtung computervermittelter Kommunikationsformen einführt. Die drei Stichproben werden in Abschnitt 8.4 vorgestellt. Die Abschnitte 8.5 bis 8.7 präsentieren die Ergebnisse der Studie.

8.2 Ziele und Bezüge zum Hypertextsortenmodell

Mit dieser Analyse werden verschiedene Ziele verfolgt. Zunächst wird untersucht, ob auf der Textoberfläche subgenerische Varietäten in den Exemplaren zweier Varianten des Hypertexttyps *persönliche Homepage* ermittelt werden können. Die Analyse bezieht sich insbesondere auf Merkmale für konzeptionelle Mündlichkeit (vgl. Abschnitt 2.2.7). Verschiedene Ausprägungen derartiger Merkmale, mit denen der Produzent die Erzeugung kommunikativer Nähe intendiert, wurden für die Kommunikationsdienste E-Mail, Usenet und insbesondere den *Internet Relay Chat* berichtet. Walker (2000, S. 107 f.) zeigt, dass Personen, die das IRC verwenden, häufig private Homepages anfertigen, so dass sich Fremde zumindest ein rudimentäres Bild von ihrem Gegenüber innerhalb der virtuellen Gesprächsrunde machen können –

[1] Dieses Kapitel basiert auf einer Studie, die bereits in Auszügen publiziert wurde (Rehm, 2002a).

Walker bezeichnet sie als intrinsische Homepages, da sie primär für Rezipienten erstellt werden, die dem Produzenten unbekannt sind (vgl. Abschnitt 4.6.3). Aus diesem Grund kann von der Hypothese ausgegangen werden, dass Personen, die die Kommunikationsmedien IRC, E-Mail und Usenet sehr häufig einsetzen und folglich mit den sprachlichen Besonderheiten dieser Medien vertraut sind, bei der Erstellung ihrer privaten Homepages ebenfalls derartige sprachliche Spezifika einsetzen (z. B. um Medienkompetenz, die Beherrschung der sprachlichen Konventionen und Codes und Gruppenzugehörigkeit zu einem bestimmten, vom Produzenten häufig frequentierten IRC-Kanal zu signalisieren).

Viele Merkmale für konzeptionelle Mündlichkeit, die in der CMC-Literatur beschrieben werden, besitzen die Eigenschaft, mit einer hohen Präzision maschinell erkannt werden zu können. Falls die Analysen zeigen, dass in sehr vielen Homepages von Studierenden derartige Merkmale eingesetzt werden, könnte es hinsichtlich der maschinellen Identifizierung von Hypertextsorten möglich sein, korrespondierende Textexemplare unter anderem auf der Grundlage dieser Eigenschaft zu erkennen und diejenigen Hypertextsorten auszuschließen, in deren Textexemplaren diese Merkmale in der Regel nicht eingesetzt werden.[2]

Im Hinblick auf das Hypertextsortenmodell beziehen sich die Dokumente der Stichproben jeweils auf die Ebene der Instanz einer Hypertextsorte. Mit Hilfe der Korpusdatenbank wurden Webserver ermittelt, die ausschließlich private Homepages von Studierenden oder persönliche Homepages von Mitarbeitern anbieten, woraufhin *alle* Dokumente in die korrespondierenden Stichproben integriert wurden, die auf diesen Webservern in den für persönliche Homepages reservierten Adressbereichen verfügbar sind. Eine Differenzierung zwischen spezifischen Hypertextknotensorten kann in der Analyse in quantitativer Hinsicht nicht stattfinden, weil die Zuordnung von einem Dokument zum jeweiligen Hypertext nicht verfügbar ist (vgl. Abschnitt 14.6.1). Der Vorgehensweise liegt die Hypothese zugrunde, dass die genannten Hypertextsortenvarianten existieren und Unterschiede zwischen ihnen bereits auf dieser globalen Ebene erfasst werden können. Die qualitative Auswertung der Vorkommen der genannten Merkmale gibt zudem Hinweise auf die Hypertextsorten, die in die jeweiligen Instanzen persönlicher bzw. privater Homepages eingebettet sind.

8.3 Konzeptionelle Mündlichkeit in der computervermittelten Kommunikation

Koch und Oesterreicher (1994) differenzieren zwischen medialer und konzeptioneller Mündlichkeit und Schriftlichkeit (vgl. Abschnitt 2.2.7). Obwohl diese beiden Ebenen unabhängig voneinander zu betrachten sind, besteht doch eine gewisse Affinität zwischen medialer und konzeptioneller Mündlichkeit sowie medialer und konzeptioneller Schriftlichkeit. Diese Affinität gilt jedoch, wie in zahlreichen Arbeiten gezeigt wurde, für digitale, medial schriftlich realisierte Kommunikationsmedien nicht in diesem Ausmaß, denn in E-Mails, IRC-Sitzungen oder Usenet-Beiträgen können sehr häufig (para)linguistische Phänomene beobachtet werden, die der konzeptionellen Mündlichkeit nahe stehen (vgl. etwa Haase et al.,

[2] Bayerl (2002) und Bittner (2003) untersuchen Stichproben privater Homepages und ermitteln verschiedene Merkmale für konzeptionelle Mündlichkeit (vgl. Abschnitt 4.6.3).

1997).[3] Lenke und Schmitz (1995) untersuchen in einem der ersten im deutschsprachigen Raum verfassten Beiträge zur computervermittelten Kommunikation (CMC, *Computer-Mediated Communication*) die Internet-Dienste E-Mail, Usenet und IRC: Neben einer im Allgemeinen informellen, „kollegiale[n] und zwanglosen Form" (ebd., S. 118) der Kommunikation mittels E-Mail werden Smileys und verbale Beschreibungen realer und fiktiver Handlungen als Mittel eingesetzt, um das Fehlen non-verbaler Signale zu kompensieren.[4] Feldweg et al. (1995) konzentrieren sich auf deutschsprachige Newsgruppen. Anhand von Frequenzanalysen des mehr als 430 000 Beiträge umfassenden 1993er Jahrgangs beinahe aller deutschen Newsgruppen, die mit dem Korpus einer deutschen Tageszeitung kontrastiert werden, weisen sie hochfrequente Wörter nach (z. B. „ich", „man", „du", „mal", „einfach", „ziemlich" und „irgendwie"), die für den Einfluss gesprochener Sprache auf dieses Kommunikationsmedium kennzeichnend sind. Günther und Wyss (1996) untersuchen ein Korpus von in der Schweiz produzierten E-Mails und ermitteln sprachliche Phänomene, die als „Elemente der Mündlichkeit" bezeichnet werden. Die Autorinnen berichten Regionalismen und dialektale Ausdrücke, produktionsbedingte Normabweichungen, Dialogizität, insgesamt sehr kurze Texte und eine „Bildlichkeit" (ebd., S. 75), die sich durch den Einsatz von Smileys, individuellen Formatierungen, ASCII-Art und durch Abkürzungen wie z. B. ROTFL („rolling on the floor laughing") oder IMHO („in my humble opinion") konstituiert. Quasthoff (1997) untersucht die Mailing-Listen „Linguist-List" und „Ethno-List" (vgl. Gruber, 1997, sowie Abschnitt 4.2.1) und entdeckt viele orthografische Fehler, durchgängige Kleinschreibung, Abkürzungen und „tageszeitorientierte Grußformel[n]" (ebd., S. 42). Durch die hierdurch hervorgerufene „Flüchtigkeit der Botschaft" wird „eher der Rahmen einer schnell hingeworfenen Notiz als eines Briefes erzeugt." (ebd.). Haase et al. (1997) verknüpfen eine Betrachtung der sprachlichen Spezifika von E-Mail, Usenet und IRC mit dem Modell von Koch und Oesterreicher und kommen zu dem Schluss, dass in diesen medial schriftlich realisierten Kommunikationsformen verschiedenste Merkmale konzeptioneller Mündlichkeit existieren. Unter anderem werden Ideogramme (Smileys), Zustands- und Gefühlsäußerungen mittels prädikativ eingebetteter und häufig durch Sonderzeichen markierter Verbstämme, Deiktika, emuliertes Flüstern und emulierte Prosodie, die Iterationen von Satzzeichen und einzelnen Buchstaben, Abkürzungen, Akronyme und aus dem Internet- und UNIX-Jargon stammenden Slangausdrücke sowie deren Übergeneralisierung diskutiert. Pansegrau (1997) beschäftigt sich mit der „Dialogizität und Degrammatikalisierung in E-mails", wobei Anredesequenzen und die in der E-Mail-Kommunikation vorhandene erhöhte Toleranz in Bezug auf Orthografie-, Interpunktions- und Grammatikfehler untersucht werden. Aufgrund dieser Merkmale sowie der insgesamt feststellbaren sprachlichen Kreativität „wird argumentiert, daß

[3] Einen Überblick liefern Beißwenger (2001) für die Chat-Kommunikation sowie Ziegler und Dürscheid (2002) für die E-Mail-Kommunikation. Einen weiteren Schwerpunkt stellt die Untersuchung der Kommunikation in MUDs und MOOs (*Multi User Dungeon*-Abenteuerspiele) dar (vgl. z. B. Grigar, 2002, und Renner, 2003). Dürscheid (2004) erweitert das Modell von Koch und Oesterreicher (1994) um eine Ebene, die zur Erfassung synchroner, quasi-synchroner und asynchroner Kommunikation eingesetzt wird.

[4] In ihrem Ratgeber zum effektiven Einsatz des Mediums E-Mail erläutern Angell und Heslop (1994, S. 111) Smileys wie folgt: „Too often the lack of inflection or facial expression can cause a typed phrase in an e-mail message to be interpreted incorrectly. A visual shorthand using *smileys* or *emoticons* has emerged to help the reader decipher the writer's original intent. Smileys are the equivalent of e-mail slang and should not be used in formal business e-mail messages. Also keep in mind that overuse of smileys marks you as a beginner."

E-mails nicht einen defizitären Stil, sondern eine zweckmäßige und kreative Anpassung an veränderte Kommunikationskanäle repräsentieren“ (ebd., S. 95). Runkehl et al. (1998) nehmen sehr umfangreiche Analysen der E-Mail-, Usenet- und IRC-Kommunikation vor. Sie untersuchen anhand verschiedener Korpora Merkmale wie durchgängige Kleinschreibung, Bigraphen, unterschiedliche Typen von Fehlern, Akronyme, Smileys, Assimilationen, Reduktionen, Iterationen, Anreden und Verabschiedungen, Signaturen, Dialektismen, Diskurspartikeln und Interjektionen: „Je stärker die Kommunikation dialogischer und synchroner erfolgt [E-Mail → Usenet → IRC, G. R.], desto häufiger lassen sich mündliche Aspekte des Sprachgebrauchs in der Internet-Kommunikation feststellen.“ (ebd., S. 116). Grzega (1999) untersucht auf der Grundlage traditioneller und elektronischer Briefe die Differenz zwischen herkömmlicher und digitaler Post sowie deren gegenseitige Beeinflussung. Es wird betont, dass *kein* genereller „e-style“ (ebd., S. 15) existiert: „The boundaries between formality and informality (private letter vs. business letter) appear much more fuzzy, but it seems entirely unjustified to speak of overall present features.“ (ebd., S. 16). Dürscheid (1999) untersucht, welche linguistischen Merkmale für die Internetkommunikation kennzeichnend sind. Neben der Zuordnung von Smileys, unflektierten Verbformen (oftmals Verbletztkonstruktionen), orthografischen Fehlern und Interjektionen wird festgestellt, dass auch wiederholte Ausrufe- und Fragezeichen sowie iterierte Buchstaben und konstante Großschreibung Merkmale für konzeptionelle Mündlichkeit sind, wobei die Authentizität im Vordergrund steht: „Das Motto [im Chat, G. R.] scheint zu sein: ›Schreib, wie Du sprichst‹ und ›Schreib so schnell, wie Du kannst‹.“ (ebd., S. 21; vgl. auch Hoffmann, 2004). Storrer (2000a, S. 153 f.) fasst die für die Internet-Kommunikation wesentlichen Charakteristika konzeptioneller Mündlichkeit zusammen: Bezüglich der Lexik herrscht offenbar eine Präferenz für einfache und kurze Wörter, sowie umgangssprachlich markierte (vgl. Walters, 1996) und dialektale Ausdrücke. Auf der syntaktischen Ebene findet sich häufig ein parataktischer und teils fehlerhafter Satzbau, der oftmals sprechsprachliche Konstruktionen aufweist. Weiterhin ist für Storrer eine „freie, assoziative, dialogisch gesteuerte Themenentwicklung“ (ebd., S. 154) charakteristisch, die an das alltägliche Gespräch zwischen vertrauten Gesprächspartnern erinnert und auf kurze Planungs- und Verarbeitungszeiten hindeutet.

8.4 Die Stichproben

Für diese Analyse wurden mit Hilfe der in Kapitel 7 dargestellten Korpusdatenbank drei Stichproben mit insgesamt 49 728 HTML-Dokumenten erhoben; Listen dieser Dokumente befinden sich in Anhang E auf der beiliegenden CD ROM. Die Webseiten der ersten beiden Stichproben (S1, S2) sind Varianten des Hypertexttyps *persönliche Homepage* zugehörig. Die Dokumente der dritten Stichprobe dienen als Vergleichsdaten (S3).

S1: Private Homepages von Studierenden – Die erste Stichprobe enthält 25 481 HTML-Dokumente, die von manuell ausgewählten Webservern stammen und ausschließlich studentische Homepages umfassen. Hierzu gehören etwa Webserver von Studentenwohnheimen (z. B. `www.wohnheim.uni-ulm.de`) oder die an der Justus-Liebig-Universität Gießen eindeutig als solche gekennzeichneten studentischen Homepages. Diese befinden sich auf dem Server `wwwstud.uni-giessen.de` und die Pfadkomponente der URL beginnt jeweils mit `/~s`.

S2: Persönliche Homepages von Mitarbeitern – Die zweite Stichprobe beinhaltet insgesamt 14 247 HTML-Dokumente, die ebenfalls von manuell ausgewählten Webangeboten stammen. So bieten z. B. die Webserver `www.physik.uni-ausgburg.de`, `www.linguistik.uni-erlangen.de` und `www.imise.uni-leipzig.de` auf dem Muster, das der üblichen Adressierungskonvention für persönliche Webangebote entspricht (`http://www.../~.../`), mit nur sehr wenigen Ausnahmen die persönlichen Homepages von Mitarbeitern der jeweiligen Universitäten bzw. Forschungseinrichtungen an.

S3: Tief eingebettete Dokumente – Die dritte Stichprobe enthält 10 000 HTML-Dokumente, die mittels der Web-Oberfläche der Korpusdatenbank zufällig ausgewählt wurden (vgl. Abschnitt 7.3.4). Die Adressen dieser Dokumente besitzen eine Einbettungstiefe von mindestens drei und maximal zehn Verzeichnissen[5] und ihre jeweilige Pfadkomponente beginnt *nicht* mit `/~`. Durch diese Einschränkungen sollte verhindert werden, dass sich Instanzen des Hypertexttyps *Homepage einer Person* in dieser dritten Stichprobe befinden. Stattdessen enthält sie, wie die manuelle Inspektion der Daten gezeigt hat, z. B. HTML-Versionen von technischen Berichten und administrative Dokumente, d. h. Instanzen der unterschiedlichsten Hypertextsorten.

8.5 Allgemeine Charakteristika

Die quantitative Auswertung der Stichproben erfolgte mit Hilfe eines in der Programmiersprache *Perl* (Wall et al., 2000) implementierten Skripts, das das Modul `HTML::Parser` (vgl. Burke, 2002) einsetzt, um den Zugriff auf HTML-Elemente, deren Attribute und den Inhalt von Elementen – den in HTML-Dokumenten enthaltenen Text – zu ermöglichen. Nachfolgend werden zunächst allgemeine Ergebnisse im Hinblick auf Wortfrequenzen und den Umfang der Stichproben aufgeführt. Abschnitt 8.6 geht auf Merkmale für konzeptionelle Mündlichkeit ein, woraufhin Abschnitt 8.7 weiterführende Ergebnisse vorstellt.

8.5.1 Umfang der Stichproben und HTML-bezogene Merkmale

Im Hinblick auf den Umfang der Stichproben fällt auf, dass bezüglich der Anzahl von Token[6] pro Dokument nur sehr geringe Schwankungen existieren (vgl. Tabelle 8.1). Hinsichtlich des Einsatzes von Framesets zur visuellen Strukturierung eines Webangebots bildet S1 diejenige Stichprobe mit den prozentual häufigsten Vorkommen. Die Webdesign- und Usability-Literatur rät aus den verschiedensten Gründen von ihrem Gebrauch ab. Dieses einfach zu realisierende Strukturierungs- und Navigationsmittel wird im eher professionell ausgerichteten Bereich (S2, S3) und auch im gesamten Korpus weniger häufig eingesetzt. Ein Dokument, das ein `frameset` Element enthält, referenziert mehrere HTML-Dateien, die die Inhalte der

[5] Beispielsweise das Dokument `http://www.tu-harburg.de/v/studinf/mb/hs_eng/bruchme.htm`. Hierbei handelt es sich um den Kommentar zu der Lehrveranstaltung „Bruchmechanik und Schwingfestigkeit“ des Studiengangs Maschinenbau an der TU Hamburg-Harburg.

[6] Ein Token ist dabei definiert als eine aus beliebigen Zeichen (außer Zwischenraum, d. h. Leerzeichen, Tabulatorzeichen oder Zeilenwechsel) bestehende Zeichenkette (üblicherweise ein Wort, jedoch subsumiert dieser Begriff auch Abkürzungen, Zahlen, Ketten von Sonderzeichen etc.), die eine Länge von mindestens einem Zeichen umfasst. Als regulärer Ausdruck (vgl. etwa Friedl, 1997) in *Perl*: `[^\s]{1,}`.

Merkmal	S1	S2	S3	Korpus
Anzahl Dokumente	25 481	14 247	10 000	3 956 692
Gesamtanzahl Token	8 531 088	4 968 339	3 125 195	1 138 794 715
Token pro Dokument	ø 367 (Med.: 94)	ø 373 (Med.: 114)	ø 321 (Med.: 91)	ø 310 (Med.: 95)
Min./Max. Anzahl Token	1 / 53 203	1 / 28 280	1 / 32 341	1 / 73 872
Gesamtanzahl Bilder	116 617	49 969	53 948	24 327 871
Mindestens ein Bild	15 369 (63,0%)	8 967 (65,4%)	6 477 (66,7%)	2 603 103 (66,8%)
Bilder pro Dokument	ø 7,6 (Med.: 3)	ø 5,6 (Med.: 3)	ø 8,3 (Med.: 5)	ø 9,4 (Med.: 5)
Min./max. Anzahl Bilder	0 / 617	0 / 199	0 / 219	0 / 4 097
Gesamtanzahl Hyperlinks	257 047	136 074	96 777	45 673 131
Mindestens ein Hyperlink	19 906 (81,6%)	11 634 (84,9%)	8 641 (89%)	3 375 313 (86,6%)
Bilder als Hyperlinks	40 251	22 199	26 058	k. A.
Hyperlinks pro Dokument	ø 12,9 (Med.: 6)	ø 11,7 (Med.: 6)	ø 11,2 (Med.: 6)	ø 13,5 (Med.: 7)
Min./max. Anzahl Links	0 / 951	0 / 1 544	0 / 473	0 / 4 629
Token pro Hyperlinkanzeiger	ø 2 (Med.: 1)	ø 2,3 (Med.: 2)	ø 2,1 (Med.: 1)	k. A.
Verwendung von Framesets	1 075 (4,2%)	538 (3,8%)	293 (2,9%)	132 151 (3,4%)

Tabelle 8.1: Umfang und HTML-Merkmale der drei Stichproben

einzelnen Frames darstellen. Da Dateien, die ein `frameset` Element enthalten, fast ausnahmslos keinen sprachlichen Inhalt besitzen, werden sie bei der Auswertung nicht berücksichtigt. Bezüglich des Einsatzes eingebetteter Bilder (d. h. Vorkommen des HTML-Elements `img`), fallen nur wenige Unterschiede zwischen den Stichproben auf. Die Vermutung, dass bei der Gestaltung studentischer Homepages sehr extensiv mit dem Einsatz von Grafiken und Fotos umgegangen wird, kann nicht bestätigt werden. Der arithmetische Durchschnitt und der Mittelwert eingebetteter Bilder pro Dokument ist in S3 höher als in S1. Auch hinsichtlich der Verknüpfung ähneln sich die Stichproben: Zwischen 81,6% (S1) und 89% (S3) der Dokumente besitzen mindestens einen Hyperlink.[7] In allen drei Stichproben sind durchschnittlich etwa 12 Hyperlinks pro Dokument enthalten (Median: 6). Auch dieser Wert weicht nicht vom Durchschnitt des Gesamtkorpus ab (13,5, Median: 7).[8] Die Inhalte der jeweiligen Hyperlinkanzeiger werden nachfolgend genauer dargestellt.

8.5.2 Wortfrequenzen

Tabelle 8.2 zeigt die 50 hochfrequenten deutschsprachigen Token der drei Stichproben mit ihren jeweiligen prozentualen Anteilen bezüglich der Gesamtanzahl Token einer Stichprobe

[7] Dass zwischen 10% und 20% der Dokumente keinerlei Hyperlinks enthalten, ist auf zwei Umstände zurückzuführen: Einerseits stellen viele Dokumente Blattknoten dar, d. h. die in ihnen enthaltene Information ist z. B. ein Gedicht, eine Geschichte oder eine Vorlesungsankündigung, die auch ohne Hyperlinks les- und benutzbar ist, wenngleich die Navigationsmöglichkeiten hierunter merklich leiden. Andererseits stellen, gerade in S1 und S2, viele Einstiegsseiten privater und persönlicher Homepages nur provisorische Dokumente ohne jeglichen Hyperlink dar (projektierte Homepages im Sinne von Döring, 2001a), die z. B. von Rechenzentrumsmitarbeitern angelegt wurden, so dass der Webserver beim Aufruf der URL keine Fehlermeldung liefert.

[8] Amitay (2000a, S. 3) untersucht ca. 1 000 HTML-Dokumente (primär persönliche Homepages) und stellt höhere Werte fest: Durchschnittlich enthalten die Dokumente der beiden Stichproben 35,9 (Median: 14–15) bzw. 17,7 (Median: 13–14) Hyperlinks. Aufgrund dieser Übereinstimmung fasst Amitay die Anzahl von etwa 14 Hyperlinks pro Dokument als etablierte Konvention auf.

S1	S2	S3	*tageszeitung*
der (2.24) und (2.07) die (1.93) in (1.17) von (0.77) zu (0.77) den (0.74) mit (0.65) das (0.60) für (0.59) ist (0.58) des (0.55) auf (0.52) im (0.51) nicht (0.49) sich (0.48) Die (0.47) eine (0.46) ein (0.44) dem (0.40) auch (0.39) es (0.35) **ich** (0.33) an (0.32) als (0.32) oder (0.28) Sie (0.26) bei (0.26) **man** (0.26) sie (0.24) sind (0.24) einer (0.24) daß (0.24) nach (0.24) zur (0.24) zum (0.24) aus (0.23) wird (0.22) noch (0.22) nur (0.21) **Seite** (0.21) Der (0.21) wie (0.20) Das (0.20) einen (0.20) werden (0.20) so (0.19) über (0.19) aber (0.18) um (0.18)	der (2.70) und (2.56) die (1.84) in (1.37) von (0.95) des (0.81) für (0.77) den (0.72) zu (0.68) im (0.63) mit (0.62) das (0.51) Die (0.51) ist (0.48) auf (0.46) eine (0.41) sich (0.41) dem (0.39) ein (0.34) nicht (0.34) an (0.33) als (0.33) zur (0.31) oder (0.29) auch (0.28) werden (0.25) bei (0.25) einer (0.25) es (0.24) zum (0.24) durch (0.24) wird (0.22) aus (0.22) sind (0.22) **Dr.** (0.20) nach (0.20) am (0.20) Sie (0.20) Der (0.19) über (0.19) Das (0.18) daß (0.17) sie (0.17) einen (0.17) einem (0.16) wie (0.16) nur (0.15) so (0.14) noch (0.14) um (0.14)	der (2.80) und (2.38) die (2.01) in (1.30) von (0.96) des (0.81) den (0.73) für (0.73) zu (0.68) mit (0.63) im (0.60) Die (0.54) das (0.53) ist (0.50) auf (0.48) eine (0.46) sich (0.40) dem (0.39) zur (0.38) nicht (0.37) ein (0.35) als (0.33) oder (0.32) an (0.30) auch (0.30) werden (0.28) einer (0.28) bei (0.28) wird (0.28) Sie (0.23) durch (0.23) sind (0.23) aus (0.23) zum (0.23) es (0.22) nach (0.22) über (0.20) Der (0.20) Das (0.18) einem (0.17) daß (0.17) einen (0.17) wie (0.16) nur (0.16) sie (0.15) **Dr.** (0.15) am (0.15) kann (0.14) um (0.14) eines (0.14)	der (3.29) die (3.20) und (2.19) in (1.71) den (1.22) von (0.98) zu (0.94) das (0.89) mit (0.81) sich (0.80) nicht (0.80) ist (0.76) für (0.72) auf (0.71) des (0.69) im (0.68) Die (0.66) dem (0.66) ein (0.63) eine (0.57) es (0.50) als (0.49) auch (0.48) an (0.44) daß (0.44) sie (0.42) aus (0.41) werden (0.39) **hat** (0.38) er (0.37) nach (0.34) noch (0.33) Der (0.33) einer (0.32) wie (0.31) wird (0.31) sind (0.31) um (0.31) am (0.30) bei (0.29) vor (0.29) so (0.29) nur (0.28) Das (0.28) über (0.28) haben (0.27) einem (0.26) einen (0.26) zum (0.25) **war** (0.24)

Tabelle 8.2: Die je 50 häufigsten deutschsprachigen Token aus S1–S3 sowie aus dem ersten Halbjahr des Jahrgangs 1994 der *tageszeitung* (7 654 357 Token).

(vgl. Tabelle 8.1). Alle englischsprachigen Begriffe, nur einen Buchstaben umfassende Zeichenketten (z. B. „a“) sowie nicht druckbare Sonderzeichen wurden aus den Listen entfernt; zum Vergleich werden ebenfalls die häufigsten Token aus dem ersten Halbjahr 1994 der *tageszeitung* dargestellt. Bei einem Vergleich der vier Listen setzt sich der im vorangegangenen Abschnitt konstatierte Trend der Ähnlichkeit auf den ersten Blick fort. Die jeweils etwa 20 häufigsten Wörter sind in den vier Stichproben beinahe identisch (definite und indefinite Artikel, Präpositionen sowie Konjunktionen und Disjunktionen). Bereits an Position 23 findet sich in der Stichprobe der studentischen Homepages das Pronomen „ich“ (S2: 62, S3: 88, *taz*: 67), das gemeinsam mit dem ebenfalls nur dort vorkommenden „man“[9] (S1: 29, S2: 52, S3: 51, *taz*: 54) einen eher informellen und in der direkten Anrede gehaltenen sprachlichen Stil der studentischen Homepages andeutet. Dieses Ergebnis bestätigt die Untersuchung von Amitay (2000a), die 155 englischsprachige persönliche Homepages analysiert. In ihnen taucht „I“ bereits an siebter und „you“ an 14. Stelle auf; in S1 besitzt „Du“ die Position 106 und in S2 die Position 211 (S3 und *taz*: nicht unter den 300 häufigsten Wörtern). Die Produzenten studentischer Homepages sprechen ihre Leser also sehr viel häufiger mit dem informellen „Du“ an als die Autoren der Dokumente der anderen Stichproben. Neben „ich“ befindet sich mit „Seite“ (Position 41) eines der beiden in Tabelle 8.2 vertretenen Nomen in der Liste der 50 häufigsten Token aus S1, das als Synonym von „Webseite“ verwendet wird. Ausschließlich in S2 (Position 35) und S3 (Position 46) befindet sich das zweite Nomen,

[9] Feldweg et al. (1995) berichten, dass diese Lexeme in einem Korpus von mehr als 430 000 deutschsprachigen Newsartikeln ebenfalls hochfrequent sind.

die Abkürzung „Dr.“. Zu den Produzenten der in S2 enthaltenen Dokumente zählen fast ausschließlich wissenschaftliche Mitarbeiter, Assistenten und Professoren, daher ist ein hochfrequentes Vorkommen von „Dr.“ in dieser Stichprobe nicht überraschend („Prof.“ befindet sich in S2 an Position 53). Die ebenfalls sehr häufigen Vorkommen in S3 sind vermutlich durch Listen der Mitarbeiter von Arbeitsgruppen oder Fachbereichen oder durch Kommentare zu Lehrveranstaltungen zu erklären („Prof.“ in S3: Position 73).

In der Liste der 50 häufigsten Wörter der *tageszeitung* sind einige konjugierte Verben enthalten („hat“, „war“), die in S1–S3 vollständig fehlen. Da Zeitungen vornehmlich über vergangene Ereignisse berichten, verwundert die Präsenz dieser Wörter nicht. Die Abwesenheit in den ersten drei Stichproben belegt die Befunde von de Saint-Georges (1998), Bayerl (2002) und Bittner (2003), die für die von ihnen untersuchten privaten Homepages angeben, dass sie vornehmlich im Präsens verfasst werden (vgl. Abschnitt 4.6.3); zusätzlich gilt dieser Umstand jedoch auch für die in S3 enthaltenen Dokumente, die *keine* persönlichen Homepages darstellen. Amitay (2000a, S. 7) kommt nach einem Vergleich der Frequenzlisten ihrer Stichproben und dem *British National Corpus* (BNC) zu einem ähnlichen Schluss und vermutet, dass die Präferenz für das Präsens möglicherweise auf den Umstand zurückzuführen ist, dass Produzenten die von ihnen betreuten Dokumente immer dann aktualisieren, wenn die Inhalte als veraltet angesehen werden, so dass die Leser zur Rezeptionszeit annehmen können, dass es sich um aktuelle Fakten handelt.

8.5.3 Trigrammfrequenzen in Hyperlinkanzeigern

Eine Frequenzanalyse der in Hyperlinkanzeigern verwendeten Trigramme – Sequenzen von drei Token – liefert Hinweise darauf, wie Autoren von HTML-Dokumenten den Inhalt des verknüpften Dokuments signalisieren und welche sprachlichen Navigationshilfen sie dem Benutzer geben. Amitay (2000a) untersucht diesen Aspekt ebenfalls, sie bezieht sich jedoch nicht auf Tri-, sondern auf Bigramme. In ihren Korpora findet sie unter den 20 häufigsten in Linkanzeigern enthaltenen Bigrammen unter anderem „home page“, „return to“, „back to“, „more info“, „click here“ und „to the“ (vgl. auch Amitay, 1998).

Tabelle 8.3 zeigt die 40 häufigsten Trigramme aus S1–S3. Diese enthalten ebenfalls räumliche Deiktika wie z. B. „Zurück zum Anfang“, „Zurück zur Homepage“ und „Zurück nach oben“ (vgl. de Saint-Georges, 1998, sowie Abschnitt 4.6.3). Die insgesamt 31 unterschiedlichen Vorkommen von „zurück“ in den 120 Trigrammen sind besonders auffällig. Von einer sehr abstrakten Perspektive aus betrachtet besitzt ein HTML-Dokument entweder den Charakter einer Übersichtsseite (also einer Liste von Hyperlinks) oder den eines Dokuments, das vornehmlich Textinhalt umfasst (Kleinberg, 1998). Die Rückwärtsnavigation von einer Inhaltsseite zur Übersicht, von der aus der Rezipient initial zur Inhaltsseite gelangt ist, wird metaphorisch durch das Zurückspringen oder Zurückgehen ausgedrückt. In Bezug auf die Vorkommen der anadeiktischen Trigramme, die „zurück“ enthalten, werden die Übersichtsseiten mit einem verhältnismäßig homogenen Vokabular wie z. B. „Leitseite“, „Homepage“, „Anfang“, „Startseite“, „Übersicht“, „Seitenanfang“, „Inhaltsverzeichnis“ und „Hauptseite“ bezeichnet. Die meisten anderen Trigramme stammen aus katadeiktischen Hyperlinkanzeigern, die also von einer Übersichtsseite zum Inhalt führen wie z. B. „Traditionelle Chinesi-

S1	S2	S3
zurück zur KaWo-Leitseite (758) vorherigen Block laden (754) nächsten Block laden (751) HTML-Dateien selbst erstellen (475) Anfang der Seite (194) **Zurück** zur Homepage (192) **zurück** zum Anfang (163) zum Anfang der (150) Traditionelle Chinesische Medizin (136) **Zurück** zur Startseite (135) für Traditionelle Chinesische (132) Arbeitskreis für Traditionelle (127) **zurück** zur Übersicht (125) **Zurück** zur Übersicht (125) der Frankfurter Rundschau (124) Homepage der Frankfurter (122) zur ersten Folie (116) **Zurück** zur ersten (115) **Zurück** zu meiner (112) **zurück** zum Seitenanfang (107) **Zurück** nach oben (107) zu meiner Homepage (105) **Zurück** zum Inhaltsverzeichnis (102) AEGEE in Europe (95) AEGEE in Heidelberg (95) **Zurück** zur Protokolliste (93) Einführung in die (88) **Zurück** zum Anfang (87) Word-Datei zum Downloaden (82) Die Welt der (76) Welt der Worte (76) **Zurück** zur Hauptseite (75) Protokoll als Word-Datei (70) eMail an Redaktion (70) als Word-Datei zum (69) Absatztypen und Textgestaltung (64) Big picture index (63) Small picture index (63) der Universität Heidelberg (61) **zurück** zur Homepage (61)	**Zurück** zum Anfang (351) zum Anfang des (249) **Zurück** zur Homepage (246) ein Partizip II (234) Institut für Geographie (224) Mitglied bei page (191) am Institut für (181) Institut für Ernährungswissenschaft (170) Informations- und Dokumentationsstelle (169) der Justus-Liebig-Universität Gießen (166) Dokumentationsstelle am Institut (161) und Dokumentationsstelle am (157) **Zurück** zur Startseite (156) für Ernährungswissenschaft der (152) Anfang des Dokuments (146) Ernährungswissenschaft der Justus-Liebig-Universität (146) Was ist die (138) ist die GfÖ (135) Mitglieder der GfÖ (135) **Zurück** zum Inhaltsverzeichnis (134) **Zurück** nach oben (123) **Zurück** zur Übersicht (121) in hexadezimaler Angabe (120) Partizip II und (115) Konventionelle syntaktische Analyse (113) der Infinitiv von (109) und der Infinitiv (108) Einführung in die (97) Partizip II von (95) das Partizip II (92) Verlag gesund essen (78) **zurück** zur Hauptseite (77) Arbeitssicherheit und Umweltschutz (77) Bereich American Football (77) zum Bereich American (76) und ein Partizip (76) **Zurück** zum Bereich (74) **zurück** zum Anfang (66) **Zurück** zur Hauptseite (64) II und der (63)	**Zurück** zur ersten (662) zur ersten Folie (662) über alle SMT-Bilder (107) Überblick über alle (107) **Zurück** zur Übersicht (94) German News Team (92) Mail Thread Index (90) **Zurück** zum Anfang (85) Einführung in die (83) Anfang des Dokuments (70) zum Anfang des (70) Liste der Leitlinien (68) Erziehungswissenschaft und Psychologie (63) Überblick über SMT-A (48) Fachbereich Erziehungswissenschaft und (47) für komplexe Zahlen (46) to first slide (46) Back to first (46) Klasse für komplexe (46) View graphic version (45) **zurück** zur Übersicht (44) Überblick über SMT (44) **Zurück** zur Startseite (36) HTML-Dateien selbst erstellen (36) Freie Universität Berlin (36) Der rote Faden (35) Garten und Botanisches (33) und Botanisches Museum (33) Index Leitlinien der (32) Botanischer Garten und (32) Kurzinfo mit Signatur (32) Botanisches Museum Berlin-Dahlem (32) **Zurück** zum Inhaltsverzeichnis (31) zur Übersicht über (30) Zum Starten hier (29) Starten hier klicken (29) Technische Universität Chemnitz (28) Übersicht über Kapitel (28) Übertragungstechnik und Bitübertragungsschicht (27) of Industrial Relations (27)

Tabelle 8.3: Die in Hyperlinkanzeigern vorkommenden Trigramme in S1–S3

sche Medizin“, „Absatztypen und Textgestaltung“ und „[Zum] Starten hier klicken“.[10] Lediglich das zuletzt genannte Trigramm enthält „klicken“. Der in Webdesign- und Usability-Ratgebern häufig kritisierte Imperativ „klicken Sie hier“ oder „hier klicken“ wird in den drei Stichproben somit nicht hochfrequent eingesetzt (S1: „Hier klicken“, Position 88 in den Bigramm-Häufigkeiten mit 101 Vorkommen, S2: kein Vorkommen in den Bi- oder Trigrammen, S3: „hier klicken“, Position 101 in den Trigramm-Frequenzen mit 37 Vorkommen).[11]

[10] „Zum Starten hier klicken“ ist die von *Microsoft Powerpoint* verwendete Phrase, um nach HTML konvertierte Foliensätze anzuzeigen; dieser Hyperlinkanzeiger wird in der Übersichtsseite des konvertierten Foliensatzes eingesetzt. 1 706 der in S3 enthaltenen HTML-Dateien (S1: 209, S2: 50) wurden von *Powerpoint* generiert (vgl. Abschnitt A.4.7 sowie Storrer, 2001b, S. 94). Auf diese Anwendung gehen auch die in S3 hochfrequenten Trigramme „Zurück zur ersten“ und „zur ersten Folie“ zurück.

[11] Storrer (2001b) spricht in diesem Zusammenhang vom „metakommunikativen Verfahren“ zur Einbettung von Hyperlinks (vgl. Abschnitt 3.5.5). Ricardo (1998, S. 149) fasst die „descriptive depth of a link“, d. h. die Qualität des Hyperlinkanzeigers, als ein Gütemaß der „authorial maturity“ auf, woraus wiederum geschlossen werden kann, wie gut der Autor das Medium Hypertext durchdrungen hat. In diesem Zusammenhang wäre ein Hyperlinkanzeiger wie „click here“ Ricardo zufolge wohl als „unreif“ zu bezeichnen.

8.6 Merkmale für konzeptionelle Mündlichkeit

Abschnitt 8.3 ist auf Merkmale eingegangen, die von der CMC-Literatur, die sich vornehmlich mit den Kommunikationsformen E-Mail, Usenet und IRC beschäftigt, als Kennzeichen für konzeptionelle Mündlichkeit aufgefasst werden. Diejenigen Vorkommen dieser Merkmale, die sich maschinell erkennen lassen, werden in den folgenden Abschnitten analysiert.

8.6.1 Smileys

Smileys evozieren zweifelsfrei eine Nähe zur konzeptionellen Mündlichkeit und werden in sämtlichen Internet-basierten Kommunikationsformen genutzt (vgl. Schlobinski, 2000a). Zur maschinellen Erkennung wurde eine Smiley-Liste mit etwa 2 100 Einträgen[12] in das Analysesystem integriert. Etwa 160 dieser Smileys wurden nicht berücksichtigt, da sie Zeichen(ketten) darstellen, die nicht eindeutig als Smileys zu identifizieren sind.[13]

Smileys werden in den Stichproben mit unterschiedlichen Häufigkeiten eingesetzt (vgl. Tabelle 8.4). In den studentischen Homepages können 1 353 Smileys (Maximum: 61) in 806 Dokumenten (3,2%) belegt werden. Die Homepages der Universitätsmitarbeiter enthalten 298 Smileys (Maximum: 16) in 178 Dokumenten (1,2%) und die tiefen Dokumente 93 Smileys (Maximum: 6) in 58 Dokumenten (0,6%). Der im Vergleich zu S2 und S3 häufige Einsatz von Smileys in S1 unterstützt die bei der Betrachtung der Wortfrequenzen aufgestellte These, dass die Dokumente dieser Stichprobe einen eher informellen, konzeptionell mündlichen Stil aufweisen. Die Dokumente, in denen sehr viele Smileys verwendet werden, sind z. B. Smiley-Listen, ein Gästebuch (vgl. Abschnitt 4.6.8), ein Reisebericht (alle S1), eine Aufstellung Computer-bezogener Anekdoten, die persönliche Homepage eines wissenschaftlichen Mitarbeiters, ein Reiseführer (alle S2), die HTML-Version der Diplomarbeit einer Psychologin (zum Thema MUD-Nutzung) und ein HTML-Archiv einer Mailing-Liste (alle S3).[14] Einerseits enthalten Dokumente, die auf den privaten Homepages von Studierenden angeboten werden (z. B. der Reisebericht oder die humoristische Darstellung der eigenen Computersucht), sehr viele Smileys. Sie werden in längeren Texten verwendet, um den saloppen und nicht zu ernst gemeinten Stil zu unterstreichen. Andererseits sind Smileys darüber hinaus in Dokumenten enthalten, die von Universitätsangehörigen (meist wissenschaftlichen Mitarbeitern) angeboten werden. Wenn es dabei um einen Themenbereich wie die Internet-Kommunikation geht (z. B. Arbeiten zur MUD-Kommunikation), tauchen Smileys zwangsläufig in Beiträgen oder Zitaten der MUD- oder IRC-Benutzer auf, jedoch enthält eine in S3 enthaltene Projektevaluation vier Smileys, die vom Projektleiter selbst eingefügt worden sind. Das Dokument besteht aus einer Liste von Anmerkungen, die aus Evaluationsbögen zu einem E-Mail-Projekt im Bereich „Deutsch als Fremdsprache“ stammen. Der Leiter hat die Anmerkungen der Studierenden kommentiert und beantwortet den informellen Stil ihrer Fragen ebenso informell, was durch den Einsatz von Smileys noch verstärkt wird.

[12] Die Liste ist erhältlich unter `http://www.astro.umd.edu/~marshall/`.

[13] Beispielsweise `0`, das in dieser Liste als der aus der Science Fiction-Filmreihe *Star Wars* bekannte Todesstern geführt wird und auch eher dem Bereich der (sehr minimalistischen) ASCII-Art zugehörig ist (vgl. Haase et al., 1997, S. 78). Im *World Wide Web* wird ASCII-Art nur sehr selten als Stilmittel eingesetzt, wie z. B. in der Überschrift „_oO(Ein kleines Werk)Oo._“ (`http://www.tu-chemnitz.de/~kirst/`).

[14] Die Smiley-Listen erklären die Vorkommen der in Tabelle 8.4 enthaltenen eher unüblichen Smileys.

S1	;-) (412) :-) (345) ;) (85) :) (79) :-)) (54) :-((37) ;-)) (33) ;)) (21) :)) (17) :o) (13) :((13) ;o) (11) :-))) (9) =:-) (8) :))) (7) -:-) (7) :-(((6) =) (5) (c: (5) ;:-) (5) =:-((5) :-D (4) |o| (4) ;-((4) :o((4) 8-) (4) (-: (3) [* (3) (: (3) :-> (3) ;((3) :- (3) :-o (3) :-X (3) :-)))) (3) ;-))) (3) :-P (3) {(:-) (3) :D (2) ;-} (2) (|) (2) :'-((2) C= (2) 8-| (2) :-@ (2) [*] (2) :=) (2) :-() (2) :-{) (2) :-{} (2) :-] (2) :-[(2) ;)))))) (2) :-| (2) >:-> (2) >;-> (2) [:-) (2) :-Q (2) (@ (2) {:-) (2) =0) (2) (:-) (2) :-, (2) :-/ (2) --. (1) :*) (1) :,((1) :-# (1) :-& (1) :-0 (1) :-7 (1) :-C (1) :-o (1) :0) (1) :^D (1) :~) (1) :-x (1) >:) (1) o-) (1) :'-) (1) :-p (1) X-((1) [#] (1) "< (1) [:] (1) '! (1) :* (1) :-e (1) :-)8 (1) :-)= (1) :-)~ (1) :< (1) :> (1) ^-^ (1) :-~) (1) :-|| (1) c[] (1) ;))) (1) {*} (1) |-) (1) |-| (1) :[(1) :-} (1) :)))) (1) <3 (1) :| (1) *<:-) (1) <:-I (1) :} (1) :-W (1) >:-((1) <|-) (1) (^-^) (1) #-) (1) %-) (1) %-6 (1) :-)-8 (1) %-| (1) :-)... (1) :-{)} (1) B:-) (1) %-} (1) 8:-) (1) '-) (1))8-) (1) *:o) (1) +:-) (1) (:I (1) :-)))))) (1) +-:-) (1) ;-)))))) (1) ,-) (1)
S2	:-) (111) ;-) (63) :) (25) :-((17) ;) (16) :-)) (11) ;-)) (7) :((5) :o) (5) ;-))) (4) -((3) ._) (2) :-o (2) :// (2) :-x (2) :<= (2) :-| (2) <:>== (2) >-^);> (2) *<:-) (2) %*@:-((2) ;-D (1) d:-) (1) (-: (1) ;))) (1) :] (1) :)))) (1) 8*) (1) ;o) (1) ;)))) (1) H-) (1) (: (1)
S3	:-) (37) ;-) (26) :) (7) :-{ (4) ;) (4) :-((3) :-)) (2) :Q (1) %-) (1) ;-} (1) :^) (1) :o) (1) =0) (1) H:) (1) :-(((1) :((1) :-)))))))) (1)

Tabelle 8.4: In S1–S3 enthaltene Smileys

8.6.2 Buchstabeniterationen

Tabelle 8.5 stellt die – oftmals an die Comic-Sprache angelehnten (vgl. Sassen, 2000, und Schmidt, 2000, S. 123) – Vorkommen von Iterationen in den Stichproben dar.[15] Dabei handelt es sich um Reduplikationen einzelner Buchstaben eines Wortes (z. B. „soooo“ und „Tschüßiiiii“) zum Zwecke der Emulierung von Prosodie, insbesondere der Sprechgeschwindigkeit (Haase et al., 1997, S. 67 f.). Wie bei den Smileys enthält S1 mit 834 Iterationen in 435 Dokumenten (1,7%) die mit Abstand meisten Vorkommen. S2 enthält 84 Iterationen in 55 Dokumenten (0,4%) und S3 lediglich 20 Vorkommen in 15 Dokumenten (0,2%).

Die Dokumente, in denen Iterationen verwendet werden, umfassen eine Sammlung von Plattenkritiken, ein Dokument mit Texten und Gitarrenakkorden der Gruppe „Tocotronic“, die ins Deutsche übersetzten Geschichten des „Bastard Operator from Hell“, ein Chat-Protokoll, „Die einzig offiziöse Spezifikation für BTML (und ehemals JTML)“[16] (alle S1), die bereits erwähnte Liste von Anekdoten („Dümmste anzunehmende User – Die Sammlung“), eine Webseite mit Bildern des Labors eines Mitarbeiters aus dem Bereich der angewandten Physik, ein Dokument, das lediglich ein Foto einer Bar sowie die Satzteile[17] „Die Rehlein beten zur Nacht,“ (oberhalb des Bildes) und „hab acht!“ (unterhalb) zeigt (alle S2), eine Informationsseite eines Arbeitskreises der Fachschaft Jura an der Universität Heidelberg, ein zehn Einträge umfassendes Gästebuch und eine Studiengangsbroschüre[18] (alle S3). Wie Smileys werden auch Iterationen vornehmlich in den privaten Homepages von Studieren-

[15] Aus Platzgründen werden die in S1 lediglich einmal verwendeten Iterationen nicht in Tabelle 8.5 aufgeführt.

[16] BTML und JTML stehen für „Bizarre“ bzw. „Joke Talk Markup Language“. Es handelt sich um Listen XML- bzw. SGML-ähnlicher Tags, die erstmals im Usenet – in den Gruppen `de.talk.bizarre` und `de.talk.jokes` – verwendet wurden. Einige Elemente werden sogar mit einer XML-Syntax definiert, wie z. B. `<beschwichtigung>`, das die Attribute `ups`, `wieKonnteIchNur` und `immerPassiertMirDas` besitzen darf.

[17] Das zur Analyse verwendete *Perl*-Skript erkannte in dem Dokument die beiden Iterationen „Aaaaaaaaaabstand“ sowie „Aaaaaaaaaaaaaaaaaaaaaaaaaaaaabstand“, die – in der gleichen Farbe wie der Hintergrund formatiert – als Blindtext eingesetzt wurden, um die beiden Satzteile links- bzw. rechtsbündig anzuordnen.

[18] Es handelt sich um ein von *Microsoft Word* nach HTML konvertiertes Dokument von etwa 60 Papierseiten Länge für Anfänger des Studiengangs Bauingineurwesen an der TU Dresden mit insgesamt vier Iterationen.

S1	sooo (34) Hmmm (17) soooo (11) hmmm (11) aaaaa (11) Gäääähn (10) gaaaanz (9) gaaanz (9) Mmmm (9) Ahhh (8) Naaah (8) gaaaaanz (6) hmmmm (6) aaaaaa (6) grrr (6) seeeehr (6) Ficköööön (5) naaaaaaaaaaaaaaaaaaaaaaaaaaa (5) öööööööööööööööööö (5) pssst (5) Sooo (5) Aahhh (4) Äääätsch (4) Pffft (4) Auuuu (4) Tadudeldidudeldidudeldiduuuuu (4) BOOOM (4) schööön (4) cooool (4) blahaaaa (4) Hmmmm (4) Pfüüüüüp (4) Alteeer (4) viiiel (4) Waaaas (3) Seeesel (3) Uuuups (3) soooooo (3) SOOO (3) Ahhhh (3) öööööööööööööööööööööööööööööööööööö (3) Prfffz (3) Jaaaaa (3) Ohhh (3) neee (3) BWLIII (3) Ooooh (3) Pssst (3) uuuu (2) Gaaanz (2) Ooohhh (2) Äquatorrr (2) noooch (2) Gääähn (2) hmmmmmmmm (2) Halllihallllllllllllllllooooooooooooooo (2) grrrrr (2) ahhhhh (2) gaaaaaaanz (2) ohhh (2) Sorrrry (2) AjrRawkFGwwAAAAACgkQQTauo (2) räääh (2) brrrrrilliant (2) aaaaaaaa (2) ähhh (2) WAAAAHNSSINNS (2) Aaahhh (2) Wuuduuu (2) URZsssssssthhhh (2) Ahhhahhhahhh (2) Määäh-Synthesizer (2) hiiiiiieer (2) HMMM (2) Ahhohhhh (2) SCCCCCCCCCCCCCHMAAAAAAAAAATTZZ (2) grrrrreat (2) groooooßen (2) lieeebe (2) seeehr (2) aaaa (2) Pieeep (2) tatatataaa (2) sooooo (2) Snnnniggers (2) Jaaa (2) Tüttüüüt (2) Fuuuunnnnnnn (2) coool (2) rrrrrrrrroor (2) Mmmmh (2) herrrliche (2) Pffffft (2) roorooroorrorruuuuum (2) JUBELLLL (2) aaaaaaaaaaa (2) GAGNEEEE (2) vieeel (2) aaaaff (2) aaaaaaaaaa (2) FZZZZT (2) AAAARRGGHH (2) Naaa (2) nööö (2) GAAANZ (2) Aaaah (2) Chhrrrr (2) Mööööööööööp (2) muuuussss (2) youuuuuuuuu (2) wunderschööön (2) Hmmmmm (2) psssst (2) firrrst-a-place (2) herrrlich (2) vieeeel (2) Mööööönsch (2) Ähhh (2) Allsssooo (2) meeeeeeeeeeehr (2)
S2	sooo (4) gaaanz (4) gaaaaaaanz (4) Mussssssss (3) Mussssssss (3) Klaaasse (2) daaaaa (2) Hoschiiiii (2) Jooooooooo (2) Suuuper (2) uuuuh (2) blubbblubb (2) Hardmannnnshennnnnnnnnnnnnn (2) gaaaaanz (2) Aaaaaaaaaabstand (2) Aaaaaaaaaaaaaaaaaaaaaaaaaabstand (2) AHV-xxxx (2) tiiieerisch (2) AAAaaaarrrghgghghgh (1) uuuuuuhhhhhhhhhooaaahhhhh (1) Aaaaaaaaahhhhhhhhhhhhh (1) Ahhhh (1) brennnt (1) Haaa (1) schaaaaa (1) uuuuuiiiiIIIIIII (1) AAAAT (1) natüüürlich (1) sooooo (1) gaaasanz (1) Tschüßiiiiii (1) Jaaa (1) Hoschiiiiiiii (1) Hoschiiiiii (1) gaaaaannz (1) süüüüüß (1) SOORRRYYY (1) seeeeeehr (1) hmmm (1) hiiiiiiilfe (1) wiesooooo (1) gaaaanz (1) vistaaa (1) loginnnnnnnnnn (1) WAAAS (1) Soooo (1) Uffff (1) Gaaaaz (1) Tierraaa (1) AIII (1) Hoschiiiiii (1) Reeeeegen (1) Sooooo (1) Duuuh (1) Hmmmm (1) AAAAEH (1) Brrr (1) Giiiiiiiiiiiiießen (1)
S3	Hmmm (4) RRRrrrr (2) huuuderts (1) viiieele (1) rassschen (1) GROOOOOSS (1) ZUUU (1) Jaaaaaaaaaaaa (1) tiriliii (1) AAAaaarrrrggghhhhhhhh (1) AAAAAAAAAAAATTTTTTTTTTT (1) mmmmmh (1) yäääääääääh (1) BRRR (1) TSCHÜÜÜÜÜÜÜÜSS (1) Duuuuuuu (1)

Tabelle 8.5: In S1–S3 enthaltene Iterationen

den verwendet. Viele der in S1 enthaltenen Dokumente, die mehrere Iterationen beinhalten, wurden nicht primär für das *World Wide Web* angefertigt: Die Sammlung von Plattenkritiken basiert auf Fan-Magazinen; Dateien mit Texten und Gitarrenakkorden werden schon seit den achtziger Jahren im Usenet getauscht. Aus dieser Zeit stammen auch die Geschichten des „Bastard Operator from Hell“. Durch derartige zusätzliche Veröffentlichungen werden demnach unweigerlich sprachliche Phänomene in das WWW transportiert. Dieser Prozess tritt bei der „Spezifikation für BTML“ besonders deutlich hervor: Ohne den Erfolg des WWW wäre XML nicht entwickelt worden, so dass ein humoristischer Einsatz von XML-Elementen, denen ad hoc Namen zugewiesen werden, die den Kontext oder die Aussage einer Nachricht unterstreichen, undenkbar gewesen wäre. Die technischen Spezifika des Mediums sind in der Lage, neue und kreative sprachliche Ausdrucksformen zu provozieren.

8.6.3 Emphasen

Unter anderem Verbstämme und Iterationen werden oftmals in einer typografisch markierten Form verwendet, die als Emphase bezeichnet wird (Haase et al., 1997, S. 67 ff.). Die Markierung erfolgt durch die Einschließung eines Wortes oder einer Wortsequenz in Sonderzeichen. Vornehmlich werden der Asterisk (*), der Unterstrich (_) und der Slash (/) benutzt.

Tabelle 8.6 zeigt die Vorkommen von Emphasen in den Stichproben. In S1 wurden 182 Vorkommen in 92 HTML-Dokumenten (0,4%) ermittelt, wohingegen S2 mit 38 Empha-

S1	⋆FERNSEHEN⋆ (14) ⋆grins⋆ (10) ⋆hihi⋆ (8) ⋆HÖRFUNK⋆ (8) ⋆verbeug⋆ (6) ⋆die⋆(3) ⋆gg⋆ (3) ⋆seufz⋆ (3) _nicht_ (3) ⋆PLONK⋆ (3) ⋆lol⋆ (3) /Rechtsberatung/ (3) ⋆Der⋆ (2) ⋆Patsch⋆ (2) /SGML/ (2) ⋆smile⋆ (2) ⋆frechgrins⋆ (2) ⋆bg⋆ (2) ⋆fg⋆ (2) _alle_ (2) ⋆so⋆ (2) ⋆SABBER⋆ (2) ⋆DER⋆(2) ⋆sehr⋆ (2) ⋆PATSCH⋆ (2) /10/ (2) ⋆ggg⋆ (2) ⋆knuddel⋆ (2) _sprache_ (2) ⋆wink⋆ (2) ⋆stutz⋆ (2) ⋆drei⋆ (1) ⋆FUSSBALL⋆ (1) /talktalktalk/ (1) ⋆Symmetrieachse⋆ (1) ⋆liebe⋆ (1) _snabel_ (1) ⋆dickesLob⋆ (1) ⋆etwas⋆ (1) ⋆schmaatz⋆ (1) /reise/ (1) ⋆fühlemichgebauchpinselt⋆ (1) ⋆gggg⋆ (1) ⋆Gähn⋆(1) ⋆verkaufen⋆ (1) ⋆aufgeregt⋆ (1) /laxt/ (1) /issues/ (1) /eng/ (1) ⋆bell⋆(1) ⋆heul⋆ (1) ⋆schmunzel⋆ (1) ⋆06⋆ (1) ⋆are⋆ (1) ⋆räckel⋆ (1) ⋆etwa⋆ (1) ⋆kein⋆ (1) ⋆bis⋆ (1) ⋆GGG⋆ (1) ⋆schleck⋆ (1) ⋆schluck⋆ (1) ⋆trommel⋆ (1) ⋆alle⋆ (1) ⋆grübel⋆ (1) ⋆inSicherheitfuehl⋆ (1) ⋆KEUCH⋆ (1) /Kleinhirnblutungen/ (1) ⋆freu⋆ (1) _underlined_ (1) ⋆grinst⋆ (1) ⋆ätsch⋆(1) /selfhtml/ (1) ⋆giggel⋆ (1) ⋆BLINDMAIL⋆ (1) ⋆tnx⋆ (1) ⋆hexhex⋆ (1) ⋆Lecker⋆ (1) ⋆YES⋆ (1) /11/ (1) ⋆springinLuft⋆ (1) ⋆ruhe⋆ (1) ⋆ubs⋆ (1) ⋆flame⋆ (1) ⋆winkewinke⋆ (1) /ei/ (1) ⋆SUPERSKIPAUSCHALE⋆ (1) ⋆nicht⋆ (1) ⋆freudig⋆ (1) /wg/ (1) ⋆keinen⋆ (1) ⋆PRIVAT⋆ (1) ⋆hoff⋆ (1) /ops/ (1) /winamp/ (1) ⋆GreatIsTheLord⋆ (1) /flyer/ (1) _NICHT_ (1) /darkzone/ (1) ⋆übs⋆ (1) ⋆lächergaaaaaaanzbreit⋆ (1) _null_ (1) ⋆neu⋆ (1) /ski/ (1)
S2	⋆Andy⋆ (5) ⋆andy⋆ (3) ⋆grins⋆ (2) _nicht_ (2) _eingeschränkt_ (2) /Kurs/ (2) ⋆such⋆ (2) ⋆gaaaaannz⋆ (1) ⋆gacker⋆ (1) ⋆difference⋆ (1) ⋆besten⋆ (1) ⋆nicht⋆ (1) _einige_ (1) ⋆Beraterin⋆ (1) ⋆grummel⋆ (1) /vor/ (1) /neues/ (1) ⋆Uffff⋆ (1) /Inhalt/ (1) ⋆funktionierende⋆ (1) ⋆aufgeregt⋆ (1)
S3	⋆kein⋆ (1) ⋆lach⋆ (1) ⋆lall⋆ (1) _impulsiv_ (1) ⋆kostenlos⋆ (1) ⋆zu⋆ (1) ⋆sollte⋆ (1) _japanisches_ (1) _first_ (1) _ohne_ (1) _ganze_ (1) ⋆erforderliche⋆ (1) /agrep/ (1) /95/ (1) /aU/ (1) ⋆bindend⋆ (1) ⋆ohne⋆ (1) ⋆jammer⋆ (1) ⋆vollständig⋆ (1) _neue_ (1) ⋆lol⋆ (1) /dev/ (1) ⋆bitte⋆ (1) ⋆not⋆ (1)

Tabelle 8.6: In S1–S3 enthaltene Emphasen

sen in 16 Dokumenten (0,1%) und S3 mit 24 Vorkommen in 14 Dokumenten (0,1%) jeweils weniger als ein Drittel enthalten. Die Webseiten in denen Emphasen benutzt werden, umfassen einen Hypertext, der aus einer großen Anzahl von Videotextseiten[19] besteht, die mit Hilfe einer TV-Karte und entsprechender Software nach HTML konvertiert wurden, die bereits erwähnte „offiziöse Spezifikation für BTML“, eine tagebuchartige Homepage („⋆schleck⋆“), ein Gästebuch („⋆grins⋆“; alle S1), die bereits angesprochene Anekdotensammlung, eine Anleitung zur Herstellung einer Internet-Verbindung (beide S2), eine HTML-Version der FAQ-Datei der Newsgruppe `bln.markt`, ein Archiv einer Mailing-Liste sowie ein Gästebuch. Emphasen werden zwar nicht annähernd so häufig benutzt wie Iterationen, doch ist die Bandbreite dieses Stilmittels sehr umfassend: Die Vorkommen in S1 beinhalten isolierte Verbstämme („⋆grins⋆“, „⋆hoff⋆“, „⋆flame⋆“), spezielle Abkürzungen (vgl. Ziegler, 2001) und Slangausdrücke („⋆PLONK⋆“, „⋆bg⋆“, „⋆gg⋆“, „⋆tnx⋆“), eine dialektale Variation („_snabel_“), iterierte Wortsequenzen („⋆lächergaaaaaaanzbreit⋆“, „⋆fühlemichgebauchpinselt⋆“) und vollständige Großschreibung, die häufig als emuliertes Schreien interpretiert wird („⋆KEUCH⋆“). Darüber hinaus resultieren einige der gefundenen Iterationen aus einer nicht vollständigen Übertragung des jeweiligen Dokuments nach HTML. Die FAQ-Liste der Gruppe `bln.markt` macht zwar durchaus Gebrauch von der technischen Möglichkeit, Überschriften als solche auszuzeichnen und Fettdruck zu benutzen, doch eine Emphase wie in „… sondern den Artikel erneut ⋆vollständig⋆ (mit Korrektur) posten.“ bleibt erhalten, obwohl ein kursiver Schriftschnitt aus typografischer Sicht angebrachter wäre.

[19] Da die Möglichkeiten der typografischen Auszeichnung im Teletext sehr begrenzt sind (es existiert kein Fettdruck und es steht lediglich eine geringe Anzahl Farben zur Verfügung), werden auch in diesem Medium Asteriske zur Hervorhebung wichtiger Wörter verwendet, wie z. B. ⋆FERNSEHEN⋆.

8.6.4 Isolierte Verbstämme

In Abschnitt 8.6.3 wurden isoliert verwendete Verbstämme angesprochen, die vor allem in der Chat-Kommunikation eingesetzt werden, um Zustände oder Gefühlsregungen des Produzenten auszudrücken (vgl. Haase et al., 1997, S. 65). Speziell markierte Zeichenketten wie z. B. Emphasen können automatisch mit Hilfe regulärer Ausdrücke erkannt werden. Um auch die nicht markierten Vorkommen isolierter Verbstämme zu identifizieren, wurde eine Liste von 5 438 Stammformen in den Erkenner integriert und die jeweiligen Treffer manuell verifiziert. Die wenigen Vorkommen isolierter Verbstämme befinden sich in S1 und können fast ausnahmslos dem Bereich der Comic-Sprache zugeordnet werden (vgl. Tabelle 8.7).

S1	grins (5) grab (2) laber (2) quietsch (2) würg (2) stöhn (2) schnief (2) seufz (2) sniff (1) röchel (1) nörgel (1) hust (1) schmatz (1) knuddel (1)

Tabelle 8.7: In S1 enthaltene isolierte Verbstämme

Zwei der 25 Vorkommen stammen aus einem HTML-Dokument, das eine Grußbotschaft an die Schwester des Autors enthält, zwei weitere Vorkommen befinden sich in einem fiktiven Dialog, der die Schwächen automatischer Dialogsysteme verdeutlichen soll. Nach einigen verzweifelten Versuchen endet der nur schleppend stattfindende Dialog mit „Anrufer: .... stöhn ... röchel ....!“. In der Bildergalerie eines Strandurlaubs wird ein Foto, das den Bau einer Sandburg zeigt, unterschrieben mit „grab grab ...“. Nur in etwa der Hälfte der Dokumente, die isolierte Verbstämme enthalten, sind auch Hinweise darauf zu finden, dass die jeweiligen Autoren häufiger Chat-Systeme (IRC oder Web-Chats, zu letzteren vgl. Storrer, 2001a) benutzen. Dies könnte einerseits die Vermutung nahelegen, dass sich isolierte Verbstämme derzeit – auch unter den Nicht-Chattern – verbreiten, oder aber, dass es sich lediglich um ein sehr seltenes sprachliches Phänomen handelt, das auch unbewusst bei der oftmals sehr raschen Produktion von HTML-Dokumenten eingesetzt wird.

8.6.5 Slangausdrücke

Viele Internet-Benutzer, Studierende der Informatik, *Linux*-Anwender und Benutzer von Chat-Systemen setzen ein umfangreiches Inventar von Abkürzungen, Akronymen und speziellen Slangausdrücken ein. Zur Ermittlung der Vorkommen wurde eine etwa 1 000 Einträge umfassende Liste von Slangausdrücken[20] in das Analyse-Skript integriert. Da viele dieser Begriffe auch mit anderen Lesarten verwendet werden, wurden die Treffer manuell verifiziert, wobei sich etwa 120 Vorkommen als fehlerhaft herausgestellt haben.[21] Tabelle 8.8 zeigt die Vorkommen in den Stichproben. S1 enthält 90, S2 75 und S3 94 Slangausdrücke.[22] Nahezu alle betroffenen HTML-Dokumente wurden bereits angesprochen. Hierzu gehören z. B. die Anekdotensammlung und die „Bastard Operator from Hell“-Geschichten (S1), wissenschaftliche Arbeiten zur MUD-Kommunikation (S2) und eine E-Mail aus dem Archiv

[20] Diese Liste ist erhältlich unter `http://www.astro.umd.edu/~marshall/`.

[21] Die Abkürzung ASAP kann z. B. neben „as soon as possible“ auch „Application Service Access Point“ bedeuten.

[22] Einige der ermittelten Vorkommen, etwa „*gg*“ oder „*bg*“, werden in Abschnitt 8.6.3 diskutiert und somit an dieser Stelle nicht gesondert berücksichtigt (vgl. Tabelle 8.6).

S1	linx (21) dau (20) imho (7) rpg (6) motd (6) bofh (5) rl (5) btw (3) lol (3) snafu (2) ppl (2) cfv (2) rtfm (2) rotfl (2) b4n (1) cu2 (1) irl (1) jff (1)
S2	dau (51) rl (9) linx (5) afk (4) rpg (2) rtfm (2) imho (2)
S3	rl (71) linx (20) imho (2) rtfm (1)

Tabelle 8.8: In S1–S3 enthaltene Slangausdrücke

einer Mailing-Liste (S3). Im Vergleich zu S2 und S3 liegt zwar in S1 eine breite Streuung von Belegen vor, jedoch erreichen die ermittelten Vorkommen weder die Bandbreite noch die Quantität, die in Analysen von Chat-Systemen berichtet werden (vgl. Abschnitt 8.3).

8.7 Weitere Charakteristika

Verschiedene weitere Charakteristika liefern Hinweise auf den Grad der konzeptionellen Mündlichkeit der drei Stichproben (vgl. Tabelle 8.9).[23] Hierzu gehören die Vorkommen von Bigraphen, Auslassungspunkten, iterierten Interpunktionszeichen und Assimilationen.

Runkehl et al. (1998, S. 36 f.) stellen fest, dass die Verwendung von Bigraphen „sehr stark mit der E-Mail“ korreliert: Es tauchen deutlich mehr dieser „Übertragungsfehler“ (ebd.) in E-Mails als in mit der Textverarbeitung oder der Schreibmaschine geschriebenen Briefen auf. Die Ermittlung der Vorkommen in S1–S3 wurde mittels einer Suche nach Token, die Bigraphen enthalten, sowie einer sukzessive erstellten Negativliste von 12 715 Wörtern durchgeführt, die korrekte Sequenzen von „ae“, „ue“ und „oe“ enthalten, (z. B. „aktuell“ und „Abenteuer“). Die meisten Bigraphen sind in S3 enthalten (0,36% aller Token; S1: 0,12%, S2: 0,09%). Diese Vorkommen sind durch den Umstand zu erklären, dass diese Stichprobe etwa 180 E-Mails enthält, die aus HTML-Archiven von Mailing-Listen stammen.[24] Etwa bis zur Mitte der neunziger Jahre – zu dieser Zeit etablierte sich der Standard MIME (*Multipurpose Internet Mail Extensions*, RFC 2045, RFC 2046, RFC 2047, RFC 2048) – war es nicht problemlos möglich, Umlaute in E-Mails zu benutzen, da deren Übertragung an den Empfänger nicht sichergestellt werden konnte. Aus diesem Grund wurden Zeichen, die Diakritika enthalten, mittels einer Umschrift kodiert und diese Umschrift wird von vielen Benutzern auch heute noch aus Gewohnheitsgründen eingesetzt.[25] Ein weiterer Ursprung könnten Dokumente wie die in Abschnitt 8.6.3 betrachtete FAQ-Liste sein, die ursprünglich für das textbasierte Usenet verfasst wurde und aufgrund der ähnlichen Produktionsbedingungen weitere Bigraphen enthält. Das prozentual etwas häufigere Auftreten von Bigraphen in S1 als in S2

[23] Tabelle 8.9 stellt die jeweils zehn häufigsten Bigraphen, das häufigste Vorkommen von Auslassungspunkten, die 15 häufigsten Iterationen von Interpunktionszeichen und die zehn häufigsten Vorkommen von Assimilationen in S1, S2 und S3 dar. Die nach den Vorkommen angegebenen Prozentzahlen sind deren prozentualer Anteil im Verhältnis zu den in der jeweiligen Stichprobe enthaltenen Token.

[24] Die HTML-Dokumente, in denen sich eingebettete E-Mails befinden, machen 1,4% aller Token in S3 aus.

[25] Eine alternative und noch heute häufig verwendete Umschrift basiert auf dem Textsatzsystem TEX/LATEX, in dem Umlaute durch dem Vokal vorangestellte Anführungszeichen realisiert werden (z. B. `Beh"orde` oder `"Ubung`). Die Ligatur ›ß‹ kann ebenfalls auf diese Weise (`gro"s`) oder mittels `\3` erzeugt werden (`gro\3`).

Merkmal	S1	S2	S3
Bigraphen	10 190 Vorkommen, 0,12% fuer (1017) koennen (172) zurueck (151) Universitaet (143) Jahrgaenge (96) natuerlich (93) wuerde (81) Fuer (76) waere (68) Zurueck (65)	4 413 Vorkommen, 0,09% fuer (402) Muenchen (217) Universitaet (125) Buecher (60) Natuerliche (57) koennen (48) Joerg (38) Juergen (36) Jaeger (36) Einfuehrung (34)	11 305 Vorkommen, 0,36% fuer (1250) koennen (119) Boerse (96) Gruenen (90) Muenchen (90) Fuer (85) Universitaet (82) zurueck (74) erklaerte (74) muesse (73)
Punkte	18 287 Vorkommen, 0,21% ... (6073)	2 586 Vorkommen, 0,05% ... (1775)	679 Vorkommen, 0,02% ... (633)
Interpunktion	5 031 Vorkommen, 0,06% !!! (1908) ‼ (1227) ??? (451) ?! (341) ?? (322) !? (196) !!!! (188) !!!!! (86) ???? (42) !!!!!! (35) ?!? (32) !!!!!!! (26) ????? (25) !?! (17) ?!! (15)	1 270 Vorkommen, 0,03% !!! (446) ‼ (360) ?? (96) ??? (89) ?! (86) !!!! (46) !!!!! (23) !?! (21) ?!? (21) !? (16) ???? (15) !!? (6) ??????? (5) !!??? (4) !!!!!!! (4)	437 Vorkommen, 0,01% !!! (138) ??? (83) ‼ (81) ?? (46) ?! (26) !!!!! (18) !!!! (12) !?! (9) ????? (6) ???? (3) !!!!!! (3) !? (2) ?????? (2) !!!!!!!! (1) ?!!! (1)
Assimilationen	13 048 Vorkommen, 0,15% mit Großbuchstaben: 5 326 mit Kleinbuchstaben: 5 297 mit Apostroph: 2 425	4 203 Vorkommen, 0,08% mit Großbuchstaben: 2 285 mit Kleinbuchstaben: 1 277 mit Apostroph: 641	1 662 Vorkommen, 0,05% mit Großbuchstaben: 669 mit Kleinbuchstaben: 418 mit Apostroph: 575
	I'm (106) Don't (90) Murphy's (88) It's (85) Arslandemir's (60) Bernd's (58) Chewy's (52) Can't (46) PC's (40) CD's (38)	Official's (122) Sa'dan (61) Wöll's (59) QB´s (54) B's (47) Attila's (44) Ferber'schen (41) A's (40) ATL´s (38) Drum'n (32)	Women's (23) CD's (16) Beck'sche (16) PC's (13) User's (12) Men's (11) AG's (9) Hill's (9) Bauer's (8) Labor's (8)
	gibt's (703) geht's (341) wird's (318) sieht's (242) gibt´s (193) don't (117) c't (117) war's (100) für's (99) ich's (87)	gibt's (93) geht's (55) c't (38) auf's (27) in's (24) gibt´s (21) d'un (21) don't (20) it's (19) d'imprimerie (16)	sieht's (21) gibt's (18) wird's (17) geht's (16) sprach's (12) don't (10) doesn't (10) c't (9) für's (9) d'enregistrements (8)

Tabelle 8.9: In S1–S3 enthaltene Bigraphen, Assimilationen und Interpunktionszeichen

lässt sich vermutlich ebenfalls durch Produktionsbedingungen erklären, da Mitarbeiter einer Universität eher Wert auf Fehlerfreiheit legen.

Auslassungspunkte („. . .“) werden von Storrer (2001a) als Zeichen für das Innehalten in einem Turn im Rahmen der Chat-Kommunikation aufgefasst. Auslassungspunkte können weitere Funktionen umfassen, z. B. das Signalisieren eines im Fluss befindlichen kognitiven Prozesses. Der Autor verdeutlicht, dass ein Text nicht vollständig geplant, sondern lediglich hastig notiert wurde. Die Vorkommen von Auslassungspunkten in S1 sind mit 18 287 (0,21% aller Token) im Vergleich zu S2 (0,05%) und S3 (0,02%) extrem[26] hoch, was darauf hindeutet, dass viele Homepages von Studierenden ohne genaue Planung des Textes, jedoch mit dem Hintergedanken der direkten Kommunikation mit einer oder mehreren Personen, ähnlich rapide und unreflektiert produziert werden wie die E-Mail an einen Kommilitonen oder die einzelnen Zeilen eines flüchtigen, im Chat stattfindenden Smalltalks.

Im Usenet werden iterierte Interpunktionszeichen häufig zum Anlass genommen, den Autor eines Artikels zu rügen (vgl. Haase et al., 1997, S. 69), in den studentischen Homepages wird dieses Stilmittel zur Betonung einer wichtigen Aussage häufiger als in den anderen Stichproben verwendet (S1: 0,06%; S2: 0,03%; S3: 0,01%). Die Dokumente mit den meisten

[26] Durch Vorkommen von Auslassungspunkten auf verschiedenen nach HTML konvertierte Videotextseiten wird diese Statistik negativ beeinflusst.

Vorkommen sind Gästebücher, als HTML-Dokumente aufbereitete Chat-Protokolle und die bereits angesprochenen Sammlungen konvertierter Videotextseiten.

Runkehl et al. (1998) finden bei der Untersuchung verschiedener Korpora von E-Mails zwischen 2% und 4% Assimilationen (etwa „war's"), die zwar als sprechsprachliche Mittel angesehen, aber nicht rekurrent verwendet werden. Mit Hilfe eines regulären Ausdrucks filtert das *Perl*-Skript Vorkommen von Assimilationen, die ein Hochkomma enthalten und entweder mit einem Groß- oder Kleinbuchstaben oder einem Hochkomma beginnen.[27] Auch hier liegt S1 mit insgesamt 13 048 Vorkommen (0,15% aller Token) wieder vor S2 (0,08%) und S3 (0,05%). Es ist auffällig, dass „gibt's" und „geht's" in allen drei Stichproben zu den zwei (S1, S2) bzw. vier (S3) häufigsten Assimilationen gehören.

8.7.1 Konstante Groß- oder Kleinschreibung

Runkehl et al. (1998, S. 36 f.) geben an, dass in 7% bis 16% der in verschiedenen Korpora enthaltenen E-Mails konsequente Kleinschreibung benutzt wird, die auf die oftmals zeitsparende Produktionsweise zurückzuführen ist (vgl. Günther und Wyss, 1996, Quasthoff, 1997, Pansegrau, 1997). Die maschinelle Erkennung konstanter Groß- oder Kleinschreibung ist technisch einfach zu realisieren, doch wurde die in einem Dokument enthaltene Anzahl von Wörtern nicht beachtet: Eine HTML-Datei, die lediglich ein Foto und darunter den mit einem Hyperlink versehenen Text „zurück" enthält, wird ebenfalls berücksichtigt.

S1 enthält insgesamt sechs Dokumente, die ausschließlich Groß- und 330 Dokumente, die ausschließlich Kleinbuchstaben enthalten. S2 umfasst zwei groß- und 23 kleingeschriebene Dateien, während S3 insgesamt 18 groß- und 53 kleingeschriebene HTML-Dokumente beinhaltet. Einige stichprobenartig betrachtete Webseiten zeigen, dass die meisten dieser Dateien lediglich Bildunterschriften enthalten oder als Platzhalter fungieren. Mehrere Dokumente scheinen Teile von Framesets zu sein, da ihr Zweck nicht auf den ersten Blick deutlich wird. Hierzu gehört z. B. ein in S3 enthaltenes Dokument, das lediglich die Wortfolge „IRAN: WERBUNG, SPONSORING, AUSSCHREIBUNGEN" enthält. Im Hintergrund befindet sich eine schematische Darstellung der Umrisse des Iran, im unteren Bereich existiert eine mit „Zurück zur Länderseite" beschriftete Grafik als Hyperlink. Aller Wahrscheinlichkeit nach umfasst dieses Dokument schlicht noch keinen Inhalt, wurde jedoch bereits vom Autor angelegt, um den eigentlichen Inhalt zu einem späteren Zeitpunkt einfacher einfügen zu können. Eine konstante Kleinschreibung wird auch in einigen HTML-Dokumenten als gestalterisches Mittel eingesetzt; die vollständige Großschreibung wird hierfür seltener verwendet.

8.7.2 Begrüßungen und Verabschiedungen

Nach Ansicht von de Saint-Georges (1998) werden Floskeln wie „Welcome to my very own home-on-the web page" oder „Welcome to my fast-paced life" eingesetzt, um das deiktische Zentrum – das virtuelle Zuhause – zu fixieren und den Leser, der sich zu Beginn der Rezeptionszeit noch außerhalb dieses Zuhauses befindet, einzuladen: „Come on in and make

[27] Unmarkierte Formen wie „wars" können nur mit einer robusten POS- oder Syntaxanalyse maschinell detektiert werden, weshalb diese Vorkommen nicht in die Analyse eingeflossen sind. Die Filterung falscher Treffer wie etwa „don't", „d'un" und „c't" ist maschinell nur schwierig zu realisieren. Da sie das Ergebnis nicht wesentlich beeinflussen, wurde auf eine manuelle Filterung verzichtet.

yourself at home“ (vgl. Abschnitt 4.6.3). Zur Analyse der Stichproben im Hinblick auf Vorkommen von Begrüßungen und Verabschiedungen wurden zwei Listen mit Floskeln erstellt. Sie wurden durch in der CMC-Literatur genannte Beispiele erweitert und durch die manuelle Analyse einer 1 000 Dokumente umfassenden Stichprobe privater Homepages ergänzt. Die Listen beinhalten 39 Begrüßungen und 42 Verabschiedungen.

S1 enthält 3 287 Begrüßungen und 1 375 Verabschiedungen.[28] S2 enthält 1 005 Begrüßungs- und 634 Verabschiedungsfloskeln. S3 umfasst erwartungsgemäß lediglich 363 Begrüßungen und 228 Verabschiedungen. Die Begrüßungen ähneln einander sehr: S1 enthält 976 Vorkommen von „Willkommen“, gefolgt von „Hallo“ (824), „Welcome“ (450), „hi“ (278), „Herzlich Willkommen“[29] (270) und „Hey“ (235). Die häufigsten Vorkommen in S2 lauten „Willkommen“ (421), „Hallo“ (224), „Herzlich Willkommen“ (123), „Welcome“ (112) und „Hi“ (31). S3 enthält „Willkommen“ (97), „Hallo“ (80), „Herzlich Willkommen“ (43), „Hi“ (41) und „Hello“ (34). Bezüglich der Verabschiedungen enthält S1 die Floskeln „Viel Spaß“ (332), „Grüße“ (193), „bis dahin“ (179), „Gruß“ (98), „CU“ (74) und „Ade“ (58). In S2 werden die Verabschiedungen „CU“ (220), „Viel Spaß“ (83), „bis dahin“ (73), „Gruss“ (46), „Grüße“ (34) und „Gruß“ (25) verwendet, während in S3 die Floskeln „bis dahin“ (50), „Viel Spaß“ (36), „Grüße“ (20), „Gruß“ (19), „CU“ (17) und „Gruss“ (17) belegt sind.[30] Die Verabschiedungen vermitteln einen weniger konventionalisierten Eindruck, jedoch zeigen die Listen, dass konzeptionell schriftliche Formulierungsmuster wie z. B. „mit freundlichen Grüßen“ vollständig fehlen. Möglicherweise haben die Produzenten den Eindruck, dass ihre Verwendung zu eng mit den Kommunikationsformen Brief und E-Mail verknüpft ist. Da das WWW keinen inhärenten Dialogcharakter aufweist, könnte den Produzenten der Einsatz formeller Verabschiedungsfloskeln nicht angemessen erscheinen.[31] Eine derartige Vermeidungsstrategie nicht adäquat erscheinender Merkmale aus anderen Kommunikationsformen wird jedoch nicht konsistent eingesetzt: Einige studentische Homepages beinhalten Textstrukturmuster, die einer generischen E-Mail bzw. einem traditionellen Brief sehr ähnlich sind. Auch ein Postscriptum wird gelegentlich verwendet. Die zweite Stichprobenanalyse geht genauer auf dieses Phänomen ein (vgl. Kapitel 9).

[28] Es wird jeweils nach den entsprechenden Zeichenketten gesucht. Eine genauere Einschätzung, ob es sich bei einem Vorkommen von z. B. „Hallo“ tatsächlich um eine Begrüßung des Lesers handelt, könnte mit Hilfe einer Überprüfung der Position des Vorkommens innerhalb eines Dokuments (Anfang vs. Ende) realisiert werden.

[29] Bei der Erkennung einer Zeichenkette wie „Herzlich Willkommen“ wird lediglich hierfür ein Treffer gezählt; nur ein isoliertes Vorkommen von „Willkommen“ inkrementiert dessen Häufigkeit.

[30] Nach Schütte (2000, S. 167) wird die Grußformel in Mailing-Listen häufig mit einem „Wetterbericht“ flankiert: „Gruss aus Hamburg – schon wieder dunkel“. Im WWW ist unklar, zu welchem Zeitpunkt ein Dokument rezipiert wird, weshalb derartige, „Kopräsenz im gemeinsamen Wahrnehmungs- und Erfahrungsraum“ (ebd.) signalisierende Verabschiedungen und auch die von Quasthoff (1997) berichteten tageszeitorientierten Grußformeln im WWW nicht adäquat erscheinen. In S1–S3 können sie nicht belegt werden.

[31] Feldweg et al. (1995, S. 147) berichten, dass „mfg“ als Abkürzung von „mit freundlichen Grüßen“ aufgrund „des verwendeten Briefstil[s]“ neben eher umgangssprachlichen Abschiedsformeln in dem von ihnen untersuchten Korpus von mehr als 430 000 Newsartikeln hochfrequent ist.

8.8 Fazit

Die Analysen deuten an, dass Studierende, die mit den Internet-Diensten E-Mail, IRC oder Usenet und ihren sprachlichen Spezifika vertraut sind, diese bei der Anfertigung ihrer privaten Homepages verwenden.[32] S1 enthält z. B. mehr Smileys, Iterationen, Emphasen und isolierte Verbstämme als S2 und S3. Die Befunde zur Verwendung von Auslassungspunkten, reduplizierten Interpunktionszeichen, Assimilationen, Wortfrequenzen, Begrüßungen und Verabschiedungen deuten an, dass die studentischen Homepages von einem saloppen, dialogbetonten, d. h. konzeptionell mündlichen Stil geprägt sind. Gleichwohl handelt es sich lediglich um Tendenzen, da die untersuchten Signale in keiner Stichprobe rekurrent verwendet werden. In S1 werden z. B. Smileys in lediglich 3,2%, Buchstabeniterationen in 1,2% und Emphasen nur in 0,4% aller Dokumente eingesetzt. Tabelle 8.10 zeigt die Verwendung im Überblick: Nur 37,3%, 24,6% und 17,2% enthalten mindestens ein Merkmal für konzeptionelle Mündlichkeit. Dabei werden im Durchschnitt nur 1,65, 1,47 und 1,35 der zehn betrachteten Merkmale benutzt.[33] Die Gesamtanzahl verwendeter Merkmale ist mit durchschnittlich 5,66, 4,18 und 8,7 Vorkommen ebenfalls sehr gering.

Merkmal		S1	S2	S3
Anzahl Dokumente		25 481	14 247	10 000
Proz. Anteil der Dokumente mit Merkmalen für konzeptionelle Mündlichkeit		37,29	24,62	17,2
Verwendung von maximal 10 Merkmalen für konzeptionelle Mündlichkeit:	Durchschnitt	1,65	1,47	1,35
	Modus	1	1	1
	Median	1	1	1
	Maximum	10	10	8
Gesamtverwendung von Merkmalen für konzeptionelle Mündlichkeit:	Durchschnitt	5,66	4,18	8,7
	Modus	1	1	1
	Median	2	2	2
	Maximum	1 093	496	261

Tabelle 8.10: Merkmale für konzeptionelle Mündlichkeit in S1–S3

Der insbesondere für S1 ermittelte dialogbetonte Stil widerspricht der von Runkehl et al. (1998, S. 116) aufgestellten These „Je stärker die Kommunikation dialogischer und synchroner erfolgt [E-Mail → Usenet → Chat, G. R.], desto häufiger lassen sich mündliche Aspekte [...] in der Internet-Kommunikation feststellen." Das WWW wurde nicht primär als Kommunikationsmedium konzipiert, es sollte vielmehr zum effektiven Informationsaus-

[32] Auch Eckkrammer (2001, S. 54) ermittelt in Korpora von traditionell und digital publizierten Kochrezepten, Kontakt- sowie Stellenanzeigen „eine erhöhte Frequenz an emotiven Textsequenzen", die aufgrund der Tendenz zur konzeptionellen Mündlichkeit „ein verändertes Näheverhältnis zwischen Textemittent und -adressat" andeuten: „In diesem Zusammenhang kommt wiederum intertextuellen Faktoren, wie der Erfahrung mit anderen Online-Kommunikationsformen (z. B. E-Mail, Chat, News-Groups) eine wichtige Rolle zu."

[33] Hierbei handelt es sich um Assimilationen, Smileys, Auslassungspunkte, reduplizierte Interpunktionszeichen, Emphasen, isolierte Verbstämme, Slangausdrücke, Bigraphen, Begrüßungen und Verabschiedungen.

tausch eingesetzt werden; die Kommunikationsmöglichkeiten des WWW besitzen den Status eines Nebenprodukts und nicht den einer bewusst integrierten Funktionalitätsschicht. Daher befindet sich das WWW in der von Runkehl et al. angegebenen Sequenz der Internet-Kommunikation aus technischer Sicht deutlich jenseits der elektronischen Post – dennoch belegen die Analysen den Einsatz sprechsprachlicher Merkmale in studentischen Homepages. Der von den Autoren dieser Dokumente gleichsam präsupponierte Dialogcharakter (vgl. Bucher, 2004, und Abschnitt 3.5.3) könnte zum einen ein Resultat mangelnder Erfahrung mit der Produktion von Inhalten für das WWW sein, zum anderen von einem Transfer und der Adaption von Spezifika ihnen bekannter Kommunikationsdienste hervorgerufen werden.[34]

Im Hinblick auf die maschinelle Identifizierung von Hypertextsorten sollte das Potenzial der in der CMC-Literatur berichteten Merkmale für konzeptionelle Mündlichkeit bestimmt werden, weil diese präzise automatisch erfasst werden können. Die Analysen belegen zwar, dass diese Mittel häufiger in studentischen Homepages als in den persönlichen Webseiten von Hochschulmitarbeitern verwendet werden, doch wird zugleich deutlich, dass sie keinesfalls zur alleinigen Erkennung eingesetzt werden können. Vielmehr erscheint es notwendig, ein übergreifendes Merkmal für den Grad der konzeptionellen Mündlichkeit eines HTML-Dokuments bzw. eines Hypertextes anzunehmen, das eine maschinelle Klassifizierung in Bezug auf das Kontinuum zwischen konzeptioneller Mündlichkeit und Schriftlichkeit voraussetzt. Zusätzlich sind darüber hinaus linguistische Befunde einzubeziehen (vgl. hierzu auch die Dimension „Involved vs. Informational Production“ bei Biber, 1988). Kapitel 14 geht ausführlich auf dieses Thema ein.

Die Analysen basieren auf der Annahme, dass *jedes* HTML-Dokument aus S1 und S2 eine Teilinstanz der korrespondierenden Hypertextsortenvariante darstellt. Da die Zuordnung eines Dokuments zu einem übergreifenden Hypertext *nicht* vorliegt, verfälschen diejenigen untergeordneten Knoten die Ergebnisse, die auf anderen Hypertextsorten basieren. Dass zahlreiche eingebettete Hypertext(knoten)sorten in den Stichproben privater und persönlicher Homepages enthalten sind, belegen die Dokumente, die exemplarisch in den Abschnitten 8.5 bis 8.7 genannt wurden (z. B. *Gästebuch*, *Reisebericht*, *Reiseführer*, *Plattenkritik*, *Studiengangsbroschüre* und *Fotogalerie*; vgl. auch Fußnote 127, S. 226). Diese Beispiele zeigen, dass eine große Bandbreite eingebetteter Hypertextsorten vorliegt, deren tatsächliche Ausprägung nur durch weitere Stichprobenuntersuchungen ermittelt werden kann. Die beiden nachfolgenden Analysen gehen in detaillierter Form auf die Hypertextsorten *private Homepage eines Studierenden* (Kapitel 9) und *persönliche Homepage eines Wissenschaftlers* ein (Kapitel 10).

[34] Runkehl et al. (1998, S. 116) vertreten die folgende Auffassung: „Die sprachliche Variation zwischen den Diensten [...] zeigt, daß Aussagen [...] über ›*die* Sprache des Internet‹ weit entfernt sind von der sprachlichen Realität [...].“ Dürscheid (2004) demontiert in diesem Zusammenhang die Kernaussage von Crystal (2001), der die Existenz von „netspeak“ postuliert, und gelangt zu der Schlussfolgerung: „*Den* Sprachgebrauch im Internet gibt es nicht, *die* Netzsprache gibt es nicht.“ Schließlich ist es unmöglich, „pauschale Aussagen über den Sprachgebrauch im Chat oder in der E-Mail oder gar im Internet zu machen.“ (ebd.).

9 Analyse 2: Die private Homepage eines Studierenden

9.1 Einleitung

Die zweite Analyse betrifft die Untersuchung studentischer Homepages und beschäftigt sich unter anderem mit der Frage, in welcher Relation diese Hypertextklasse zur Hypertextsorte *private Homepage* steht. Die nachfolgende Analyse geht zur Kontrastierung der Ergebnisse auf die Hypertextsorte *persönliche Homepage eines Wissenschaftlers* ein.

Nach der Darstellung der Ziele dieser Analyse sowie der verwendeten Stichprobe (Abschnitt 9.3) diskutiert Abschnitt 9.4 zunächst die Funktionen studentischer Homepages und von den jeweiligen Produzenten explizit thematisierte Konventionen, woraufhin Abschnitt 9.5 eine detaillierte Inhalts- und Makrostrukturanalyse vorstellt. Abschnitt 9.6 geht auf das Phänomen der E-Mail-ähnlichen Textstrukturmuster, Grüße an Freunde und Bekannte sowie heterogene Formen der Rezipientenadressierung ein.

9.2 Ziele und Bezüge zum Hypertextsortenmodell

Diese Untersuchung verfolgt verschiedene Ziele. Bates und Lu (1997), Dillon und Gushrowski (2000) und Bittner (2003) beschäftigen sich mit persönlichen bzw. privaten Homepages (vgl. Abschnitt 4.6.3), so dass sich die Frage stellt, ob private Homepages, die von Studierenden angefertigt werden, spezifische Eigenschaften aufweisen, die sie von anderen privaten Homepages unterscheiden und es möglicherweise sogar gestatten, von einer subgenerischen Varietät, d. h. einer Hypertextsortenvariante zu sprechen. Von Relevanz ist auch die Arbeit von Bayerl (2002), die eine linguistische Analyse studentischer Homepages vorstellt.

In Bezug auf das Hypertextsortenmodell sind verschiedene Aspekte zu untersuchen. Zunächst betrifft dies die kommunikative Funktion der Hypertextsorte, die von einigen in der

Stichprobe enthaltenen Homepages explizit thematisiert wird. Dies gilt ebenfalls für verschiedene allgemeine Konventionen, die darüber hinaus durch eine Analyse der Inhalte und Themen sowie der generellen Makrostruktur weiter spezifiziert werden können. Auf diese Weise wird ein Profil studentischer Homepages angefertigt, das insbesondere auf die Einstiegsseite als Hypertextknotensorte sowie auf die Hypertextsortenmodule Bezug nimmt, die aus der makrostrukturellen Analyse abgeleitet werden können.

9.3 Die Stichprobe

Diese Analyse knüpft unmittelbar an die vorangegangene Untersuchung an. Aus der in Abschnitt 8.4 dargestellten Stichprobe studentischer Homepages wurden mittels eines Zufallsverfahrens 20 Hypertextexemplare ausgewählt. Fünf Dokumente wurden nicht in die Stichprobe aufgenommen, da sie nicht von Studierenden angefertigt wurden oder keine studentischen Homepages darstellen.[1] Tabelle 9.1 zeigt eine Aufstellung der 15 Homepages (HP 1–15), ihren Umfang (zwischen lediglich einem und 224 HTML-Dokumenten) und Datumsangaben bezüglich der Integration in das Korpus sowie der letzten Änderung (vgl. Kapitel 7).[2] Von einer Studentin wurde lediglich eine der Homepages angefertigt, was die häufig geäußerte Vermutung bestätigt, dass private Homepages meist von männlichen Personen erstellt werden. Oftmals wird auch beobachtet, dass die Autoren privater Homepages einen naturwissenschaftlichen Hintergrund besitzen. In der hier untersuchten Stichprobe ist die Verteilung gleichmäßig: Dem natur- oder ingenieurswissenschaftlichen Bereich können sechs Homepages zugeordnet werden, während fünf Homepages von Studierenden geisteswissenschaftlicher Fächer erstellt wurden.[3]

9.4 Funktionen und Konventionen

Private und somit auch studentische Homepages können mit zahlreichen potenziellen Funktionen erstellt und publiziert werden. Abschnitt 4.6.3 ist diesbezüglich auf verschiedene funktionale Typologisierungsversuche eingegangen, die nachfolgend auf die 15 in der Stichprobe enthaltenen Webangebote angewendet werden. Zuvor werden explizit in den Hypertexten aufgeführte Äußerungen zu ihrer Funktion und zu generellen Konventionen diskutiert.

[1] Eines der verworfenen HTML-Dokumente enthält z. B. lediglich das Wort „Hallo“, ein weiteres ist offenbar das praktische Ergebnis eines HTML-Kurses und enthält mit Blindtext beschriftete Überschriften und Listen.

[2] Bayerl (2002) berichtet für die von ihr untersuchte Stichprobe von 40 studentischen Homepages, dass diese einen Durchschnittsumfang von 16,7 HTML-Dokumenten besitzen. Mit durchschnittlich 41,5 Dokumenten sind die in dieser Analyse betrachteten Hypertexte deutlich umfangreicher. Es ist zusätzlich anzumerken, dass diese Daten nicht manuell erhoben, sondern der Korpusdatenbank entnommen wurden: Da eine HTML-Datei in bestimmten Kontexten über mehrere URLs adressiert werden kann (z. B. falls auf der Ebene des Dateisystems ein symbolischer Link, der in einer URL referenziert wird, auf eine Datei verweist), kann die Anzahl tatsächlich physikalisch verfügbarer HTML-Dateien von diesen Zahlen abweichen.

[3] Vier Homepages konnten mangels expliziter Angabe des Studiengangs nicht zugeordnet werden. Dem natur- oder ingenieurswissenschaftlichen Bereich sind HP 3 (Mathematik/Physik), HP 7 (Elektrotechnik), HP 4 (Umwelt- und Verfahrenstechnik), HP 15 (Bauingenieurwesen) sowie HP 6 und HP 11 (Medizin) zugehörig. Zu den geisteswissenschaftlichen Studiengängen gehören HP 5 (Pädagogik), HP 10 (Politologie) sowie HP 8, HP 13 und HP 14 (Betriebswirtschaftslehre).

	URL	Anzahl Dokumente	Download	Letzte Änderung
HP 1	http://wwwstud.uni-giessen.de/~st2137/	22	16.1.2001	06.10.2000
HP 2	http://wwwstud.uni-giessen.de/~s2498/	5	16.1.2001	21.11.1998
HP 3	http://www.rzuser.uni-heidelberg.de/~cellsaes/	27	18.4.2001	18.12.2000
HP 4	http://studserv.rzpool.tu-cottbus.de/~gepperj/	53	07.2.2001	17.12.1999
HP 5	http://wwwstud.uni-giessen.de/~st2763/	64	16.1.2001	24.10.2000
HP 6	http://www.dorf.rwth-aachen.de/~postap/	31	25.7.2001	k. A.
HP 7	http://www.uni-ulm.de/~s_tschne/	7	02.2.2001	14.03.1998
HP 8	http://www.student.uni-augsburg.de/~schwinge/	9	03.4.2001	10.07.2000
HP 9	http://www.dorf.rwth-aachen.de/~ostauffer/	86	26.7.2001	k. A.
HP 10	http://www.rzuser.uni-heidelberg.de/~tstiepak/	10	18.4.2001	28.01.2000
HP 11	http://www.uni-ulm.de/~s_sstrob/Strobel_start.html	58	02.2.2001	22.06.1998
HP 12	http://www.rzuser.uni-heidelberg.de/~ctschamb/	10	18.4.2001	10.05.2000
HP 13	http://wwwstud.uni-giessen.de/~s5401/	1	16.1.2001	29.07.2000
HP 14	http://wwwstud.uni-giessen.de/~s6534/	16	16.1.2001	14.01.2001
HP 15	http://www.dorf.rwth-aachen.de/~marco/	224	25.7.2001	05.08.2000

Tabelle 9.1: Die untersuchten studentischen Homepages

9.4.1 Metadiskursive Äußerungen

Der Anlass zur Anfertigung und Veröffentlichung einer studentischen Homepage betrifft nicht notwendigerweise die weltweite Publikation von Texten oder Informationen zu bestimmten Themengebieten. In zehn Homepages gehen die jeweiligen Verfasser explizit auf den Sinn persönlicher Homepages im Allgemeinen und die spezifische Funktion ihres Webangebotes im Speziellen ein.[4] Diese metadiskursiven Äußerungen belegen eine intensive Auseinandersetzung mit den Regeln und Vorgaben der korrespondierenden Hypertextsorte (vgl. Eckkrammer, 2001, S. 54) und verdeutlichen zugleich ihre zentralen Charakteristika.

Der Verfasser von HP 4 listet in einem von der Einstiegsseite mit Hilfe des Hyperlinks „about me“ erreichbaren HTML-Dokument mit dem Dateinamen `iche.htm` [sic] verschiedene biografische Angaben auf. Anschließend enthält diese Webseite einen Text, der den Verfasser neben der tabellarischen Präsentation auch in einer narrativen und somit individualisierten Weise vorstellt. Nach der Begrüßung des Lesers thematisiert der Autor die Funktion seiner Website: „Seit nun mehr als 3 Jahren existiert nun schon meine Homepage, aber mir faellt immernoch nichts Grossartiges zu dieser Seite ein.“ Dieser Äußerung zufolge hat der Verfasser offenbar keine präzise Vorstellung von der Funktion seiner Homepage.[5]

[4] Einige Homepages der Stichprobe enthalten generelle Bedienungshinweise für das *World Wide Web* und spezifische Benutzungshinweise für eine Website, die ebenfalls als metadiskursive Äußerungen aufgefasst werden können: „Um einen besseren Überblick über meine Web-Seiten zu erhalten ist hier das Inhaltsverzeichnis zu finden, von dem aus es mit Hyperlinks weiter geht.“ (HP 11). Nach der Begrüßung des Lesers merkt der Verfasser von HP 6 an: „Um zu den eigentlichen Inhalten dieser Seite zu gelangen, bitte die Buttons drücken.“

[5] Umso erstaunlicher ist, dass der Autor von HP 4 mit 53 HTML-Dokumenten ein im Vergleich zu anderen studentischen Homepages sehr umfangreiches Angebot pflegt, das vornehmlich aus Informationen, Texten und Fotos von Auftritten seiner Band „Sandow“ besteht. Diese hat sich jedoch aufgelöst, weshalb die mit „Projekte“ betitelte Webseite einen Vermerk enthält, der die vom Verfasser eingeschätzte Funktionslosigkeit seiner Homepage zu erklären hilft: „So kann ich zum Beispiel nicht mehr ueber meine Sandow-Page schreiben, die ich immer gerne und sorgfältig geflegt habe. Tja da löst sich die Band auf und ich meine Page.“ Die Datei `iche.htm` wurde etwa ein Jahr nach dem Datum der letzten Änderung des Dokuments „Projekte“ aktualisiert.

Im Gegensatz hierzu verbindet der Autor von HP 11 eine klare Zielsetzung mit seiner persönlichen Homepage, die im dritten Absatz einer E-Mail-ähnlichen und in der Einstiegsseite dargestellten Begrüßung des Lesers und Erläuterung seines Webangebots angesprochen wird: „Ich hoffe, daß ich den Besuchern dieser Seite ein paar nette Hyperlinks präsentieren kann, und Ihnen somit die tägliche Arbeit erleichtern und das Surfvergnügen maximieren kann." Hyperlinks zu externen Angeboten führen den Rezipienten naturgemäß von demjenigen Hypertext fort, der diese Verknüpfungen anbietet, weshalb kommerziell ausgerichtete Websites in der Regel keine externen Hyperlinks beinhalten. Der Produzent von HP 11 sieht sich vermutlich nicht in der Lage, einen substanziellen inhaltlichen Eigenbeitrag zum WWW leisten zu können, weshalb er stattdessen auf die effizient realisierbare Strategie ausweicht, zumindest eine Reihe von Hyperlinks anzubieten, so dass seine Homepage nicht als informationelle Sackgasse, sondern vielmehr als Sprungbrett zu anderen Websites fungieren kann und hierdurch den Rezipienten zu einem erneuten Besuch einlädt.[6] Mit einer ähnlichen Äußerung leitet der Verfasser von HP 1 eine sehr umfangreiche Liste externer Hyperlinks ein: „Wer meiner Homepage überdrüssig geworden ist, kann sich hier ja was besseres suchen."

Diese Beispiele deuten an, dass die Autoren privater Homepages mit einem Dilemma konfrontiert werden. Nahezu alle betrachteten Homepages belegen in unterschiedlichster Weise, dass es ihre Verfasser einerseits als einen sehr reizvollen Gedanken betrachten, mit einfach zu erlernenden technischen Mitteln und einem vertretbaren Zeitaufwand eine weltweite Öffentlichkeit zu erreichen, andererseits liegt in den meisten Fällen zwar sehr viel Engagement, jedoch keine konkrete Idee vor, welche spezifischen Inhalte publiziert werden sollen (vgl. auch Abschnitt 4.6.10). Der Verfasser von HP 13 präsentiert in der Einstiegsseite zunächst einen externen Hyperlink zu einer Website, die sich mit der Kampfsportart Kendo beschäftigt und fügt als Kommentar an: „Das wäre auch schon das 1. Thema. Ihr werdet schon erraten haben warum. Richtig, es ist mein Hobby!" Nach der Nennung seines Studiengangs folgt eine metadiskursive Äußerung, die verdeutlicht, dass es dem Autor schwer fällt, neben der präferierten Freizeitbeschäftigung weitere Inhalte für seine Homepage ausfindig zu machen: „Hierzu [gemeint ist der Studiengang, G. R.] wird mir auch noch was einfallen! Vielleicht eine Skripttauschbörse? Bei Interesse einfach melden!" Der Verfasser diskutiert mögliche zukünftige Inhalte seiner Homepage und bittet die ihm unbekannte Leserschaft um Rückmeldungen, ob die präsentierte Idee als nützlich erachtet wird. Zudem zeigt dieses Beispiel das Bemühen des Autors, Inhalte anzubieten, die über das Sammeln und Präsentieren externer Hyperlinks hinausgehen. Durch die Publikation der Homepage sieht er sich veranlasst, nach weiterführenden Inhalten zu suchen, die sein Angebot aufwerten und von anderen abheben. Nach der

[6] Diese Interpretation wird durch eine Absichtserklärung im sechsten Absatz des Textes gestützt. Der Autor von HP 11 bezieht sich auf zukünftige Inhalte: „Sehr bald werde ich auch hier auf dieser Site ein [sic] Auswahl kleiner aber feiner Share- und Freewareprogramme [...] direkt zum Download anbieten, so daß Ihr nicht das ganze Web abgrasen müsst um an ein gutes Prögrämmchen zu kommen. Also noch ein bischen [sic] Geduld!" Der informelle und joviale Duktus sowie die umgangssprachliche Lexik („abgrasen", „Prögrämmchen") sollen den Rezipienten vermutlich prophylaktisch beschwichtigen, so dass dieser nicht zu einer negativen Einstellung gegenüber dem Produzenten gelangt, falls sich bei einem erneuten Besuch herausstellen sollte, dass die Absichtserklärung nicht eingehalten werden konnte. Das Anbieten ausgewählter und nicht selbst implementierter Programme kann als weitere Strategie gewertet werden, einerseits vermeintlich sinnvolle Inhalte anzubieten, andererseits den notwendigen Realisierungsaufwand zu minimieren. Zusätzlich besitzt diese Äußerung die Funktion, den Rezipienten zu einem wiederholten Besuch des Angebots zu motivieren.

Begrüßung des Lesers befindet sich in HP 13 eine weitere Absichtserklärung, die den rudimentären Status dieser Homepage anspricht: „Noch befindet sich diese Seite im Aufbau, aber sie wird einige interessante Themen enthalten."[7]

Es stellt sich die Frage, *weshalb* Studierende persönliche Homepages anfertigen und veröffentlichen, obwohl ihnen den genannten Beispielen zufolge zumindest gelegentlich durchaus bewusst ist, dass sie keinen inhaltlichen Mehrwert einbringen können. Besonders ausführlich geht der Verfasser von HP 7 auf dieses Thema ein. Im Begrüßungstext diskutiert er den Grund zur Erstellung dieses sehr rudimentären Webangebots: „Ich bin echt mal gespannt, was herauskommt, wenn ich hiermit fertig bin. Zweck des Ganzen ist es mich ein wenig [in] HTML und Java (ja es wird noch ein wenig Java geben) einzuarbeiten und herumexperimentieren [sic]." Die Motivation zur Erstellung betrifft demnach das Erlernen der *Hypertext Markup Language* und der Programmiersprache *Java* sowie parallel das Erstellen eines initialen Webangebots, das jedoch nicht nur zu Lernzwecken auf dem heimischen Rechner erstellt und betrachtet, sondern zugleich einer weltweiten Öffentlichkeit präsentiert wird. Daraufhin spricht der Autor die verfügbaren Inhalte seiner Homepage an: „Viel gibt es noch nicht zusehen [sic], diese Seite hier, einen kurzen Lebenslauf, ein Photo und natürlich meine Email-Adresse. Aber [für] den Anfang müsste das reichen." Diese Äußerung belegt, dass der Verfasser der Ansicht ist, sein Hypertext genüge den Anforderungen bzw. Konventionen der Hypertextsorte *private* bzw. *studentische Homepage*. Anschließend geht er im Kontext einer Erläuterung der künftigen Inhalte auf die generelle Funktion dieser Hypertextsorte ein: „Die Links die es in Zukunft geben wird, werden wahrscheinlich sowieso nur mir was nützen. Ich fr[a]ge mich so oder so nach dem Sinn von privaten Homepages. Ist doch eigentlich nur eine Spielerei, oder? Aber dafür eine sehr schöne und nette Spielerei." Zweifelsohne empfindet der Verfasser die Erstellung von HTML-Dokumenten als interessante und herausfordernde Beschäftigung, die Existenz eines eigentlichen Zwecks, einer *kommunikativen* Funktion privater Homepages wird jedoch in Frage gestellt. Der Begrüßungstext schließt mit der Feststellung: „Mit anderne [sic] Worten: ich denke, daß ich noch genügend zu tun habe. Mal schauen wo ich rauskomme." Dem Autor ist bewusst, dass Möglichkeiten der Erweiterung seines Webangebots vorliegen, er macht jedoch keine Angaben, zu welchem Zeitpunkt welche spezifischen Inhalte oder Funktionen integriert werden sollen. Da keine Verpflichtung zur Publikation oder fortwährenden Erweiterung eines privaten Webangebots existiert, ist die inhaltliche und funktionale Gestaltung vollständig vom individuellen Engagement und Ehrgeiz des Produzenten abhängig, dem darüber hinaus – abgesehen von stagnierenden Werten des Zugriffszählers – keine sanktionierenden Maßnahmen drohen, falls ein rudimentäres Angebot trotz gegenteiliger Absichtserklärung nicht aktualisiert und erweitert werden sollte.

Der Verfasser von HP 10 bezieht sich in besonders drastischer Form auf die Konventionen der Hypertextsorte *private Homepage*. In einem Knoten, der von der Einstiegsseite über den Hyperlinkanzeiger „Wer ist Tobias Stiepak?" erreichbar ist, antwortet der Produzent: „Nun, ich heiße Tobias Stiepak, und der Rest, der geht euch gar nichts an." Gefolgt wird diese Antwort von der Überschrift „Warum ich hier keinen 10-seitigen Lebenslauf mit passenden Fotos präsentiere?" Der sich anschließende Text liefert eine aussagekräftige Antwort:

[7] Dieses Dokument wurde vom Verfasser zuletzt am 29. Juli 2000 modifiziert und am 16. Januar 2001 in das Korpus übertragen. Die beabsichtigten Änderungen wurden bis zum 16. November 2004 nicht durchgeführt.

> Weil ich absolut keinen Bock habe, mich genauso zum Affen zu machen, wie all' die anderen Handlanger im Netz, die meinen, ihre Lebensgeschichte müsse unbedingt der ganzen Welt zugänglich gemacht werden. Und sich deshalb tagelang mit dem Einscannen von noch so tollen Bildern und der Formulierung ihres erbärmlichen Lebenslaufes beschäftigen. [...] Nur um dann jeden Tag die eigene Homepage anzusurfen, und zu überprüfen, wieviele Bemitleidenswerte den literarischen Erguss bisher gelesen haben. [...] Nur, es interessiert doch überhaupt niemanden; wem sollen diese Informationen denn nutzen? (Und sagt jetzt bloß nicht: „jemandem, der sich über mich informieren will", denn wer, außer dem BKA, will das schon ???) Nee, nee, das muß ich mir nicht geben. So, nu is genug. [HP 10]

Diese verkürzt reproduzierte Antwort verdeutlicht, dass der Verfasser das Präsentieren eines Lebenslaufes – zumindest jedoch einer narrativen Darstellung der wichtigsten biografischen Angaben – zwar als konstitutiv für die Hypertextsorte *private Homepage* erachtet, jedoch keinen triftigen Grund sieht, dieser Konvention Folge zu leisten.[8] Stattdessen wird das Fehlen des Lebenslaufes thematisiert, weil die angesprochene Konvention bewusst gebrochen wurde. Der Autor verzichtet also auf die schlichte Option, *keinen* Lebenslauf anzubieten und diesen Umstand *nicht* zum Thema zu machen. Da er jedoch an verschiedenen Stellen seiner Homepage Bemühungen zeigt, seine Medienkompetenz herauszustellen, sieht er sich vermutlich gezwungen, diesen Konventionsbruch erklären zu müssen.[9] Eine weitere potenzielle Konvention der Hypertextsorte thematisiert der Verfasser von HP 5 in einem HTML-Dokument, das mit „Fotos" überschrieben ist. Neben einigen Bemerkungen zu den Bildern, die auf den Folgeseiten eingesehen werden können, merkt er den Grund für die Existenz der „Foto-Sektion meiner Site" an: „JEDE(R) hat auf seiner persönlichen Homepage diese Bilder die die Welt nicht braucht 8-))". Der Autor von HP 7 fasst diese Konvention spezifischer auf und kommentiert „ein ziemlich aktuelles Foto von mir" mit der metadiskursiven Äußerung „Tja,

[8] Das Wissen über die Existenz dieser Konvention kann verschiedene Ursprünge besitzen (vgl. auch Amitay und Oberlander, 1997, Amitay, 2001). Es ist als wahrscheinlich zu erachten, dass der Autor von HP 10 diese Konvention als generalisierende Beobachtung nach der Rezeption zahlreicher privater Homepages aufgestellt hat (vgl. ausführlich hierzu Abschnitt 4.3.2). Alternative Möglichkeiten betreffen die Lektüre von Webdesign-Ratgebern, die bezüglich der Anfertigung einer privaten Homepage oftmals den Hinweis geben, einen Lebenslauf zu integrieren. Die Konvention könnte auch in einem HTML-Kurs vermittelt worden sein. Unmittelbar vor dem Abschluss dieses Dokuments, der durch den Vor- und Nachnamen des Verfassers konstituiert wird, befolgt der Autor die Konvention interessanterweise dennoch, wenngleich in sehr rudimentärer Weise: „Nur soviel sei gesagt: Ich studiere in Heidelberg Pol, Vwl und Geschichte – also schaut mal vorbei."

[9] In weiteren HTML-Dokumenten („Informationen zur Musik" und „Allgemeine Informationen") schildert der Verfasser in stilistisch vergleichbarer Form seine Meinungen zum Thema Webdesign sowie zu Unterhaltungsmusik der Genres „Schlager" und „Neue Deutsche Welle". Insbesondere die Ausführungen zum Webdesign, die ebenfalls Medienkompetenz signalisieren sollen, können in vergleichbarer Form in vielen persönlichen Homepages belegt werden: „Wie der geneigte Leser dieser Seiten schon bemerkt hat, findet ihr auf diesen meinen Seiten keinerlei Animationen, Java-applets und sonstige »Verschönerungen« meiner Seiten. Ich bin der Ansicht, dass der Inhalt einer Seite stimmen muß, und die Aufmachung nur ein kleiner unwesentlicher Teil ist. [...] Nun, einige werden jetzt sicherlich sagen, der kann das bestimmt gar nicht und hat deshalb seine Homepage so schlicht gestaltet, aber ihr wisst ja alle selbst, wie einfach es mittlerweile ist, eine einigermaßen gute (vom Design-Gesichtspunkt) Homepage zu entwerfen. Viel zu einfach vielleicht, denn dann hätten wir jetzt nicht diese Flut von übelsten Seiten im Netz (Ob meine dazugehört – entscheidet selbst)." Bittner (2003, S. 99 f.) beobachtet „Manifestationen persönlicher, politischer oder weltanschaulicher Überzeugungen" in nur zwei der 25 untersuchten Homepages (vgl. Fußnote 130, S. 228). Neben den genannten HTML-Dokumenten können HP 8 und HP 14 als (gemäßigte) Beispiele derartiger „Manifestationen" angeführt werden.

was wäre eine Homepage ohne ein Foto des Homepagebesitzers ?“ Der Verfasser von HP 12 bietet auf seiner Homepage eine große Anzahl thematisch organisierter Fotos an, die bei einem Aufenthalt in Australien entstanden sind. In der alten Einstiegsseite seiner Homepage wird eine kontinuierliche Liste der durchgeführten Änderungen gepflegt. Einer der Einträge lautet: „Es gibt schon wieder eine neue Sektion (wenn jemandem ein besseres Wort dafür einfällt, bitte melden). [...]: Der Strand.“ Diese Äußerung zeigt, dass dem Autor die Existenz der Konvention bewusst ist, größere Mengen von Fotos in Webangeboten nur in sortierter Form anzubieten. Das Lexem, mit dem derartige Kategorien etikettiert werden, kennt er jedoch nicht, weshalb er sich an die Rezipienten seines Angebots wendet, um die ad hoc gewählte Bezeichnung „Sektion“ nach einer entsprechenden Rückmeldung möglicherweise durch einen adäquateren Terminus ersetzen zu können.

Der Produzent von HP 14 besitzt im Gegensatz zu fast allen anderen Autoren der untersuchten Homepages eine sehr konkrete Vorstellung, wer zum Rezipientenkreis seines Webangebots gehört: „Auf diesen Seiten werdet ihr ziemlich viel von mir erfahren, wenn ihr wollt. Von meiner Art zu denken, glauben, die Welt zu sehen, etc, von dem, was ich mache, gemacht habe und machen will. Diese Seiten sind (bzw. werden – da alles noch im Aufbau ist) sehr persönlich und sind vor allem für Freunde gedacht. Aber wenn sie anderen auch gefallen – so soll es sein!“ Dieser Text beschließt einen E-Mail-ähnlichen Begrüßungstext, der mit „Euer Jan Wagner (:“ unterzeichnet ist. Er fungiert als inhaltliche Einleitung der Homepage.[10]

9.4.2 Anwendung funktionaler Typologien

Miller (1995) und Walker (2000) schlagen Typologien privater bzw. persönlicher Homepages vor. Miller (1995) betrachtet vor allem die sozialen Rollen und die Spezifika der jeweils dargestellten Personen (vgl. Abschnitt 4.6.3). Für die Typen (2) „Hi, this is me (as a member of an organisation)“, (3) „Hi, this is us“, (5 a) „Cool style“ sowie (5 c) „An advertisement for the service I can provide“ sind in der Stichprobe keine Belege enthalten. Insgesamt sechs der untersuchten Dokumente können eindeutig Typ (1) „Hi, this is me (as an individual)“ zugeordnet werden (HP 3, HP 4, HP 5, HP 7, HP 11, HP 14), der die primäre Funktion der Selbstdarstellung besitzt und biografische Angaben einer Person, Informationen zum Studiengang, Hobbys und Fotos beinhaltet. Zwei weitere Homepages (HP 2, HP 9) können diesem Typ ebenfalls zugerechnet werden, jedoch sind einige der von Miller (1995) genannten Merkmale in diesen Webangeboten nicht ausgeprägt. Beispielsweise fehlen in HP 2 der Name des Verfassers und biografische Angaben, jedoch enthält diese Homepage mehrere Fotos sowie Grüße und Hyperlinks zu den Webangeboten von Freunden. HP 9 enthält zwar zahlreiche Kontaktinformationen, jedoch keine Fotos oder Angaben zur Herkunft des Au-

[10] Trotz der prophylaktisch integrierten Beschwichtigungsfloskel („da alles noch im Aufbau ist“) besitzt HP 14 einen großen Umfang. Der Autor präsentiert Gedichte und lyrische Texte, zahlreiche kommentierte Fotos, Listen von Hyperlinks und ein Gästebuch. Das Kernstück der Homepage, „Jan's Tagebuch“, befindet sich jedoch unmittelbar in der Einstiegsseite (vgl. Abschnitt 4.6.7). Es enthält 13, teils umfangreiche Einträge, die zwischen dem 23. Juli 2000 und dem 14. Januar 2001 erstellt wurden und Erfahrungen aus dem Alltag sowie sehr persönliche Gedanken des Verfassers thematisieren. Ein ebenfalls mit einem Zeitstempel (23. Juli 2000) versehenes „Vorwort“ kommentiert den Beginn des Tagebuchs: „Nachdem ich die Seiten hier nun ein Jahr lang ziemlich vernachlässigt habe (was den Effekt hat, daß ich nun hier ziemlich ungestört bin und keiner mehr vorbeischaut), werde ich nun soetwas wie ein Tagebuch hier beginnen. Mal schaun, wie's wird!“

tors. Stattdessen verweist der Verfasser auf eine „Bewohnerdatenbank“, in der alle Bewohner seines Studentenwohnheims erfasst sind. Vier Homepages (HP 1, HP 6, HP 10, HP 13) können eindeutig Typ (4) „This is what I think is cool“ zugeordnet werden (vgl. Fußnote 108). Dieser Typ, der nur rudimentäre persönliche Informationen umfasst, enthält vornehmlich Hyperlinks zu spezifischen Themen (z. B. Kinofilme, Sportarten und Musikgruppen oder -stile). Darüber hinaus liegen zwei Überschneidungen vor: HP 8 kann den Typen (1), (4) und (5 b) zugeordnet werden, wohingegen HP 15 den Typen (1) und (5 b) zugehörig ist; Typ (5 b) wird von Miller (1995) „The electronic curriculum vitae“ genannt und als ernstgemeinter und direkter Versuch, eine Anstellung zu finden, interpretiert. Die genannten Homepages enthalten zwar sehr umfangreiche tabellarische Lebensläufe, jedoch wird nicht explizit betont, dass die Verfasser auf der Suche nach einer neuen Anstellung sind. Die Lebensläufe geben darüber Auskunft, dass einer der Produzenten als Doktorand tätig ist, während der andere seine erste Anstellung nach dem Studienabschluss angetreten hat. Diese beiden problematischen Beispiele zeigen, dass eine Konzeptualisierung von Hypertextsorten als flexible Konstrukte, deren Instanzen mit Hypertextsortenmodulen bestückt werden, die unter anderem als kommunikativ-funktional markierte Bausteine fungieren, einem statischen Typensystem vorzuziehen ist, da in Hypertexten nicht notwendigerweise eindeutige Ausprägungen vorliegen. Mit HP 12 existiert eine Homepage, die keinem der von Miller (1995) aufgestellten Typen zugeordnet werden kann: Der Autor präsentiert seinen Namen, seine E-Mail-Adresse und Fotos von einem Aufenthalt in Australien.

Walker (2000) geht von drei Typen aus (vgl. Abschnitt 4.6.3). Der erste Typ umfasst nur rudimentäre Informationen wie z. B. den Namen des Autors, Alter und Wohnort, Angaben zu Hobbys und sonstigen Interessen. Der zweite Typ ist als eine Erweiterung aufzufassen und stellt den Werdegang des Produzenten als narrativen Text dar. Homepages des dritten Typs konzentrieren sich auf ein spezifisches Interesse des Produzenten, wobei biografische Informationen lediglich in rudimentärer Form angegeben werden und somit in den Hintergrund rücken. Auch im Hinblick auf diese Typologie kann keine eindeutige Zuordnung vorgenommen werden. Der erste Typ umfasst 14 Homepages, dem zweiten Typ können drei Dokumente zugeordnet werden und der dritte Typ beinhaltet neun Webangebote. Überlappungen liegen in acht Homepages vor, wobei eine gleichwertige Zuordnung zum ersten und dritten Typ dominiert. Zwei Dokumente (HP 14 und HP 15) können allen drei Typen zugeordnet werden. Eine eindeutige Differenzierung in Bezug auf intrinsische und extrinsische Homepages kann ebenfalls nicht für jedes Dokument vorgenommen werden: Eine intrinsische Homepages ist Walker (2000) zufolge primär an eine dem Autor unbekannte Rezipientengruppe gerichtet. Korrespondierende Merkmale sind in insgesamt 12 Homepages enthalten. Extrinsische Homepages dienen vornehmlich der Ansprache von Lesern, die dem Autor bekannt sind. Entsprechende Merkmale können in acht Homepages belegt werden. In sechs Homepages liegen entweder Merkmale für beide Typen vor oder eine Zuordnung zu einem der beiden Typen kann nicht vorgenommen werden, da keine Indikatoren für die vom Verfasser intendierte Rezipientengruppe vorhanden sind.

Abschließend kann festgehalten werden, dass diese Typologien zur Bestimmung der Typenausprägungen bei einer konkreten Stichprobe von Homepages keine trennscharfen Differenzierungskriterien bereitstellen. Sowohl Miller als auch Walker geben an, dass in vielen Fällen Überlappungen vorliegen, so dass die skizzierten Typen vornehmlich als „primäre“

und „sekundäre Typen“ aufgefasst werden sollten, jedoch können hierdurch diejenigen Kategorien, die in den Typologien nicht vorgesehen sind, nicht reflektiert werden. Die Beispiele verdeutlichen, dass zur Charakterisierung bzw. Typologisierung von Instanzen nicht ein atomares und statisches Inventar von Typen, sondern vielmehr konzeptuelle Subtypen auf einer spezifischeren Ebene (d. h. Hypertextsortenmodule) vorzusehen sind, so dass unterschiedliche Ausprägungen einer Hypertextsorte als variierende Vertreter eines prototypischen Kerns konzeptualisiert werden können, die zwar auf der Peripherie einander ähneln, jedoch unterschiedliche Konfigurationen von Hypertextsortenmodulen besitzen (vgl. Abschnitt 5.8.1).

9.5 Inhalte und makrostrukturelle Komponenten

Die Hyperlinkanzeiger, die in den zentralen, als Inhaltsverzeichnisse fungierenden Navigationshilfen der jeweiligen Einstiegsseiten der 15 studentischen Homepages enthalten sind, zeigen das sehr breit gefächerte Themenspektrum dieser Hypertextsorte auf. Zusätzlich werden bereits auf dieser Ebene etablierte Konventionen in Bezug auf die sprachliche Signalisierung spezifischer Inhaltsknoten deutlich. Tabelle 9.2 stellt die in den Navigationshilfen von HP 1–15 enthaltenen Hyperlinkanzeiger in den ursprünglichen Reihenfolgen dar.[11]

Bei einer Betrachtung der Hyperlinkanzeiger wird die grundlegende und in den 15 Hypertexten in unterschiedlichem Maße ausgeprägte Funktion der privaten bzw. studentischen Homepage deutlich: Die in Abschnitt 4.6.3 diskutierte Selbstdarstellung der Person des Autors manifestiert sich z. B. in der Präsentation von Informationen zu eigenen Interessen und Hobbys wie etwa bestimmten Film- oder Literaturgenres („Science Fiction“), alternativer Rockmusik („New Model Army“) und Haustieren („Berner Sennenhunde-Links“). Durch relevante Hyperlinks oder von den Autoren erstellte Informationsangebote („Windsurfpage“) werden diese Informationen mit zusätzlichen Details versehen. Von besonderer Bedeutung sind biografische Informationen, die in drei Homepages explizit als „Lebenslauf“ verknüpft werden, in drei weiteren Webangeboten wird dieser Link „about me“ bzw. „Über mich“ genannt, zusätzlich existieren die Varianten „Das Ego“, „Persönliche Angaben“ und „Wer ist Tobias Stiepak?“.[12] Möglichkeiten der Kontaktaufnahme werden nur in wenigen zentralen Navigationshilfen aufgeführt; es überwiegen als `mailto:`-Hyperlinks realisierte Verknüpfungen wie z. B. „E-Mail“ oder „Comments“ sowie Gästebücher (in drei Fällen als „Gästebuch“, in einem Fall als „Guestbook“ bezeichnet). Der Hyperlinkanzeiger „Kontakt“ bzw. „Contact“ verweist in drei Homepages auf ein HTML-Dokument mit weiteren Kontaktinformationen. Einen wichtigen Stellenwert besitzt auch die Präsentation von Fotos, die in vier Homepages

[11] Die in Tabelle 9.2 dargestellten Hyperlinkanzeiger wurden den Navigationshilfen in nahezu allen Fällen wortwörtlich entnommen. Selten existiert, wie z. B. in HP 12, keine primäre Navigationshilfe, so dass sich die Hyperlinkanzeiger im Fließtext befinden und daher diesem entnommen wurden. Weitere von den Einstiegsseiten aus verknüpfte Dokumente, deren Hyperlinkanzeiger von den Verfassern isoliert positioniert worden sind, wurden nicht in die Tabelle aufgenommen. Die mit einem Stern markierten Dokumente wurden in der Phase der maschinell durchgeführten Datensammlung (vgl. Abschnitt 7.2) nicht in das Korpus integriert, weshalb diese Zieldokumente – sofern es nicht möglich war, über den Hyperlinkanzeiger auf ihren Inhalt zu schließen – nicht in die Analyse einfließen konnten, da sie zum Zeitpunkt der Durchführung dieser Untersuchung beinahe ausnahmslos nicht mehr auf den ursprünglichen Webservern verfügbar waren.

[12] In Tabelle 9.2 werden diejenigen Hyperlinkanzeiger in Fettdruck dargestellt, die in Bezug auf diese Stichprobe als Konvention der Hypertextsorte aufgefasst werden können.

HP 1	Seitenübersicht (D_1), **Links** (D_2), Musik (D_3), Spiele (D_4), **Bilder** (D_5), Gitarre (D_6), E-Mail (`mailto:`-Link), **Gästebuch** (D_7)
HP 2	**Pics** (D_1), **Linx** (D_2), Badnerlied (D_3), Comments (`mailto:`-Link)
HP 3	**Lebenslauf** (D_1), Mathestudium (D_2), GRIPS (D_3), **Bilder** (D_4), Sprüche (D_5), **Links** (D_6)
HP 4	**Links** (D_1), Freunde (D_2), Projekt (D_3), **about me** (D_4)
HP 5	**About Me** (D_1), **Contact** (D_2), Fotos (D_3), SMS-Versand (D_4^*; externer Link), **Guestbook** (D_5), **Linx** (D_6), Mailingliste (D_7), Uni (D_8), Webcam/Chat (D_9^*; externer Link), WWW-Projekte (D_{10})
HP 6	Medizin (D_1), Science Fiction (D_2^*), New Model Army (D_3^*), Gedichte (D_4^*), Filez'n'Stuff (D_5), The Mad God (D_6), Credits (D_7^*), **Links** (D_8^*)
HP 7	Das Ego (D_1), **Bilder** (D_2), **Kontakt** (D_3)
HP 8	**Lebenslauf** (D_1), Berner Sennenhunde-Links (D_2), Berner Sennenhunde-Buch-Tips (D_3), **Gästebuch** (D_4^*; externer Link), Smileys (D_5), **Links** (D_6)
HP 9	**Links** (D_1^*), Projekte (D_2), Persönliche Angaben (D_3)
HP 10	Wer ist Tobias Stiepak? (D_1), **Links zu anderen coolen Seiten** (D_2), Informationen zur Musik (D_3), Allgemeine Informationen (D_4), Windsurfpage (D_5), Die Luxemburg-Seite (D_6), Informationen aus der Stadt Kuppenheim (D_7), Informationen zum Daimler-Chrysler Snow-Team (D_8)
HP 11	Site Index (D_1), **WWW-Link Index** (D_2), My Bookmarks (D_3), **WWW-Links** (D_4^*), Lehrbücher/-Buchtips (D_5), Mein PGP-Public-Key (D_6), Compliance/Non-Compliance (PPT-Präsentation) (D_7)
HP 12	Australien-Seite (D_1; in den Text eingebetteter Hyperlinkanzeiger: „diesen Link“)
HP 13	—
HP 14	Gedichte/Lyrik (D_1), **Bilder** (D_2), **Über mich** (D_3), **Links** (D_4), **Gästebuch** (D_5^*; externer Link)
HP 15	**Lebenslauf** (D_1), **Photos** (D_2), Info (D_3), **Kontakt** (D_4), Impressum (D_5)

Tabelle 9.2: Die Hyperlinkanzeiger der primären Navigationshilfen

als „Bilder“ und in einem als „Photos“ referenziert werden. Nahezu omnipräsent sind jedoch „Links“, auf die in dieser Form in neun Navigationshilfen Bezug genommen wird. „Linx“ wird – als „Soundalike Slang“ (Haase et al., 1997) – in zwei Fällen verwendet.

Die alleinige Betrachtung der Hyperlinkanzeiger primärer Navigationshilfen ist für eine detaillierte Charakterisierung der konventionalisierten Bestandteile studentischer Homepages nur bedingt geeignet. Daher wurden die Einstiegsseiten sowie diejenigen HTML-Dokumente, auf die von den Navigationshilfen verwiesen wird, in Bezug auf ihre Inhalte und weitere Eigenschaften analysiert. Die Tabellen 9.3 bis 9.5 stellen die Ergebnisse dieser Untersuchung dar und umfassen Gruppen von Merkmalen und ihre Verwendung in der Einstiegsseite (E) oder einem verknüpften Dokument (D_n).[13] Die Ergebnisse gliedern sich in die Bereiche biografische Angaben, Kontaktinformationen, Studienprofil, persönliche Interessen und Hobbys, Navigationshilfen sowie Metainformationen und sonstige Merkmale.

9.5.1 Biografische Angaben und Informationen über den Autor

Die Autoren der Homepages integrieren verschiedene biografische Angaben in ihre Webangebote, wobei jedoch unmittelbar deutlich wird, dass unterschiedliche Auffassungen darüber existieren, wie detailliert die Angaben zur eigenen Person ausfallen sollten: Neben Homepages, die zahlreiche biografische Informationen enthalten (HP 5, HP 8, HP 15), existieren

[13] Die Dokumente D_n beziehen sich auf die in Tabelle 9.2 dargestellten Indizes. Die in HP 5 verwendete Splash-Seite wird als ›S‹ notiert. Die analysierten Merkmale lehnen sich an Bates und Lu (1997), Dillon und Gushrowski (2000) sowie Bittner (2003) an.

	HP 01	HP 02	HP 03	HP 04	HP 05	HP 06	HP 07	HP 08	HP 09	HP 10	HP 11	HP 12	HP 13	HP 14	HP 15	%
Biografische Angaben und Informationen über den Verfasser																
Vor- und Nachname			E, D_1	D_4	S, D_1	E, $D_{1,6}$	D_3	E, $D_{1...6}$	D_3	E, $D_{1...8}$	E, $D_{1,6}$	D_1	E	E, $D_{1...4}$	E, $D_{1...5}$	86,7
Foto des Autors			E, D_1	D_4	D_1	E	D_2	D_1				D_1		E, D_3	D_1	60,0
Geburtsort			D_1	D_4	D_1		D_1	D_1		D_7				D_3	E, D_1	53,3
Alter/Geburtsdatum			D_1	D_4	D_1		D_1	D_1						D_3	D_1	46,7
Tätigkeit/Beruf					D_1		D_1	D_1	D_3						E, D_1	33,3
Wohnort					D_1	E	D_1		D_3						E, D_1	33,3
Spitz- oder Rufname	E				D_3		E							D_1		26,7
Lebenslauf (als Etikett)			D_1					D_1							D_1	20,0
Praktika			D_1					D_1							D_1	20,0
Familienstand				D_4										D_3		13,3
Zivil- bzw. Wehrdienst					D_1			D_1								13,3
Nur Vorname	E															6,7
Konfession														D_3		6,7
Inform. zur Familie							D_1									6,7
Preise, Auszeichnungen			D_1													6,7
Kontaktinformationen																
E-Mail-Adresse	E, D_6	E	E, $D_{1...6}$	D_4	$D_{1,2}$	E, D_1	D_3	D_1	D_3	E	E, $D_{3,6,7}$	E, D_1	E	E, $D_{1...4}$	E, $D_{4,5}$	100,0
Bitte um Kontaktaufn.	$D_{2,6}$		E		$D_{1,8}$	E	D_3	$D_{2,3}$			E	D_1			E	60,0
Gästebuch	D_7				D_5			D_4	D_3					D_5		33,3
Post-/Straßenadresse				D_4				D_1	D_3						D_5	26,7
Telefonnummer				D_4	D_2				D_3							20,0
E-Mail-Formular					D_2										D_4	13,3
ICQ-UID					$D_{1,2}$										D_1	13,3
PGP-Key					D_2						D_6					13,3
Mobiltelefonnummer					D_2											6,7
Pers. Mailing-Liste					D_7											6,7
Faxnummer					D_2											6,7
Hinweis auf Scall											E					6,7
Studienprofil																
Studiengang			E, $D_{1,2}$	D_4	D_1	E	D_1	D_1		D_1	D_1		E	D_3	E, D_1	73,3
Universität			D_1	D_4	D_1	E	D_1	D_1		D_1					E, D_1	53,3
Semester/Studienphase			E		D_1	E	D_1	D_1							D_1	40,0
Studienschwerpunkte			E, D_1		D_1			D_1							E, D_1	26,7
Auslandsaufenthalte			E, D_1									D_1			E, D_1	20,0

Tabelle 9.3: Ergebnisse der Makrostrukturanalyse – Biografische Angaben, Kontaktinformationen und Studienprofil

auch solche, die nur ein einziges (HP 13) oder gar keines dieser Merkmale umfassen (vgl. Tabelle 9.3). Keinerlei biografische Angaben sind in HP 2 enthalten. Der Autor verzichtet sogar darauf, seinen Vor-, Nach- oder Rufnamen zu nennen, sein vollständiger Name kann jedoch der E-Mail-Adresse entnommen werden, die in der Statuszeile des Browsers dargestellt wird, sobald der Mauszeiger über den Hyperlinkanzeiger „Comments" bewegt wird; diese Homepage richtet sich vermutlich in erster Linie an Freunde und Bekannte des Autors, da die Einstiegsseite Grüße an etwa 60 Personen enthält (vgl. Fußnote 30, S. 416). Während der Verfasser von HP 1 lediglich seinen Vornamen angibt, sind in den 13 verbleibenden Homepages (86,7%) jeweils Vor- und Nachname enthalten. Zu den hochfrequenten Merkmalen dieser Kategorie zählt auch ein Foto des Produzenten (60%) sowie sein Geburtsort (53,3%).

In drei Homepages verweist der jeweils erste Hyperlink der primären Navigationshilfe zu HTML-Dokumenten, deren Überschriften sie explizit als „Lebenslauf" ausweisen (HP 3: „Lebenslauf von Carmen Ellsässer", HP 8: „Mein Lebenslauf", HP 15: „LEBENSLAUF"). Diese tabellarischen Lebensläufe ähneln einander in mehreren Aspekten: Sie werden von einem Porträtfoto flankiert und enthalten die obligatorischen Konstituenten dieser Textsorte (z. B. Geburtsdatum, Geburtsort, Adresse, Informationen zur Schul- und Hochschulausbildung, jedoch nur in einem Fall die Anschrift). Zusätzlich werden in den drei Lebensläufen Hyperlinks verwendet, z. B. um die Einstiegsseiten der Webauftritte von Städten, Schulen, Universitäten, Studiengängen oder denjenigen Unternehmen zu verknüpfen, an denen ein Praktikum absolviert wurde. Die Thematisierung von privaten Interessen und Hobbys in Lebensläufen, die für Bewerbungsunterlagen gedacht sind, gilt mittlerweile nicht mehr als Normverstoß. Von den drei Autoren machen nur zwei von der Möglichkeit Gebrauch, Informationen über die persönlichen Interessen und ihre Freizeitgestaltung in den Lebenslauf zu integrieren. In beiden Fällen werden die Angaben in gleicher Weise gestaltet wie z. B. die Auflistungen zur Schulausbildung („Private Aktivitäten" in HP 3, „Andere Aktivitäten" in HP 15). Während der in HP 15 enthaltene Lebenslauf hinsichtlich Inhalt und Gestaltung einem typischen gedruckten Lebenslauf unmittelbar entspricht, können in den beiden anderen Textexemplaren Individualisierungstendenzen festgestellt werden, die im Hinblick auf das gedruckte Pendant als Verstöße gegen die Norm aufgefasst werden können. Da die Veröffentlichung derartiger Lebensläufe in studentischen Homepages jedoch in der Regel nicht mit dem Zweck durchgeführt wird, eine Anstellung zu finden, spielt die Befürchtung von Sanktionen in dieser Kommunikationssituation keine Rolle. Die Verfasserin des in HP 3 enthaltenen Lebenslaufes unterteilt z. B. den Abschnitt „Sprachkenntnisse" in „zuerst die »klassischen« – Englisch [...], Französisch [...] und die »modernen!« – C, Pascal, HTML, AWK". Der Autor des zweiten Lebenslaufes (HP 8) verwendet mehrere grafische Dekorationsobjekte, z. B. horizontale Trennelemente, die die einzelnen Abschnitte teilen oder einen stilisierten Briefkasten, der neben der Anschrift abgebildet ist. Neben den drei explizit als „Lebenslauf" markierten Exemplaren existieren auch mehrere narrative Lebensläufe, in denen die Autoren wesentliche biografische Angaben, Fakten und Ereignisse ihres Lebens in Form eines Fließtextes formulieren. Narrative Lebensläufe werden in mehreren Fällen unter Zuhilfenahme eines Textstrukturmusters gestaltet, das an eine generische E-Mail erinnert (vgl. Abschnitt 9.6.1). In HP 15 existiert sowohl ein tabellarischer als auch ein narrativer Lebenslauf.

Die Verteilung biografischer Angaben erfolgt zwar nicht nach eindeutigen Mustern, doch können verschiedene Korrelationen festgestellt werden. Der Name des Produzenten wird in

der Regel schon in der Einstiegs- oder Splash-Seite genannt und die weiterführenden Informationen werden in ein separates HTML-Dokument integriert (vgl. die Angaben zu HP 4, HP 5, HP 7 und HP 14 in Tabelle 9.3). Weiterhin ist der Name in vielen Fällen in mehreren HTML-Dokumenten einer Homepage enthalten (z. B. in der Fußzeile). Biografische Informationen werden entweder in Form eines narrativen oder tabellarischen Lebenslaufes oder als Liste dargestellt. Die zuletzt genannte Variante dominiert die Stichprobe. Als mögliche Ursache kann der Umstand angeführt werden, dass die unstrukturierte Auflistung derjenigen Fakten, die der Autor im WWW zu publizieren bereit ist, sehr effektiv umgesetzt werden kann, wohingegen die Anfertigung eines narrativen Lebenslaufes mit mehr Aufwand verbunden ist. Dies gilt ebenfalls für den tabellarischen Lebenslauf, dessen Präsenz der Homepage darüber hinaus einen sehr formalen und offiziellen Charakter gäbe, den einige Autoren zu vermeiden versuchen und stattdessen authentische, persönliche und individuelle Themen und Gestaltungsmerkmale präferieren.

9.5.2 Kontaktinformationen

Die zur Gruppe der Kontaktinformationen gehörenden Merkmale beziehen sich auf Möglichkeiten, mit dem Autor in Kontakt zu treten. Eine E-Mail-Adresse wird in allen 15 Homepages angegeben, in einigen Fällen wird sie in mehr als einem, gelegentlich auch in allen HTML-Dokumenten einer Homepage präsentiert (vgl. Tabelle 9.3). Den Grund für die Allgegenwärtigkeit der E-Mail-Adresse liefert der Autor von HP 7: „Da Du Dich ja bereits schon im Internet befindest, ist der einfachste Weg eine kleine Email an mich zu senden: [...]“. Darüber hinaus dokumentiert die Omnipräsenz der E-Mail-Adresse den dominanten Stellenwert der Kontaktfunktion studentischer Homepages (vgl. Bayerl, 2002). Die Adresse wird in der Regel als `mailto:`-Hyperlink mit einem Linkanzeiger wie „Mail“, „Contact“ oder der Adresse selbst realisiert. Weniger häufig werden Icons für diesen Zweck eingesetzt.

In neun Homepages existiert jeweils mindestens ein expliziter Appell zur Kontaktaufnahme (60%).[14] Es liegen drei Varianten vor, die gelegentlich auch in Kombination verwendet werden: Erstens bitten die Produzenten häufig um eine Rückmeldung, falls der Rezipient auf einen nicht mehr existenten Hyperlink treffen sollte (wie z. B. in HP 1: „Falls ein toter Link dabei sein sollte, wäre ich sehr dankbar für eine kurze Benachrichtigung per mail, da ich mittlerweile ein wenig die Übersicht verloren habe :)“; vgl. Fußnote 170, S. 253). Die zweite Variante betrifft das Einsenden von Hinweisen auf andere, relevante Webangebote (z. B. in HP 8: „!!Nicht mehr funktionierende Links oder neue Linkvorschläge bitte melden!!“). Die dritte Bitte um Kontaktaufnahme ist sehr genereller Natur und bezieht sich auf allgemeine Kritik, Fragen und Anregungen (beispielsweise in HP 6: „So, wer mehr wissen will, kann mir ja mal eine Mail schreiben.“, HP 12: „Wie gehabt, wenn irgendwer Vorschläge oder sonstige Anregungen oder Anmerkungen oder gar Lob (kommt vor) loswerden möchte, so wende er oder sie sich doch bitte vertrauensvoll an mich.“, HP 15: „Ich freue mich jederzeit über konstruktive Kritik, Anregungen und sonstiges Feedback. Schreiben Sie mir einfach per E-Mail, die ich auch gerne beantworten werde.“).

[14] Dieses Merkmal wurde der Gruppe der Kontaktinformationen zugeordnet, weil Vorkommen in aller Regel in Kombination mit der Präsentation der E-Mail-Adresse oder einem `mailto:`-Hyperlink auftreten.

Fünf Homepages geben dem Leser die Möglichkeit, einen Gästebucheintrag zu hinterlassen (vgl. Abschnitt 4.6.8). In drei Fällen wird auf die Website eines externen Dienstleisters verwiesen, zwei Produzenten betreiben ihre Gästebücher auf universitären Webservern. Das in HP 9 enthaltene Gästebuch ist in zweierlei Hinsicht unkonventionell: Erstens werden nicht ausschließlich die einzelnen Einträge dargestellt, stattdessen befinden sich in diesem Dokument eingangs die private und die geschäftliche Adresse sowie weitere Kontaktinformationen, anschließend folgen die Gästebucheinträge. Zweitens existiert *keine* Möglichkeit, einen Eintrag zu verfassen. Entweder hat sich der Autor dazu entschieden, zwar dieses HTML-Formular (bzw. den Hyperlink darauf) zu entfernen, nicht jedoch die bereits vorhandenen Einträge, oder es handelt sich nicht um ein genuines Gästebuch: Der Produzent könnte z. B. interessante E-Mails, die ihm als Rückmeldungen zu seiner Homepage zugeschickt wurden, manuell in diese Webseite integriert haben.[15]

Eine Post- bzw. Straßenadresse wird in vier Homepages (26,7%), und die Nummer eines Festnetzanschlusses in lediglich drei Dokumenten aufgeführt (20%). Einerseits belegt dieser Befund, dass die Produzenten das Kommunikationsmedium E-Mail präferieren, andererseits handelt es sich bei der Anschrift um eine Information, die missbraucht werden könnte, so dass auf die Angabe vermutlich sehr bewusst verzichtet wird. Die weiteren in Tabelle 9.3 aufgeführten Möglichkeiten der Kontaktaufnahme – z. B. eine Mobiltelefon- oder Faxnummer oder die ICQ-Benutzernummer – werden in nur einer bzw. zwei Homepages angeboten.

9.5.3 Studienprofil

Ein vollständiges Studienprofil – bestehend aus den Angaben Studiengang, Universität, Semesteranzahl und Studienschwerpunkte – ist in lediglich vier Homepages enthalten (26,7%). Vier Autoren verzichten vollständig auf die Nennung ihres Studiengangs oder der Universität, an der sie studieren.[16] In 11 Homepages (73,3%) wird der Studiengang erwähnt, wobei in 53,3% der Webangebote zusätzlich der Name der Hochschule enthalten ist. Die Semesteranzahl bzw. die aktuelle Studienphase wird in lediglich fünf HTML-Dokumenten angegeben (40%). Auch diese Informationen werden in der Regel in den E-Mail-ähnlichen Text integriert, der sich üblicherweise in der Einstiegsseite des Dokuments befindet.

9.5.4 Hobbys, Interessen, Dienstleistungen und weitere Charakteristika

Hobbys und Freizeitinteressen werden in allen Homepages thematisiert, 73,3% bieten hierzu mindestens Hotlists an (vgl. Tabelle 9.4).[17] Es dominieren Informationen zu den The-

[15] Die visuelle Gestaltung der einzelnen Einträge, die verfügbaren Daten („City:“, „Country:“) sowie die Zeitstempel (Sommer/Herbst 1996) legen die erstgenannte Interpretation nahe.

[16] Da es sich um studentische Homepages handelt, die auf universitären Webservern gepflegt werden, ist die Universität, an der die Verfasser studieren, zusätzlich aus der URL des Webangebots ablesbar. Es ist zu vermuten, dass einige Produzenten aus diesem Grund den Namen ihrer Hochschule nicht explizit in den Text aufgenommen haben. Alternativ könnten sie auch davon ausgehen, dass lediglich ihre Kommilitonen die Website besuchen, weshalb die Angabe der Universität nicht notwendig erscheint.

[17] Der Umstand, dass für HP 9 keine Belege in Tabelle 9.4 aufgeführt werden, liegt darin begründet, dass das Dokument („Links, Links, Links …“) in der Phase der Datensammlung nicht in das Korpus integriert wurde, da seine URL in der Einstiegsseite einen Tippfehler enthält. Wird dieser korrigiert, werden zahlreiche externe Hyperlinks präsentiert, von denen auf die persönlichen Interessen des Autors geschlossen werden kann.

men Musik (spezifische Musikrichtungen, -stile und Künstler), Sport, Computer, Internet, Computer- und Rollenspiele, Kino, Reisen sowie vom Autor verfasste Gedichte. Die Aufstellung der Dokumente, in denen Hobbys und Freizeitaktivitäten besprochen werden, deutet an, dass diese Themen nur in den seltensten Fällen ausführlich vorgestellt werden (z. B. in HP 6 und HP 10). Meist beschränken sich die Verfasser auf die reine Aufzählung ihrer persönlichen Interessen, die durch externe Hyperlinks zusätzlich angereichert wird. Die Nennung der eigenen Interessen erfolgt oftmals innerhalb eines Textes, der simultan der Begrüßung des Rezipienten und der Vorstellung des Produzenten dient (vgl. Abschnitt 9.6.1).

Neben einem Porträt des Autors sind in acht Homepages weitere Fotos enthalten (53,3%). In nur wenigen Fällen dienen diese Bilder ausschließlich als Dekorationselemente. Vornehmlich werden sie in eigens für diesen Zweck angelegten Webseiten präsentiert, um z. B. Freunde vorzustellen (HP 14) oder Reisen zu dokumentieren (HP 5, HP 12). Kommentare oder ergänzende Hinweise zu Fotos werden lediglich in drei Homepages angeboten.

Der Bereich Dienstleistungen und Nützliches ist im Vergleich zu anderen Studien (vgl. Abschnitt 4.6.3) in der Stichprobe unterrepräsentiert.[18] Zwar werden Informationen, die dieser Kategorie – im weitesten Sinne – zugehörig sind, in 12 der 15 Homepages angeboten, doch können sie in den meisten Fällen lediglich als rudimentär bezeichnet werden. Umfangreiche Inhalte, die von den Produzenten selbst erstellt worden sind, befinden sich z. B. in HP 1 (eine Sammlung transkribierter Gitarren-Tabulaturen, die als ASCII-Dateien vorliegen), HP 5 (verschiedene Hausarbeiten), HP 6 (Skripte und Lernmaterialien zum Medizinstudium).

Die studentischen Homepages enthalten weitere Charakteristika, die ebenfalls in Tabelle 9.4 aufgeführt sind. Eine explizite Begrüßung des Lesers erfolgt in allen Webangeboten, was wiederum den von den Autoren präsupponierten Dialogcharakter dieses Mediums reflektiert. Dieser manifestiert sich auch in den hochfrequenten E-Mail-ähnlichen Textstrukturmustern. In sechs Homepages verwenden die Autoren ein Diktum oder ein Epitheton, um ihrer Homepage eine Art Motto – in HP 14 ist explizit vom „Lebensmotto“ des Verfassers die Rede – zu geben und sie hierdurch zusätzlich zu individualisieren.[19]

9.5.5 Navigationshilfen und Hotlists

Der in Tabelle 9.5 zusammengefasste Bereich der Navigationshilfen subsumiert das in der Einstiegsseite enthaltene Hauptmenü (die primäre Navigationshilfe), Listen externer und interner Hyperlinks sowie verschiedene weiterführende Navigationsmöglichkeiten.

Die primäre Navigationshilfe besteht aus einer Sequenz von Hyperlinks, die entweder untereinander (meist am linken Rand) oder nebeneinander (z. B. zentriert im oberen Bereich der Webseite) dargestellt werden. Alternativ oder zusätzlich werden Verweise unmittelbar in den Text integriert, der in der Einstiegsseite enthalten ist; HP 11 verwendet ausschließlich diese zuletzt dargestellte Option. Die alleinige Präsentation der zentralen Hyperlinks nebeneinander findet in zwei Homepages Verwendung. Eine am linken Rand dargestellte primäre

[18] Hyperlinks zu externen Websites werden (anders als z. B. bei Bayerl, 2002) nicht zu dieser Kategorie gezählt. Es wurden diejenigen Angebote notiert, die physikalisch innerhalb der jeweiligen studentischen Homepage gespeichert sind und von dem jeweiligen Autor selbst angefertigt wurden.

[19] Zu Beginn der Einstiegsseite von HP 6 ist eine Strophe eines Liedes von „New Model Army“ aufgeführt, in HP 8 befindet sich eine „Weissagung“, die „vermutlich von den Cree Indianern [stammt]“ und in HP 9 hat der Verfasser einen „Tip [sic] des Tages“ integriert: „Sehen ist gut, aber klicken ist besser . . .“

	HP 01	HP 02	HP 03	HP 04	HP 05	HP 06	HP 07	HP 08	HP 09	HP 10	HP 11	HP 12	HP 13	HP 14	HP 15	%
Hobbys des Autors																
Hyperlinks zu Hobbys	D_2	D_2	$D_{1,3,6}$	D_3	$D_{1,6}$			$D_{2,3}$		$D_{2,3}$	D_2		E	D_4	E	73,3
Musik	E, D_3, 6			D_3	D_1	D_3	D_1			D_3				D_3	E	53,3
Sport				D_4	D_1		D_1			D_5			E		E, D_1	40,0
Computer/Internet					D_1	D_5	D_1			D_4				D_3	D_1	40,0
Computerspiele	D_4				D_1	D_5										20,0
Rollenspiele					D_1	$D_{5,6}$							E			20,0
Kino							D_1								E	13,3
Reisen				D_4								D_1				13,3
Vom Autor verfasste Gedichte						D_4								D_1		13,3
WWW-Projekte des Autors				D_3	D_{10}											13,3
Literatur								D_3								6,7
Sonstige Hobbys			D_1		D_1			E, $D_{2,3}$		$D_{6,8}$			E		E	40,0
Fotos																
Ohne Kommentare	D_5	D_1	D_4		$D_{1,3}$					$D_{6,7}$						33,3
Mit Kommentaren								D_2						D_2	D_2	20,0
Dienstleistungen/Nützliches																
Informationen zu einer Region		D_3								D_7		D_1				20,0
Liste der Homepages von Freunden			$D_{3,6}$	D_2										D_4		20,0
Hausarbeiten/Papiere					D_8	D_1										13,3
Studentische Vereinigungen			D_3												D_1	13,3
Foliensätze											D_7					6,7
Materialien/Skripte zum Studium						D_1										6,7
Weiteres Lernmaterial						D_1										6,7
Sonstiges	$D_{4,6}$					D_5		D_5		D_6						26,7
Weitere Charakteristika																
Explizite Begrüßung	E	E	E	D_4	S, D_1	E	E	E	E	D_3	E	E, D_1	E	E	E	100,0
E-Mail-ähnliche Textstruktur	E	E	E	D_4	D_1	E	E	E	E		E	E, D_1	E	E, D_3	E	93,3
Explizite Verabschiedung	E				$D_{1,3}$	E	E	E	E		E	E, D_1	E	E	E	73,3
Zitat/Motto					E	E		E	D_3		E			D_3		40,0
Grüße an Freunde und Bekannte		E								E, D_3						13,3
Tagebuch														E		6,7
Sprüchesammlung			D_5													6,7

Tabelle 9.4: Ergebnisse der Makrostrukturanalyse – Inhalte

	HP 01	HP 02	HP 03	HP 04	HP 05	HP 06	HP 07	HP 08	HP 09	HP 10	HP 11	HP 12	HP 13	HP 14	HP 15	%
Navigationshilfen																
Primäre Navigationshilfe																
Links vertikal		E	E	E	E	E, $D_{1,6}$	E	E	E	E		D_1			E, $D_{1...5}$	73,3
Links horizontal	E								E					E, $D_{1...4}$		20,0
In Text eingebettet											E	E			E, D_1	20,0
Hotlist																
Thematisch sortiert	D_2		$D_{2,3,4}$		$D_{6,8}$	D_8	$D_{2,3,6}$	$D_{1,2}$	D_2	$D_{3,4}$			D_4			60,0
Unsortiert		D_2	D_6	D_3	S				E		E, D_2			D_4		46,7
Liste interner Hyperlinks																
Thematisch sortiert	$D_{3,4,6}$		$D_{3,4}$	D_1		$D_{5,6}$						D_1				33,3
Unsortiert	D_4	D_1			D_8											20,0
„Zurück zur Homepage“	$D_{2...6}$	$D_{1,2,3}$	$D_{1...6}$	$D_{1...4}$						$D_{1...8}$	$D_{1...7}$					40,0
Seiteninterne Navigation	D_2				D_6									D_1		20,0
„Zum Seitenanfang“	D_2										E, D_2			D_1		20,0
Sitemap	D_1										D_1					13,3
Metainformationen und Sonstiges																
Grafische Dekorationen	E	E	E, $D_{2,6}$	E, $D_{1,3}$	E, $D_{1...10}$	E, $D_{1,6}$	$D_{1,3}$	E, $D_{1...6}$	E, D_3		E, $D_{1,2,6}$	D_1	E	E, $D_{1...4}$	E, D_1	93,3
Prophyl. Beschwichtigung	D_2	E		D_4	$D_{7,8}$		E	$D_{5,6}$		E, D_8	E, D_5	E, D_1	E	D_2		73,3
Datum der letzten Änderung					$D_{5...8}$	E		E, $D_{1...6}$	E	E	E, D_3	D_1	E	E, $D_{1...4}$	E, $D_{1...5}$	66,7
Copyright-Hinweis						E, $D_{1,6}$		E, $D_{1,2,5,6}$		E	D_7	D_1	E	E, $D_{1...4}$	D_5	53,3
Zugriffszähler				E	S, D_9	D_1		E, $D_{1...5}$	E	E						40,0
Einsatz einer Fußzeile			$D_{1...6}$			E		E, $D_{1...6}$							E, $D_{1...5}$	26,7
„Optimiert für . . .“	E				S	E										20,0
Angabe einer Versionsnummer		E			D_1		E									20,0
Haftungsausschluss					S	$D_{1,5}$									D_5	20,0
Datum der Erstellung					D_3									E, $D_{1...4}$		13,3
Benutzungshinweise										D_4					E, $D_{2,3}$	13,3
„under construction“		E		D_4												13,3
Liste der Änderungen	E, D_4											D_1				13,3
Suchmaske (interne Suchmaschine)					D_{10}											6,7
Suchmaske (externe Suchmaschine)					S, $D_{6,8}$											6,7

Tabelle 9.5: Ergebnisse der Makrostrukturanalyse – Navigationshilfen, Metainformationen und Sonstiges

Navigationshilfe wird in 11 Homepages benutzt und kann somit als Konvention bezeichnet werden (vgl. Abschnitt 4.6.2). Die Hyperlinkanzeiger werden jeweils in etwa der Hälfte dieser Navigationshilfen entweder als Icons oder in reiner Textform dargestellt.

In der Stichprobe enthalten mindestens 24 HTML-Dokumente mehr als 70 Listen externer Hyperlinks – nur die Autoren von HP 12 und HP 15 verzichten auf Hotlists.[20] Die meisten Verfasser sortieren die Hyperlinks nach thematischen Gesichtspunkten (60%), unsortierte Listen können in 46,7% der Homepages belegt werden. Zur Strukturierung von Verweisen auf interne HTML-Dokumente werden Hyperlinklisten weniger häufig eingesetzt.[21]

Ein Grundsatz des Webdesigns bezüglich eingebetteter Dokumente betrifft die Bereitstellung von Verweisen zur Einstiegsseite. Derartige, oftmals „zurück zur Homepage“ genannte Hyperlinks werden in 40% der Webangebote konsequent eingesetzt. Seiteninterne Navigationshilfen umfassen Hyperlinks, die auf Zielanker verweisen, die sich im gleichen HTML-Dokument befinden und werden zur Strukturierung sehr umfangreicher Einzeldokumente in drei Homepages eingesetzt (20%), wobei in zwei Fällen entsprechende Rückverweise zum Seitenanfang zur Verfügung gestellt werden. Sitemaps, die jedoch eher verschachtelten Listen als grafischen Karten ähneln, werden ebenfalls in zwei Homepages angeboten.

9.5.6 Metainformationen und sonstige Merkmale

Der abschließend betrachtete Bereich bezieht sich vornehmlich auf zusätzliche Informationen über die Homepage, die der Produzent üblicherweise in der Einstiegsseite zur Verfügung stellt, sowie auf einige Struktur- und Gestaltungscharakteristika. Grafische Dekorationselemente (Hintergrundmuster, animierte Icons, grafische Trennlinien etc.) werden zwar in 14 der 15 studentischen Homepages verwendet, in einigen Fällen bezieht sich ihr Einsatz jedoch lediglich auf die Einstiegsseite (z. B. HP 1) und einige weitere Dokumente (HP 3, HP 6). Eine konsistente grafische Gestaltung besitzen HP 5, HP 8 und HP 14.

Besonders auffällig sind zahlreiche metadiskursive Äußerungen, die als prophylaktische Beschwichtigungsfloskeln und allgemeine Absichtserklärungen eingesetzt werden (vgl. Abschnitt 9.4.1). Mit ihnen verfolgen die Produzenten den Zweck, dem möglicherweise von einer Homepage inhaltlich oder funktional enttäuschten Rezipienten zu verdeutlichen, dass dem Verfasser diese Probleme bewusst sind und er darüber hinaus bestrebt ist, schon in naher Zukunft zahlreiche Modifikationen und Erweiterungen vorzunehmen; der Rezipient soll zu einem wiederholten Besuch der Homepage eingeladen werden. Vollständig ausgeprägte prophylaktische Beschwichtigungsfloskeln befinden sich z. B. in HP 8 („Sorry, hier fehlen leider noch die Links! Aber ich arbeite daran auf Hochtouren! Versprochen!“), HP 11 („So leider ist es noch nicht ganz so weit. . .Also auch hir [sic] noch ein bischen [sic] Geduld, bis ich die Tipparbeit erlegigt [sic] habe.“) und in HP 13 („Noch befindet sich diese Seite im Aufbau, aber sie wird einige interessante Themen enthalten.“). Eine Variante bezieht sich auf die Thematisierung inhaltlicher oder funktionaler Probleme, z. B. in HP 2 („Noch nicht alles

[20] Diese Feststellung bedeutet nicht, dass diese Homepages keinerlei externe Hyperlinks enthalten. Die Verweise zu externen Angeboten werden jedoch nicht in Listenform, d. h. als Hotlists präsentiert (vgl. Abschnitt 4.6.6), sondern z. B. in den Fließtext oder den bereits angesprochenen Lebenslauf in HP 15 eingebettet, d. h. das Anbieten externer Ressourcen ist lediglich eine sekundäre Funktion eben dieser Komponenten.

[21] Die primären Navigationshilfen sind ebenfalls Listen bzw. Sequenzen interner Hyperlinks, die jedoch in diese Statistik nicht eingeflossen sind.

funktioniert (ich arbeite dran) – deshalb einfach ausprobieren …“), HP 11 (vor einem defekten Pull-Down-Menü befindet sich der Hinweis: „funftioniert [sic] leider noch nicht ganz so wie ich mir das vorstelle…“) und HP 14 („Es könnte etwas dauern, bis alle Bilder geladen sind […], aber ich denke es lohnt sich.“). Auf den Aspekt der Ladegeschwindigkeit bezieht sich auch der Autor von HP 12, der seine überarbeitete Homepage einleitet mit: „Alles neuer, grösser [sic], schöner, besser (sowieso) und wahrscheinlich langsamer.“

9.5.7 Fazit

Im Hinblick auf die zentralen Charakteristika privater Homepages bestätigt diese Analyse die in Abschnitt 4.6.3 vorgestellten Untersuchungen von Bates und Lu (1997), Dillon und Gushrowski (2000), Bayerl (2002) und Bittner (2003). Die Unterschiede beziehen sich auf die Frequenzen von Begrüßungen (100%, Dillon und Gushrowski: 67%, Bayerl: 58%), Hobbys (93%, Bates und Lu: 31%, Bayerl: 63%, Bittner: 68%), Hotlists (87%, Bates und Lu: 56%, Dillon und Gushrowski: 68%), Appelle zur Kontaktaufnahme (60%, Bayerl: 90%) sowie den Einsatz von Fußzeilen (27%, Bayerl: 88%). Allein aufgrund dieser Angaben können HP 1–15 als vordergründig dialogische (Begrüßungen), de facto jedoch monologische (Bitte um Kontaktaufnahme) und die persönlichen Interessen des Produzenten fokussierende (Hobbys) Homepages charakterisiert werden. Die hochfrequenten Hotlists werden zur Kompensierung des Umstandes eingesetzt, dass der Produzent sich nicht in der Lage sieht, selbst erstellte Inhalte in seine Homepage zu integrieren (vgl. Fußnote 6, S. 390).

Bayerl (2002) ermittelt für ihre Stichprobe von 40 Homepages die Themenbereiche Selbstdarstellung (mit den drei Unterbereichen Person und privates Umfeld, Freizeitaktivitäten und Rolle als Student), Kontakt zum Leser (Begrüßungen und Verabschiedungen, Aufforderung zur Kontaktaufnahme, Gästebücher), Dienstleistungen (vom Autor verfasste Texte oder externe Hyperlinks), Unterhaltung (Witze, Sammlungen von Sprüchen, ironische Kommentare) und Metainformation (Stand der Homepage). Für die meisten dieser Bereiche können zwar ähnliche Frequenzangaben ermittelt werden (vgl. Tabelle 9.6), doch existieren verschiedene Ausnahmen.[22] Diejenigen Inhalte, die Bayerl (2002) unter dem Schlagwort „Unterhaltung“ subsumiert, können in HP 1–15 nur in marginaler Ausprägung nachgewiesen werden. Witze und ironische Kommentare sind in den Homepages nicht enthalten, die Verfasserin von HP 3 bietet eine Liste der „Sprüche, die andere Leute unter ihre E-Mails setzten“ an.[23] Eine Diskrepanz besteht auch im Hinblick auf die Merkmale, die in Tabelle 9.6 als Studienprofil bezeichnet werden (80%). Bayerl (2002) zufolge enthalten lediglich 23% der von ihr untersuchten Webangebote „studienrelevante Angebote“, wobei jedoch anzumerken ist, dass Bayerl hierunter vornehmlich externe Hyperlinks zu Hochschulen, Organisationseinheiten, Forschungseinrichtungen und Sammlungen von Lehr- und Lernmaterialien versteht.

[22] Die in Tabelle 9.6 angegebenen prozentualen Anteile für die Arbeit von Dillon und Gushrowski (2000) beziehen sich auf die Nennungen der Probanden (vgl. Tabelle 4.10, S. 221). Tabelle 9.6 fasst die 55 häufigsten Merkmale der Tabellen 9.3 bis 9.5 zusammen. Die als „allgemein“ markierten Merkmale beziehen sich auf beliebige Vorkommen einzelner Charakteristika inhaltlicher Gruppen. Die in Fettschrift dargestellten Angaben stellen Korrespondenzen bezüglich der Ergebnisse anderer Studien mit einer Differenz von $\pm$ 10 dar.

[23] Bayerl (2002, S. 54) fasst auch „Links zu Spielen“ als „Elemente, die allein der Unterhaltung dienen“ auf. Diese werden in Tabelle 9.4 dem Merkmal „Hobby: Computerspiele“ zugerechnet. Bayerl (2002) gibt an, dass 22,5% der von ihr untersuchten Homepages unterhaltende Elemente besitzen.

	Merkmal	%	Bates und Lu (1997)	Dillon und Gushrowski (2000)	Bayerl (2002)	Bittner (2003)
1.	E-Mail-Adresse	**100**	**92**	86		**100**
2.	Begrüßung des Lesers	100		67	58	
3.	Biografische Angaben (allgemein)	93				
4.	E-Mail-ähnliche Textstruktur	93				
5.	Grafische Dekorationselemente	93	79			
6.	Hobbys (allgemein)	93	31		63	68
7.	Name (allgemein)	**93**	79		**88**	**100**
8.	Vor- und Nachname	**87**			**78**	
9.	Primäre Navigationshilfe (typografisch abgesetzt)	87				
10.	Hotlist	**87**	56	**68**		
11.	Dienstleistungen/Nützliches (allgemein)	80				
12.	Kontaktinformationen (außer E-Mail-Adresse)	80				
13.	Studienprofil (allgemein)	80	32		23	
14.	Navigationshilfe am linken Rand	73				
15.	Hyperlinks zu Hobbys	73				
16.	Prophylaktische Beschwichtigungsfloskel	73			23	
17.	Studiengang	73				
18.	Verabschiedung	73				
19.	Datum der letzten Änderung	**67**	43	84	**73**	
20.	Bitte um Kontaktaufnahme	**60**			90	**68**
21.	Foto des Autors	60				72
22.	Copyright-Hinweis	53	13		20	
23.	Fotos (allgemein)	**53**	35	**56**		
24.	Geburtsort	53				
25.	Hobby: Musik	53				
26.	Angabe der Universität	53				
27.	Alter/Geburtsdatum	47	11			64
28.	Hobby: Sport	40				
29.	Hobby: Computer/Internet	40				
30.	Semester/Studienphase	40				
31.	Diktum/Zitat/Motto	40				
32.	Zugriffszähler	**40**	53	4	**30**	
33.	„Zurück zur Homepage“-Verweis	40				
34.	Tätigkeit/Beruf	**33**	**33**			64
35.	Gästebuch	33	18	19	10	
36.	Hotlist (thematisch sortiert)	33				
37.	Wohnort	33				72
38.	Einsatz einer Fußzeile	27			88	
39.	Post- oder Straßenadresse	**27**	**19**			
40.	Spitz- oder Rufname	27	15		5	
41.	Studienschwerpunkte	27	4			
42.	Angabe einer Versionsnummer	20				
43.	Auslandsaufenthalte	20				
44.	Haftungsausschluss	**20**	4		**10**	
45.	Navigationshilfe (Hyperlinks nebeneinander)	20				
46.	Navigationshilfe (Hyperlinks in Text integriert)	20				
47.	Hobby: Computerspiele	20				
48.	Hobby: Rollenspiele	20				
49.	Informationen zu einer Region	20	9			
50.	Lebenslauf (explizites Etikett)	**20**	**11**	56	5	
51.	Liste der Homepages von Freunden	**20**	7		**25**	
52.	„Optimiert für …“-Hinweis	**20**			**12**	
53.	Praktika/vergangene Beschäftigungen	**20**	**19**			
54.	Seiteninterne Navigationshilfen	20				
55.	Telefonnummer	**20**	**16**			**28**

Tabelle 9.6: Ergebnisse der Makrostruktur- und Inhaltsanalyse im Vergleich

Auch Bittner (2003) führt Gruppen von Kerninhalten an, die in den von ihm untersuchten 25 privaten Homepages maßgeblich enthalten sind (persönliche Informationen, Interessen und Hobbys, Berufliches, Nützlichkeit und Kommunikationsplattform). Bereits diese Aufstellung zeigt den wesentlichen Unterschied zwischen arbiträren privaten Homepages und studentischen Homepages auf: Angaben zum Beruf des Verfassers werden in letztgenannten durch die Angabe studienbezogener Informationen ersetzt, so dass zur Charakterisierung *beliebiger* privater Homepages nicht eine einschränkende Kategorie wie z. B. „Beruf“ oder „Berufliches“ sondern vielmehr „Tätigkeit“ verwendet werden sollte, um sowohl den Status einer Person als Student als auch eine möglicherweise ausgeübte berufliche Tätigkeit (z. B. als studentische Hilfskraft) subsumieren zu können. Dennoch stellt sich nun die Frage, ob studentische Homepages als eigenständige subgenerische Varietäten – als Hypertextsortenvarianten – der privaten Homepage konzeptualisiert werden können. Abschnitt 9.7 geht genauer auf dieses Thema ein.

9.6 Weitere Besonderheiten

Nachfolgend werden weitere sehr auffällige Merkmale in den untersuchten studentischen Homepages diskutiert, die sich auf die Verwendung E-Mail-ähnlicher Textstrukturmuster (Abschnitt 9.6.1), Grüße an Freunde und Bekannte (Abschnitt 9.6.2) und heterogene Formen der Leseranrede beziehen (Abschnitt 9.6.3).

9.6.1 E-Mail-ähnliche Textstrukturmuster

Die quantitative Analyse hat verdeutlicht, dass viele der Einstiegsseiten Begrüßungen und Verabschiedungen enthalten, es kann somit in Bezug auf die Textstruktur eine Ähnlichkeit zur E-Mail-Kommunikation festgestellt werden (vgl. Abschnitt 8.7). Auch de Saint-Georges (1998, S. 72) weist darauf hin, dass einige der von ihr untersuchten persönlichen Homepages hinsichtlich des Textaufbaus – initiale und informelle Begrüßung, Text, Verabschiedung – eher einem Brief als einem publizierten Text ähneln. Das Textstrukturmuster einer generischen E-Mail besteht aus einer Begrüßung, dem eigentlichen Text, einer Verabschiedung, einer Signatur (d. h. dem Vor- und eventuell Nachnamen des Autors)[24] und einem Postscriptum; lediglich der Text stellt hierbei eine obligatorische Konstituente der Makrostruktur dar, alle anderen Teile sind prinzipiell optional.[25]

Nachfolgend werden die Vorkommen dieser Komponenten in der Stichprobe diskutiert, in der Regel können sie unmittelbar innerhalb der Einstiegsseite nachgewiesen werden; in vier Fällen befinden sie sich in einem von dort erreichbaren HTML-Dokument. Abbildung 9.1 zeigt die Einstiegsseite von HP 8, die ein sehr prägnantes Beispiel für einen Textbaufbau darstellt, der eindeutig als eine – vermutlich nicht bewusst realisierte – Adaption des generischen

[24] Mit diesem Begriff ist nicht die aus der E-Mail-Kommunikation bekannte *signature* gemeint, bei der es sich um eine Datei handelt, die z. B. Kontaktinformationen enthält und vom E-Mail-Client automatisch an eine neue Mail angefügt wird.

[25] Es ist zu beachten, dass sich dieses Textstrukturmuster auf eine sehr abstrakte Ebene bezieht. Da innerhalb der E-Mail-Kommunikation unterschiedliche Textsorten identifiziert werden können (vgl. Orlikowski und Yates, 1994, Ziegler, 2002, Fußnote 68, S. 59, sowie Abschnitt 4.2.4), existieren somit auch spezifischere und konventionalisierte Ausprägungen dieses generischen Musters.

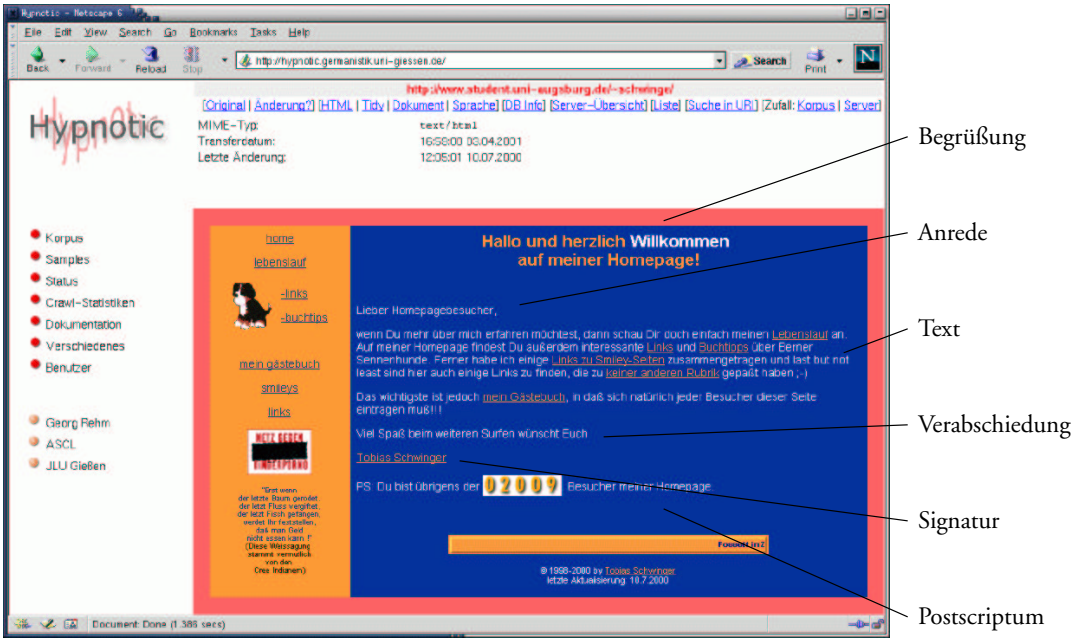


Abbildung 9.1: E-Mail-ähnliche Textstruktur in einer studentischen Homepage

Aufbaus einer E-Mail aufzufassen ist. Tabelle 9.7 stellt die Analyseergebnisse und zusätzlich die Vorkommen von Grüßen an Freunde und Bekannte (Abschnitt 9.6.2) sowie die unterschiedlichen Formen der Leseranrede (Abschnitt 9.6.3) im Überblick dar.

Die Begrüßung

In 14 der 15 Homepages (93,3%) wird der Leser mit einer Begrüßungsfloskel willkommen geheißen. Hierzu gehören unter anderem „Willkommen auf meiner neuen alten Homepage.“ (HP 1), „WILLKOMMEN . . . endlich: meine Homepage ist da!!“ (HP 2), „Guten Tag . . . Herzlich Willkommen auf meiner Internetseite!“ (HP 3), „Herzlich Willkommen auf meiner Homepage!“ (HP 6, HP 11), „Halloechen und Willkommen auf meiner Homepage !!!“ (HP 7), „Hallo und herzlich willkommen“ (HP 8, HP 12), „Willkommen auf der Homepage von Martin S. Vogel“ (HP 13) und „Willkommen in Jan’s Home“ (HP 14). Auffallend ist die Ähnlichkeit der einzelnen Begrüßungen: Nach der eigentlichen Begrüßung („Guten Tag“, „Hallo“, „Halloechen“), die an der Initialposition in sechs der 15 Homepages enthalten ist, folgt ein expliziter Willkommensgruß (11 Homepages, vgl. de Saint-Georges, 1998). Jeweils zwei Formen der Begrüßung werden in zwei studentischen Homepages verwendet.

Die Anrede

Zusätzlich zu der Begrüßung befindet sich in HP 8, den eigentlichen Text einleitend, mit „Lieber Homepagebesucher“ eine direkte Anrede des Rezipienten (vgl. Abbildung 9.1). Auch HP 10 enthält eine solche direkte Anrede („Reisender – Du bist auf der Homepage von Tobias

	Begrüßung	Anrede	Text	Verabschiedung	Signatur	P. S.	Grüße	*Du*, *Sie*, ...
HP 1	✓	—	✓	✓	✓	—	—	passiv
HP 2	✓	—	✓	—	—	—	✓	passiv
HP 3	✓	—	✓	—	—	—	—	passiv
HP 4	✓	—	✓	—	—	—	—	*Ihr*, *Euch*
HP 5	✓	—	✓	✓	✓	—	—	passiv
HP 6	✓	—	✓	✓	—	—	—	passiv
HP 7	✓	—	✓	✓	✓	—	—	*Du*
HP 8	✓	✓	✓	✓	✓	✓	—	gemischt
HP 9	✓	—	—	✓	✓	—	—	passiv
HP 10	—	✓	—	✓	—	—	✓	passiv
HP 11	✓	—	✓	—	—	✓	—	gemischt
HP 12	✓	—	✓	—	✓	—	—	*Ihr*
HP 13	✓	—	✓	✓	✓	—	—	*Ihr*, *Euer*
HP 14	✓	—	✓	—	✓	—	—	*Ihr*, *Euer*
HP 15	✓	—	✓	✓	✓	—	—	*Sie*
Anteil	93,3%	13,3%	86,7%	60%	60%	13,3%	13,3%	

Tabelle 9.7: Merkmale E-Mail-ähnlicher Textstrukturmuster in der Stichprobe

Stiepak gelandet."); auf eine Begrüßungsfloskel wird in diesem Fall jedoch verzichtet. Gerade diese direkte Form der Anrede des Lesers in Verbindung mit einer separaten Begrüßungsfloskel tritt in HTML-Dokumenten anderer Hypertextsorten nur sehr selten in Erscheinung und ist äußerst signifikant für das Textstrukturmuster einer generischen E-Mail.

Die Verabschiedung

Verabschiedungen im weitesten Sinne sind in neun der 15 Homepages (60%) enthalten: „Viel Spaß damit!" (HP 1), „cU" (HP 5), „Ansonsten noch viel Spaß beim Surfen!" (HP 6), „In diesem Sinne: Viel Spaß im Netz !!!!" (HP 7), „Viel Spaß beim weiteren Surfen wünscht Euch" (HP 8), „Also dann noch viel Spaß!" (HP 9), „Bis morgen." (HP 10), „Noch einen schönen Tag, und bis bald im Netz der tausend Möglichkeiten!" (HP 13) und „Ansonsten wünsche ich noch viel Spaß auf meiner Homepage und beim Surfen im Web" (HP 15).[26] Die Verabschiedungsfloskeln fallen im Hinblick auf ihre Varietät heterogener aus als die Begrüßungen (vgl. Abschnitt 8.7). Neben expliziten Verabschiedungen werden auch Formulierungen verwendet, die als implizite Verabschiedungen fungieren und eine Kontaktaufnahme zum Ziel haben, z. B. „Ach übriges [sic], wie jeder andere, freue auch ich mich über eine Mail", wobei das Wort „Mail" als `mailto:`-Hyperlink mit der E-Mail-Adresse der Autorin verknüpft wurde (HP 3), „So bleibt mir noch zu sagen das [sic] ich Euch schnellen Seitenaufbau wünsche, und habt ihr Kummer oder Sorgen, schreibt gleich morgen, an den Jens, for my best friends..." (HP 4), „Tja, mehr fällt mir so auf die Schnelle nicht ein, Feedback ist jedoch jederzeit willkommen und exxtrem [sic] gern gesehen!" (HP 5), „So, wer mehr wissen will, kann mir ja mal eine Mail schreiben. Meine email-Adresse ist: [...]" (HP 6), „Und falls Ihr nun mich irgendwie erreichen wollt, so könnt Ihr das auch machen, und zwar so: [...]" (HP 11) und „Bis dahin könnt ihr mir ja schreiben." (HP 12). Insgesamt deuten die Verabschiedungen und die

[26] Besonders interessant ist die Verabschiedungsfloskel „Bis morgen" (HP 10). In der Einstiegsseite, in der sich auch diese Floskel befindet, gibt der Produzent den Hinweis „Die Seite ist ständig im Umbau, jeden Tag neue Uploads !!!" Offenbar geht der Autor davon aus, dass seine Freunde die Homepage mindestens täglich rezipieren. Unmittelbar vor der Verabschiedungsfloskel richtet sich der Produzent an einen Teil des von ihm intendierten Rezipientenkreises: „Spezielle Grüsse [sic] an Astrid und Erik. Bleibt sauber ;-)".

hochfrequente Floskel „Viel Spaß“ darauf hin, dass die Produzenten der in der Stichprobe enthaltenen Dokumente ihre eigenen Homepages als Unterhaltungsangebote verstehen: Der Leser gelangt ohne spezifische Intention, möglicherweise zum Zwecke des unterhaltenden Zeitvertreibs auf die Homepage, rezipiert einen kleineren oder größeren Teil des Angebots und verlässt die Website anschließend über eine der in zahlreichen Listen von Hyperlinks präsentierten Verknüpfungen. Zusätzlich wird deutlich, dass zumindest einige Produzenten ihre Angebote nicht ernst nehmen und sich nicht vorstellen können, dass andere Personen an ihren Homepages ein Interesse besitzen könnten: „Ich weis [sic] zwar nicht wie Du hierher gekommen bist, aber ist ja auch egal.“ (HP 7).

Der innerhalb der E-Mail-ähnlichen Textstruktur präsentierte Text

Neben den bislang vorgestellten Konstituenten wurde auch der Text untersucht, der sich zwischen Begrüßungs- und Verabschiedungsfloskel befindet. Hierbei handelt es sich um den eigentlichen Textkörper des jeweiligen an die E-Mail angelehnten Textstrukturmusters. Zwei Homepages enthalten unter diesen Gesichtspunkten *keinen* Text. Besonders hervorzuheben ist HP 9: Zu Beginn der Einstiegsseite befindet sich vor drei thematischen Kategorien („Links, Links, Links …“, „Projekte“, „persönliche Angaben“) in großer Schrift lediglich eine Begrüßung, eine Verabschiedung und die aus einem Vornamen bestehende Signatur:

> Hallo ,
> herzlich willkommen auf meinen WWW-Seiten.
> Also dann noch viel Spaß!
> Olaf [HP 9]

Die 13 Dokumente, in denen auf Grundlage der eingangs vorgestellten Terminologie Text enthalten ist, umfassen vornehmlich Abschnitte von Fließtext. Die meisten Texte enthalten einen einzelnen Absatz, der im Extremfall aus nur einem Satz besteht wie z. B. in HP 2: „Noch nicht alles funktioniert (ich arbeite dran) – deshalb einfach ausprobieren …“. HP 7 und HP 15 enthalten jeweils sieben Absätze (Durchschnitt: 3,1). HP 7 besteht primär aus metadiskursiven Äußerungen, wohingegen HP 15 einen vollständig durchkomponierten Text darstellt, der als narrativ gestalteter Lebenslauf bezeichnet werden kann. Neben Abschnitten von Fließtext enthalten HP 2, HP 3 und HP 15 in den Text eingebettete Listen, die vier (HP 2, HP 15) bzw. sechs (HP 3) Hyperlinks umfassen. Der Autor von HP 15 fügt zusätzlich kurze Kommentare an, die über den Inhalt des Linkziels informieren.

Zugleich enthält HP 15 den einzigen Text, der *keine* Merkmale für konzeptionelle Mündlichkeit aufweist. Zu diesen Merkmalen (vgl. auch die Abschnitte 4.6.3 und 8.3) gehören z. B. zahlreiche Ellipsen und Auslassungspunkte, die eine unvollständige Textplanung und spontane Textformulierung andeuten (HP 1, HP 2, HP 4, HP 5). Diese Feststellung wird durch zahlreiche orthografische Fehler zusätzlich gestützt. Darüber hinaus werden reduplizierte Interpunktionszeichen, typografisch markierte Verbstämme und Smileys verwendet (HP 8, HP 14, „klick’ mal auf die Pix. :-)“, HP 5, „*hüstel*“, HP 7). Die Lexik ist nur marginal elaboriert und umgangssprachlich geprägt („super Sachen“, „eimerweise Links“, HP 1, „tja“, HP 5, HP 13), was in gleicher Weise für den Satzbau festzustellen ist: „ähmm wie peinlich!“ (HP 11), „Mal schauen wo ich rauskomme.“ (HP 7), „Ich weis [sic] zwar nicht wie Du hierher

gekommen bist, aber ist ja auch egal. Auf alle Faelle bist Du auf Tobi's Homepage gelandet" (HP 7). In beinahe allen Dokumenten kann ein Bemühen um Kreativität bei der Textgestaltung festgestellt werden, z. B. schließt der Autor von HP 4 seinen Text mit einem Reim: „habt ihr Kummer oder Sorgen, schreibt gleich morgen, an den Jens, for my best friends . . .". Dieser unmittelbare Adressatenbezug als Form der „starke[n] Personifizierung" (Eckkrammer, 2001, S. 55) ist ebenfalls eine Eigenschaft, die in allen Texten beobachtet werden kann.

Neben den Merkmalen für konzeptionelle Mündlichkeit wurde eine Analyse der Inhalte und Themen durchgeführt, die in den Texten diskutiert werden. Tabelle 9.8 stellt die Ergebnisse dar; bezüglich der in der linken Spalte aufgeführten Inhalts- bzw. Themenbereiche wurde eine bestmögliche Abdeckung angestrebt.[27] Die Texte lassen sich grob in zwei Klassen einteilen, die einerseits primär metadiskursive Äußerungen und andererseits vornehmlich biografische Angaben umfassen. Es ist jedoch zu betonen, dass keine trennscharfe Grenze gezogen werden kann, weshalb nachfolgend zunächst die zentralen Inhaltsbereiche und daraufhin spezifische Textcharakteristika – hierzu zählen auch die erwähnten metadiskursiven Äußerungen – diskutiert werden. Sechs der Texte werden vom Autor durch die initiale Nennung seines Namens eingeleitet, der in vier Fällen ebenfalls in Form einer abschließenden Signatur wiederholt wird. In drei Texten wird lediglich der Vorname („Hallo, ich bin der Jens.", HP 4), in drei weiteren Texten werden Vor- und Nachname verwendet („Gestatten, mein Name ist Stephan Mosel, [. . .]", HP 5, „Mein Name ist Marco Wegener und ich [. . .]", HP 15). Diese initiale Vorstellung des Autors ähnelt der eher formellen Kontaktaufnahme per E-Mail an einen Adressaten, der dem Produzenten unbekannt ist. Übereinstimmungen zwischen den Inhalten sind insbesondere im Hinblick auf studienbezogene Informationen zu finden: In vier Texten (30,8%) wird die Universität, an der der Autor studiert, explizit genannt, sieben Texte (53,8%) enthalten die Angabe des Studiengangs und drei Texte (23,1%) das Semester, in dem sich der Autor befindet. Darüber hinaus wird in zwei Texten auf Auslandsaufenthalte hingewiesen (15,4%). Bates und Lu (1997), Heiber (2001) und Killoran (2003a, 2003b, 2004) ermitteln in den von ihnen untersuchten Korpora Vorkommen von Lebensläufen in zahlreichen unterschiedlichen Ausprägungen (vgl. Abschnitt 4.6.3). Innerhalb der Texte, die in an die E-Mail angelehnte Textstrukturmuster eingebettet werden, kann diesbezüglich zwischen zwei Typen von Lebensläufen differenziert werden. Zum einen verweisen Hyperlinks in drei Texten (23,1%) auf genuine Lebensläufe, die sich in separaten Dokumenten befinden. Zum anderen können drei Texte als narrative Lebensläufe aufgefasst werden. Der Autor von HP 5 stellt sich zunächst vor, nennt seinen Geburtsort und geht auf die Entstehungsgeschichte seiner Homepage ein. Anschließend schildert er seine schulische Laufbahn und seine Zeit als Zivildienstleistender. Daraufhin thematisiert er seinen Studiengang, spezifische Interessengebiete und seine Tätigkeit als studentische Hilfskraft. Im letzten Absatz geht der Autor auf verschiedene Hobbys ein. Informationen zu außeruniversitären Interessen und Hobbys bilden in sieben Texten einen Schwerpunkt. Interessanterweise werden hierbei nicht der Computer oder das Internet am häufigsten aufgeführt, sondern verschie-

[27] Die Dokumente HP 9 und HP 10 werden in Tabelle 9.8 nicht aufgeführt, da sie, wie eingangs erwähnt, keinen Text als Rumpf der generischen E-Mail-Struktur besitzen. Die Tabelle stellt in der rechten Spalte zusätzlich den prozentualen Anteil des jeweiligen thematischen Bereichs in Bezug auf die 13 Texte dar. Es ist zu beachten, dass sich die in der Tabelle aufgeführten Merkmale auf *Themen*, also – mit Ausnahme des letzten Merkmals – nicht auf Hyperlinks beziehen, die in diesen Texten diskutiert werden.

Inhalte	HP 01	HP 02	HP 03	HP 04	HP 05	HP 06	HP 07	HP 08	HP 11	HP 12	HP 13	HP 14	HP 15	%
Biografische Angaben und Fakten über den Autor														
Name														
Vorname				✓			✓					✓		23,1
Vorname und Nachname			✓		✓								✓	23,1
Alter				✓	✓									15,4
Geburtsort					✓								✓	15,4
Ehemalige Schule					✓									7,7
Zivildienst					✓									7,7
Studium														
Universität				✓	✓	✓							✓	30,8
Studiengang			✓	✓	✓	✓			✓		✓		✓	53,8
Semesterzahl			✓		✓	✓								23,1
Auslandsaufenthalte			✓										✓	15,4
Beruf													✓	7,7
Lebenslauf														
Genuiner Lebenslauf			✓					✓					✓	23,1
Narrativer Lebenslauf					✓		✓						✓	23,1
Fotos			✓				✓			✓			✓	30,8
Hobbys/Interessen														
Computer allgemein					✓									7,7
Internet					✓								✓	15,4
Computerspiele	✓				✓									15,4
Musik	✓				✓								✓	23,1
Kino													✓	7,7
Rollenspiele					✓						✓			15,4
Sport				✓	✓						✓		✓	30,8
Reisen				✓						✓				15,4
Sonstige				✓	✓			✓			✓		✓	38,5
Kontaktinformationen														
E-Mail-Adresse						✓			✓					15,4
Gästebuch								✓						7,7
PGP-Key									✓					7,7
ICQ-UID					✓									7,7
Medienkompetenz					✓		✓		✓					23,1
Textcharakteristika														
Metadiskursive Äußerungen		✓		✓	✓	✓	✓	✓	✓	✓		✓	✓	76,9
Hinweis auf Hyperlinks	✓		✓				✓	✓	✓					38,5
Explizite Funktion				✓			✓		✓		✓	✓		38,5
Bitte um Kommentare			✓		✓				✓	✓			✓	38,5
Liste von Hyperlinks		✓	✓										✓	23,1

Tabelle 9.8: Inhalte und Charakteristika der Texte mit E-Mail-ähnlichen Textstrukturen

dene Sportarten (vier Texte, 30,8%) gefolgt von der Nennung unterschiedlicher Musikstile bzw. -gruppen (drei Texte, 23,1%). Angaben zu Kontaktinformationen werden in der Regel außerhalb des Textes aufgeführt, doch sind in zwei Texten die E-Mail-Adressen der Autoren enthalten. Diese werden zusätzlich von einer (indirekten) Bitte um Kontaktaufnahme flankiert („So, wer mehr wissen will, kann mir ja mal eine Mail schreiben. Meine email-Adresse ist: [...]", HP 6). In HP 8 wird der Rezipient aufgefordert, eine Nachricht im Gästebuch zu hinterlassen: „Das wichtigste ist jedoch mein Gästebuch, in daß [sic] sich natürlich jeder Besucher dieser Seite eintragen muß!!!" (vgl. Abbildung 9.1). In drei Texten sind die Verfasser bestrebt, Medienkompetenz zu signalisieren (vgl. auch Thelwall, 2003). Der Autor von HP 5 stellt dar, mit welchen Anwendungen er sein Webangebot pflegt, welche Browser er verwendet und gibt an, dass er sich „PC-mässig für Audiobearbeitung" und „SMS aus dem Web" interessiert. Der Produzent von HP 11 führt eine metadiskursive Absichtserklärung an, die eine geplante Dienstleistung erläutert: „Sehr bald werde ich auch hier auf dieser Site ein Auswahl [sic] kleiner aber feiner Share- und Freewareprogramme (Lizenzbestimmungen beachten!) direkt zum Download anbieten, so daß Ihr nicht das ganze Web abgrasen müsst um an ein gutes Prográmmchen zu kommen. Also noch ein bischen [sic] Geduld!"

Eckkrammer (2001, S. 54) stellt im Rahmen einer Analyse von traditionellen und digitalen Exemplaren verschiedener Textsorten „eine deutliche Zunahme humoristischer Komponenten sowie ein Plus an metadiskursiven Äußerungen" fest, „die auf eine intensive Auseinandersetzung mit der Textsorte hinweisen". Derartige Äußerungen können in zahlreichen Abstufungen für insgesamt zehn Texte (76,9%) belegt werden, sie stellen das auffälligste und am häufigsten verwendete inhaltliche Merkmal der untersuchten Texte dar (vgl. Tabelle 9.8).[28] Einige Texte (z. B. HP 2, HP 7, HP 14) bestehen fast ausschließlich aus metadiskursiven Äußerungen, die die Funktion der Homepage, den intendierten Rezipientenkreis oder für die Zukunft geplante Inhalte thematisierten. Zu Beginn schreibt der Autor von HP 2: „...endlich: meine Homepage ist da !!! Noch nicht alles funktioniert (ich arbeite dran) – deshalb einfach ausprobieren ...". Nach einer langen Liste von Grüßen an Freunde und Bekannte wird angemerkt: „Falls ich jemanden vergessen habe, nicht sauer sein – einfach eine böse Mail an mich, und ich erledige das !!!" Der Verfasser von HP 11 thematisiert eingangs die jüngste Revisionsphase seiner Homepage: „Ich habe in der Zwischenzeit ein paar Reperaturen [sic] an meiner Site vorgenommen und auch ein paar Rechtschreibfehler (ähmm wie peinlich!), die mir in der Eile unterlaufen sind beseitigt. Das Versenden einer e-mail an mich direkt von meiner HP aus sollte nun endlich auch funktionieren." Der Produzent von HP 7 geht auf defizitäre Aspekte seiner Homepage ein, die er in der Zukunft zu beheben gedenkt: „Das »to do«: ein paar nette Icons fehlen natürlich noch und dann möchte ich mein Umfeld noch etwas mehr beschreiben oder »abbilden«. Und da man es ja weltweit lesen kann: das Ganze noch in Englisch." In fünf Texten befinden sich darüber hinaus explizite Hinweise auf Hyperlinks, die eine Homepage beinhaltet, ohne jedoch die entsprechenden Erläuterungen als Hyperlinkanker zu realisieren. Diese Textabschnitte fungieren als eine Art unvollständige und nicht mediengerechte Kurzzusammenfassung der inhaltlichen Aspekte eines Angebots: „Hier

[28] Die Tabelle führt in der Zeile „Explizite Funktion" diejenigen metadiskursiven Äußerungen auf, die sich innerhalb der hier untersuchten Texte befinden. HP 10 enthält keinen derartigen Text, jedoch können in drei verknüpften Dokumenten metadiskursive Äußerungen zu unterschiedlichen Themengebieten festgestellt werden (vgl. Abschnitt 9.4.1).

kann man nach wie vor super Sachen finden, wie z. B. Levels für diverse 3D-Shooter (Quake3, Duke3D, ...), Tabulaturen von tollen Songs (Fu Manchu, Kyuss, ...), Informationen über beliebte Bands (A Perfect Circle, Doors, Led Zeppelin, ...) und natürlich eimerweise Links." (HP 1), „Viel gibt es noch nicht zusehen [sic], diese Seite hier, einen kurzen Lebenslauf, ein Photo und natürlich meine Email-Adresse." (HP 7). Explizite Aufforderungen, mit dem Autor per E-Mail in Kontakt zu treten, befinden sich in fünf Texten, z. B. „Feedback ist jedoch jederzeit willkommen und exxtrem [sic] gern gesehen!" (HP 5) und „Ich freue mich jederzeit über konstruktive Kritik, Anregungen und sonstiges Feedback. Schreiben Sie mir einfach per E-Mail, die ich auch gerne beantworten werde." (HP 15).

Abschließend kann festgehalten werden, dass in sämtlichen Texten Anleihen an Textstrukturmuster belegt werden können, deren Ursprünge meiner Vermutung zufolge in den persönlichen Erfahrungen der Autoren mit der E-Mail- und in Ausnahmefällen auch der Chat-Kommunikation zu finden sind. Dieser Umstand sowie die zahlreichen metadiskursiven Äußerungen – insbesondere die bereits in Abschnitt 9.4.1 dargestellten Beispiele – belegen einen Befund, den Eckkrammer (2001) unter anderem für digitale Kontaktanzeigen ermittelt hat: Die meisten Produzenten studentischer Homepages setzen sich in den von ihnen erstellten Webangeboten explizit mit spezifischen Merkmalen und kommunikativen Funktionen dieser Hypertextsorte auseinander. Dies ist wiederum als ein eindeutiger Beleg für die sehr unscharfen Konturen und nur marginal ausgeprägten Normen dieser Hypertextsorte zu interpretieren, weshalb die Autoren veranlasst werden, die korrespondierenden Regeln der Hypertextsorte explizit zu thematisieren. Ein weiteres Spezifikum betrifft den Umstand, dass die Verfasser derjenigen studentischen Homepages, die lediglich rudimentär ausgeprägt sind, kaum substanzielle Inhalte umfassen oder kein professionell wirkendes oder ästhetisches Webdesign besitzen, prinzipiell keine Maßnahmen der Sanktionierung fürchten müssen. Nicht die Qualität der äußeren Form scheint für die Autoren von besonderer Relevanz zu sein, sondern vielmehr der persönliche Ausdruck und die Authentizität der präsentierten Inhalte.

Die Signatur

In einer Signatur gibt der Autor der Homepage nur seinen Vornamen, Vornamen und Nachnamen oder Spitz- bzw. Rufnamen an. Signaturen sind in neun Dokumenten der Stichprobe enthalten (60%): „Christoph", „Stephan", „Dein Tobi", „Tobias Schwinger", „Olaf", „Carsten", „Euer M. Vogel", „Euer Jan Wagner :)" sowie „Marco Wegener im August 2000".[29] Verschiedene Aspekte fallen auf: Es dominieren Vornamen, wodurch der informelle Charakter einer Homepage betont wird. Der Autor von HP 7 „unterschreibt" seine Homepage mit „Dein Tobi", wohingegen in anderen Seiten *alle* potenziellen Rezipienten angesprochen werden sollen: „Euer M. Vogel" oder „Euer Jan Wagner :)". Es existieren also heterogene Formen der Leseranrede. Produzenten, die ungeübt in der Anfertigung von HTML-Dokumenten sind, sehen sich mit dem Umstand konfrontiert, den Leser ansprechen zu müssen. Einerseits kann die Gesamtheit aller potenziellen Leser angeredet werden, andererseits nur der jeweils aktuelle Rezipient. Abschnitt 9.6.3 geht genauer auf dieses Thema ein. Der Beleg „Marco

[29] Als Signaturen wurden nur diejenigen Vorkommen gewertet, die Bestandteil einer E-Mail-ähnlichen Textstruktur sind. In einigen Homepages, die keine Signatur enthalten, befindet sich eine typografisch abgesetzte Fußzeile, die den Namen des Autors sowie seine E-Mail-Adresse umfasst.

Wegener im August 2000“ scheint sich an Signaturen in Vorwörtern von Buchpublikationen (insbesondere Monografien) anzulehnen, die von den jeweiligen Verfassern in sehr vielen Fällen nach eben diesem Schema, jedoch unter zusätzlicher Angabe des Aufenthalts- oder Wohnortes verfasst werden.

Das Postscriptum

Ein abschließendes Postscriptum als Bestandteil der E-Mail-ähnlichen Textstruktur wird in HP 8 (vgl. Abbildung 9.1) sowie in HP 11 verwendet (Hervorhebungen im Original):

> P.S. Zu guter letzt wollte ich Euch noch einen guten Tipp mit auf den Weg geben: Wear sunscreen. . .
>
> P.S.S. [sic] Das Leben ist *kein Kindergeburtstag*, immer schön auf den *Blutdruck* achten und nicht so viel Aufregen! [sic] [HP 11]

Zusätzlich befindet sich in HP 10 ein HTML-Dokument, das die Heimatstadt des Autors vorstellt. „Die Kuppenheim-Seite“ schildert Schäden, die der Sturm „Lothar“ 1999 am Schwimmbad der Stadt verursacht hat und schließt mit: „PS: Mittlerweile hat der Gemeinderat beschlossen, das Bad neu aufzubauen (Kosten ca. 700 000.-) um die Arbeitsplätze der Angestellten zu erhalten. Man hofft es noch bis zur Freibad-Saison 2000 zu schaffen. Ich werde euch natürlich auf dem Laufenden halten. Also – auf ein Neues, Lothar !“ Das Postscriptum wurde offenbar nach der initialen Erstellung dieses Dokuments vom Autor eingefügt, um den angesprochenen Beschluss des Gemeinderates zu integrieren. Somit kann dieses Postscriptum als strategisch eingesetztes Stilmittel aufgefasst werden, da dem Autor die Umformulierung des initial vorhandenen Textes zur Reflektierung dieses aktuellen Ereignisses möglicherweise zu aufwändig oder zu umständlich erschien.

Fazit

Die Stichprobe enthält zahlreiche Konstituenten, die als Elemente des Textstrukturmusters einer generischen E-Mail aufgefasst werden können. Es wird deutlich, dass die Autoren keine Erfahrung mit der Anfertigung von HTML-Dokumenten haben; einige geben an, dass sie ihre persönlichen Angebote erstellen, um HTML zu erlernen. Es ist jedoch davon auszugehen, dass alle Autoren Kenntnisse im passiven, rezipierenden Umgang mit dem WWW sowie im aktiven, produzierenden und rezipierenden Umgang mit dem Internet-Dienst E-Mail und möglicherweise auch dem IRC und dem Usenet besitzen (vgl. Abschnitt 4.3.2). Der Verfasser von HP 14 betreibt in der Einstiegsseite seiner Homepage mit „Jan's Tagebuch“ eine Art Weblog (vgl. Abschnitt 4.6.7) und notiert im ersten Eintrag:

> . . .wollt jetzt schon zum Abschluß „Gruß Jan (:“ schreiben, wie unter alle meine Mail[s], aber dies ist ja keine. Dennoch grüße ich den geneigten Leser, der sich irgendwie auf diese Seiten verirrt hat – wie auch immer. [HP 14]

Vorhandene Formulierungsmuster, bestehendes Textsortenwissen sowie Kenntnisse und Erfahrungen, die in Bezug auf andere Kommunikationsformen existieren, werden bei der

Realisierung einer kommunikativen Handlung bewusst oder unbewusst integriert (vgl. Abschnitt 4.3.2). Der Verfasser präsupponiert in dieser speziellen Kommunikationssituation einen Dialogcharakter, obwohl das WWW nicht – wie E-Mail, IRC oder das Usenet – als genuines Kommunikations-, sondern als Informationsmedium aufzufassen ist (vgl. auch Crystal, 2001, S. 204). Der allgegenwärtige Dialogcharakter wird ebenfalls anhand der angebotenen Gästebücher und interaktiven Foren, Grüße an Freunde und Bekannte und Aufforderungen zur Kontaktaufnahme deutlich.

Viele der in die E-Mail-ähnlichen Textstrukturen eingebetteten Texte beinhalten Darstellungen der Person des Autors und kommen nicht über den Status hinaus, den Miller (1995) als „extended lonely-hearts advert" bezeichnet (vgl. Abschnitt 4.6.3), da die Produzenten keine Erfahrung im Umgang mit der Erstellung selbstdarstellender Texte besitzen und sich infolgedessen auf die Nennung einiger prägnanter Fakten beschränken. Eckkrammer (2001) untersucht verschiedene Textsorten in traditionellen und digitalen Medien. Interessanterweise stellt sie gerade für die Kontaktanzeige fest, dass korrespondierende Textexemplare, die im und für das WWW geschrieben werden, ebenfalls „starke formale und strukturelle Annäherung an Strukturen der E-Mail-Kommunikation" besitzen (ebd., S. 54). Die von Eckkrammer untersuchten Kontaktanzeigen dürften – zumindest zum Teil – ebenfalls von Personen verfasst worden sein, die keine Erfahrungen mit der Anfertigung von HTML-Dokumenten besitzen, so dass sie zu denjenigen Textmustern greifen, die ihnen vertraut sind: „Das etablierte Textwissen und Denken fließt deutlich in die Vertextung der digitalen Varianten ein." (Eckkrammer, 2001, S. 56). Zusätzlich zu dieser „Verankerung in traditionellen Konventionen" (ebd., S. 59) merkt die Verfasserin an, „daß die Textproduzenten [...] auf bewährte Modelle zurückgreifen. Auf formal-struktureller Ebene resultieren die häufigsten Modifikationen offensichtlich [...] aus intertextuellen Phänomenen, da das Medium dem Textproduzenten eine starke Anlehnung an die Vertextungsstrategien der Textsorte ›E-Mail‹ suggeriert. Im speziellen gilt es hier die in zahlreichen Studien beschriebene Annäherung an Modelle der Mündlichkeit zu nennen [...]." (ebd., S. 53). Die angesprochene „Suggerierung", die vom Medium WWW ausgeht, bezieht sich auf die produzentenseitige Ausfüllung von Eingabemasken, die aufgrund der spezifischen Kommunikationssituation (Anfertigung einer Kontaktanzeige) durchaus mit der Erstellung einer E-Mail vergleichbar ist.

9.6.2 Grüße an Freunde und Bekannte

Zwei Homepages enthalten Grüße an explizit genannte Freunde und Bekannte (vgl. Tabelle 9.7 sowie Fußnote 121, S. 224). Auf HP 10 liest man „Spezielle Grüsse an Astrid und Erik. Bleibt sauber ;-)", und in einem weiteren Dokument seiner Homepage richtet der Verfasser „Grüße an Emme+Auto (Fiat Panda), Hirsch+Auto(laut)." In HP 2 werden nach den vier Kategorien „Pics", „Linx (!)", „Badnerlied" und „Comments" sowie der Einleitung „...und noch schnell ein paar Grüße, und zwar an:" die teilweise kommentierten Vor- und Spitznamen von insgesamt 59 Personen aufgeführt.[30] Der Autor beschließt die Seite mit „(Falls ich

[30] Die Formulierung „und noch schnell ein paar Grüße" kann als Hinweis dafür gedeutet werden, dass der Produzent nicht ausgesprochen viel Zeit in die Erstellung seiner Homepage investiert hat. Einige der angesprochenen Grüße lauten: „Carsten (wieder in Kuwait !!!), Yvi (jetzt auch unter den Medi's in Homburg/Saar), Katharina (einen DICKEN Schmatz), Oli V., Jochen (besser ??), Dieter, CK & SN (1 x in M und 1 x in S), Döb, Maike,

jemanden vergessen habe, nicht sauer sein – einfach eine böse Mail an mich, und ich erledige das !!!)". Gerade derartige Belege geben einen deutlichen Hinweis auf die primär von den Autoren dieses Typs studentischer Homepages intendierte Gruppe von Rezipienten, die sich aus Freunden, Bekannten, Verwandten und Kommilitonen zusammensetzt (extrinsische Homepages im Sinne von Walker, 2000). Weitere Homepages (z. B. HP 3, HP 14) enthalten zwar keine Grüße, aber Listen von Hyperlinks zu den persönlichen Webangeboten von Freunden und Bekannten (vgl. Erickson, 1996). Der Autor von HP 4 bietet ein HTML-Dokument mit dem Dateinamen `freunde.htm` an, das in Tabellenform den „Nickname", die „e-mail" und die „Pages" von neun Personen enthält. Die Funktion dieser Webseite wird explizit thematisiert: „Wunderbar ist wenn man Freunde hat, ich habe den Meinigen eigentlich schon immer eine Seite gewidmet und so möchte ich auch diesmal die Tradition fortsetzten und somit aus meinem Leben schwafeln. Denn Freunde sagen auch ziemlich viel ueber einen selber aus. Die meisten haben auch Seiten im Netz, tja wer etwas auf sich hält oder gerade hipp [sic] sein möchte der braucht eine Homepage, ist doch claro." Es sei dahin gestellt, ob diese Aussage ironisch oder ernst gemeint ist. Zumindest belegt sie, dass dynamische Prozesse in sozialen Gruppen als Motivation zur Erstellung einer privaten Homepage einzubeziehen sind.

9.6.3 Heterogene Formen der Leseranrede

Besonders auffällige Unterschiede in den 15 Homepages existieren im Hinblick auf die jeweilige Form der Anrede des Lesers. In sieben Homepages werden vornehmlich passive Konstruktionen in Verbindung mit einigen Imperativen benutzt, d. h. eine Festlegung in der pronominalen Anrede auf Singular- („Du", „Sie") oder Pluralformen („ihr", „sie") wird bewusst vermieden, obwohl diese Homepages zahlreiche Merkmale der kommunikativen Nähe aufweisen. Die intime Singularform („Du") wird – ebenso wie das distanzierte und formalere „Sie" – von nur einem Autor konsequent eingesetzt, wohingegen die Pluralform („ihr", „euch" etc.) von vier Autoren benutzt wird. Zwei Produzenten (HP 8, HP 11) vermischen intime und distanzierte Anredeformen, was möglicherweise durch Unachtsamkeit bei der Erstellung oder Unsicherheiten bezüglich der intendierten Zielgruppe zu erklären ist. Abbildung 9.1 (S. 408) stellt HP 8 dar: Der Leser („Lieber Homepagebesucher") wird zunächst mit „Du" angesprochen („wenn Du mehr über mich erfahren möchtest, dann schau Dir [...] meinen Lebenslauf an."), woraufhin das Informationsangebot erläutert wird. Die Verabschiedung lautet „Viel Spaß beim weiteren Surfen wünscht Euch [...]". Im Postscriptum verschiebt sich die Anredeform erneut – bedingt durch die Technik des automatischen Aufrufzählers, der nur *einen* Zugriff von *einer* Personen zählt, weshalb gezwungenermaßen nur *eine* Person angesprochen werden kann: „Du bist übrigens der 02009 Besucher meiner Homepage". Bittner (2003, S. 86) stellt in der von ihm untersuchten Stichprobe privater Homepages ebenfalls „Uneinheitlichkeiten der Anrede" fest und führt sie auf „die Unentscheidbarkeit zwischen Privatheit und Öffentlichkeit" zurück. Nach Bittner (2003, S. 90) stellt die Sprecherrolle „aufgrund der Virtualisierung" den einzigen fixierten Referenzpunkt

Meike, Volker (2x), [...] Bengi, Andrea, Manu (ah, – na toll ...), Julia Schwabe (grins)". Derartig umfangreiche Listen von Grüßen ähneln vergleichbaren Aufstellungen in Abiturzeitungen und den Danksagungen an Freunde und Verwandte, die die Mitglieder semiprofessioneller Bands in CD-Booklets publizieren und die sich nur wenige Merkmale mit formalisierteren Danksagungen in akademischen Texten teilen.

dar, weshalb in zahlreichen privaten Homepages heterogene Formen der Anrede beobachtet werden können: „Das »Hier« und das »Jetzt« sind nur im Rahmen des virtuellen »Orts« der Homepage [...] und im virtuellen »Jetzt« der konkreten Rezeptionssituation zu verorten und in dieser Form zu Variablen geworden, die sich von der realen Welt und der Produktionssituation abgelöst haben. Die Behandlung dieser Variablen durch die Autoren zeugt dabei von individuell unterschiedlichen Fähigkeiten, mit diesen umzugehen, d. h. die Variablen entweder als solche zu belassen oder sie durch entsprechende Referenzierungen zu kompensieren." (ebd.). Der in HP 11 enthaltene Text richtet sich vornehmlich an *alle* potenziellen Besucher dieser Homepage: „Ich hoffe, daß ich den Besuchern dieser Seite ein paar nette Hyperlinks präsentieren kann, und Ihnen somit die tägliche Arbeit erleichtern und das Surfvergnügen maximieren kann. [...] Falls Ihr mir ein e-mail schicken wollt [...]. Und falls Ihr nun mich irgendwie erreichen wollt, so könnt Ihr das auch machen, und zwar [...]". Die Großschreibung von „Ihnen", die eine Interpretation als Anredepronomen der Distanz nahe legt, ist vielmehr als Orthografiefehler zu interpretieren. Es handelt sich um eine Anapher, die auf die Nominalphrase „den Besuchern dieser Seite" Bezug nimmt. Diese Interpretation wird durch die beiden folgenden Sätze gestützt, in denen der Autor alle potenziellen Leser mit dem informellen „Ihr" anspricht.[31]

Mit intimen oder distanzierten Pluralformen der Anrede wendet sich der Autor einer Homepage an die Gesamtheit der potenziellen Leser bzw. die von ihm intendierte Rezipientengruppe seines Webangebots. Der Produzent sieht sich demnach primär als Anbieter von Informationen, Texten und nützlichen Hyperlinks und nur sekundär als Partner einer dialogorientierten kommunikativen Situation. Dass aber dennoch eine allgemeine Orientierung an einer generischen Dialogsituation festzustellen ist, zeigen die Befunde zum E-Mail-ähnlichen Aufbau vieler Einstiegsseiten, so dass „eine Art »Face-to-Face« Situation entsteht" (Heiber, 2001, S. 18). Das eigentliche Paradoxon besteht in dem Umstand, dass sich zur tatsächlichen Rezeptionszeit in der Regel lediglich ein einzelner Leser vor einem Bildschirm befindet und das Webangebot rezipiert, doch wird dieser einzelne Leser nur in den seltensten Fällen unmittelbar als solcher angesprochen. Diese Anrede der Gesamtheit aller potenziellen Rezipienten kann auch häufig bei Anrufbeantworteransagen beobachtet werden: „Ihr könnt mir nach dem Signalton eine Nachricht hinterlassen."

9.7 Fazit – Das Hypertextsortenprofil

Die 15 in der Stichprobe enthaltenen Homepages reflektieren nahezu vollständig die generellen Eigenschaften, die privaten Homepages von Storrer (1999b, S. 5) zugeschrieben werden: Die Verfasser bemühen sich um Spontaneität sowie eine authentische und individuelle Ausdrucks- und Gestaltungsweise und richten sich in vielen Fällen direkt an den Rezipienten.

[31] Der in sehr großer Schrift dargestellte Titel dieser Homepage lautet „Soo Hallo erstmal.....also ich weiß nicht ob Sie's schon wussten.....ich hab jetzt auch 'ne Homepage..." und spielt auf das Markenzeichen des Kabarettisten Rüdiger Hoffmann an. Diese auch gelegentlich in der Umgangssprache verwendete Floskel ist einer von zahlreichen Belegen, in denen die Autoren Elemente der kontemporären Popkultur zur Individualisierung ihrer Homepages einsetzen. Die Individualisierung erfolgt vornehmlich durch das Text- bzw. Webdesign, z. B. spezielle horizontale Trennlinien, besonders auffällige oder humorvolle, animierte Icons, Hintergrundgrafiken und verspielte Schriftarten (vgl. Roberts, 1998, S. 79, und Walker, 2000, S. 110).

Hiermit gehen, wie die Beispiele zeigen, zahlreiche Merkmale konzeptioneller Mündlichkeit einher, wodurch ein ungezwungenes, informelles und gelegentlich sehr persönliches Ambiente verbreitet wird. In Bezug auf die häufig berichteten Merkmale privater Homepages existieren insbesondere im Bereich der eigentlichen Inhalte verschiedene Defizite, so bieten die Autoren von HP 1–15 nur in Ausnahmefällen umfangreiche und selbst geschriebene Texte, Artikel, Gedichte, Tagebücher oder sonstige Ressourcen an. Meist beschränken sie sich auf einige wenige Sätze oder Abschnitte zur eigenen Person, Aufzählungen verschiedener Kontaktmöglichkeiten und biografischer Angaben sowie auf zahlreiche Hotlists. Entsprechend können in sieben Homepages Hinweise darauf gefunden werden, dass die Autoren keine explizite Vorstellung davon haben, welche Funktion ihre Homepage besitzt und welche Inhalte sie publizieren können oder sollen. Dieser Befund bestätigt die Studie von Walters (1996), die zu dem Schluss kommt, dass nur die wenigsten Homepages einen spezifischen Zweck besitzen. Während Walters jedoch die Ansicht vertritt, dass sich von Studierenden angefertigte Homepages nie zu einem eigenständigen Genre entwickeln werden, belegen z. B. HP 5, 8, 14 und 15, dass durchaus Konventionen existieren und von den Autoren privater Homepages befolgt werden.[32] Basierend auf der hier untersuchten Stichprobe und den zahlreichen Übereinstimmungen mit den Ergebnissen von z. B. Bates und Lu (1997), Bayerl (2002) und Bittner (2003) kann die Beobachtung von Dillon und Gushrowski (2000) bestätigt werden, dass sich die „personal home page“ als „unique digital genre“ etabliert hat.

Es stellt sich jedoch nun die Frage, ob von Studierenden angefertigte private Homepages eine subgenerische Variante der Hypertextsorte *private Homepage* darstellen oder ob keine prinzipiellen Unterschiede zwischen diesen beiden Klassen von Hypertexten existieren. Die Analyse hat diesbezüglich deutlich gemacht, dass unterschiedliche Ausprägungen studentischer Homepages beobachtet werden können: Neben sehr rudimentären Webangeboten, die lediglich den Namen des Autors, einen sehr kurzen Begrüßungstext, eine E-Mail-Adresse, mehrere externe Hyperlinks, Grüße an Freunde und Bekannte und einige grafische Dekorationselemente umfassen, sind sehr umfangreiche, professionell gestaltete und zahlreiche Themen aufgreifende Homepages in der Stichprobe enthalten. Die Annahme der Existenz eines einzelnen, sämtliche Hypertextexemplare abdeckenden Typs erscheint somit nicht gerechtfertigt, da nicht jede studentische Homepage ein „carefully constructed portrayal of a person“ (Erickson, 1996, S. 15) darstellt. Die Klasse der nur rudimentär ausgeprägten Homepages kann in zweierlei Hinsicht konzeptualisiert werden: Zunächst stellen sie eine Art evolutionäre Vorstufe ›vollständiger‹ privater Homepages dar. In den rudimentären Angeboten der Stichprobe sind verschiedene Hinweise enthalten, die belegen, dass die Autoren ihre Homepages zum Erlernen von HTML angefertigt haben.[33] Entsprechend besitzen sie zahlreiche Merkmale, die zwar auf eine große Freude am Experimentieren mit den neuen technischen Möglichkeiten hinweisen, gleichzeitig wird jedoch ebenso deutlich, dass die Autoren Schwierigkeiten bei der Auswahl und Umsetzung geeigneter Inhalte und Themen haben. Diese Schwierigkeiten manifestieren sich unter anderem in heterogenen Formen der Rezipientenanrede und der Adaption von Textstrukturmustern und sprachlichen Spezifika, die den Au-

[32] Zur Erklärung des Entstehungsprozesses und der Befolgung dieser Konventionen kann das in Abschnitt 4.3.2 dargestellte Modell herangezogen werden.

[33] Buten (1996) zufolge ist das Erlernen von HTML für 48% der Autoren der wesentliche Grund, eine persönliche Homepage anzufertigen (vgl. Fußnote 114, S. 218, sowie Abschnitt 9.4, S. 388 ff.).

toren aus anderen Kommunikationsformen bekannt sind. Weiterhin scheinen einige Autoren den zeitlichen und konzpetionellen Aufwand zu scheuen, Inhalte zu erstellen, weshalb sie sich darauf beschränken, teils sehr umfangreiche Listen externer Hyperlinks anzubieten, die – aller Wahrscheinlichkeit nach – nahezu vollständig auf den eigenen Bookmark-Listen basieren und somit effizient in die Homepages integriert werden können. Rudimentäre Homepages werden publiziert, obwohl – wie die Diskussion der metadiskursiven Äußerungen zeigt – den Autoren bewusst ist, dass sich die jeweiligen Angebote in einem Status der Unvollständig- und Vorläufigkeit befinden. Abhängig vom Ehrgeiz und Engagement der Produzenten (vgl. Bayerl, 2002, S. 22 f.) können diese Homepages zu späteren Zeitpunkten sukzessive aktualisiert und um zusätzliche Inhalte erweitert werden, wodurch sie sich allmählich dem Status annähern, den vollständig ausgeprägte studentische Homepages besitzen.[34] Die zweite Möglichkeit der Konzeptualisierung betrifft die Beschreibung des aktuellen Zustands rudimentärer studentischer Homepages: Diese können als Ausprägung des minimalen prototypischen Kerns des Hypertexttyps *Homepage einer Person* betrachtet werden, der unter anderem den Namen des Autors und seine E-Mail-Adresse umfasst.

Dass ein solcher prototypischer Kern existiert, zeigen die in Abschnitt 4.6.3 diskutierten Korrespondenzen zwischen den Analysen von Bates und Lu (1997), Dillon und Gushrowski (2000) und Bittner (2003). Auch die auf S. 225 reproduzierte Definition der Hypertextsorte „personal home page“ von de Saint-Georges (1998, S. 76) entspricht prinzipiell der Beschreibung eines prototypischen Kerns und verschiedenen fakultativen Ausprägungen hinsichtlich der Peripherie eines korrespondierenden Hypertextexemplars. Basierend auf dieser Definition stellt die von einem Autor angefertigte Vorstellung der eigenen Person, die mindestens seinen Namen oder ein Foto enthält, den Kern der Hypertextsorte dar. Weiterhin können aktuelle Aktivitäten, die berufliche Tätigkeit oder persönliche Interessen thematisiert werden. Bayerl (2002, S. 16) ersetzt in ihrer Definition der *Homepage eines Studierenden* (vgl. ebenfalls Abschnitt 4.6.3) die berufliche Tätigkeit durch „studentische Aktivitäten“ und fügt „Dienstleistungen (in textueller Form oder als Links auf externe Seiten)“ hinzu, d. h. der prototypische Kern der übergeordneten Hypertextsorte *private Homepage* bleibt unverändert bestehen. Die Ausprägungen der Peripherie erfahren jedoch kontextuell bestimmte Veränderungen, die von dem Status des Produzenten als Studierendem sowie der mit einem Hypertext verfolgten Funktion abhängig sind. Da sie „vor allem [eine] inhaltlich geprägte Subklasse“ (Heinemann und Heinemann, 2002, S. 143) darstellt, kann die *private Homepage eines Studierenden* somit als Hypertextsortenvariante der *privaten Homepage* aufgefasst werden (vgl. Abschnitt 2.3.2).

Die im Sinne des Hypertextsortenmodells (vgl. Kapitel 5, insbesondere Abbildung 5.4, S. 293) wesentlichen Merkmale der Hypertextsorte *private Homepage* wurden bereits in Kapitel 4 sowie in den vorangegangenen Abschnitten ausführlich diskutiert: Die *private Homepage eines Studierenden* kann als Hypertextsortenvariante konzeptualisiert werden, die mehrere kommunikative Funktionen umfassen kann, z. B. die Kontaktaufnahme mit anderen Anwendern, das Informieren über die eigene Person, das Anbieten spezifischer Dienstleistungen und Ressourcen oder das Erlernen von HTML. Die kontextuellen Faktoren betreffen den Status des Produzenten als Person, die an einer Universität als Student immatrikuliert ist, weshalb

[34] Ein vollständig ausgeprägtes Hypertextexemplar umfasst mindestens diejenigen Hypertextsortenmodule, die die Mehrzahl der Hypertextexemplare besitzen, die in einer korrespondierenden Stichprobe enthalten sind.

eine institutionelle Bindung existiert. Die Inhalte und Themen, die in Exemplaren dieser Hypertextsorte enthalten sind, beziehen sich auf biografische Angaben, Kontaktinformationen sowie persönliche Interessen und Hobbys. Oftmals werden auch Dienstleistungen (etwa Lernmaterialien oder Informationen über studentische Vereinigungen) und ein Studienprofil angeboten. Genuine interaktive Komponenten können nur in Einzelfällen nachgewiesen werden, diese umfassen vornehmlich Möglichkeiten der Kommunikation wie z. B. Gästebücher oder E-Mail-Formulare. Bezüglich der Strukturierung können in der Stichprobe nur wenige Gemeinsamkeiten ermittelt werden. Die Inhalte werden primär hierarchisch strukturiert und in abgeschlossenen Knoten behandelt. Eingebettete Hypertexte (im Sinne einer Gruppe von HTML-Dokumenten, die eine funktional-thematisch markierte kommunikative Ganzheit darstellen) zu spezifischen Themen stellen die Ausnahme dar. Im Hinblick auf das Merkmal der Dekoration können ebenfalls keine Generalisierungen vorgenommen werden. In der Stichprobe sind einige Exemplare vorhanden, die eine ansprechende, konsistente und professionell wirkende Gestaltung besitzen. Andere Exemplare weisen ein sehr heterogenes, inkonsistentes und verspieltes Webdesign auf, das klar erkennen lässt, dass ihre Produzenten keine Erfahrung im Umgang mit der Anfertigung von Webangeboten besitzen.

Die wesentlichen Hypertextsortenmodule der Hypertextsorte *private Homepage eines Studierenden*, die aus der Inhalts- und Makrostrukturanalyse (vgl. Abschnitt 9.5) abgeleitet werden können, sind in Tabelle 9.9 dargestellt. Mehr als 85% der in der Stichprobe vertretenen Homepages besitzen eine *explizite Begrüßung des Rezipienten*, *Kontaktinformationen*, eine *Identifikation* der Homepage, eine *E-Mail-ähnliche Textstruktur*, *grafische Dekorationselemente*, *biografische Angaben* (vgl. auch Bates und Lu, 1997), Informationen zu *Hobbys und Interessen* des Produzenten sowie *Hotlists*. Neben den Bezeichnungen atomarer und komplexer Hypertextsortenmodule beinhaltet Tabelle 9.9 verschiedene weitere Informationen. Diese beziehen sich zunächst auf ihren primären und sekundären Typ (vgl. Abschnitt 5.6.4): Diejenigen Hypertextsortenmodule, die nicht einem der Typen Interaktion, Kommunikation, Navigation, Metainformation, Dekoration oder Textstrukturmuster zugeordnet werden können, werden als vornehmlich inhaltlich-thematisch markiert aufgefasst (z. B. die *explizite Begrüßung*, der *Name des Homepage-Besitzers* oder die *Bitte um Kontaktaufnahme*). Hypertextsortenmodule dieses primären Typs können – neben inhärenten Merkmalen wie z. B. der kommunikativen Funktion – zusätzliche Aspekte beinhalten, die als sekundärer Typ aufgefasst werden, so wird z. B. ein *Diktum/Zitat/Motto* oftmals als dekoratives Element eingesetzt, und die *prophylaktische Beschwichtigungsfloskel* kann zusätzlich als Metainformation angesehen werden. Der primäre Typ Textstrukturmuster bezieht sich insbesondere auf die Hypertextsortenmodule *E-Mail-ähnliche Textstruktur* und *Lebenslauf*. Eine *E-Mail-ähnliche Textstruktur* kann insbesondere in den Einstiegsseiten beobachtet werden, wodurch einige thematisch-strukturelle Konstituenten vorgegeben werden (Anrede, Begrüßung, Textkörper, Verabschiedung, Postscriptum), weshalb dieses Hypertextsortenmodul zusätzlich sekundär inhaltlich-thematisch markiert ist und als übergeordnete Konstituente für Hypertextsortenmodule des primären Typs Inhalt/Thema fungiert (z. B. die *Begrüßung*, *Verabschiedung*, *Identifikation*, das *Studienprofil* sowie *Hobbys und Interessen*).

Der Status eines Hypertextsortenmoduls bezieht sich auf Aspekte der Typologisierung: Wenn die *private Homepage eines Studierenden* als subgenerische Variante der Hypertextsorte *private Homepage* aufgefasst wird, die wiederum eine spezifische Ausprägung des Hypertext-

Bezeichnung des Hypertextsortenmoduls	Ebene	Primärer Typ	Sekundärer Typ	Status	Verwendung	Typische Ausprägung	Frequenz
Explizite Begrüßung	atomar	Inhalt/Thema		generell	obligatorisch	Einstiegsseite	100
Kontaktinformationen	komplex	Kommunikation	Interaktion	generell	obligatorisch	intern	100
E-Mail-Adresse	atomar	Kommunikation	Interaktion	generell	obligatorisch	beliebig	100
Straßenadresse	atomar	Inhalt/Thema		generell	optional	intern	33
Telefonnummer	atomar	Inhalt/Thema		generell	optional	intern	20
E-Mail-Formular	atomar	Kommunikation	Interaktion	generell	optional	intern	13
ICQ-UID	atomar	Kommunikation	Interaktion	spezifisch	optional	intern	13
PGP-Key	atomar	Interaktion	Kommunikation	generell	optional	intern	13
Identifikation	komplex	Inhalt/Thema		generell	obligatorisch	Einstiegsseite	93
Name des Homepage-Besitzers	atomar	Inhalt/Thema		generell	obligatorisch	Einstiegsseite	93
. . .begleitet von einem Foto	atomar	Dekoration	Inhalt/Thema	generell	obligatorisch	intern	60
E-Mail-ähnliche Textstruktur	atomar	Textstrukturmuster	Inhalt/Thema	spezifisch	obligatorisch	Einstiegsseite	93
Grafische Dekorationselemente	atomar	Dekoration	Inhalt/Thema	universal	obligatorisch	beliebig	93
Biografische Angaben	komplex	Inhalt/Thema		generell	obligatorisch	intern	93
Geburtsort	atomar	Inhalt/Thema		generell	obligatorisch	intern	53
Alter/Geburtsdatum	atomar	Inhalt/Thema		generell	optional	intern	47
Tätigkeit	atomar	Inhalt/Thema		generell	optional	intern	33
Wohnort	atomar	Inhalt/Thema		generell	optional	intern	33
Lebenslauf	atomar	Inhalt/Thema	Textstrukturmuster	generell	optional	intern	20
Praktika	atomar	Inhalt/Thema		spezifisch	optional	intern	20
Familienstand	atomar	Inhalt/Thema		spezifisch	optional	intern	13
Zivil- bzw. Wehrdienst	atomar	Inhalt/Thema		spezifisch	optional	intern	13
Hobbys und Interessen	komplex	Inhalt/Thema		generell	obligatorisch	intern	93
Externe Hyperlinks	atomar	Navigation	Inhalt/Thema	generell	obligatorisch	intern	73
Eigene Inhalte	atomar	Inhalt/Thema		spezifisch	obligatorisch	intern	73
Hotlist (Liste externer Hyperlinks)	atomar	Navigation	Inhalt/Thema	generell	obligatorisch	intern	87
Dienstleistungen	komplex	Inhalt/Thema	Navigation	spezifisch	obligatorisch	intern	80
Informationen zu einer Region	atomar	Inhalt/Thema		spezifisch	optional	intern	20
Liste der Homepages von Freunden	atomar	Navigation	Inhalt/Thema	spezifisch	optional	intern	20
Hausarbeiten/Papiere	atomar	Inhalt/Thema		spezifisch	optional	intern	13
Studentische Vereinigungen	atomar	Inhalt/Thema		spezifisch	optional	intern	13
Primäre Navigationshilfe	atomar	Navigation	Inhalt/Thema	generell	obligatorisch	Einstiegsseite	80
Studienprofil	komplex	Inhalt/Thema		spezifisch	obligatorisch	intern	80
Studiengang	atomar	Inhalt/Thema		spezifisch	obligatorisch	intern	73
Universität	atomar	Inhalt/Thema		spezifisch	obligatorisch	intern	53
Semester/Studienphase	atomar	Inhalt/Thema		spezifisch	optional	intern	40
Studienschwerpunkte	atomar	Inhalt/Thema		spezifisch	optional	intern	27
Auslandsaufenthalte	atomar	Inhalt/Thema		spezifisch	optional	intern	20
Prophyl. Beschwichtigungsfloskel	atomar	Inhalt/Thema	Metainformation	spezifisch	obligatorisch	intern	73
Explizite Verabschiedung	atomar	Inhalt/Thema		generell	obligatorisch	Einstiegsseite	73
Datum der letzten Änderung	atomar	Metainformation		universal	obligatorisch	beliebig	67
Bitte um Kontaktaufnahme	atomar	Inhalt/Thema		generell	obligatorisch	intern	60
Fotos	komplex	Dekoration	Inhalt/Thema	generell	obligatorisch	intern	53
Foto	atomar	Dekoration	Inhalt/Thema	generell	optional	intern	33
. . .begleitet von einem Kommentar	atomar	Inhalt/Thema	Dekoration	generell	optional	intern	20
Copyright-Hinweis	atomar	Metainformation	Inhalt/Thema	universal	obligatorisch	beliebig	53
Rückverweis zur Einstiegsseite	atomar	Navigation		universal	optional	intern	40
Diktum/Zitat/Motto	atomar	Inhalt/Thema	Dekoration	spezifisch	optional	Einstiegsseite	40
Zugriffszähler	atomar	Metainformation		universal	optional	beliebig	40
Gästebuch	atomar	Kommunikation	Interaktion	generell	optional	intern	33
Fußzeile	atomar	Metainformation	Inhalt/Thema	generell	optional	beliebig	27
Seiteninterne Navigationshilfe	atomar	Navigation	Inhalt/Thema	generell	optional	intern	20
Sitemap	atomar	Navigation	Inhalt/Thema	generell	optional	intern	13
Grüße an Freunde	atomar	Inhalt/Thema		spezifisch	optional	Einstiegsseite	13

Tabelle 9.9: Die Hypertextsortenmodule der *privaten Homepage eines Studierenden*

typs *Homepage einer Person* ist, liegen sowohl gemeinsame als auch unterschiedliche Merkmale vor. Generelle Hypertextsortenmodule sind dieser sowie den übergeordneten Hypertextsorten zugehörig, wohingegen spezifische Hypertextsortenmodule distinktive Merkmale einer Hypertextsortenvariante darstellen. Universale Hypertextsortenmodule können in Hypertextexemplaren vollständig arbiträrer Hypertexttypen und -sorten eingesetzt werden.

Die Verwendung bezieht sich auf die empirische Frequenz eines Hypertextsortenmoduls in einer Stichprobe korrespondierender Hypertextexemplare. Sobald ein Hypertextsortenmodul in mehr als der Hälfte aller Exemplare eingesetzt wird, kann es als obligatorisch oder auch typisch für diese Hypertextsorte betrachtet werden. Hochfrequente Hypertextsortenmodule, die in mindestens ca. 85% aller Exemplare eingesetzt werden, werden als Bestandteile des (proto)typischen Kerns einer Hypertextsorte aufgefasst.[35]

Die vorletzte Spalte von Tabelle 9.9 bezieht sich auf die typische Ausprägung eines Hypertextsortenmoduls in einem Hypertextexemplar. Die Tabelle unterscheidet diesbezüglich zwischen der Realisierung in der Einstiegsseite, in einem eingebetteten HTML-Dokument oder in beliebigen Knoten. Bezüglich des Merkmals der typischen Ausprägung kann auch eine detailliertere Analyse vorgenommen werden, die sich z. B. auf die Position innerhalb eines Dokuments oder die Form der typischen Ausprägung, also auf ein Hypertextmodul bezieht. Auf diese Weise könnte z. B. ausgedrückt werden, dass das Hypertextsortenmodul *explizite Begrüßung* in der Regel als typografisch abgesetztes Textfragment am oberen Rand der Einstiegsseite oder das Hypertextsortenmodul *biografische Angaben* typischerweise als Abschnitt von Fließtext innerhalb des mittleren Bereichs der Einstiegsseite realisiert wird (vgl. Ihlström und Åkesson, 2004, sowie Abschnitt 4.6.4). Für diese Stichprobe wurde diesbezüglich nur eine rudimentäre Analyse vorgenommen, da die beobachtete Varianz in Bezug auf die Ausprägungen von Hypertextsortenmodulen zu umfangreich ist, was wiederum die unscharfen Grenzen dieser Hypertextsortenvariante im Hinblick auf ihre Peripherie verdeutlicht.[36] Die typische Ausprägung betrifft aber neben einer detaillierteren Binnendifferenzierung auch unterschiedliche Möglichkeiten der Realisierung, die sich auf die Ebene des Knotens oder eines eingebetteten Hypertextes beziehen. Dieser Umstand kann anhand des Hypertextsortenmoduls *Hobbys und Interessen* veranschaulicht werden. Einige Autoren beschränken sich auf die aufzählende Nennung ihrer Hobbys innerhalb eines Begrüßungstextes. Weitere Autoren gehen innerhalb dieses Textes in mehreren Absätzen auf jeweils ein Interesse ein. Weiterhin sind Homepages in der Stichprobe enthalten, in denen mehrere unterschiedliche Interessen in eigenständigen HTML-Dokumenten vorgestellt werden, und schließlich kann ein Hobby auch in einem vollständigen Hypertext thematisiert werden (vgl. allgemein hierzu Kapitel 5). Umfassende Varianzen dieser Form können in der Stichprobe nur für die Hypertextsortenmodule *Hobbys und Interessen* sowie *Dienstleistungen* beobachtet werden, wobei festzuhalten ist, dass die Realisierung als vollständiger Hypertext die Ausnahme darstellt. Typisch ist für das erstgenannte Hypertextsortenmodul die aufzählende Nennung. Die unter Dienstleistungen zu-

[35] Hochfrequente Merkmale werden von Rezipienten häufiger wahrgenommen, weshalb sie sich unter Bezug auf das zyklische Modell der Entwicklung von Hypertextsorten (vgl. Abschnitt 4.3.2) naturgemäß schneller verbreiten als niedrigfrequente Merkmale.

[36] Weiterhin ist bei der Berücksichtigung eines derart detaillierten Analysekriteriums wie der Positionierung und typischen Ausprägung eines Hypertextsortenmoduls die Untersuchung zahlreicher weiterer Hypertextexemplare notwendig, um generalisierbare Ergebnisse erzielen zu können.

sammengefassten Hypertextsortenmodule werden in der Regel durch eigenständige Knoten realisiert, auf die von einem übergeordneten Knoten verwiesen wird, der Instanzen weiterer Hypertextsortenmodule umfasst. Aus Darstellungsgründen wurden derartige optionale Realisierungsmöglichkeiten einzelner Hypertextsortenmodule nicht in Tabelle 9.9 aufgeführt. Dies gilt auch für den Umstand, dass einige Hypertextsortenmodule als rekurrente Instanzen realisiert werden können. Das atomare Hypertextsortenmodul *Name des Homepage-Besitzers* wird z. B. in mehreren studentischen Homepages in *jedem* HTML-Dokument als Bestandteil der Fußzeile verwendet, die häufig zusätzlich einen *Copyright-Hinweis*, das *Datum der letzten Änderung* dieses Dokuments sowie die *E-Mail-Adresse des Autors* und gelegentlich einen *Zugriffszähler* umfasst. Es werden also oftmals diejenigen Hypertextsortenmodule mehrfach instanziiert, die die primären Typen Metainformation, Navigation, Textstrukturmuster oder Dekoration besitzen. In Bezug auf inhaltlich-thematisch und kommunikativ markierte Hypertextsortenmodule kann eine rekurrente Instanziierung stattfinden, sofern sie für eine Hypertextsorte von besonderer Relevanz sind und über einen sehr kurzen Durchschnittsumfang verfügen (z. B. der *Name des Homepage-Besitzers* oder seine *E-Mail-Adresse*).

Neben den Hypertextsortenmodulen zeigt Tabelle 9.9 die typische Vorbelegung der Einstiegsseite einer studentischen Homepage als Hypertextknotensorte auf. Diese umfasst in der Regel eine *E-Mail-ähnliche Textstruktur*, die wiederum eine *Begrüßung* und optional eine *Verabschiedung* beinhaltet. Zusätzlich enthält die Einstiegsseite den *Namen des Homepage-Besitzers* und seine *E-Mail-Adresse*. Eine *primäre Navigationshilfe*, die sich in der Regel am linken Rand des Dokuments befindet, listet interne Hyperlinks auf, die zu weiteren Knoten führen. Ein *Diktum/Zitat/Motto* sowie *Grüße an Freunde und Bekannte* sind weitere fakultative Bestandteile dieser Hypertextknotensorte. Bates und Lu (1997) gelangen in ihrer Stichprobe von 114 privaten Homepages zu drei Typen von Einstiegsseiten, die ausschließlich Listen von Hyperlinks (9%), ausschließlich Fließtext und nur wenige Verknüpfungen (14%) sowie Hyperlinks und ein Inhaltsverzeichnis enthalten (77%). Ob auch in der Hypertextsorte *private Homepage eines Studierenden* im Hinblick auf die Einstiegsseite multiple Hypertextknotensorten existieren, kann nur durch zusätzliche Analysen mit zahlreichen weiteren Hypertextexemplaren ermittelt werden.

10 Analyse 3: Die persönliche Homepage eines Wissenschaftlers

10.1 Einleitung

Die dritte Analyse betrifft die Untersuchung persönlicher Homepages von Hochschulangehörigen, geht unter anderem auf die Ursprünge einiger zugehöriger Hypertextsortenmodule ein und kontrastiert die Ergebnisse mit der in Kapitel 9 dargestellten Analyse.[1]

Im Anschluss an die Darstellung der Ziele dieser Analyse sowie der verwendeten Stichprobe (Abschnitt 10.3) diskutiert Abschnitt 10.4 zunächst metadiskursive Äußerungen und leitet aus ihnen unterschiedliche Funktionen der Hypertextsorte *persönliche Homepage eines Wissenschaftlers* ab. Abschnitt 10.5 stellt die Ergebnisse der Inhalts- und Makrostrukturanalyse vor, aus denen ein Hypertextsortenprofil sowie eine Typologie des Hypertexttyps *Homepage einer Person* konstruiert werden kann (Abschnitt 10.6).

10.2 Ziele und Bezüge zum Hypertextsortenmodell

Diese Untersuchung einer Stichprobe der Homepages von Hochschulangehörigen versucht die Frage zu beantworten, ob die *persönliche Homepage eines Wissenschaftlers* ebenso wie die *private Homepage eines Studierenden* als Hypertextsortenvariante der privaten bzw. beruflichen Homepage aufgefasst werden kann. Die Erstellung eines empirisch fundierten Hypertextsortenprofils verfolgt diesbezüglich das Ziel, den Grad der Konventionalisierung dieser Hypertextsorte zu bestimmen. Eine weitere Motivation zur Durchführung dieser Analyse betrifft die in der Literatur bislang nicht diskutierte Frage, auf welche Weise unterschiedlich ausführlich realisierte Homepages konzeptualisiert und typologisiert werden können.

[1] Initiale Ergebnisse der in diesem Kapitel dargestellten Analyse wurden in Rehm (2004c) publiziert. Darüber hinaus geht Rehm (2002b) auf eine Vorstudie ein, in der vier persönliche Homepages von Wissenschaftlern in Bezug auf ihre Hypertextsortenmodule untersucht wurden.

In methodischer Hinsicht orientiert sich diese Analyse an der Vorgehensweise, die bereits in Kapitel 9 zur Untersuchung der studentischen Homepages eingesetzt wurde, d. h. es werden zunächst metakommunikative Äußerungen und explizit in den Webangeboten enthaltene Angaben zur Funktion einer Homepage untersucht, woraufhin die einzelnen makrostrukturellen Komponenten ermittelt und diskutiert werden.

10.3 Die Stichprobe

Die verwendete Stichprobe basiert auf 100 persönlichen Homepages von Hochschulangehörigen, die semiautomatisch mit Hilfe der Korpusdatenbank zusammengestellt wurden. Zunächst wurde eine nach der Anzahl persönlicher Homepages sortierte Liste derjenigen Webserver generiert, auf denen derartige Webangebote verfügbar sind (vgl. Abschnitt 7.3.4). Auf Grundlage der Häufigkeiten konnten mehrere Webserver ermittelt werden, die nahezu ausschließlich von Mitarbeitern der jeweiligen Hochschule gepflegte persönliche Homepages enthalten. Von diesen Webangeboten wurde wiederum eine Liste ihrer URLs angefertigt und per Zufallsverfahren angeordnet. Aus dieser Liste wurden iterativ 100 Homepages in die Stichprobe aufgenommen, sofern sie die folgenden Bedingungen erfüllten: Das HTML-Dokument ist (a) die Einstiegsseite der persönlichen Homepage einer Person, die an einer Hochschule angestellt ist, (b) deutschsprachig, (c) eindeutig der Person zuzuordnen und (d) beschäftigt sich unter anderem mit der Tätigkeit des Emittenten an der Universität.[2]

Die in der Stichprobe enthaltenen Homepages sind in den Tabellen 10.1 und 10.2 aufgeführt. Von Hochschullehrern werden 34 Webangebote gepflegt, auf Personen mit Doktortitel fallen 29 Homepages, 16 Autoren besitzen ein Diplom, und 12 Produzenten machen keine Angabe bezüglich ihrer akademischen Titel. Sechs Homepages stammen von Privatdozenten, zwei Autoren besitzen einen Magisterabschluss, eine Person verfügt über ein abgeschlossenes Staatsexamen.[3] Mit nur wenigen Ausnahmen werden alle Homepages von Personen angeboten, die in Forschung und Lehre tätig sind, weshalb sie auch als *persönliche Homepages von Wissenschaftlern* bezeichnet werden können. Die Ausnahmen beziehen sich unter anderem auf HP 22 (Diplomand in einem physikalischen Institut), HP 32 (Mitarbeiterin in einem Prüfungsamt), HP 34 (Mitarbeiter in einem Rechenzentrum) und HP 69 (Mitarbeiter in einem zentralen Entwicklungslabor). Von weiblichen Produzenten wurden 28 der 100 Homepages angefertigt.[4] Insgesamt beinhalten die in der Stichprobe vertretenen Homepages 1 835 HTML-Dokumente, was einem Durchschnitt von 18,4 Webseiten entspricht (Median: 6); die hier betrachteten Homepages sind weniger umfangreich als die in Kapitel 9 untersuchten studentischen Homepages, die im Schnitt 41,5 HTML-Dokumente enthalten. Etwa 500 dieser HTML-Dateien wurden nicht in das Korpus aufgenommen, was ein Hin-

[2] Aufgrund dieser Kriterien mussten mehrere vermeintlich persönliche Homepages verworfen werden, beispielsweise Webangebote von Instituten und Arbeitsgruppen, die an der Universität Gießen oftmals innerhalb des Adressbereichs persönlicher Homepages (d. h. `http://www.uni-giessen.de/~g...`) angeboten werden.

[3] In vier Homepages werden die Titel nicht angegeben, können jedoch aus dem Kontext geschlossen werden; sie werden in den Tabellen 10.1 und 10.2 in Klammern dargestellt. Einige Autoren führen ihre vollständigen akademischen Titel bzw. Abschlüsse auf (z. B. „Prof. Dr. Dr. h. c.“, „Dr. rer. nat.“ oder „Dipl.-Psych.“).

[4] Dabei handelt es sich um HP 1, 10, 13, 14, 32, 33, 42, 44, 50, 51, 54, 56, 58, 59, 60, 65, 68, 72, 75, 77, 80, 89, 91, 95, 96, 97, 98 und 100.

	URL	Titel	Anzahl Dokumente	Download	Letzte Änderung
HP 1	http://www.uni-giessen.de/~g91060/	Prof. Dr.	18	16.01.2001	k. A.
HP 2	http://www.uni-giessen.de/~gb21/	Prof. Dr.	5	16.01.2001	k. A.
HP 3	http://www.uni-giessen.de/~g51027/	k. A.	17	16.01.2001	k. A.
HP 4	http://www.uni-giessen.de/~gb1026/	Prof. Dr.	7	16.01.2001	k. A.
HP 5	http://www.uni-giessen.de/~gde9/	k. A.	35	16.01.2001	k. A.
HP 6	http://www.uni-giessen.de/~ga46/	Prof. Dr.	4	16.01.2001	28.06.2000
HP 7	http://www.uni-giessen.de/~gf1020/	PD Dr.	8	16.01.2001	10.01.2001
HP 8	http://www.uni-giessen.de/~gf1003/	Dr.	31	16.01.2001	24.02.2000
HP 9	http://www.uni-giessen.de/~g91062/	Prof. Dr.	146	16.01.2001	29.11.2000
HP 10	http://www.uni-giessen.de/~g91003/	Prof. Dr.	6	16.01.2001	24.11.2000
HP 11	http://www.uni-giessen.de/~gg1015/	(Dr.)	10	16.01.2001	15.12.2000
HP 12	http://www.uni-giessen.de/~g51010/	Prof. Dr.	9	16.01.2001	13.12.2000
HP 13	http://www.uni-giessen.de/~gc1050/	Dr.	9	16.01.2001	18.10.2000
HP 14	http://www.uni-giessen.de/~ga1022/	Dipl.	13	16.01.2001	16.10.2000
HP 15	http://www.uni-giessen.de/~g81020/	Prof. Dr.	26	16.01.2001	18.12.2000
HP 16	http://www.uni-giessen.de/~g11026/	Prof. Dr.	10	16.01.2001	11.11.1998
HP 17	http://www.uni-giessen.de/~ghc6/	Prof. Dr.	3	16.01.2001	08.01.1997
HP 18	http://www.uni-giessen.de/~g91022/	Prof. Dr.	9	16.01.2001	18.10.2000
HP 19	http://www.uni-giessen.de/~ge47/	Dr.	2	16.01.2001	12.05.1999
HP 20	http://www.uni-giessen.de/~gb10/	Prof. Dr.	7	16.01.2001	31.10.2000
HP 21	http://www.uni-giessen.de/~gkp2/	Dipl.	1	16.01.2001	06.08.1996
HP 22	http://www.uni-giessen.de/~gd1123/	Dipl.	3	16.01.2001	27.10.1998
HP 23	http://www.uni-giessen.de/~g51051/	Dipl.	7	16.01.2001	19.10.2000
HP 24	http://www.uni-giessen.de/~g61048/	Dr.	1	16.01.2001	28.02.2000
HP 25	http://www.uni-giessen.de/~gh1223/	Dr.	12	16.01.2001	08.12.2000
HP 26	http://www.uni-giessen.de/~g91016/	Dr.	1	16.01.2001	09.05.2000
HP 27	http://www.tu-chemnitz.de/~jarn/	Dipl.	9	12.02.2001	08.01.2001
HP 28	http://www.tu-chemnitz.de/~rar/	Dr.	4	12.02.2001	04.03.1998
HP 29	http://www.tu-chemnitz.de/~bekl/	Prof. Dr.	1	12.02.2001	05.10.2000
HP 30	http://www.tu-chemnitz.de/~blk/	k. A.	19	12.02.2001	18.05.2000
HP 31	http://www.tu-chemnitz.de/~tboh/	Dipl.	13	12.02.2001	26.01.2001
HP 32	http://www.tu-chemnitz.de/~bda/	k. A.	1	12.02.2001	17.11.2000
HP 33	http://www.tu-chemnitz.de/~anbuc/	Dipl.	3	12.02.2001	18.05.1999
HP 34	http://www.tu-chemnitz.de/~meh/	Dipl.	48	12.02.2001	24.11.2000
HP 35	http://www.tu-chemnitz.de/~hader/	Dipl.	9	12.02.2001	02.01.2001
HP 36	http://www.tu-chemnitz.de/~rhaf/	Dr.	57	12.02.2001	19.12.2000
HP 37	http://www.tu-chemnitz.de/~huebner/	Prof. Dr.	120	12.02.2001	27.12.1996
HP 38	http://www.uni-bonn.de/~dbuncic/	Staatsex.	446	26.10.2001	16.10.2001
HP 39	http://www.uni-bonn.de/~upp30009/	Dr.	12	26.10.2001	22.08.2001
HP 40	http://www.uni-bonn.de/~uph60016/	Dr.	28	26.10.2001	01.10.2001
HP 41	http://www.uni-bonn.de/~ntrunte/	Dr.	18	26.10.2001	03.07.2001
HP 42	http://www.uni-bonn.de/~ckinitz/	(Dipl.)	1	26.10.2001	27.09.2001
HP 43	http://www.uni-bonn.de/~ute401/	Prof. Dr.	4	26.10.2001	08.12.2000
HP 44	http://www.uni-bonn.de/~cschroed/	Dr.	5	26.10.2001	08.11.2000
HP 45	http://staff-www.uni-marburg.de/~albrechm/	Dr.	6	28.11.2001	k. A.
HP 46	http://staff-www.uni-marburg.de/~altwasse/	k. A.	48	28.11.2001	k. A.
HP 47	http://staff-www.uni-marburg.de/~anz/	Prof. Dr.	7	28.11.2001	k. A.
HP 48	http://staff-www.uni-marburg.de/~asbach/	Dr.	17	28.11.2001	k. A.
HP 49	http://staff-www.uni-marburg.de/~barthh-m/	Prof. Dr.	2	28.11.2001	k. A.
HP 50	http://staff-www.uni-marburg.de/~bauerb/	Prof. Dr.	22	28.11.2001	k. A.

Tabelle 10.1: Die untersuchten persönlichen Homepages von Wissenschaftlern (HP 1–50)

	URL	Titel	Anzahl Dokumente	Download	Letzte Änderung
HP 51	http://staff-www.uni-marburg.de/~behrensl/	PD Dr.	28	28.11.2001	k. A.
HP 52	http://staff-www.uni-marburg.de/~beise/	Dr.	3	28.11.2001	k. A.
HP 53	http://staff-www.uni-marburg.de/~berns/	Prof. Dr.	9	28.11.2001	k. A.
HP 54	http://staff-www.uni-marburg.de/~bertelsm/	Prof. Dr.	13	28.11.2001	k. A.
HP 55	http://staff-www.uni-marburg.de/~blaser/	Dipl.	1	28.11.2001	k. A.
HP 56	http://staff-www.uni-marburg.de/~brake/	Dipl.	13	28.11.2001	k. A.
HP 57	http://staff-www.uni-marburg.de/~brandtw/	Prof. Dr.	4	28.11.2001	k. A.
HP 58	http://staff-www.uni-marburg.de/~bschmidt/	(PD Dr.)	4	28.11.2001	k. A.
HP 59	http://staff-www.uni-marburg.de/~buechne2/	k. A.	1	28.11.2001	k. A.
HP 60	http://staff-www.uni-marburg.de/~buerger/	k. A.	1	28.11.2001	k. A.
HP 61	http://staff-www.uni-marburg.de/~conradi/	Dipl.	6	28.11.2001	k. A.
HP 62	http://staff-www.uni-marburg.de/~foersthd/	Prof. Dr.	13	28.11.2001	k. A.
HP 63	http://staff-www.uni-marburg.de/~friedri3/	Prof. Dr.	2	28.11.2001	k. A.
HP 64	http://staff-www.uni-marburg.de/~funkg/	Dr.	2	28.11.2001	k. A.
HP 65	http://staff-www.uni-marburg.de/~gaehler/	Dr.	1	28.11.2001	k. A.
HP 66	http://staff-www.uni-marburg.de/~garn/	Dr.	1	28.11.2001	k. A.
HP 67	http://staff-www.uni-marburg.de/~gemeinha/	(Dr.)	8	28.11.2001	k. A.
HP 68	http://staff-www.uni-marburg.de/~gimmler/	Dr.	28	28.11.2001	k. A.
HP 69	http://staff-www.uni-marburg.de/~gladisch/	k. A.	1	28.11.2001	k. A.
HP 70	http://staff-www.uni-marburg.de/~greiner/	Prof. Dr.	6	28.11.2001	k. A.
HP 71	http://staff-www.uni-marburg.de/~gruener/	Dr.	66	28.11.2001	k. A.
HP 72	http://staff-www.uni-marburg.de/~hartlieb/	Dr.	7	28.11.2001	k. A.
HP 73	http://staff-www.uni-marburg.de/~hartmann/	Dipl.	1	28.11.2001	k. A.
HP 74	http://staff-www.uni-marburg.de/~haspelm/	Dr.	4	28.11.2001	k. A.
HP 75	http://staff-www.uni-marburg.de/~hessed/	k. A.	2	28.11.2001	k. A.
HP 76	http://staff-www.uni-marburg.de/~hoeffken/	PD Dr.	5	28.11.2001	k. A.
HP 77	http://staff-www.uni-marburg.de/~huehnm/	k. A.	1	28.11.2001	k. A.
HP 78	http://staff-www.uni-marburg.de/~huelst/	PD Dr.	2	28.11.2001	k. A.
HP 79	http://staff-www.uni-marburg.de/~janich/	Prof. Dr.	29	28.11.2001	k. A.
HP 80	http://staff-www.uni-marburg.de/~jentges/	M. A.	10	28.11.2001	k. A.
HP 81	http://staff-www.uni-marburg.de/~jungclas/	Prof. Dr.	3	28.11.2001	k. A.
HP 82	http://staff-www.uni-marburg.de/~kesslerr/	Prof. Dr.	6	28.11.2001	k. A.
HP 83	http://staff-www.uni-marburg.de/~korsch/	Prof. Dr.	8	28.11.2001	k. A.
HP 84	http://staff-www.uni-marburg.de/~krafft/	Prof. Dr.	15	28.11.2001	k. A.
HP 85	http://staff-www.uni-marburg.de/~kriegerw/	Prof. Dr.	1	28.11.2001	k. A.
HP 86	http://staff-www.uni-marburg.de/~kuhlmans/	Dipl.	29	28.11.2001	k. A.
HP 87	http://staff-www.uni-marburg.de/~kunath/	k. A.	1	28.11.2001	k. A.
HP 88	http://staff-www.uni-marburg.de/~kunih/	Prof. Dr.	22	28.11.2001	k. A.
HP 89	http://staff-www.uni-marburg.de/~leupold/	M. A.	3	28.11.2001	k. A.
HP 90	http://staff-www.uni-marburg.de/~luhr/	Dr. des.	1	28.11.2001	k. A.
HP 91	http://staff-www.uni-marburg.de/~margrafs/	Dr.	3	28.11.2001	k. A.
HP 92	http://staff-www.uni-marburg.de/~mittendv/	Dipl.	4	28.11.2001	k. A.
HP 93	http://staff-www.uni-marburg.de/~kuesterm/	Prof. Dr.	50	28.11.2001	22.11.2001
HP 94	http://staff-www.uni-marburg.de/~bonacket/	Dr.	16	28.11.2001	18.10.2001
HP 95	http://staff-www.uni-marburg.de/~mennehar/	PD Dr.	83	28.11.2001	09.11.2001
HP 96	http://staff-www.uni-marburg.de/~arend/	Dr.	4	28.11.2001	17.10.2001
HP 97	http://staff-www.uni-marburg.de/~niehaus/	k. A.	1	28.11.2001	04.05.1998
HP 98	http://staff-www.uni-marburg.de/~baumann/	Dr.	6	28.11.2001	21.03.2000
HP 99	http://staff-www.uni-marburg.de/~stemmler/	Prof. Dr.	3	28.11.2001	17.01.1997
HP 100	http://staff-www.uni-marburg.de/~rausch/	Prof. Dr.	4	29.11.2001	17.10.2001

Tabelle 10.2: Die untersuchten persönlichen Homepages von Wissenschaftlern (HP 51–100)

weis auf den Umstand ist, dass persönliche Homepages von Wissenschaftlern häufig auch in einer englischsprachigen Version angeboten werden und darüber hinaus HTML-Versionen englischsprachiger Veröffentlichungen enthalten. Geisteswissenschaftlichen Fächern können 61 Homepages zugeordnet werden, während dem natur- und ingenieurswissenschaftlichen Bereich 39 Webangebote zugehörig sind. Die vertretenen Disziplinen umfassen Germanistik (17 Homepages), Theologie (11), Medizin (6), Psychologie (5), Soziologie, Physik, klassische Philologie, Mathematik, Informatik, Chemie und Biologie (jeweils 4) sowie Sportwissenschaft, Romanistik, Pharmazie und Anglistik (jeweils 3).

10.4 Funktionen und Konventionen

Analog zu den Ausführungen in Abschnitt 9.4 werden im Folgenden zunächst von den Produzenten explizit thematisierte Äußerungen bezüglich der Funktion ihrer Homepage diskutiert, woraufhin eine Verortung der Homepages hinsichtlich verschiedener Typologien persönlicher Webangebote vorgenommen wird.

10.4.1 Metadiskursive Äußerungen

Im Gegensatz zu den studentischen Homepages sind in der vorliegenden Stichprobe nur wenige metadiskursive Äußerungen enthalten. Während die in Kapitel 9 diskutierten Vorkommen meist längere Textabschnitte umfassen, beschränken sich die hier thematisierten metadiskursiven Äußerungen im Regelfall auf eine Phrase oder einen kurzen Satz.

Lediglich sieben Homepages umfassen metadiskursive Anmerkungen zum Status einer Homepage; prophylaktische Beschwichtigungsfloskeln sind in der Ausprägung, die in Kapitel 9 diskutiert wurde, in dieser Stichprobe nicht enthalten. Der Verfasser von HP 17 bietet in der Einstiegsseite seiner Homepage die drei Themenbereiche „Forschungsschwerpunkte“, „Lehre“ und „Veröffentlichungen“ an, von denen nur das letzte Nomen einen Hyperlinkanzeiger zu einem weiteren HTML-Dokument darstellt. Anschließend wird – in kleinerer Schrift und durch einen horizontalen Strich abgetrennt – eine weitere Sektion namens „Sonstiges“ eingeführt, die lediglich die Absichtserklärung „(kommt noch)“ umfasst. Der Produzent von HP 21 verwendet die traditionelle Floskel zur Markierung eines vorläufigen Status seines Webangebots: „this homepage is still under construction“ (vgl. Abschnitt 4.6.10), der Verfasser von HP 61 wählt mit „diese Seite befindet sich derzeit noch im Aufbau“ in drei HTML-Dokumenten eine deutschsprachige Variante. Die Autorin von HP 13 bezieht sich mit einer metadiskursiven Äußerung lediglich auf einen spezifischen Teil ihrer Homepage: „Die Vorankündigung mit den Folien zur Vorlesung ist zwar noch nicht fertig, kann aber trotzdem schon eingesehen werden.“ Der Verfasser von HP 5 bietet in der Einstiegsseite seiner Homepage mehrere längere Textabschnitte an. Zum Schluss führt er auf, „Was Sie zudem demnächst auf dieser Seite finden werden: Gedanken und Diskussionspunkte zu meinen laufenden Forschungsarbeiten“. Auf ein ähnliches Thema bezieht sich ein Kommentar in HP 23, der den einzigen Inhalt in einem Dokument darstellt, das von der Einstiegsseite über den Hyperlinkanzeiger „Forschungsinteressen“ erreichbar ist: „in Bearbeitung“. Auch der Autor von HP 86 leitet das Dokument „Forschung“ mit einer derartigen Äußerung ein: „Für alle, die es interessiert und für alle, die trotzdem fragen, beabsichtige ich hier ein paar Infos über meine

momentane Arbeit zusammenzustellen. Sorry, wenn es im Moment noch etwas dünn ist, aber ich arbeite dran. . .". Sofern sie nicht mit einem globalen Skopus in der Einstiegsseite eingesetzt werden, finden sich metadiskursive Äußerungen insbesondere im inhaltlichen Bereich der persönlichen Forschungsinteressen. Die drei letztgenannten Beispiele stammen aus Homepages, die von Personen angefertigt wurden, die dem Mittelbau zugehörig und (noch) nicht promoviert sind. Es ist davon auszugehen, dass die drei Autoren zum Zeitpunkt der Erstellung ihrer Homepages unter anderem an ihren Dissertationen gearbeitet haben, weshalb sie möglicherweise antizipierten, dass der Bereich der persönlichen Forschungsinteressen gerade in ihren Fällen von besonderer Wichtigkeit ist, weil er den Rezipienten über das jeweilige Dissertationsvorhaben informieren sollte.

In insgesamt sechs Webangeboten werden die Funktionen und Inhalte einer Homepage thematisiert. Der Verfasser von HP 8 merkt in der Einstiegsseite seiner Homepage an: „Diese Seite informiert über mich und meine Arbeit. Von besonderem Interesse ist für einige Besucher vielleicht die Möglichkeit zu Wissenschaftlichen [sic] Literaturrecherchen (Botanik/Ökologie), die ich unten anbiete [. . .]." Wissenschaftler verbinden mit ihrer Homepage demzufolge unter anderem eine generelle Informationsfunktion, die sich sowohl auf fachfremde Personen als auch auf Kollegen bezieht, die möglicherweise ein Interesse an spezifischen Ressourcen haben. Diese Funktion wird auch in HP 11 deutlich: „Auf dieser Seite möchte ich Ergebnisse meiner Arbeit vorstellen und allgemein zugänglich machen." Im Anschluss an diesen Satz geht der Autor auf weitere Inhalte und Funktionen seiner Homepage ein: „Sie enthält zudem Informationen über meine Person und die ein [sic] oder andere Kleinigkeit, die im Laufe der Jahre entstanden ist. Einfach mal unter Hobbies [sic] nachschauen und überraschen lassen." Weitere zentrale Bestandteile dieser Hypertextsorte sind somit neben den persönlichen Forschungsinteressen auch Angaben über die eigene Person. Zusätzlich weist der Produzent von HP 11 auf Informationen zu seinen Freizeitaktivitäten hin, die jedoch nur in vier Homepages (HP 11, HP 22, HP 46, HP 86) der Stichprobe thematisiert werden und daher – im Gegensatz zur Homepage eines Studierenden (vgl. Abschnitt 9.5.4) – in dieser Hypertextsorte als Randerscheinung aufzufassen sind. Zwei weitere Kernbereiche der persönlichen Homepage eines Wissenschaftlers werden vom Verfasser von HP 23 angesprochen: „Sie können hier weniges über meinen Lebenslauf erfahren, die bescheidene Liste meiner Publikationen lesen und sich über meine Forschungsinteressen und Lehrveranstaltungen informieren." In ähnlicher Weise äußert sich der Autor von HP 94: „Hier finden Sie unter anderem Informationen zu aktuellen Terminen, zu meinen Arbeitsschwerpunkten, Lehrveranstaltungen und Projekten." Der Kern der hier betrachteten Hypertextsorte umfasst diesen Angaben zufolge nahezu sämtliche Inhaltsbereiche, mit denen sich Hochschulangehörige, die in Forschung und Lehre tätig sind, in ihrem Arbeitsalltag beschäftigen und deren Veröffentlichung im WWW sowohl für den Produzenten als auch für unterschiedliche Gruppen von Rezipienten einen potenziellen Mehrwert darstellt. Im Hinblick auf Lehrveranstaltungen handelt es sich dabei um Informationen, Kommentare und weiterführende Materialien zu Vorlesungen, Seminaren und Übungen. In Bezug auf die Forschung werden vornehmlich Interessengebiete und Projekte thematisiert. Parallel werden Listen von Veröffentlichungen angeboten, mit denen unterschiedliche Intentionen verbunden sind. Ein Lebenslauf umfasst darüber hinaus die wesentlichen biografischen Angaben des Autors. Der Verfasser von HP 18 geht in einem separaten und mit der Überschrift „Seminar- und Abschlußarbeiten" versehe-

nen HTML-Dokument auf einen weiteren Inhaltsbereich ein: „Dieser Platz ist für ausgezeichnete Seminar- oder Abschlußarbeiten vorgesehen. Nutzen Sie diese Publikationsmöglichkeit, Ihre Arbeit einem breiteren Publikum bekannt zu machen." Obschon in diesem Dokument keine derartigen Arbeiten aufgeführt werden, belegen diese metadiskursive Äußerung und die bereits dargestellten Hinweise zu lehrveranstaltungsbezogenen Informationen, dass sich die Autoren mit ihren persönlichen Homepages unter anderem an die Teilnehmer ihrer Lehrveranstaltungen wenden. Diese Kernbestandteile deuten bereits an, dass Listen externer Hyperlinks in dieser Hypertextsorte einen anderen Stellenwert besitzen als in studentischen Homepages. Der Verfasser von HP 86 geht in einem HTML-Dokument mit dem Titel „Links" explizit auf diesen Umstand ein:

> An dieser Stelle möchte ich den geneigten Surfer mit der Erkenntnis verschonen, daß es so unglaublich praktische Dinge wie Suchmaschinen im Internet gibt, und daß diese Yahoo, Google oder Altavista heißen. Suchen – so denk ich mir – kann jeder selbst, deswegen ist dies eine kleine Liste der Seiten, die ich inzwischen gefunden habe und obendrein auch noch ganz interessant finde. [HP 86]

Diese metadiskursive Äußerung belegt, dass der Bedarf an Hotlists, die lediglich allgemeine Hyperlinks zu Suchmaschinen und Katalogen enthalten, zumindest aus Sicht des Verfassers von HP 86 nicht mehr existiert. Zusätzlich verdeutlicht die Bemerkung, dass Wissenschaftler mit ihren persönlichen Homepages spezifische Intentionen verfolgen, die über das Experimentieren mit dem Medium WWW und das Anbieten genereller Hyperlinklisten hinausgehen. Externe Hyperlinks werden stattdessen in nahezu allen Homepages ausschließlich zu sehr spezifischen Informationen angeboten. In HP 86 beziehen sich diese aller Wahrscheinlichkeit nach auf das Dissertationsthema des Autors. Der entsprechende Abschnitt der Linkliste wird eingeleitet mit dem Kommentar: „Gute Lignan-Seiten mit biochemischem Schwerpunkt habe ich bislang nicht finden können, das scheint noch eine Marktlücke zu sein. Die meisten Seiten beschäftigen sich mit [...]". Diese Hotlist besitzt die implizite Funktion, den Rezipienten über die Forschungsinteressen des Autors zu informieren, die in einem weiteren Dokument namens „Forschung" explizit vorgestellt werden. Da der Verfasser der Ansicht ist, in Bezug auf das angesprochene Thema im WWW eine „Marktlücke" ausgemacht zu haben, ist er offenbar bestrebt, in Zukunft weiterführende Informationen und auch externe Hyperlinks anzubieten, weshalb dieses Dokument, das drei Hotlists mit insgesamt neun Einträgen umfasst, mit der Äußerung „to be continued ..." beendet wird.

10.4.2 Anwendung funktionaler Typologien

Eine eindeutige Kategorisierung der Stichprobe im Hinblick auf die von Miller (1995) aufgestellte Typologie privater und persönlicher Homepages (vgl. Abschnitt 4.6.3) kann nicht vorgenommen werden, da die meisten Homepages multiplen Typen zugeordnet werden können. Den Typ „Hi, this is me (as an individual)" bezieht Miller zwar vornehmlich auf private Homepages, sehr viele Homepages der Stichprobe weisen jedoch Elemente auf, die die jeweilige Person als Individuum vorstellen, z. B. biografische Informationen und ein Porträtfoto – gerade Informationen dieses Typs stellen den prototypischen Kernbereich des Hypertexttyps *Homepage einer Person* dar. Die zweite Kategorie, die Miller „Hi, this is me (as a member of

an organisation)“ nennt und für die die Homepage eines Wissenschaftlers als Beispiel aufgeführt wird, dominiert die Stichprobe. Die Thematisierung der Mitgliedschaft einer Person innerhalb einer spezifischen Organisation oder Institution besitzt jedoch für das Gros der 100 Webangebote nur einen sekundären Stellenwert. Obwohl nahezu alle Emittenten auch auf ihre Tätigkeit innerhalb der jeweiligen Hochschule eingehen, stellen sich die Autoren primär in ihrer beruflichen Rolle als Wissenschaftler vor, weshalb die Bezeichnung „Hi, this is me (as a scientist/scholar/academic)“ die Stichprobe treffender charakterisiert. Die Einbindung der Autoren in spezifische Organisationseinheiten einer Universität werden insbesondere bei der Gestaltung und Strukturierung einer Homepage sowie bei den Hypertextsortenmodulen *Angaben zu Lehrveranstaltungen* und *Forschungsschwerpunkte* deutlich (vgl. Abschnitt 10.5). Der dritte Typ, Homepages von Familien, kann in der Stichprobe nicht nachgewiesen werden, der Typ „This is what I think is cool“ ist ebenfalls nicht vertreten. Die fünfte Kategorie „An advertisement for myself“ umfasst die drei Subtypen „Cool style“, „The electronic curriculum vitae“ und „An advertisement for the service I can provide“. Während der erste und der dritte Subtyp nicht belegt werden können, enthalten sehr viele der 100 Homepages Elemente, die der zweiten Subkategorie sowie der Oberkategorie zugeordnet werden können.[5] Die korrespondierenden Inhalte beziehen sich zunächst auf verschiedene Ausprägungen tabellarischer und narrativer Lebensläufe, die in den Webangeboten enthalten sind. Zudem werden oftmals Informationen publiziert, die die durch eine Homepage repräsentierte Person als aktives und produktives Mitglied einer spezifischen Forschungs-Community positionieren sollen (z. B. Publikationslisten, Forschungsschwerpunkte, Mitgliedschaften in Fachverbänden sowie externe Hyperlinks zu Gesellschaften und Verbänden). Derartige Inhalte dienen gleichzeitig der Positionierung der eigenen Person als Wissenschaftler sowie in generellerer Hinsicht der Werbung, da sich die Emittenten auf diese Weise gegenüber Journalisten, den Organisatoren von Konferenzen sowie den Kollegen und Organisationseinheiten der eigenen Institution implizit als Experten ausweisen können. Viele Webangebote der Stichprobe – insbesondere diejenigen von Professoren – enthalten unter anderem Bestandteile, die üblicherweise den Unterlagen für eine Bewerbung als Hochschullehrer beigefügt werden (Lebenslauf, Aufstellung der Forschungsschwerpunkte, Publikationsliste, Herausgebertätigkeiten, die Namen der betreuten Doktoranden, Mitgliedschaften und Funktionen in Verbänden etc.). Meiner Einschätzung nach haben diese obligatorischen Konstituenten von Bewerbungsunterlagen die Struktur und den Inhalt der Hypertextsorte *persönliche Homepage eines Wissenschaftlers* maßgeblich beeinflusst, da die weltweite Zugänglichkeit der Informationen, die sie umfassen, mit verschiedenen potenziellen Vorteilen verbunden ist.

Von den drei Typen privater Webangebote, die Walker (2000) vorschlägt, kann die dritte Kategorie, die sich auf spezifische Freizeitinteressen des Autors bezieht, nicht in der Stichprobe belegt werden. Die beiden anderen Typen beziehen sich einerseits auf sehr rudimen-

[5] An vielen Universitäten existieren Zentren oder eigenständige Institutionen zur Durchführung von Projekten im Bereich des Wissens- und Technologietransfers zwischen Hochschule und Privatwirtschaft. Institute, Seminare oder einzelne Professuren haben oftmals die Möglichkeit, auf diesem Wege Drittmittel aus dem industriellen Umfeld einzuwerben, das entsprechende Dienstleistungen in Auftrag gibt. Die explizite Nennung der Mitarbeit in einem Transferzentrum, die als Werbung für dieses Zentrum und die eigene Dienstleistung aufgefasst werden kann, ist lediglich in HP 9 und HP 47 zu finden. Als Werbung können zudem auch Hinweise auf Webangebote konzeptualisiert werden, die innerhalb von Forschungsprojekten entstanden sind.

täre Informationen zur eigenen Person, andererseits auf eine Erweiterung dieses Typs, die detailliertere Angaben enthält, z. B. einen narrativ gestalteten Lebenslauf. Obwohl sich Walker lediglich auf private Homepages bezieht, kann eine derartige Differenzierung in ihren Grundzügen erfolgreich auf die Dokumente der Stichprobe angewendet werden: Viele Autoren stellen lediglich ihren Namen, einige Kontaktinformationen und ihre Funktion innerhalb der jeweiligen Organisationseinheit zur Verfügung. Andere Produzenten stellen eine sehr umfassende Menge von Informationen bereit, wozu interessanterweise in einigen Fällen auch narrativ gestaltete und in Ausnahmefällen sehr private Aspekte thematisierende Lebensläufe unterschiedlichen Umfangs gehören (HP 8, HP 41, HP 63, HP 72, HP 74, HP 80). Da mehrere Abstufungen existieren, kann jedoch nicht von einer dichotomischen Unterscheidung gesprochen werden, wodurch wiederum die Einführung eines prototypischen Kerns motiviert wird, der auf der Peripherie unterschiedliche Ausprägungen besitzen kann. Walker (2000) unterscheidet zusätzlich zwischen extrinsischen privaten Homepages, die sich an Personen richten, die dem Produzenten bekannt sind, sowie intrinsischen Homepages, die sich vornehmlich an eine unbekannte Leserschaft wenden. Diese Differenzierung in Bezug auf die vom Emittenten intendierte Zielgruppe muss im Hinblick auf die *persönliche Homepage eines Wissenschaftlers* überarbeitet werden, da hochfrequente Hypertextsortenmodule existieren, die sich an unterschiedliche Gruppen von Rezipienten richten.

10.5 Inhalte und makrostrukturelle Komponenten

Für die Stichprobe wurde eine Analyse der Inhalte und der Makrostruktur vorgenommen, deren Ergebnisse nachfolgend vorgestellt werden. Im Vergleich mit den studentischen Homepages besitzen die Dokumente einen in vielerlei Hinsicht unterschiedlichen und nahezu ausnahmslos seriöseren sowie offizielleren Charakter, der sich insbesondere in überaus konventionalisierten Inhaltskomponenten, einer gemäßigten und eher funktional als rein dekorativ eingesetzten grafischen Gestaltung sowie einer vorherrschenden konzeptionellen Schriftlichkeit manifestiert. Aus der E-Mail-, IRC- oder Usenet-Kommunikation stammende Merkmale für konzeptionelle Mündlichkeit können in der Stichprobe nicht belegt werden, d. h. Emphasen, Buchstabenreduplikationen etc. werden ebenso wie Smileys in ihrer getippten Form (:-)) nicht verwendet, die Homepage eines Rechenzentrumsmitarbeiters enthält jedoch vier Smileys als kleinformatige Grafiken (HP 34). Entsprechend der vorherrschenden konzeptionellen Schriftlichkeit sind in den Dokumenten nur sehr wenige ortografische oder grammatikalische Fehler enthalten, die Inhalte scheinen nicht – wie bei der Mehrzahl der studentischen Homepages – spontan erzeugt worden zu sein. Ein weiterer Unterschied betrifft die Strukturierung der Homepages sowohl auf der Knoten- als auch auf der Hypertextebene, die in dieser Stichprobe eine nahezu standardisierte Beschaffenheit besitzt. Auf der Ebene des Einzelknotens bezieht sich die Strukturierung insbesondere auf den Aspekt der Positionierung einzelner Hypertextsortenmodule innerhalb der Einstiegsseite. Die studentischen Homepages besitzen oftmals Begrüßungstexte, die an E-Mail-ähnliche Textstrukturmuster angelehnt sind. In den 100 Homepages von Hochschulangehörigen kann dieses Phänomen nicht belegt werden. Stattdessen enthalten die Einstiegsseiten weniger Absätze von Fließtext, vorherrschend sind isolierte Textfragmente und insbesondere verschiedene Formen der listenartigen oder tabel-

larischen Anordnung von Informationen. Interessanterweise bezieht sich die Präsentation von Listen nur in Ausnahmefällen auf Sammlungen externer Hyperlinks, die in lediglich 12 Homepages enthalten sind. Dieser Befund ist, wie bereits angesprochen, ein deutliches Indiz für den Umstand, dass diese Hypertextsorte konventionalisierte Inhalte umfasst und beinahe alle Produzenten sehr spezifische Intentionen mit ihren Homepages verbinden.

10.5.1 Identifizierende Informationen

In jeder Homepage befindet sich innerhalb der Einstiegsseite der Name des Emittenten, der in 69 Fällen zusätzlich den akademischen Titel enthält.[6] Der Name wird in fast allen Fällen in einer größeren Schrifttype zu Beginn des HTML-Dokuments dargestellt und von den umgebenden Objekten visuell separiert – in HP 3 wird darüber hinaus am unteren Rand der Einstiegsseite der Spitzname des Produzenten aufgeführt. Der Name fungiert somit als Überschrift des Hypertextes und markiert die Person des Autors als dominierenden Referenzpunkt der präsentierten Inhalte. In 27 Einstiegsseiten wird innerhalb dieses Überschriftmoduls oder in unmittelbarer Nachbarschaft zusätzlich die berufliche Tätigkeit genannt; im Falle von Hochschullehrern handelt es sich dabei um die Bezeichnung der Professur (z. B. in HP 16: „Professur für Bürgerliches Recht, Arbeitsrecht und Zivilprozeßrecht“, HP 43: „Universitätsprofessor für Kirchengeschichte“ oder HP 56: „Wissenschaftliche Mitarbeiterin im Bereich Empirische Pädagogik“). Eine explizite Nennung der Affiliation innerhalb dieses Hypertextmoduls kann in 34 Einstiegsseiten belegt werden, wobei in der Regel die übergeordnete Organisationseinheit sowie der Name der Hochschule genannt werden (z. B. in HP 5: „Berthold Suchan [*Zeilenumbruch*] Wissenschaftlicher Assistent [*Zeilenumbruch*] Zentrum für Philosophie der Justus-Liebig-Universität Giessen“).[7] Weiterhin wurden die Frequenzen isoliert positionierter Affiliationsmarkierungen ermittelt. In 75 Einstiegsseiten wird der Name der Universität, an der der Emittent tätig ist, im Klartext aufgeführt, 16 Einstiegsseiten enthalten das Logo oder das Siegel der Hochschule als eingebettete und meist in der oberen linken oder rechten Ecke des Dokuments angeordnete Grafik (vgl. Fußnote 16, S. 400).

Ein weiterer hochfrequenter Bestandteil ist das Lexem „Homepage“, das in 18 Einstiegsseiten benutzt wird.[8] Es können drei Verwendungstypen unterschieden werden: In neun Einstiegsseiten wird der Rezipient mit einer Begrüßungsfloskel empfangen: „Willkommen auf der Homepage von Mathias Reiser“ (HP 23; identisches Formulierungsmuster in HP 40,

[6] Cronin et al. (1998) ermitteln durch eine Analyse der Ergebnisse von Suchmaschinenanfragen, in denen fünf Namen von Wissenschaftlern recherchiert wurden, eine Hierarchie der Nennungstypen von Namen, die 11 Supertypen umfasst, zu denen auch die „Personal Home Page: The person's own page“ gehört. Die weiteren Typen lauten „Abstract“, „Article“, „Conference“, „Current Awareness“, „External Home Page“, „Listserv“, „Resource Guide“, „Book Review“, „Syllabus“ sowie „Table of Contents“ und enthalten jeweils bis zu neun Subtypen. Cronin et al. (1998, S. 1326) gehen davon aus, dass diese „proto-typology of invocations“ einen vorläufigen Charakter besitzt und modifiziert werden muss, sobald durch „new social practices“ im WWW „fresh categories“ entstehen. Die von Cronin et al. ermittelten „11 dimensions, or contexts, of a scholar's virtual presence“ können als Hypertext(knoten)typen bzw. Hypertext(knoten)sorten aufgefasst werden.

[7] Der Autor von HP 5 führt zusätzlich seinen vorgesetzten Hochschullehrer auf: „Assistent von Prof. Dr. Bernulf Kanitscheider (Philosophie der Naturwissenschaften)“. Eine derartige Nennung ist ebenfalls in HP 13 (innerhalb einer Liste von Hyperlinks: „Homepage von Professor Albrecht Beutelspacher“) sowie in HP 90 enthalten („Wissenschaftlicher Assistent am Institut Neuere deutsche Literatur und Medien Prof. Dr. Thomas Anz“).

[8] Der Verfasser von HP 61 benutzt eine alternative Bezeichnung: „Website von Manuel Conradi“.

HP 67, HP 70, HP 72, HP 96, HP 97), „Herzlich Willkommen auf der Homepage von Prof. Dr. Rainer Kessler“ (HP 82) und „Pax & bonum! Willkommen auf der homepage [sic] von Prof. Dr. Hans-Martin Barth“ (HP 49). Darüber hinaus wird die possessive Markierung („Ulrich Horstmanns Homepage“, HP 6, „Homepage von Albert Jeltsch“, HP 7, „Rolf Deubner's Homepage“, HP 21, „Homepage von Prof. Dr. rer. nat. Dipl.-Hist. Christoph Friedrich“, HP 63, „Dies ist die Homepage von Dr. Holger Garn“, HP 66, „Homepage von E. Grüner“, HP 71) sowie die schlichte Reihung („Stefan Künzell – Homepage“, HP 3, „Homepage Klaudia Seibel“, HP 14, „Homepage Renate Rausch“, HP 100) verwendet.

Ein Foto des Emittenten gehört in 54 Homepages ebenfalls zu den identifizierenden Informationen. Dieses wird in den meisten Fällen im oberen Bereichs der Einstiegsseite abgebildet (vgl. Abbildung 10.6, S. 459). Die formale Bandbreite der Darstellungen ist sehr umfangreich: Neben Porträtaufnahmen, die aller Wahrscheinlichkeit nach von ambitionierten Fotoamateuren (HP 2, HP 48) und professionellen Fotografen (HP 41, HP 74) angefertigt wurden, sind Passfotos (HP 3, HP 61, HP 84), Urlaubsschnappschüsse (HP 64), Ausschnitte von Gruppenbildern (HP 19) sowie Fotos des Produzenten am Arbeitsplatz (HP 22, HP 69) in der Stichprobe vertreten. Gerade Porträts der beiden erstgenannten Typen vermitteln dem Rezipienten in einigen grafisch schlicht und typografisch traditionell gestalteten Homepages den bereits angesprochenen Eindruck, dass eine inhaltliche und funktionale Verwandtschaft zwischen den Textsorten, die üblicherweise Bestandteile von Bewerbungsunterlagen sind und der Hypertextsorte *persönliche Homepage eines Wissenschaftlers* besteht. Eine besonders interessante Form der grafischen Identifizierung hat der Emittent von HP 6 gewählt: In der linken oberen Ecke der Einstiegsseite ist statt einer Fotografie eine digitalisierte Bleistifzeichnung enthalten, die eine Karikatur des Produzenten darstellt.

10.5.2 Einleitungstexte, Begrüßungen und Verabschiedungen

Es wurde bereits erwähnt, dass in keiner der 100 Homepages von Hochschulangehörigen E-Mail-ähnliche Textstrukturmuster in Texten eingesetzt werden, die in studentischen Homepages die Funktion besitzen, den Rezipienten zu begrüßen, den Produzenten vorzustellen und in das thematische Angebot der Homepage einzuführen. Während studentische Homepages nur sehr wenige konventionalisierte Hypertextsortenmodule umfassen, können in der vorliegenden Stichprobe mehrere hochfrequente Komponenten ermittelt werden. Vermutlich aus eben diesem Grund wird auf eine narrative Einführung in das Themenspektrum eines solchen Webangebots in nahezu allen Fällen verzichtet, da sie als redundant aufgefasst wird. Lediglich fünf Einstiegsseiten enthalten kurze Einleitungstexte wie z. B.:

> Ich bin Diplombiologe und arbeite im Fachgebiet Naturschutz an der Universität Marburg. Zusammen mit Prof. Dr. H. Plachter leite ich das Projekt „[...]“. Innerhalb dieses Projektes untersuche ich im Rahmen einer Doktorarbeit den durch die Beweidung herbeigeführten strukturellen und funktionalen Wandel und [...]. Ausgehend von dieser Startseite möchte ich mich und meine Arbeit vorstellen. [HP 61]

Der Verfasser geht in knapper Form auf seinen Beruf und das derzeit betreute Forschungsprojekt ein, d. h. der Text fungiert als einleitende Kurzzusammenfassung der innerhalb der Homepage thematisierten Inhalte. Drei weitere Einleitungstexte wurden bereits im Kontext der metadiskursiven Äußerungen diskutiert (vgl. Abschnitt 10.4.1): Sie können als Hinweise

auf den Umstand interpretiert werden, dass sich auch Hochschulangehörige, die eine persönliche Homepage anbieten, intensiv mit den Inhalten und Funktionen ihres Webangebots auseinander setzen.[9] Der fünfte Begrüßungstext ist in der Einstiegsseite von HP 94 enthalten, zwischen dem Namen und diesem Text befindet sich jedoch die Instanz eines weiteren Hypertextsortenmoduls. Der Autor begrüßt den Rezipienten und spricht ihn im Folgenden mit dem distanzierten „Sie" an. Der Text geht kurz auf die Inhalte der Homepage und die Aufgaben des Verfassers an der Philipps-Universität Marburg ein.

Es stellt sich nun die Frage, welche Inhalte die Autoren der 95 verbleibenden Homepages unmittelbar nach der Nennung ihres Namens anführen. Eine tabellarisch strukturierte Kurzbiografie präsentieren fünf Produzenten (HP 6, HP 41, HP 67, HP 91, HP 99), in jeweils einer Einstiegsseite wird das Arbeitsgebiet (HP 30), eine Liste aktueller Lehrveranstaltungen (HP 40), ein Porträtfoto (HP 44) oder – als einziges weiteres Hypertextsortenmodul der Einstiegsseite – eine Navigationshilfe angeboten (HP 12). Die institutionelle Funktion des Emittenten, die Anschrift sowie weiterführende Kontaktmöglichkeiten werden in insgesamt 87 Einstiegsseiten präsentiert (dies ist auch in HP 94 der Fall). Aufgrund der prominenten Platzierung kann davon ausgegangen werden, dass diesen Informationen von den Autoren ein besonderer Stellenwert zugeschrieben wird. Die Frequenzangaben belegen, dass diese Positionierung der beiden Hypertextsortenmodule als Konvention aufgefasst werden kann.

Zu den Bestandteilen E-Mail-ähnlicher Textstrukturmuster gehören auch Begrüßungen und Verabschiedungen. Die 100 Homepages enthalten insgesamt 15 Begrüßungsfloskeln, von denen einige bereits erwähnt wurden. Die verbleibenden Begrüßungen wurden von den Autoren unabhängig von der Nennung ihres Namens positioniert, weshalb sie als eigenständige Instanzen fungieren. Dementsprechend beinhalten sie eine umfassendere Bandbreite: „Willkommen!" (HP 38), „Willkommen auf meiner persönlichen Homepage" (HP 12), „Herzlich willkommen im dienstlichen Internet-Angebot von Dietmar Osthus" (HP 39). Der Autor von HP 41 ist Linguist und heißt den Rezipienten in insgesamt 11 verschiedenen Sprachen „Willkommen!" (vgl. Abbildung 10.6, S. 459). HP 46 enthält die Begrüßung „Elmar Altwasser [...] ... freut sich über Ihren Besuch auf seiner Homepage!" und als einzige Homepage eine Verabschiedungsfloskel.[10] Diese befindet sich neben einem Zugriffszähler und lautet „Vielen Dank für Ihr Interesse und auf Wiedersehen!" Anschließend folgen neben den Daten der Erstellung sowie der letzten Änderung verschiedene weitere Metainformationen. Auch bezüglich dieser einzigen Verabschiedungsfloskel stellt sich die Frage, ob in den 99 verbleibenden Homepages Elemente existieren, die – vor oder nach einer Fußzeile – am unteren Ende der Einstiegsseite platziert wurden und als eine Art indirekte Verabschiedung des Rezipienten fungieren. An dieser Position enthalten 38 Homepages Hyperlinks, die den Rezipienten zum zentralen Verzeichnis persönlicher Homepages oder zum Webauftritt der Organisationseinheit des Produzenten führen. In der Stichprobe können vier Realisierungstypen

[9] Hierbei handelt es sich um HP 8: „Diese Seite informiert über mich und meine Arbeit.", HP 11: „Auf dieser Seite möchte ich Ergebnisse meiner Arbeit vorstellen und allgemein zugänglich machen. Sie enthält zudem Informationen über meine Person [...]" und HP 23: „Sie können hier weniges über meinen Lebenslauf erfahren, die bescheidene Liste meiner Publikationen lesen und sich über meine Forschungsinteressen und Lehrveranstaltungen informieren." Weder in diesen vier Texten noch in anderen Teilen der 100 Homepages können heterogene Formen der Leseranrede belegt werden (vgl. Abschnitt 9.6.3).

[10] Die Begrüßungsfloskel wurde als isoliert positioniert gewertet, da sich zwischen dem Namen und der Begrüßung ein Foto des Autors befindet.

differenziert werden, die sich ausnahmslos auf Nominalphrasen oder Präpositionalphrasen beziehen. Sechs Autoren verweisen auf die Einstiegsseite des Webauftritts der übergeordneten Organisationseinheit, ohne jedoch deren Bezeichnung aufzuführen:[11]

1. Zurück zum Fachbereich [HP 16]
2. Zurück zur Homepage der Abteilung [HP 24]
3. zurück zur Fachgebietsseite [HP 83]
4. zur Institutsseite [HP 57]
5. Zum Institut hier: [*Foto des Instituts*] [HP 58]
6. Fachbereich-Homepage [HP 98]

Diesen Belege zufolge gehen die ersten drei Autoren davon aus, dass die Leser zunächst das Webangebot der übergeordneten Organisationseinheit rezipieren und von der dort präsentierten Liste der Mitarbeiter per Hyperlink zu der persönlichen Homepage gelangen. Die drei weiteren Hyperlinkanzeiger beziehen sich hingegen nicht auf den ursprünglichen Navigationspfad. In den als implizite Verabschiedungsfloskel oder auch als vom Produzenten angebotene Fortsetzung des Navigationspfades interpretierbaren Hyperlinkanzeigern fünf weiterer Homepages wird der Name der verknüpften Organisationseinheit genannt:

1. Zurück zur Homepage des Instituts für Romanische Philologie [HP 2]
2. zurück zu Betriebsinformatik [sic] [HP 17]
3. Homepage des Zentralen Entwicklungslabor für Elektronik [HP 69]
4. Institut für Geschichte der Pharmazie [HP 84]
5. Justus-Liebig-Universität Giessen [HP 21]

Auch bei der spezifischen Nennung der Bezeichnungen präsupponieren zwei Autoren, dass der Rezipient über die Webseiten der übergeordneten Organisationseinheit zu der persönlichen Homepage gelangt ist und bieten den korrespondierenden Navigationsrückweg an. Der Autor des fünften Hyperlinkanzeigers führt ausschließlich eine Verknüpfung zur Einstiegsseite des Webauftritts der Universität auf.[12] Im finalen Bereich der Einstiegsseiten von 16 Einstiegsseiten wird mehr als eine Organisationseinheit referenziert, in einigen Fällen werden mehrere Hierarchieebenen aufgeführt:

1. [*Teil einer Navigationshilfe:*] Uni Giessen Fachbereich Germanistik [HP 1]
2. Homepage der Justus-Liebig-Universität [HP 7]
   Homepage des Instituts für Biochemie, FB 08
3. Homepage der Universitaet Giessen [HP 13]
   Homepage des Fachbereichs Mathematik
   Homepage der KryptoAG [...]
4. • Homepage der Kernchemie Marburg [HP 81]
   • Homepage der Nuklearmedizin Marburg [*Teil einer Liste*]
5. Zur Homepage Anglistik/Amerikanistik [HP 93]
   Zur Homepage der Universität Marburg

[11] Die in den Listen unterstrichen dargestellten Bestandteile stellen Hyperlinkanzeiger dar. In den HTML-Dokumenten enthaltene Zeilenwechsel wurden nahezu ausnahmslos übernommen; bei sehr langen Hyperlinkanzeigern wurden aus typografischen Gründen Zeilenwechsel eingefügt.

[12] Diese Homepage besitzt einen sehr rudimentären Charakter und enthält lediglich den Namen des Autors, visuelle und textuelle „under construction"-Hinweise sowie mehrere Kontaktinformationen. Bei dem genannten Hyperlink handelt es sich um die einzige Verknüpfung in diesem HTML-Dokument.

6. Professor am Lehrstuhl Mathematische Optimierung an der Fakultät für Mathematik der TU Chemnitz. [HP 29]
7. Homepages H. Lobin, Arbeitsbereich Angewandte Sprachwissenschaft und Computerlinguistik, Universität Gießen [HP 9]
8. Philipps-Universität – Fachbereich Germanistik und Kunstwissenschaften – Institut für neuere Deutsche Literatur und Medien [HP 50, HP 53]
9. [Homepage Prof. Greiner] [Homepage AK Greiner] [FB Chemie] [Links AK Greiner] [HP 70]
10. [Institut für Soziologie] [Fachbereich 03] [Philipps-Universität Marburg] [HP 94]
11. Fachgebiet Systematische Theologie Fachbereich Evangelische Theologie [HP 72]
12. [*am linken Rand: stilisierte Abbildung des Hauptgebäudes*] Slavistisches Seminar der Universität Bonn [*am rechten Rand:*] Slavisches Institut der Universität zu Köln [*Siegel der Universität*] [HP 38]
13. [*am linken Rand:*] Zur Homepage der Universität [*am rechten Rand:*] Zur Homepage der Fakultät [HP 43]
14. zurück: • Abt. DaF • IGS (Institut für Germanistische Sprachwissenschaften) • FB 09 • Uni Marburg [*Liste*] [HP 80]
15. weiter zum Sonderforschungsbereich 434 „Erinnerungskulturen“ weiter zum Fachbereich 04 der JLU Giessen [HP 15]

Innerhalb dieses dritten Typs liegen bezüglich der Auswahl und Sequenzierung von Organisationseinheiten diverse Varietäten vor. Einige Rückverweise enthalten Hyperlinks zu den Einstiegsseiten der Hochschule sowie des Fachbereichs oder Instituts. Andere beziehen sich lediglich auf die unmittelbar über- oder beigeordneten Organisationseinheiten. Die Verweise werden primär entweder als Listen oder als sequenziell angeordnete Nominalphrasen dargestellt, wobei gelegentlich Kommata, eckige Klammern oder Gedankenstriche als Separatoren fungieren. In zwei Homepages werden zusätzlich eingebettete Grafiken verwendet. Besonders auffällig ist der Umstand, dass in diesem dritten Typ lediglich eine Gruppe von Hyperlinks vorliegt, die „zurück“ führen (HP 80). Interessant ist in diesem Zusammenhang die in HP 15 dargestellte Verknüpfungsgruppe, die den Rezipienten nicht „zurück“, sondern „weiter“ führt. Möglicherweise präsupponiert der Produzent, dass die Rezipienten sein Webangebot primär über Suchmaschinen auffinden, weshalb er ihnen zwei Navigationspfade zur weiterführenden Exploration der Webangebote seines Sondersforschungs- und Fachbereiches anbietet. Der vierte Realisierungstyp bezieht sich ausschließlich auf Homepages von Wissenschaftlern, die an der Philipps-Universität Marburg tätig sind:

1. Zurück zur Liste der Homepages [HP 55, HP 60, HP 64, HP 87]
2. Zurück zur Liste der Homepages Kardiovaskuläre Nuklearmedizin [HP 76]
3. Zurück zur Liste der Homepages [HP 45]
4. Zurück zur Liste der Homepages an der Philipps-Universität [HP 49]
5. Zur Liste der Homepages der Uni Marburg [HP 73]
6. Hier geht's zur Liste der Homepages der Professor/innen und Mitarbeiter/innen der Philipps-Universität Marburg [HP 67]
7. [*Teil einer Navigationshilfe:*] [*Icon*] Liste [*der Homepages*] [HP 86]

Insgesamt 10 Homepages enthalten einen Rückverweis zur „Liste der Homepages“. Die in obiger Darstellung an erster Position aufgeführte Variante befindet sich aller Wahrscheinlich-

keit nach in einer Vorlage, die bei der erstmaligen Einrichtung einer Homepage automatisch durch ein Skript in das Benutzerverzeichnis des Anwenders kopiert wird, woraufhin sie an die eigenen Bedürfnisse angepasst werden kann; die sieben dargestellten Varianten zeigen hierfür verschiedene Beispiele.[13] Es fällt jedoch auf, dass in *keiner* der verbleibenden 90 Homepages Rückverweise auf zentral administrierte Listen persönlicher Homepages enthalten sind, so dass die vom Rechenzentrum gewählte und in dem Template realisierte Empfehlung der Integration eines Hyperlinks zu eben dieser Liste als untypisch aufzufassen ist.

10.5.3 Kontaktinformationen

Alle 100 Homepages enthalten Kontaktinformationen, die in nahezu allen Fällen in typografisch abgesetzter Listenform unmittelbar in der Einstiegsseite präsentiert werden.[14] Einige Webangebote enthalten zusätzlich Kontaktinformationen zum Sekretariat sowie zu den wissenschaftlichen Mitarbeitern. In 99 der 100 Homepages geben die Emittenten ihre E-Mail-Adresse an, die in fast allen Fällen zusätzlich als Anzeiger eines `mailto:`-Hyperlinks fungiert.[15] Die Dienstanschrift ist in 90 Homepages enthalten, in acht Webangeboten wird zusätzlich eine gesonderte Post- oder Lieferanschrift genannt. Mit der Telefonnummer wird eine weitere Kontaktmöglichkeit, die nicht unmittelbar über das Internet verwendet werden kann, in 86 Homepages angeboten, eine Faxnummer ist in 66 Homepages enthalten.[16]

Die Kontaktinformationen werden in 14 Homepages durch eine Überschrift eingeleitet. Den Autoren der verbleibenden 86 Homepages erscheint eine derartige Kennzeichnung vermutlich redundant, da Rezipienten die Inhalte dieser Komponente aufgrund ihrer spezifischen Formate bereits auf den ersten Blick als Anschrift und Telefonnummer(n) identifizieren können. Die verwendeten Überschriften lauten „Adresse“ (HP 3, HP 24, HP 30, HP 66, HP 79), „Kontakt“ (HP 7, HP 23, HP 27, HP 35, HP 38), „Anschrift“ (HP 10, HP 26, HP 48) und „Adresse (Uni)“ (HP 73). Auch die einzelnen Komponenten dieses oftmals als einzelnes Hypertextmodul realisierten komplexen Hypertextsortenmoduls werden in einigen Fällen mit dem zugehörigen Informationstyp etikettiert. In der Regel werden lediglich potenziell mehrdeutige Kontaktinformationen mit einem solchen Etikett versehen, insbesondere die Telefon- und die Faxnummer. Interessanterweise wird auch die E-Mail-Adresse in sehr vielen Fällen explizit als solche markiert, obwohl das Format der korrespondierenden Zeichenfol-

[13] Die in der Stichprobe enthaltenen Homepages der anderen Universitäten beinhalten keine Hinweise auf den Einsatz derartiger Homepage-Templates.

[14] In vier Homepages werden die Kontaktinformationen in separaten HTML-Dokumenten angeboten, die über die Hyperlinkanzeiger „Kontakt“ (HP 38, HP 86), „Lebenslauf“ (HP 11) und „Anschrift“ (HP 12) erreichbar sind. In 89 Homepages sind detaillierte Kontaktinformationen in der Einstiegsseite enthalten.

[15] Neun Einstiegsseiten enthalten lediglich die E-Mail-Adresse des Produzenten, die entweder über den Hyperlinkanzeiger „Kontakt“, den Namen des Produzenten oder die E-Mail-Adresse selbst mit einem `mailto:`-Hyperlink verknüpft sind (HP 6, HP 11, HP 12, HP 40, HP 41, HP 44, HP 91, HP 92, HP 99).

[16] Abbildung 10.1 zeigt verschiedene Beispiele für Kontaktinformationen, die die beobachtete Bandbreite sowie die nachfolgend besprochenen Charakteristika aufzeigen: HP 10 enthält separate Hypertextmodule für die Kontaktinformationen der dargestellten Person sowie zweier Mitarbeiterinnen. HP 22 enthält sowohl private als auch dienstliche Kontaktinformationen. Sämtliche in HP 38 enthaltenen Kontaktinformationen sind etikettiert, sie umfassen zusätzlich Angaben zu Sprechstunden. HP 50 enthält eine Privatnummer. Dies gilt auch für HP 57, die zusätzlich die private Anschrift enthält. HP 94 umfasst äußerst typische Kontaktinformationen und ähnelt in der typografischen Aufbereitung einer Visitenkarte.

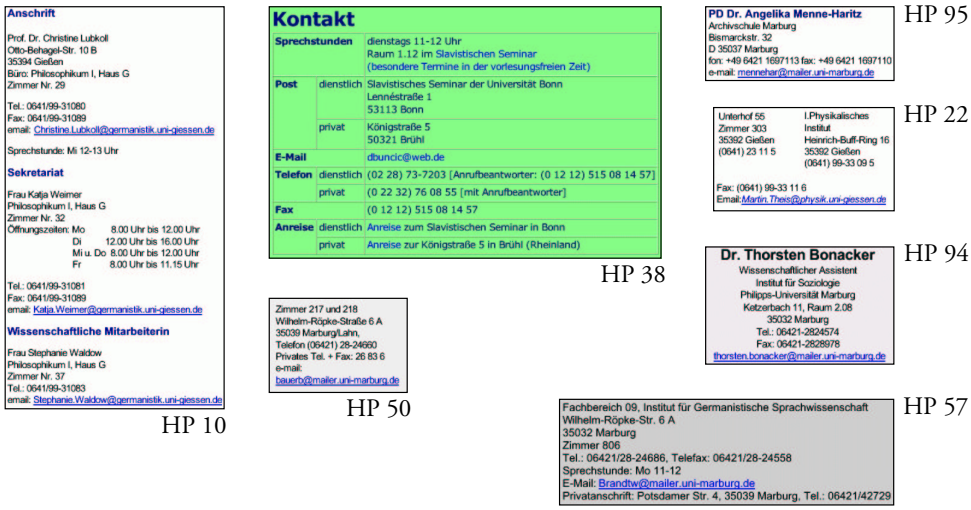


Abbildung 10.1: Beispiele für die Angabe von *Kontaktinformationen*

ge diese hinreichend identifiziert (`*@*`).[17] Eine vollständige Etikettierung aller Bestandteile kann in 17 Homepages beobachtet werden, wobei diese Markierung in sieben Einstiegsseiten durch korrespondierende Icons realisiert wird.

Mehr als 20 Homepages umfassen auch private Kontaktinformationen. Die Privatnummer wird in 22 Webangeboten aufgeführt, die private Anschrift ist in 18 Homepages enthalten.[18] Zusätzlich zur geschäftlichen wird in fünf Webpräsenzen eine private E-Mail-Adresse angeboten, z. B. „Kontakt (dienstlich)“ und „Kontakt (privat)“ (realisiert als `mailto:`-Hyperlinks) in HP 41 oder „Email (Uni): [...]“ und „Email (privat): [...]“ in HP 73. Die privaten Kontaktinformationen werden meist durch ein entsprechendes Etikett ausgewiesen.[19] In sechs Homepages werden beide Anschriften durch eine Überschrift markiert, z. B. „Dienstanschrift“ und „Privatanschrift“ (HP 8, HP 88), „Dienst“ und „Privat“ (HP 71), „dienstlich“ und „privat“ (HP 38, HP 48, vgl. Abbildung 5.3, S. 286), „Arbeit“ und „Privat“ (HP 51) sowie „dienstliche Anschrift“ und „private Anschrift“ (HP 82). In vier Fällen wird nur die Privatanschrift explizit als solche ausgewiesen (HP 26, HP 57, HP 62, HP 67).

[17] Die Etikettierung der E-Mail-Adresse erfolgt durch insgesamt 11 verschiedene Varianten: „e-mail“ (20 Homepages), „Email“ (17), „E-Mail“ (16), „email“ (10), „E-mail“ (7), „eMail“ (2), „E-Post“, „mail“, „E.mail“, „Email-Adresse“ und „E-Mail-Adresse“ (jeweils eine Homepage).

[18] Falls der Produzent die Privatanschrift angibt, wird in der Regel auch die private Telefonnummer angegeben. In einigen Fällen beschränkt sich die Angabe privater Kontaktinformationen auf die Telefonnummer. Die Zahl privater Anschriften erscheint überraschend hoch, doch enthalten auch viele gedruckte Vorlesungsverzeichnisse die Privatanschriften der Lehrenden. Aus diesem Grund besitzt die Vermeidung des Missbrauchs dieser Informationen bei einigen Produzenten möglicherweise nur einen sekundären Stellenwert.

[19] Die offiziellen Kontaktinformationen werden in fast allen Fällen zuerst aufgeführt. Eine Ausnahme stellt die Homepage eines Diplomanden dar (HP 22), der zunächst seine Privatanschrift und anschließend die Adresse des Instituts präsentiert, an dem er seine Examensarbeit schreibt (vgl. Abbildung 10.1). Der Autor von HP 46 besitzt zwar einen Lehrauftrag an der Universität Marburg, er ist dort jedoch nicht angestellt. Auch er führt zunächst seine „Privat“-Adresse und anschließend die „Büro“-Anschrift an. Während die Reihenfolge in diesen Fällen nachvollziehbar erscheint, muss HP 49 als untypisch bezeichnet werden: Der Autor ist Hochschullehrer und stellt zunächst die durch „privat“, und anschließend die mittels „dienstlich“ eingeleitete Anschrift dar.

In 27 Homepages geben die Produzenten ihre Sprechstunden an, die in nahezu allen Fällen ein Bestandteil der Kontaktinformationen sind.[20] Bis auf eine Ausnahme („Sprechzeiten" in HP 32) verwenden alle Autoren das Lexem „Sprechstunde" (bzw. „Sprechstunden") zur Einleitung dieser Komponente, anschließend folgt entweder ein Doppelpunkt und die Nennung von Tag und Uhrzeit (bzw. Beginn und Ende) oder eine tabellarische Darstellung mehrerer Termine (z. B. die Sprechzeiten in der vorlesungsfreien Zeit).[21] In drei Fällen wird das Semester angegeben, auf das sich diese Angaben beziehen, z. B. „Sprechstunde im Wintersemester 2000/2001: Mittwoch, 17.00 Uhr" (HP 18); in sieben Homepages wird zusätzlich der Raum genannt. Ein Hinweis, dass Termine ausschließlich oder zusätzlich „nach Vereinbarung" möglich sind, befindet sich in fünf Webangeboten. Es existieren lediglich vier Homepages, in denen von diesem standardisierten Schema der Darstellung von Sprechstunden abgewichen wird: In HP 94 wird zusätzlich der Termin einer „Mentorenstunde" aufgeführt. Der Autor von HP 38 stellt sämtliche Kontaktinformationen in einer Tabelle dar (vgl. Abbildung 10.1), die in der ersten Zeile die Sprechstunden enthält. In der rechten Tabellenzelle befindet sich die Angabe „dienstags 11-12 Uhr [...] Raum 1.12 im Slavistischen Seminar [...] (besondere Termine in der vorlesungsfreien Zeit)", wobei die in Klammern gesetzte Nominalphrase einen Hyperlinkanzeiger darstellt, der zu einem weiteren HTML-Dokument führt, das – ebenfalls in tabellarischer Form – sechs Termine enthält. Informationen zu Sprechstunden werden in HP 39 und HP 94 in eigenständigen HTML-Dokumenten angeboten, auf die jeweils von der Einstiegsseite über den Hyperlink „Sprechstunde" bzw. „Sprechstunden" verwiesen wird.[22]

Neben den Kontaktinformationen, die sich auf den Produzenten der Homepage beziehen, sind in insgesamt 15 Homepages Angaben zum Sekretariat und zu wissenschaftlichen Mitarbeitern enthalten. Die Telefonnummer des Sekretariats wird in sieben Einstiegsseiten innerhalb des Hypertextmoduls genannt, das die Kontaktinformationen des Emittenten umfasst (HP 9, HP 14, HP 20, HP 36, HP 43, HP 79, HP 100).[23] In fast allen Fällen wird dabei auch der Name sowie die E-Mail-Adresse der jeweiligen administrativen Mitarbeiterin bzw. des Mitarbeiters aufgeführt. In fünf Einstiegsseiten werden diese Informationen in separaten Komponenten genannt (HP 1, HP 10, HP 18, HP 47, HP 98). Der Produzent von HP 16 präsentiert ein eingebettetes HTML-Dokument, das die Namen, E-Mail-Adressen, die Anschrift, Telefon- und Faxnummer, die Öffnungszeiten und ein Foto seiner Sekretärinnen enthält. Fünf Einstiegsseiten enthalten mindestens ein separat positioniertes Hypertextmodul, das Kontaktinformationen eines oder mehrerer wissenschaftlicher Mitarbeiter umfasst (HP 10, HP 18, HP 47, HP 50, HP 53); die beiden letztgenannten Homepages enthalten lediglich die Namen der Mitarbeiter. Zwei Webangebote umfassen eingebettete HTML-Dokumente, die Namen, Fotos, E-Mail-Adressen und Telefonnummern der Mitarbeiter präsentieren (HP 16, HP 79). Durch die Aufnahme derartiger Informationen, die sich

[20] Interessanterweise stammen 24 dieser Angaben von Geisteswissenschaftlern. Eine Ausnahme bezieht sich auf die in Abschnitt 10.3 angesprochene Mitarbeiterin eines naturwissenschaftlichen Prüfungsamtes.

[21] In HP 93 wählt der Produzent eine zweisprachige Überschrift: „Sprechstunden/Office Hours".

[22] Der Autor von HP 94 nennt zunächst die Uhrzeit, die Straßenadresse und den Raum und gibt anschließend einen Hinweis: „Leistungsnachweise erhalten Sie nach persönlicher Besprechung der abgegebenen Hausarbeiten. An meiner Tür finden Sie einen Zettel mit den bereits gelesenen Arbeiten." Es ist unklar, weshalb der Produzent die Titel der „bereits gelesenen Arbeiten" nicht (zusätzlich) auf seiner Homepage veröffentlicht.

[23] Die Kontaktinformationen wissenschaftlicher Mitarbeiter werden innerhalb dieses Hypertextmoduls in keiner der 100 Homepages angeboten.

nicht primär auf die dargestellte Person, sondern z. B. die personelle Ausstattung der Organisationseinheit beziehen, nähert sich die jeweilige *persönliche Homepage eines Wissenschaftlers* der Hypertextsorte *Webauftritt einer Professur bzw. Arbeitsgruppe* an.

10.5.4 Lebenslauf und biografische Angaben

Insgesamt 60 Homepages enthalten biografische Angaben, deren Bandbreite von sehr rudimentären Informationen bis zu vollständigen Lebensläufen reicht (vgl. Heinemann, 2000a, S. 610 f.).[24] In 51 Einstiegsseiten verweisen in primären Navigationshilfen enthaltene Hyperlinks auf diese Informationen, die in zehn Homepages unmittelbar in der Einstiegsseite präsentiert werden. Zwei Homepages enthalten in der Einstiegsseite rudimentäre und in einem eingebetteten HTML-Dokument ausführliche Angaben (HP 41, HP 74).

Die Hyperlinkanzeiger, die zu biografischen Angaben führen, lauten in 13 Fällen „Lebenslauf“[25] (der Autor von HP 6 verwendet „Tabellarischer Lebenslauf und Ausbildungsgang“), der lateinische Ausdruck wird in 12 Webangeboten verwendet, wobei Unterschiede in der Schreibweise, der Bezeichnung selbst sowie dem intendierten Rezipientenkreis festzustellen sind: „Curriculum vitae“ (HP 7, HP 41), „Curriculum Vitae“ (HP 44, HP 79), „Curriculum Vita“ [sic] (HP 90), „Curriculum vitae/Biography“ (HP 84), „Curriculum vitae (deutsch)“ und „Curriculum vitae (english)“ (HP 74) sowie „Vita“ (HP 58, HP 64, HP 72, HP 78, HP 96). Hochfrequent sind auch mehrere Varianten eines deutschsprachigen Synonyms: „Biographie“ (HP 4, HP 40, HP 94, HP 98), „Kurzbiographie“ (HP 20, HP 43, HP 68), „Biographisches“ (HP 47, HP 53, HP 80). Einen eingeschränkten Skopus der dargestellten Informationen deutet der Hyperlinkanzeiger „Wissenschaftlicher Werdegang“ (HP 48, HP 91, HP 99, HP 100; „Werdegang“ wird in HP 50 und HP 92 verwendet)[26] an, wohingegen die Komponenten, auf die durch die Hyperlinkanzeiger „Zur Person“ (HP 54, HP 82, HP 89), „Personalia“ (HP 88) und „Persönliches“ (HP 8) verwiesen wird, vielschichtigere Informationen umfassen als traditionellerweise in einem Lebenslauf enthalten sind. Vier Linkanzeiger werden nur jeweils einmal verwendet: „Forschungsstationen“ (HP 51), „Wissenschaftliche Stationen“ (HP 12), „Profil“ (HP 18), „Elmar Altwasser“ (HP 46).[27]

Insgesamt können in Bezug auf die Textgestaltung und Strukturierung biografischer Angaben vier Typen unterschieden werden (vgl. Abbildung 10.2).[28] Hochfrequent ist der Einsatz meist sehr knapper Listendarstellungen (31 Homepages), in 16 Webangeboten werden detailliertere Listen präsentiert, die eine Binnenstrukturierung besitzen, d. h. einzelne Teile der Liste werden mit Überschriften etikettiert. Der dritte Typ bezieht sich auf sechs tabellarische

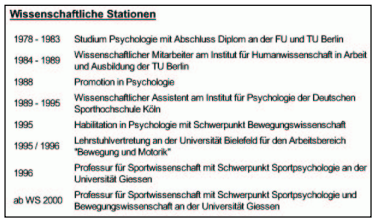
[24] In 76,5% aller Homepages von Hochschullehrern sind biografische Angaben enthalten. Dies gilt ebenfalls für 72,4% der Homepages promovierter Mitarbeiter und 66,7% der von Privatdozenten gepflegten Webpräsenzen.

[25] Der Produzent von HP 31 bietet in der Einstiegsseite seiner Homepage den Hyperlink „Meine private Seite“ an, die sich ebenfalls auf dem universitären Webserver befindet. Der Lebenslauf dieses Autor ist ausschließlich über die „private Seite“ zu erreichen.

[26] Auch der Titel „Akademische Biographie“ (HP 98) verdeutlicht die Restriktion auf wissenschaftliche Angaben.

[27] Die zuletzt genannte Variante wird in Abschnitt 4.6.3 diskutiert (vgl. Fußnote 115, S. 219). Es wurde bereits angesprochen, dass diese Homepage als untypisch zu bezeichnen ist. Das HTML-Dokument `vita.html` enthält nicht etwa ein Exemplar der Textsorte Lebenslauf, sondern einen Artikel über den Produzenten, der in einer Marburger Stadtzeitung anlässlich seines fünfzigsten Geburtstages erschienen ist.

[28] Die in separaten HTML-Dokumenten angebotenen Lebensläufe von HP 31 und HP 44 wurden bei der Datensammlung nicht in das Korpus aufgenommen und konnten bei dieser Analyse nicht berücksichtigt werden.

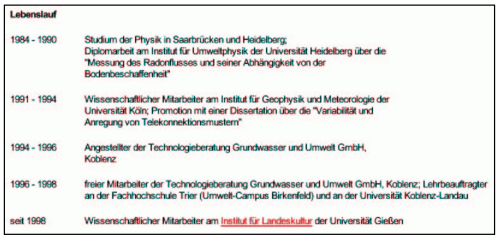


Wissenschaftliche Stationen

1978 - 1983	Studium Psychologie mit Abschluss Diplom an der FU und TU Berlin
1984 - 1989	Wissenschaftlicher Mitarbeiter am Institut für Humanwissenschaft in Arbeit und Ausbildung der TU Berlin
1988	Promotion in Psychologie
1989 - 1995	Wissenschaftlicher Assistent am Institut für Psychologie der Deutschen Sporthochschule Köln
1995	Habilitation in Psychologie mit Schwerpunkt Bewegungswissenschaft
1995 / 1996	Lehrstuhlvertretung an der Universität Bielefeld für den Arbeitsbereich "Bewegung und Motorik"
1996	Professur für Sportwissenschaft mit Schwerpunkt Sportpsychologie an der Universität Giessen
ab WS 2000	Professur für Sportwissenschaft mit Schwerpunkt Sportpsychologie und Bewegungswissenschaft an der Universität Giessen

HP 12

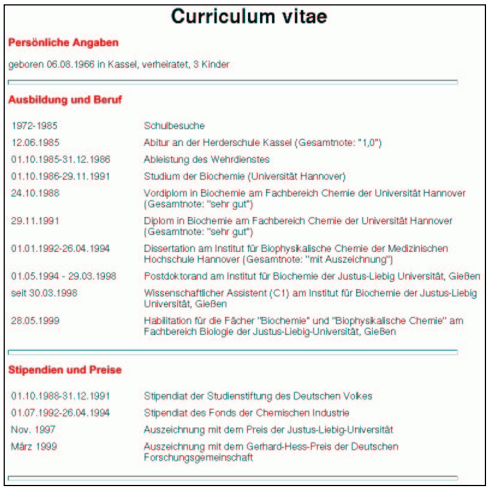
Typ 1: Liste



Lebenslauf

1984 - 1990	Studium der Physik in Saarbrücken und Heidelberg; Diplomarbeit am Institut für Umweltphysik der Universität Heidelberg über die "Messung des Radonflusses und seiner Abhängigkeit von der Bodenbeschaffenheit"
1991 - 1994	Wissenschaftlicher Mitarbeiter am Institut für Geophysik und Meteorologie der Universität Köln; Promotion mit einer Dissertation über die "Variabilität und Anregung von Telekonnektionsmustern"
1994 - 1996	Angestellter der Technologieberatung Grundwasser und Umwelt GmbH, Koblenz
1996 - 1998	freier Mitarbeiter der Technologieberatung Grundwasser und Umwelt GmbH, Koblenz; Lehrbeauftragter an der Fachhochschule Trier (Umwelt-Campus Birkenfeld) und an der Universität Koblenz-Landau
seit 1998	Wissenschaftlicher Mitarbeiter am Institut für Landeskultur der Universität Gießen

HP 25



Curriculum vitae

Persönliche Angaben

geboren 06.08.1966 in Kassel, verheiratet, 3 Kinder

Ausbildung und Beruf

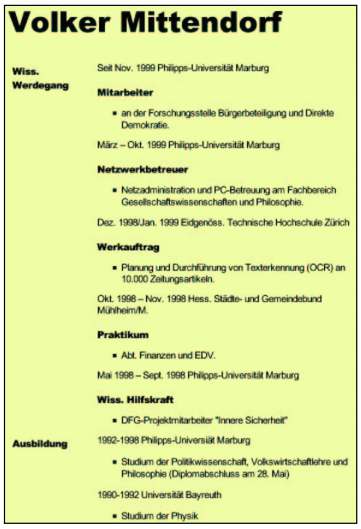
1972-1985	Schulbesuche
12.06.1985	Abitur an der Herderschule Kassel (Gesamtnote: "1,0")
01.10.1985-31.12.1986	Ableistung des Wehrdienstes
01.10.1986-29.11.1991	Studium der Biochemie (Universität Hannover)
24.10.1988	Vordiplom in Biochemie am Fachbereich Chemie der Universität Hannover (Gesamtnote: "sehr gut")
29.11.1991	Diplom in Biochemie am Fachbereich Chemie der Universität Hannover (Gesamtnote: "sehr gut")
01.01.1992-26.04.1994	Dissertation am Institut für Biophysikalische Chemie der Medizinischen Hochschule Hannover (Gesamtnote: "mit Auszeichnung")
01.05.1994 - 29.03.1998	Postdoktorand am Institut für Biochemie der Justus-Liebig Universität, Gießen
seit 30.03.1998	Wissenschaftlicher Assistent (C1) am Institut für Biochemie der Justus-Liebig Universität, Gießen
28.05.1999	Habilitation für die Fächer "Biochemie" und "Biophysikalische Chemie" am Fachbereich Biologie der Justus-Liebig-Universität, Gießen

Stipendien und Preise

01.10.1988-31.12.1991	Stipendiat der Studienstiftung des Deutschen Volkes
01.07.1992-26.04.1994	Stipendiat des Fonds der Chemischen Industrie
Nov. 1997	Auszeichnung mit dem Preis der Justus-Liebig-Universität
März 1999	Auszeichnung mit dem Gerhard-Hess-Preis der Deutschen Forschungsgemeinschaft

HP 7

Typ 2: Sortierte Liste



Volker Mittendorf

Wiss. Werdegang

Seit Nov. 1999 Philipps-Universität Marburg

Mitarbeiter

- an der Forschungsstelle Bürgerbeteiligung und Direkte Demokratie.

März – Okt. 1999 Philipps-Universität Marburg

Netzwerkbetreuer

- Netzadministration und PC-Betreuung am Fachbereich Gesellschaftswissenschaften und Philosophie.

Dez. 1998/Jan. 1999 Eidgenöss. Technische Hochschule Zürich

Werkauftrag

- Planung und Durchführung von Texterkennung (OCR) an 10.000 Zeitungsartikeln.

Okt. 1998 – Nov. 1998 Hess. Städte- und Gemeindebund Mühlheim/M.

Praktikum

- Abt. Finanzen und EDV.

Mai 1998 – Sept. 1998 Philipps-Universität Marburg

Wiss. Hilfskraft

- DFG-Projektmitarbeiter "Innere Sicherheit"

Ausbildung

1992-1998 Philipps-Universität Marburg

- Studium der Politikwissenschaft, Volkswirtschaftlehre und Philosophie (Diplomabschluss am 28. Mai)

1990-1992 Universität Bayreuth

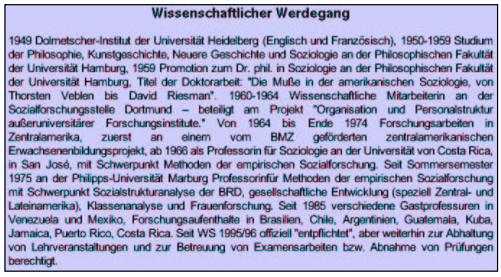
- Studium der Physik

HP 92 (Ausschnitt)

Kurzbiographie

Geboren am **7. März 1960** in Mannheim. **1978-1985** Studium der evang. Theologie und der Latinistik in Heidelberg und Lausanne; **1984** Staatsexamen; **1985** 1. theologisches Examen der Evang. Landeskirche in Baden; **1985/86** Graduate Visiting Student in Christ Church, Oxford; **1986/87** dasselbe in Trinity College, Cambridge; **1988** Promotion zum Dr. theol. in Heidelberg mit einer Dissertation über die Psalmenhomilien des Asterius; **1988-1992** Fellow of Peterhouse (Research Fellow), Cambridge; **1991** Habilitation in Historischer Theologie in Heidelberg mit einer Arbeit zum Fortschrittsdenken in der Alten Kirche; **1992-1995** Privatdozent in Heidelberg (Heisenberg-Stipendiat der DFG) und Fellow of King's College, Cambridge (Mitarbeit an einem interdisziplinären Forschungsprojekt zu den Ursprüngen des Christentums); **1. April 1996** Ernennung zum Professor (C4) für Kirchengeschichte an der Universität Bonn (Nachfolge Prof. Dr. Knut Schäferdiek).

HP 43

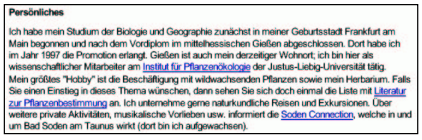


Wissenschaftlicher Werdegang

1949 Dolmetscher-Institut der Universität Heidelberg (Englisch und Französisch), 1950-1959 Studium der Philosophie, Kunstgeschichte, Neuere Geschichte und Soziologie an der Philosophischen Fakultät der Universität Hamburg, 1959 Promotion zum Dr. phil. in Soziologie an der Philosophischen Fakultät der Universität Hamburg, Titel der Doktorarbeit: "Die Muße in der amerikanischen Soziologie, von Thorsten Veblen bis David Riesman". 1960-1964 Wissenschaftliche Mitarbeiterin an der Sozialforschungsstelle Dortmund – beteiligt am Projekt "Organisation und Personalstruktur außeruniversitärer Forschungsinstitute." Von 1964 bis Ende 1974 Forschungsarbeiten in Zentralamerika, zuerst an einem vom BMZ geförderten zentralamerikanischen Erwachsenenbildungsprojekt, ab 1966 als Professorin für Soziologie an der Universität von Costa Rica, in San José, mit Schwerpunkt Methoden der empirischen Sozialforschung. Seit Sommersemester 1975 an der Philipps-Universität Marburg Professorinfür Methoden der empirischen Sozialforschung mit Schwerpunkt Sozialstrukturanalyse der BRD, gesellschaftliche Entwicklung (speziell Zentral- und Lateinamerika), Klassenanalyse und Frauenforschung. Seit 1985 verschiedene Gastprofessuren in Venezuela und Mexiko, Forschungsaufenthalte in Brasilien, Chile, Argentinien, Guatemala, Kuba, Jamaica, Puerto Rico, Costa Rica. Seit WS 1995/96 offiziell "entpflichtet", aber weiterhin zur Abhaltung von Lehrveranstaltungen und zur Betreuung von Examensarbeiten bzw. Abnahme von Prüfungen berechtigt.

HP 100

Typ 3: Als Fließtext formatierte Liste

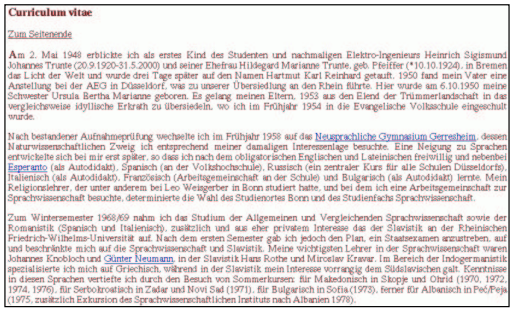


Persönliches

Ich habe mein Studium der Biologie und Geographie zunächst in meiner Geburtsstadt Frankfurt am Main begonnen und nach dem Vordiplom im mittelhessischen Gießen abgeschlossen. Dort habe ich im Jahr 1997 die Promotion erlangt. Gießen ist auch mein derzeitiger Wohnort; ich bin hier als wissenschaftlicher Mitarbeiter am Institut für Pflanzenökologie der Justus-Liebig-Universität tätig. Mein größtes "Hobby" ist die Beschäftigung mit wildwachsenden Pflanzen sowie mein Herbarium. Falls Sie einen Einstieg in dieses Thema wünschen, dann sehen Sie sich doch einmal die Liste mit Literatur zur Pflanzenbestimmung an. Ich unternehme gerne naturkundliche Reisen und Exkursionen. Über weitere private Aktivitäten, musikalische Vorlieben usw. informiert die Soden Connection, welche in und um Bad Soden am Taunus wirkt (dort bin ich aufgewachsen).

HP 8

Typ 4: Narrative Darstellung



Curriculum vitae

Zum Seitenende

Am 2. Mai 1948 erblickte ich als ersten Kind der Studenten und nachmaligen Elektro-Ingenieurs Heinrich Sigismund Johannes Trunte (20.9.1920-31.5.2000) und seiner Ehefrau Hildegard Marianne Trunte, geb. Pfeiffer (*10.10.1924), in Bremen das Licht der Welt und wurde drei Tage später auf den Namen Hartmut Karl Reinhard getauft. 1950 fand mein Vater eine Anstellung bei der AEG in Düsseldorf, was zu unserer Übersiedlung an den Rhein führte. Hier wurde am 6.10.1950 meine Schwester Ursula Bertha Marianne geboren. Es gelang meinen Eltern, 1953 aus den Elend der Trümmerlandschaft in das vergleichsweise idyllische Erfrath zu übersiedeln, wo ich im Frühjahr 1954 in die Evangelische Volksschule eingeschult wurde.

Nach bestandener Aufnahmeprüfung wechselte ich im Frühjahr 1958 auf das Neusprachliche Gymnasium Gerresheim, dessen Naturwissenschaftlichen Zweig ich entsprechend meiner damaligen Interessenlage besuchte. Eine Neigung zu Sprachen entwickelte sich bei mir erst später, so dass ich nach dem obligatorischen Englischen und Lateinischen freiwillig und nebenbei Esperanto (als Autodidakt), Spanisch (an der Volkshochschule), Russisch (ein zentraler Kurs für alle Schulen Düsseldorfs), Italienisch (als Autodidakt), Französisch (Arbeitsgemeinschaft an der Schule) und Bulgarisch (als Autodidakt) lernte. Mein Religionslehrer, der unter anderem bei Leo Weisgerber in Bonn studiert hatte, und bei dem ich eine Arbeitsgemeinschaft zur Sprachwissenschaft besuchte, determinierte die Wahl des Studienortes Bonn und des Studienfachs Sprachwissenschaft.

Zum Wintersemester 1968/69 nahm ich das Studium der Allgemeinen und Vergleichenden Sprachwissenschaft sowie der Romanistik (Spanisch und Italienisch), zusätzlich und aus eher privatem Interesse das der Slavistik an der Rheinischen Friedrich-Wilhelms-Universität auf. Nach dem ersten Semester gab ich jedoch den Plan, ein Staatsexamen anzustreben, auf und beschränkte mich auf die Sprachwissenschaft und Slavistik. Meine wichtigsten Lehrer in der Sprachwissenschaft waren Johannes Knobloch und Günter Neumann, in der Slavistik Hans Rothe und Miroslav Kravar. Im Bereich der Indogermanistik spezialisierte ich mich auf Griechisch, während in der Slavistik mein Interesse vorrangig dem Südslavischen galt. Kenntnisse in diesen Sprachen vertiefte ich durch den Besuch von Sommerkursen: für Makedonisch in Skopje und Ohrid (1970, 1972, 1974, 1976), für Serbokroatisch in Zadar und Novi Sad (1971), für Bulgarisch in Sofia (1973), ferner für Albanisch in Peć/Peja (1975, zusätzlich Exkursion des Sprachwissenschaftlichen Instituts nach Albanien 1978).

HP 41 (Ausschnitt)

Abbildung 10.2: Beispiele für die vier Typen *biografischer Informationen*

Lebensläufe, die als fortlaufender Fließtext formatiert sind. Narrative Lebensläufe stellen den vierten Typ dar, der in fünf Homepages verwendet wird. Abbildung 10.2 enthält jeweils zwei Beispiele dieser vier Typen.[29] Die Beispiele zeigen darüber hinaus, dass weiterführende Generalisierungen nur hinsichtlich sehr spezifischer Eigenschaften durchführbar sind, weshalb im Folgenden einige der sehr auffälligen und untypischen Charakteristika dargestellt werden.

Es wurde bereits angesprochen, dass bezüglich des Umfangs der biografischen Informationen eine sehr große Bandbreite vorliegt. Der Produzent von HP 24 präsentiert in der Einstiegsseite mehrere Abschnitte, die z. B. mit „Adresse", „Forschungsschwerpunkte" und „Auswahl von Veröffentlichungen" betitelt sind. Unmittelbar nach der Anschrift folgen zwei Abschnitte, die mit „Diplom" und „Promotion" überschrieben sind und lediglich das Jahr, in dem der Abschluss erworben wurde und die jeweilige Hochschule enthalten; bei der Dissertation wird zusätzlich das Thema genannt. Im Gegensatz hierzu stehen drei sehr umfangreiche Biografien, die zahlreiche Detailinformationen beinhalten: Beispielsweise gliedert die Autorin von HP 58 ihre „Vita" in die Abschnitte „Bildungsgang", „Berufserfahrung", „Feldforschungen", „Organisation/Leitung von Symposien", „Mitgliedschaft in Berufsverbänden" und „Universitäre Gremienarbeit". Ähnlich ausführliche Angaben macht der Autor von HP 74, der seine Informationen wie folgt strukturiert: „Berufliche Ausbildung und Praxis", „Wissenschaftlicher Werdegang", „Ämter und Funktionen in Kirche und Hochschule" und „Mitgliedschaften in akademischen Organisationen".[30] Die in HP 84 enthaltene Biografie ist ebenso detailliert und wird in zwei Sprachen angeboten: Die mittlere Spalte einer Tabelle enthält Jahreszahlen, während in der linken Spalte (rechtsbündig gesetzt) die deutschsprachige Version und in der rechten Spalte (linksbündig gesetzt) die englischsprachige Variante dargestellt wird. Der Autor von HP 85 bietet seinen Lebenslauf ebenfalls in zwei Sprachen an, diese werden jedoch sukzessive dargestellt („zur Person", „Curriculum Vitae"; ähnlich in HP 88); interessanterweise werden in diesen Varianten *unterschiedliche* Inhalte präsentiert. Der Emittent von HP 38 stellt in zwei HTML-Dokumenten seiner Homepage biografische Angaben dar. Die Einstiegsseite enthält eine primäre Navigationshilfe, deren Hyperlinks von knappen Beschreibungen flankiert werden, die Kurzzusammenfassungen der in den separaten Dokumenten diskutierten Themen darstellen. Der Kommentar des Hyperlinkanzeigers „Lebenslauf" kann somit als Kurzbiografie aufgefasst werden (ähnlich in HP 74). In mehreren Lebensläufen vermeiden Produzenten die Verwendung von Personalpronomina, die auf das „ich"-Verbot in tabellarischen Lebensläufen zurückzuführen ist (z. B. in HP 74, HP 80, HP 90). Eine Ausnahme stellt der in HP 63 enthaltene Lebenslauf dar, in dem der Autor in der dritten Person Singular über sich spricht („Christoph Friedrich wurde 1954 in Salzwedel/Altmark geboren. [...]"). Am Ende dieses Dokuments befindet sich ein `mailto:`-Hyperlink mit dem Hinweis „Mailversand an Prof. Friedrich". Dieser Kommentar weist darauf hin, dass die Homepage nicht von der dargestellten Person angefertigt wurde. Die un-

[29] In Abbildung 10.2 wurden zwei Einträge des in HP 92 enthaltenen Lebenslaufes nicht in den Bildschirmabzug übernommen. Der in HP 41 enthaltene Lebenslauf umfasst insgesamt 13 Textabsätze, von denen lediglich die ersten drei gezeigt werden. Die vier Typen können in Bezug auf nahezu alle Vorkommen biografischer Angaben als trennscharf bezeichnet werden, in HP 84 und HP 94 liegen jedoch Mischformen vor.

[30] Mitgliedschaften in Berufs- oder wissenschaftlichen Fachverbänden sind in insgesamt vier Biografien aufgeführt. In ebenso vielen Lebensläufen werden Angaben zur akademischen Gremientätigkeit gemacht. Forschungsschwerpunkte sind in zwei Lebensläufen enthalten (HP 2, HP 62).

gewöhnliche Präsentation des Lebenslaufs ähnelt den biografischen Angaben, die häufig in Sammelbänden und Zeitschriftenartikeln enthalten sind, Auskunft über den Hintergrund der einzelnen Beiträger geben und in der Regel von den Autoren selbst verfasst werden.[31]

10.5.5 Publikationsliste

In 71 der 100 Homepages sind Publikationslisten enthalten. Bezüglich dieser Komponente können umfassende Konventionen festgestellt werden. Diese beziehen sich zunächst auf die Hyperlinkanzeiger, die zu diesem Hypertextsortenmodul führen sowie auf die Überschriften der jeweiligen Abschnitte in der Einstiegsseite, falls es dort aufgeführt ist.[32] In 29 Fällen lautet das Etikett „Publikationen", zusätzlich werden 11 verschiedene Varianten verwendet (z. B. „Aktuelle Publikationen", „Publications", „Publikationsliste", „Publikationen und Vorträge", „Publikationsverzeichnis", „Zur Publikationsliste", „Elektronisch verfügbare Publikationen" und „Meine Publikationen"). Das Etikett „Veröffentlichungen" wird in 16 Homepages benutzt, und auch hier existieren acht Variationen (z. B. „Liste bisheriger Veröffentlichungen", „Vorträge und Veröffentlichungen", „Auswahl von Veröffentlichungen", „Links zu Veröffentlichungen" und „Verzeichnis der wissenschaftlichen Veröffentlichungen"). Die Bezeichnung „Schriftenverzeichnis" wird in drei Homepages verwendet (mit den Varianten „Schriftenverzeichnis & Herausgeberschaften" sowie „Download-Schriften"). Darüber hinaus werden die Überschriften „Original Articles and Books" sowie „Bibliographie" eingesetzt.

Die inhaltliche und strukturelle Bandbreite, die für die biografischen Informationen festgestellt wurde, trifft für die Publikationslisten nicht zu. Hinsichtlich der Nummerierung und Formatierung der Einträge bzw. Eintragstypen existieren jedoch diverse Unterschiede, die auf fachspezifische Konventionen zurückzuführen sind und an dieser Stelle nicht weiter berücksichtigt werden.[33] Bis auf zwei Ausnahmen enthalten alle Publikationslisten vollständige Literatureinträge, in einigen Fällen werden zusätzliche Informationen angeboten (z. B. Hyperlinks zu den Websites von Verlagen). Es existieren drei Strukturierungstypen: Eine Sortierung nach dem Typ der jeweiligen Publikation (Monografie, Zeitschriftenbeitrag, Rezension etc.) wird in 33 Listen angewendet, wobei Buchpublikationen in der Regel zu Beginn und Rezensionen (meist nur „in Auswahl") am Ende aufgeführt werden.[34] Eine in der Regel absteigende chronologische Sortierung nach dem Erscheinungsjahr wird in 24 Schriftenverzeichnissen benutzt. Der dritte Typ bezieht sich auf die Sortierung nach Themengebieten, die in vier Publikationslisten vorgenommen wird.[35] Zusätzlich existieren zwei Mischformen,

[31] Da die Publikationsliste des Produzenten deutlich mehr als 100 Einträge umfasst, kann davon ausgegangen werden, dass er sehr ausgeprägte Erfahrungen bei der Anfertigung derartiger biografischer Angaben besitzt.

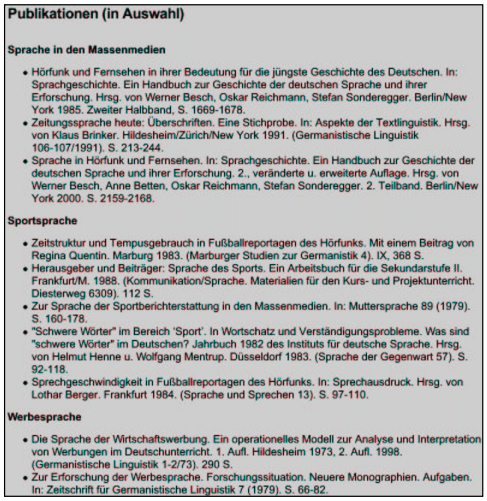
[32] Die Publikationsliste wird in 83,1% derjenigen Homepages, die dieses Hypertextsortenmodul enthalten, in einem eingebetteten HTML-Dokument präsentiert.

[33] Die meisten Autoren pflegen die Liste ihrer Veröffentlichungen vermutlich nicht in Form eines HTML-Dokuments, weshalb davon auszugehen ist, dass die fachspezifischen Formatierungskonventionen aus dem konvertierten Quelldokument ohne Änderung übernommen werden (vgl. die Abschnitte 3.3.6 und 4.3.2).

[34] In der Phase der automatisch durchgeführten Datensammlung wurden drei der HTML-Dokumente, die Publikationslisten umfassen, nicht in das Korpus integriert und konnten somit nicht in diese Statistik einfließen.

[35] Abbildung 10.3 stellt fünf beispielhafte Ausschnitte aus Publikationslisten dar. Die in HP 27 enthaltene Liste ist chronologisch sortiert und enthält zusätzlich Hyperlinks, die zu Postscript-Versionen einiger Beiträge führen. Die aus HP 57 stammende Aufstellung ist thematisch sortiert, wohingegen die Liste aus HP 44 nach dem Publikationstyp strukturiert ist.

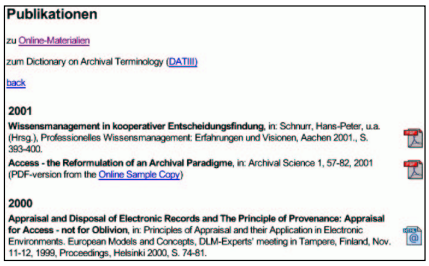
HP 27



HP 57

HP 44

HP 41



HP 95

Abbildung 10.3: Beispiele für die Angabe von *Publikationslisten*

z. B. die Sortierung nach Publikationstyp und Thema (HP 50). Eine weitere Ausnahme bezieht sich auf HP 38, in der die Publikationen des Autors primär nach ihrem Status sortiert sind: „Im Druck (bzw. in Vorbereitung)“ und „Nur online veröffentlicht“. Dieses Beispiel zeigt, dass von der Möglichkeit Gebrauch gemacht wird, Publikationen online anzubieten; drei Autoren stellen in ihren Veröffentlichungslisten ausschließlich diejenigen Artikel dar, deren Volltext online über die Homepage zugreifbar ist. Insgesamt bieten die Autoren von 16 Webangeboten einen oder mehrere wissenschaftliche Texte in digitaler Form an, in den meisten Fällen handelt es sich um kürzere Zeitschriften- oder Sammelbandbeiträge sowie Rezensionen (vgl. Abbildung 10.3).[36] Vollständige Bücher werden in keiner Homepage angeboten, doch stellen fünf Produzenten weiterführende Informationen zu Buchpublikationen zur Verfügung, z. B. das Umschlag- oder Titelbild, eine Errataliste, das Inhaltsverzeichnis oder ein Probekapitel. Vier Emittenten bieten Hyperlinks zu externen Dienstleistern an, bei denen eine Online-Bestellung vorgenommen werden kann.

[36] Lawrence (2001) zeigt, dass online publizierte wissenschaftliche Beiträge häufiger zitiert werden.

In drei Homepages wird die Publikationsliste in mehreren HTML-Dokumenten präsentiert. Es liegen eingebettete Hypertexte vor: Der Produzent von HP 41 führt in sieben Webseiten am oberen Rand jeweils drei oder vier Literatureinträge auf, Hyperlinks verweisen zu spezifischen Erläuterungen, Kommentaren oder Nachträgen, die im gleichen Dokument angeboten werden (diese sehr untypische Art der Aufbereitung wird als „Resümes [sic] und Inhaltsangaben zu den Publikationen“ bezeichnet). In den Knoten kann sich der Rezipient sequenziell mit Hilfe der Verknüpfungen „Weiterblättern“, „Zurückblättern“ sowie „Zum Seitenende“ und „Zum Seitenanfang“ bewegen (vgl. Abbildung 10.3).[37] In HP 84 werden für die Kategorien „Herausgebertätigkeit“, „Bücher“, „Aufsätze in Zeitschriften und Sammelbänden“, „Artikel in Lexika und Handbüchern“, „Rundfunkvorträge“ und „Rezensionen (in Auswahl)“ jeweils separate und teils sehr umfangreiche Dokumente angeboten. Drei Knoten werden in HP 88 verwendet, um die Liste von Veröffentlichungen in „Download-Schriften“, „Publikationen 1990–. . .“ und „Publikationen . . .–1989“ zu gliedern.

10.5.6 Forschungsschwerpunkte und Forschungsprojekte

In 50 Homepages führen die Produzenten ihre Forschungsinteressen auf, in 21 Fällen werden sie in separaten Dokumenten präsentiert. Auch dieses Hypertextsortenmodul unterliegt Konventionen, die sich – wie bereits diskutiert – auf die Etikettierung der entsprechenden Textabschnitte beziehen. Die vier Nomina „Forschung“ (insgesamt 33 Vorkommen), „Schwerpunkt“ (22), „Arbeit“ (13) und „Gebiet“ (16) werden, meist als Teil eines Kompositums, in allen Überschriften verwendet, z. B. „Forschungsschwerpunkte“ (11 Homepages), „Forschungsgebiete“ (3), „Arbeitsgebiete“ (3), „Forschungsschwerpunkte und Arbeitsgebiete“ (2), „Forschungsinteressen“ (2), „Forschung“, „Wissenschaftliche Arbeitsgebiete“, „Schwerpunkte in Forschung und Lehre“, „Schwerpunkte in der Forschung“, „Schwerpunkte“, „Meine Interessengebiete“ und „Ausgewählte Forschungsaktivitäten“ (jeweils einmal).

Die Forschungsinteressen werden in nahezu allen Homepages in Listenform dargestellt und bestehen vornehmlich aus Nominalphrasen. In der Regel wird diese Liste – unabhängig von anderen Komponenten – in der Einstiegsseite präsentiert, es existieren jedoch Ausnahmen, die sich auf die Integration dieses Hypertextsortenmoduls in die Publikationsslist e (vier Homepages) oder die biografischen Angaben beziehen (drei Homepages). In vier Webangeboten werden die Forschungsinteressen gemeinsam mit den Schwerpunkten in der Hochschullehre thematisiert. Die Autoren nehmen eine klare Trennung zwischen ihren wissenschaftlichen und privaten Aktivitäten vor, die in lediglich acht Homepages – an anderer Stelle als die beruflichen Interessengebiete – angerissen werden.

Spezifische Forschungsprojekte werden in 22 Homepages aufgeführt, doch liegen diesbezüglich zahlreiche Unterschiede vor, die auf die Unschärfe des von dem Lexem „Projekt“ bezeichneten Konzepts zurückzuführen sind. Die Konventionalisierung beschränkt sich auf die Existenz dieser Komponente sowie auf die Überschrift bzw. den korrespondierenden Hyperlinkanzeiger: In fünf Homepages wird „Forschungsprojekte“ verwendet, drei Webangebote benutzen „Projekte“, „Laufende Forschungsprojekte“ wird in zwei Webpräsenzen eingesetzt. Daneben können in jeweils einer Homepage Varianten wie z. B. „Projekte und Tagungen“,

[37] Die zwei letztgenannten Hyperlinks beziehen sich auf das erste bzw. letzte HTML-Dokument dieses monosequenziert organisierten Hypertextes.

„Laufende Projekte“, „Forschung und Projekte“ und „Aktuelles Forschungsvorhaben“ beobachtet werden. Neben diesen abstrakten Etiketten werden in drei Homepages die Titel von Projekten in Hyperlinkanzeigern der zentralen Navigationshilfe verwendet: „Projektseite EURIPIDES“ (HP 31), „EU-Projekt TRIAL-SOLUTION“ (HP 36) und „Mnemonik / Ars memorativa (DFG-Projekt)“ (HP 53). In HP 25 werden die Titel von Projekten im Inhaltsverzeichnis aufgeführt, ohne eine explizite Markierung als Projekt vorzunehmen, z. B. „Automatische Modellkalibrierung“ oder „Auswirkungen von Landnutzungsänderungen“, was für diese Hypertextsorte sehr untypisch ist. Einige Homepages enthalten in dieser Komponente lediglich Hyperlinks auf Angebote, die sich entweder auf den offiziellen Webseiten der zugehörigen Organisationseinheit oder auf externen Webservern befinden (z. B. in HP 40, HP 48).

Die Inhalte sind sehr unterschiedlich. In einigen Fällen werden nur knappe Listen aufgeführt, die Nominalphrasen als Charakterisierungen des Projektgegenstandes oder -ziels darstellen (z. B. HP 5, HP 11). Andere Aufstellungen beinhalten den Titel des Projekts, seine Laufzeit, den Drittmittelgeber, die Namen der wissenschaftlichen Mitarbeiter und von Kooperationspartnern, bereits publizierte Dokumentationen, untersuchte Fragestellungen und Zusammenfassungen (z. B. HP 4, HP 15; vgl. auch Abschnitt 12.7.4).[38] Der letzte Abschnitt in der Einstiegsseite von HP 91 ist zwar mit „Forschung und Projekte“ überschrieben, enthält jedoch keine Angaben zum zweiten Bestandteil dieses Titels. Untypisch ist auch die Webseite namens „Projekte“ in HP 94, die eine „to do“-Liste darstellt und auf einige geplante Vorhaben des Autors eingeht, z. B. die Anfertigung verschiedener Publikationen und einer Qualifikationsschrift sowie die wissenschaftliche Begleitung einer Tagung. Die Autorin von HP 100 stellt ihr „Aktuelles Forschungsvorhaben“ in narrativer Form dar.

10.5.7 Lehrveranstaltungen und Schwerpunkte in der Hochschullehre

In 51 Homepages bieten die Produzenten Informationen zu ihren Lehrveranstaltungen an. Auch bezüglich dieses Hypertextsortenmoduls können Konventionen ermittelt werden, die auf der Ebene des Hyperlinkanzeigers bzw. der Überschrift ansetzen: In 17 Homepages wird die Bezeichnung „Lehrveranstaltungen“ verwendet, neun Homepages beschränken sich auf „Lehre“. Weitere 16 Webangebote benutzen Varianten, die sich z. B. auf die Nennung des Semesters (HP 52 und HP 98: „Lehrveranstaltungen im Wintersemester 2001/2002“, HP 83: „Lehrveranstaltungen Wintersemester 2001/02“, HP 78: „Lehrveranstaltungen im Sommersemester 2001“, HP 50: „Lehrveranstaltungen im aktuellen Semester“), die Integration des Produzentennamens (HP 71: „Lehrveranstaltungen von Erwin Grüner“, HP 47: „Lehrveranstaltungen von Thomas Anz im Wintersemester 2001/2002“, HP 58: „Bettina E. Schmidt: Universitäre Lehre“) oder die Signalisierung eines umfassenderen Skopus der dargebotenen Informationen beziehen (HP 83: „Lehre, Forschung und Projekte“, HP 5: „Lehrveranstaltungen und wissenschaftliche Aktivitäten“, HP 61: „Lehre und Betreuung“).[39] Andere Autoren verwenden Veranstaltungstypen (z. B. HP 16: „Vorlesungen“, HP 92: „Seminare“, HP 45:

[38] Ein spezifischer Projekttitel und die Laufzeit eines Projekts werden in allen Fällen nur dann genannt, wenn es sich um ein drittmittelgefördertes Forschungsvorhaben handelt.

[39] Eine untypische Ausnahme stellen die in HP 82 innerhalb von Zwischenüberschriften verwendeten temporalen Deiktika dar, die von expliziten Semesterangaben flankiert werden: „laufendes Semester (SoSe 2001)“, „vergangenes Semester (WiSe 2000/2001)“ und „kommendes Semester (WiSe 2001/2002)“.

„Vorlesungen und Kurse“) oder synonyme Termini (HP 51: „Kurse“, HP 1: „Studienangebot“, HP 4: „Aktuelle Veranstaltungen“). Fast alle Übersichtsdarstellungen werden als Listen realisiert, wobei meist die Titel der Lehrveranstaltungen als Hyperlinkanzeiger dienen. Der Autor von HP 5 geht – ebenfalls in Form einer Liste – in vollständigen Sätzen auf seine Kurse ein, z. B. „Mit Ontologie beschäftigt sich ein Seminar im laufenden Wintersemester 2000/2001. Auf der entsprechenden Seite finden Sie Abstract [sic] zu den einzelnen Sitzung [sic] und (hoffentlich) wertvolle Informationen zum Thema.“

In Bezug auf den inhaltlichen Umfang können verschiedene Typen unterschieden werden. In 20 Homepages führen die Produzenten nur die organisatorischen Kerndaten ihrer Lehrveranstaltungen auf, die sich in den meisten Fällen auf den Titel, den Veranstaltungstyp, die Zeit und den Raum beschränken.[40] Die zweite Kategorie beinhaltet zusätzlich umfassendere Informationen (z. B. einen Kommentar, Literaturhinweise, einen Zeitplan, Angaben zur Arbeitsform oder die Voraussetzungen für den Erwerb eines Leistungsnachweises) und wird in 27 Homepages benutzt. Der erste Typ orientiert sich am tabellarischen Vorlesungsverzeichnis, in dem lediglich die Kerndaten aller Veranstaltungen einer Hochschule in fachneutraler Form aufgeführt werden. Das Vorbild des zweiten Typs ist hingegen zweifelsohne das kommentierte Vorlesungsverzeichnis, das in der Regel von Fachgebieten, Instituten oder Seminaren publiziert wird und fachspezifische Konventionen besitzt.[41] In sieben Homepages, die diesen Typ einsetzen, werden darüber hinaus digital verfügbare Materialien angeboten (insbesondere Foliensätze in den Formaten Postscript, PDF und *Powerpoint*). Derartige Dateien werden in den Angaben des ersten Typs nicht präsentiert. In vier Homepages wird ein dritter Typ verwendet, der lediglich einen Hyperlink zum Webangebot des zugehörigen Instituts oder der Universität enthält, wo das offizielle Vorlesungsverzeichnis zugreifbar ist; als Hyperlinkanzeiger wird typischerweise der Titel der Veranstaltung benutzt. Auf diese Weise kann es vermieden werden, die bereits innerhalb des offiziellen Webangebots verfügbaren Lehrveranstaltungsdaten in der eigenen Homepage zu reproduzieren (z. B. in HP 54).[42]

Bezüglich der Positionierung von Instanzen der beiden erstgenannten Typen existieren ebenfalls Konventionen. Der zweite Typ wird in 14 der 27 Fälle in einem und in zehn weiteren Fällen in mehreren separaten und hierarchisch organisierten HTML-Dokumenten realisiert, die in einigen Webangeboten in jeweils sehr unterschiedlicher und inkonsistenter Weise grafisch aufbereitet werden (z. B. in HP 93). In zwei Homepages werden die ausführlichen

[40] Diese Informationen werden in nahezu allen Homepages in neutraler, listenartiger und sehr knapper Form präsentiert. Die Ausnahmen beziehen sich auf zwei Hyperlinkanzeiger (HP 81: „zu meinen Unterrichtsmaterialien“, HP 93: „Materialien zu meinen Kursen“) sowie einen Kommentar in HP 7, der eine Liste externer Hyperlinks einleitet: „Zur Zeit bin ich an folgenden Vorlesungen (V) und Praktika (P) beteiligt: [...]“.

[41] Die Emittentin von HP 1 bezieht sich explizit auf den Umstand, dass die präsentierten Daten dem kommentierten Vorlesungsverzeichnis entnommen wurden. Der Autor von HP 52 bietet in einem separaten HTML-Dokument Informationen zu dem Kurs „Welttheater des 17. Jahrhunderts: Tragödie“ in Verbindung mit einer Quellenangabe an: „Auszug aus Kommentiertes Vorlesungsverzeichnis des FB 9, S. 28“.

[42] Der Autor von HP 41 stellt in einem mit „Lehrtätigkeit am Slavistischen Seminar der Universität“ überschriebenen Dokument eine sehr umfangreiche Liste thematisch sortierter Lehrveranstaltungen dar (z. B. „Kirchenslavischkurse“, „Kirchenslavische kursorische Lektüren und Übungen“ und „Westslavistische Übungen und Proseminare“), die bis ins Wintersemester 1980/1981 zurückreichen. Da hierbei lediglich das Semester und der jeweilige Veranstaltungstitel angegeben werden, ist davon auszugehen, dass der Verfasser mit diesem Dokument primär eine werbende Funktion verbindet. Ähnlich geht der Verfasser von HP 48 vor, der jedoch zu aktuelleren Veranstaltungen weiterführende Informationen anbietet.

Angaben in vernetzter Hypertextform präsentiert (HP 39, HP 40). Lediglich eine Homepage enthält detaillierte Angaben zu Lehrveranstaltungen in der Einstiegsseite (HP 47). An dieser Position wird der erste Typ in sieben der 20 Vorkommen platziert, die verbleibenden 13 Homepages, die die rudimentären Informationen in eingebetteten Webseiten enthalten, umfassen in der Regel die Aufstellungen aus mehreren Semestern, d. h. die Produzenten aktualisieren das korrespondierende Dokument offenbar vor Semesterbeginn, um den aktuellen Stand zu reflektieren. Nahezu alle Vorkommen von Lehrveranstaltungsinformationen in eingebetteten Dokumenten enthalten keine weiteren inhaltlich-thematisch markierten Hypertextsortenmodule. Eine Ausnahme stellt HP 58 dar, deren Autorin eingangs – in Form eines Auszugs ihrer Biografie – diejenigen Hochschulen aufführt, an denen sie tätig war bzw. ist, um daraufhin eine Liste aller absolvierten Lehrveranstaltungen zu präsentieren.

Abschnitt 10.5.6 ist auf die listenartige Angabe von Forschungsschwerpunkten eingegangen, die in 50 Homepages belegt werden können. Spezifische Angaben zu Schwerpunkten in der Lehre werden in lediglich vier Homepages aufgeführt. Der Autor von HP 2 geht in der Einstiegsseite seiner Homepage auf „Lehr- und Forschungsbegiete“ [sic] ein, die sich auf die „Spanisch- und portugiesischsprachige Literatur der iberischen Halbinsel und Lateinamerikas“ beziehen. In HP 17 werden unter „Lehre“ die Themen „Informatik, Unternehmensrechnung, Steuerlehre, Produktionsplanung, Betriebsanalyse“ genannt. In HP 26 gibt der Autor drei „Schwerpunkte in der Forschung“ an und nennt zuvor „Pro- und Hauptseminare in den Bereichen Sprachwissenschaft und ältere deutsche Literatur“ als „Schwerpunkte in der Hochschullehre“. Eine ähnliche Wortwahl findet sich mit „Schwerpunkte in Forschung und Lehre“ in HP 53, wo insgesamt acht verschiedene Bereiche aufgeführt werden. Keines dieser vier Webangebote enthält Angaben zu spezifischen Lehrveranstaltungen.

10.5.8 Hobbys und Hotlists

In acht Homepages gehen die Verfasser auf Hobbys und Freizeitaktivitäten ein, wobei zwei Ausprägungen unterschieden werden können: Falls Hobbys in der Einstiegsseite der Homepage angesprochen werden, geschieht dies zumeist knapp und beiläufig. Sofern sie jedoch in einem eingebetteten Knoten präsentiert werden, geht der Autor in detaillierter Form auf mehrere Interessengebiete ein. Nahezu allen Vertretern ist gemein, dass im Falle narrativer Darstellungen konsistent die erste Person Singular verwendet wird.

Zum ersten Typ gehört HP 45, in deren Einstiegsseite mehrere zentriert angeordnete Abschnitte enthalten sind, zu denen auch „Privates“ gehört. Die beiden Hyperlinkanzeiger „Promotionsumtrunk“ und „Starposter: Don Juan“ führen zu zwei privaten Schnappschüssen. Eine Aufteilung in einzelne Abschnitte wurde auch in der Einstiegsseite von HP 5 vorgenommen: In dem mit „And now for something completely different …“ überschriebenen letzten Abschnitt lädt der Produzent anlässlich der Fußballeuropameisterschaft 2000 „zu unserem legendären Tip-Spiel“ [sic] ein.[43] In dem mit „Persönliches“ überschriebenen ersten

[43] Die in der Überschrift genannte Phrase „And now for something completely different“ wurde in zahlreichen Episoden der Fernsehserie „Monty Python's Flying Circus“ des britischen Komikerensembles „Monty Python“ als Überleitung zwischen zwei Sketchen eingesetzt. Trotz des offensichtlichen popkulturellen Bezugs kann diese Überschrift als Hinweis auf den Umstand gewertet werden, dass dem Verfasser bewusst ist, dass der Inhalt dieses Abschnitts für die Hypertextsorte *persönliche Homepage eines Wissenschaftlers* untypisch ist.

Teil der Einstiegsseite von HP 8 geht der Autor auf seinen wissenschaftlichen Hintergrund ein, woraufhin er „die Beschäftigung mit wildwachsenden Pflanzen“ als sein „größtes »Hobby« “ bezeichnet. Für detailliertere Informationen über „weitere private Aktivitäten“ wird auf ein externes Webangebot verwiesen. In HP 22 werden ebenfalls „Persönliche Interessen“ aufgeführt, die verschiedene Sportarten, Literatur und Strategiespiele umfassen.

Der Autor von HP 11 stellt in dem mit der Überschrift „Meine Hobbies“ [sic] versehenen Dokument `hobbies.htm` zunächst in narrativer Weise allgemeine Interessen vor (Sport, Literatur, Kulinarisches) und präsentiert in dem Abschnitt „Musik, Musik, Musik“ eine Auswahl von fünf Eigenkompositionen, die als MP3-Dateien verfügbar sind. Der folgende Abschnitt „Der Zustand von Kultur und Zivilisation am Ende des 20. Jahrhunderts“ umfasst einen Liedtext, der vom Autor im Rahmen eines Kabarettprogramms aufgeführt wurde. Auf das Thema Musik geht auch HP 31 ein, in der der Verfasser innerhalb des Bereichs „Meine private Seite“ seine Leidenschaft für den Bau hochwertiger Verstärker schildert. Der Autor von HP 86 nutzt seine Homepage zur Präsentation von fünf Texten zu den Themen „Schreiben“, „Bahnfahrten“, „Die Bahn virtuell“, „Drei Minuten sind drei zuviel …“ und „Rollenspiel“. Die beiden Kurzgeschichten „Das Böse“ und „Die verträumte Mücke“ werden zwar in ähnlicher Form aufgelistet, sind jedoch nicht digital verfügbar. Die in HP 46 vorgenommene Diskussion privater Interessen in dem Dokument `hobbies.html` stellt einen Sonderfall dar: Unter der Überschrift „Elmar Altwasser's Hobbies [sic]“ befindet sich links ein Foto des Autors, und rechts der folgende listenartig präsentierte Fließtext:

> Falls er nicht seinem Haupt-Hobby, nämlich seiner Arbeit nachgeht, oder Schlagzeug spielt bei der Mick Schwarz Band oder mit seiner Freundin Siggi unterwegs ist und manchmal auch in ihrem Ferienhaus in Spanien Urlaub macht; [sic] ist er ein begeisterter Briefmarkensammler und sucht Tauschpartner, denn es fehlen ihm noch eine ganze Menge Marken – […]. [HP 46]

Es wird nicht die erste, sondern die dritte Person benutzt, weshalb auch dieser Text an die narrative Charakterisierung eines Autors erinnert, die sich oftmals am Ende von Sammelbänden oder Zeitschriftenartikeln befindet und typischerweise von der Person selbst verfasst wird (ähnlich in HP 63, dort jedoch in Form eines narrativen Lebenslaufes).[44] Die Darstellung von Hobbys und Freizeitinteressen stellt den einzigen übergreifenden Bereich innerhalb der 100 Homepages dar, in dem Merkmale für konzeptionelle Mündlichkeit ermittelt werden können. Diese beziehen sich jedoch nicht auf Signale, die aus Internet-basierten Kommunikationsdiensten stammen, sondern insbesondere auf die Lexik (z. B. „Na klar“) sowie den vorwiegend parataktischen und gelegentlich nicht grammatikalischen Satzbau.

Hotlists sind in insgesamt 11 Homepages vertreten und enthalten nahezu ausnahmslos Verweise auf Webangebote, die für die spezifischen Forschungsinteressen der Produzenten relevant sind.[45] Beispielsweise ist der Autor von HP 12 Sportpsychologe und bietet eine Liste

[44] Die erwähnte listenartige Darstellung wurde nicht reproduziert. Sie bezieht sich auf den Umstand, dass jeder Hyperlinkanzeiger als einzelnes Wort bzw. einzelne Phrase eines isolierten Absatzes präsentiert wird.

[45] Die entsprechenden Hyperlinkanzeiger und Überschriften besitzen eine umfangreiche Bandbreite. „Links“ (bzw. Varianten) wird in HP 6, HP 12, HP 25 und HP 86 verwendet. Der Autor von HP 22 benutzt „Hotlist“ als Überschrift, in HP 38 wird „Verweise“ verwendet. Der Produzent von HP 40 stellt drei „Favoriten“ dar, der Autor von HP 34 präsentiert „Meine Bookmarks“. Die in HP 15 enthaltene Liste ist mit „weitere nützliche Adressen im Internet …“ betitelt. Der Emittent von HP 46 offeriert „Elmar Altwasser's Lieblinkslinx“ [sic].

von Hyperlinks zu nationalen und internationalen Forschungsgruppen und Fachverbänden zu diesem Themengebiet an. Der Verfasser von HP 15 ist Geschichtswissenschaftler und präsentiert Verweise zu relevanten Zeitschriften, einem „Nachrichtendienst für Historiker“ sowie zum „English-Server [sic] für allgemeine Geschichte“. Die Ausnahme betrifft HP 45, in der der Verfasser zunächst „Medizinische Links“ und „Molekularbiologische Links“ aufführt, um daraufhin unter der Überschrift „Sonstige Links“ mehrere allgemeine Webangebote zu referenzieren (Suchmaschinen, Auktionshäuser, öffentliche Transportmittel, Zeitschriften etc.).

10.5.9 Alternative Versionen in anderen Sprachen

Zehn Homepages sind neben der deutschsprachigen Version auch in einer englischsprachigen Variante erhältlich, HP 38 liegt zusätzlich in einer russischsprachigen Version vor. Meist verweisen textuell realisierte Hyperlinkanzeiger, die in der linken oder rechten oberen Ecke positioniert sind auf diese Versionen; in drei Fällen wird zusätzlich die britische Flagge als Icon dargestellt. Als Hyperlinkanzeiger fungieren „This page in English“ (HP 70, HP 91, HP 99), „This page is also available in English.“ (HP 27, HP 35), „english version“ (HP 8), „English“ (HP 88), „Some informations [sic] in English“ (HP 3) sowie „Englische Version dieser Homepage“ (HP 7).[46] In der zentralen Navigationshilfe von HP 38 werden die drei Nomina „Deutsch“, „Russisch“ und „Englisch“ dargestellt, wobei die beiden in dem aktuellen Dokument jeweils nicht verwendeten Sprachen als Hyperlinks realisiert sind.

Der Umstand, dass nur 10% der Homepages in englischer Sprache angeboten werden, ist auf zwei wesentliche Faktoren zurückzuführen: Einerseits dürfte es vielen Autoren zu aufwändig erscheinen, neben der deutschsprachigen Homepage eine zweite Version in einer anderen Sprache zu pflegen.[47] Andererseits richtet sich ein – vermutlich geringerer – Teil der Autoren ausschließlich an Sprecher des Deutschen.

10.5.10 Dekorationsobjekte – Markierung von Traditionsbewusstsein

Grafische Dekorationsobjekte werden in den 100 Homepages vornehmlich unter funktionalen Gesichtspunkten eingesetzt (z. B. horizontale Trennlinien; vgl. Abbildung 10.6, S. 459). Auffällig ist eine Verwendungsweise, die als Markierung von Traditionsbewusstsein interpretiert werden kann und in fünf Homepages benutzt wird. In den Einstiegsseiten von HP 1, HP 50, HP 53, HP 54 und HP 89 stellen teils großformatige Grafiken digitialisierte Versionen von Gemälden, Holzschnitten und Fresken dar.[48] Diese Dekorationsobjekte dienen der Individualisierung, wobei es den Produzenten aller Wahrscheinlichkeit nach ein Anliegen ist, ein spezifisches Traditionsbewusstsein bezüglich der klassischen Medien zum Ausdruck zu bringen, da die Abbildungen unter anderem Bücher und Personen bei ihren Studien zeigen. Interessanterweise sind alle fünf Produzenten als Hochschullehrerinnen bzw. -lehrer oder wissenschaftliche Mitarbeiter in germanistischen Instituten tätig.

[46] Es ist auffällig, dass der zuletzt genannte Hyperlinkanzeiger in deutscher Sprache präsentiert wird. Rechts daneben wird jedoch zusätzlich die britische Flagge dargestellt.

[47] In HP 74, HP 84 und HP 85 liegt der Lebenslauf in englischer Sprache vor. Er wird vermutlich als Kern der persönlichen Homepage betrachtet, weshalb er einer umfassenderen Zielgruppe zur Verfügung gestellt wird.

[48] Die Hintergrundgrafiken in HP 5, HP 14, HP 20, HP 41 und HP 63 erinnern an Farbe und Textur von Büttenpapier und werden vermutlich mit einer ähnlichen Intention oder metaphorisch eingesetzt.

10.5.11 Metainformationen und Angaben zum tatsächlichen Produzenten

In 42 der 100 Homepages werden Angaben zum Datum der Erstellung bzw. der letzten Änderung des Webangebots gemacht; in den meisten Homepages beschränken sich die Verfasser auf die Hervorhebung des Modifikationsdatums. Es existieren verschiedene rekurrente Formulierungsmuster, die mit nur minimalen Variationen eingesetzt werden. In neun Einstiegsseiten wird diese Metainformation mit „zuletzt“ eingeleitet, z. B. „zuletzt geändert: 11.02.2000“ (HP 1; ähnlich in HP 5, HP 17, HP 39, HP 42), in einigen Fällen wird auch „zuletzt bearbeitet“ (HP 7), „zuletzt überarbeitet“ (HP 18) oder „zuletzt aktualisiert“ (HP 11) verwendet. Eine Variante bezieht sich auf die „letzte Änderung: 10.10.01“ (HP 94, ähnlich in HP 95) oder die „letzte Fassung: 10.9.2001“ (HP 85). Hervorzuheben ist HP 8, in der die „Letzte Aktualisierung *dieser Seite*: [...]“ (ähnlich in HP 24; meine Hevorhebung, G. R.) angegeben wird. Die Spezifizierung des Skopus der Datumsangabe erscheint zunächst redundant, verdeutlicht jedoch, dass das Problem der potenziellen Ambiguität dieser Metainformation erkannt wird, schließlich wird in den meisten Fällen nicht expliziert, ob sich die Angabe auf *alle* HTML-Dokumente des Webangebots oder nur diejenige Webseite bezieht, in der sie enthalten ist. Wenn nicht explizit angegeben wird, ob es sich um das Datum der Erstellung oder das der letzten Modifikation handelt, wird die Interpretation erschwert (z. B. in HP 31: „27. Mai 1999“; ähnlich in HP 12, HP 15, HP 34, HP 35, HP 79). In 11 Homepages führen die Produzenten zusätzlich ihren Namen an, der in einigen Fällen als `mailto:`-Hyperlink realisiert wird, z. B. „Zuletzt bearbeitet: 9. Januar 2001 von Albert Jeltsch“ (HP 7), „C. Schröder 12.10.2000“ (HP 44), „Rolf Haftmann, 22.9.2000“ (HP 36) oder „© Dr. Nikolaos Trunte. Zuletzt aktualisiert am 4.7.2001.“ (HP 41). Einige der Homepages, die auf dem Webserver der Universität Marburg angeboten werden, basieren auf einer gemeinsamen Vorlage. Hierauf beruht vermutlich auch die hochfrequente Angabe des Datums der letzten Änderung als „Stand: 06. September 2001“ (HP 84), die in dieser Form in 13 Homepages enthalten ist. In fünf dieser Homepages werden untypische Positionierungen dieses universalen Hypertextsortenmoduls benutzt, das in der Regel den Abschluss der Einstiegsseite bildet: Vier Homepages bieten diese Metainformation unterhalb der Kontaktinformationen zu *Beginn* des Einstiegsdokuments an, in einem weiteren Dokument befindet es sich – ähnlich einem Privatbrief – in der rechten oberen Ecke. Auf die Kommunikationsform Brief geht vermutlich auch die Angabe „Dipl.-Ing. Tobias Lucas, Chemnitz, d. 17.11.2000“ (HP 32) zurück.

Zugriffszähler werden in neun Homepages eingesetzt und in sieben Fällen zusammen mit dem Datum der letzten Änderung in einem Hypertextmodul realisiert. Konventionen können weder im Hinblick auf die Sequenzierung dieser Hypertextsortenmodule noch in Bezug auf die lexikalische Realisierung ermittelt werden. Neben einer englischsprachigen Version („This page was visited [...] times.“, HP 33) werden elliptische Konstruktionen („[...] Zugriffe auf dieses Dokument.“, HP 34), vollständige Sätze („Sie waren mein [...] Besucher seit dem 1. August 1999.“, HP 46) und allgemeinere Informationen verwendet („[...] Zugriffe seit 17.01.97“, HP 35, „[...] Besucher auf diesen Seiten.“, HP 84).

Abschnitt 3.5.3 ist auf den Umstand eingegangen, dass persönliche Homepages nicht notwendigerweise von der Person angefertigt werden, die von dem Webangebot dargestellt wird. Einige Homepages der Stichprobe enthalten eine explizite Nennung der Person (bzw. einer

Gruppe von Personen), die die Homepage erstellt hat und die mit der vorgestellten Person *nicht* übereinstimmt; in insgesamt 19 Homepages sind Hinweise dafür enthalten, dass sie nicht von der präsentierten Person angefertigt wurden.[49] Dieser Umstand wird meist dadurch deutlich, dass als verantwortliche Person ein anderer Name genannt wird, z. B. „webmaster: Thomas.Eckhardt@[...].de“ (HP 1; Name der dargestellten Person: Swantje Ehlers), „(c) MTW 99“ (HP 6; Ulrich Horstmann), „Anmerkungen an: Ulrich Ott“ (HP 24; Peter Kirsch), „(Stand 17.11.01) Betreuung der Seite durch C. Adam“ (HP 79; Peter Janich) und „(c) Andrea Schützenmeister/Andreas Staets – letzte Fassung: 10.9.2001“ (HP 85; Wolfgang Krieger). In vier weiteren Homepages wird ausführlich auf die Ersteller eingegangen: „Diese Homepage wurde von Dr. Eckart Fuchs unter Mitwirkung von Franz-Josef Hanke installiert und von Siegrid Schmeer programmiert und gestaltet. Wir danken dem HRZ-Marburg für die zur Verfügungstellung seiner Ressourcen.“ (HP 46) sowie „Gestaltung durch: Projektgruppe Pädagogische Psychologie Sommersemester 1996 Mark Apfelbaum, [...], Dr. Sylwia Wilberg – Beratung: Jan-Georg Wildegans[,] Thomas Wilberg“ (HP 91, ähnlich in HP 99). In HP 16 (Wolf-Dietrich Walker) befindet sich der Hyperlinkanzeiger „Info“, der zu einem HTML-Dokument führt, das den Text „Erstellt im November 1998. [*Zeilenumbruch*] Autor: Volker Bittner! [*Zeilenumbruch*] Zurück zur Homepage“ enthält. Derartige explizite Hinweise können in zehn Homepages belegt werden, d. h. die meist von studentischen Hilfskräften oder wissenschaftlichen Mitarbeitern durchgeführte Erstellung des Webangebots wird von den präsentierten Personen in ähnlicher Weise anerkannt wie vergleichbare unterstützende Korrektur- oder Textsatzarbeiten, die ebenfalls von diesem Personenkreis im Kontext der Veröffentlichung von Monografien oder Sammelbänden erledigt werden. In aller Regel werden diese jedoch im Vorwort einer Buchpublikation ausführlicher thematisiert und sind mit einer expliziten Danksagung verbunden.

10.6 Fazit – Das Hypertextsortenprofil

Die 100 Homepages unterscheiden sich grundlegend von den in Kapitel 9 untersuchten Dokumenten. Während in den studentischen Webangeboten hinsichtlich der inhaltlichen und sprachlichen Gestaltung auf Spontaneität, Authentizität, Individualität und Dialogizität Wert gelegt wird, können diese Merkmale – ebenso wie Merkmale konzeptioneller Mündlichkeit – in den Homepages von Wissenschaftlern nur in wenigen Fällen belegt werden. Diesbezüglich ist der Trend zu beobachten, dass die genannten Eigenschaften vornehmlich in den Homepages von Angehörigen des akademischen Mittelbaus (insbesondere wissenschaft-

[49] Neben der expliziten Nennung des Produzentennamens kann die HTML-Quelle mit dem Namen der dargestellten Person verglichen werden: Viele HTML-Editoren hinterlassen in einem `meta`-Element den Namen der Person, die das Werkzeug verwendet hat oder für die diese Software lizenziert worden ist. In 11 Homepages wird innerhalb des Quelltextes ein anderer Name als der der präsentierten Person angegeben. Weiterhin existiert auch die Möglichkeit, dass der Autor die vollständige HTML-Quelle einer anderen Homepage kopiert und als Grundlage der eigenen Homepage eingesetzt hat, ohne die `meta`-Elemente zu ändern. In vier der 11 Homepages (HP 1, HP 16, HP 17, HP 46) befindet sich eine explizite Äußerung, dass das Angebot nicht von der präsentierten Person angefertigt wurde. Von den 19 Homepages stellen 11 eine Hochschullehrerin bzw. einen Hochschullehrer und fünf einen promovierten Mitarbeiter vor. Allen 19 Homepages ist gemein, dass sie im Vergleich mit anderen Dokumenten der Stichprobe nur selten über gestalterische oder inhaltliche Elemente der Individualisierung verfügen und demzufolge einen allgemein „offizielleren“ Charakter besitzen.

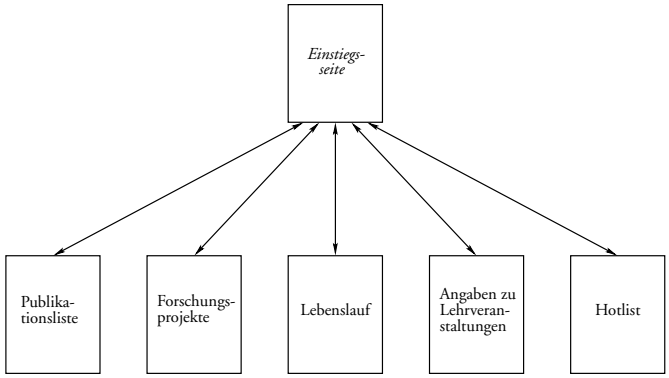


Abbildung 10.4: Typische Sequenzierung der *persönlichen Homepage eines Wissenschaftlers*

lichen Mitarbeitern) beobachtet werden können, wohingegen nahezu alle Webangebote von Hochschullehrern einen formalen, offiziellen, seriösen und durchkomponierten Charakter besitzen, der simultan manifestierte Elemente des persönlichen Ausdrucks umfasst (z. B. bezüglich der Markierung von Traditionsbewusstsein). Beiden Stichproben ist gemein, dass die HTML-Dokumente der enthaltenen Homepages vornehmlich hierarchisch strukturiert werden; eingebettete Hypertexte werden nur von sehr wenigen Produzenten realisiert. Abbildung 10.4 stellt die typische hierarchische Hypertextsequenzierung schematisch dar.[50]

Im Gegensatz zu den studentischen Homepages können zahlreiche, teils sehr ausgeprägte Konventionen (vgl. Abschnitt 10.5) und spezifische Funktionen ermittelt werden, die die Emittenten mit unterschiedlichen Hypertextsortenmodulen ihrer Webangebote verbinden (vgl. Tabelle 10.3 und Abbildung 10.5 für ein Beispiel).[51] Hierbei handelt es sich um die Vorstellung der eigenen Person, die in *allen* 100 Homepages durch die typografisch hervorgehobene Nennung des Namens sowie in 54 Fällen durch ein begleitendes *Foto* realisiert wird.[52] Auch *Kontaktinformationen* werden in allen Homepages angeboten, wobei die *E-Mail-Adresse*, die *Straßenadresse*, *Telefon-* und *Faxnummer* sowie die *Zimmer- bzw. Büronummer* hochfrequente Komponenten darstellen.[53] Die Emittenten von 85 Homepages streben die Etablierung eines *wissenschaftlichen Profils* an, indem z. B. eine *Publikationsliste*, persönliche *Forschungsinteressen*, *Forschungsprojekte* und *Mitgliedschaften in Fachverbänden* thematisiert werden, um den Produzenten als aktives und innovatives Mitglied der wissenschaftlichen Community zu positionieren. Mit Hilfe zweier weiterer komplexer Hypertextsortenmodule signalisieren 79 bzw. 59 Emittenten, dass sie Angehörige einer spezifischen Hochschule

[50] Die bidirektionalen Kanten deuten an, dass die Einstiegsseite eine primäre Navigationshilfe umfasst, die zu den einzelnen Knoten führt, die wiederum einen Rückverweis zur Einstiegsseite enthalten.

[51] Abbildung 10.5 stellt HP 42 innerhalb der Oberfläche der Korpusdatenbank dar und zeigt komplexe Hypertextsortenmodule in Fettdruck und ihre Konstituenten in normaler Schrift (vgl. auch Abschnitt 9.7).

[52] Das Hypertextsortenmodul *Identifikation* besteht aus einzelnen Komponenten, von denen der *Name des Homepage-Besitzers* den zentralen Baustein darstellt. Lediglich die *Angabe des akademischen Titels* ist spezifisch für die Hypertextsorte *persönliche Homepage eines Wissenschaftlers*, alle weiteren Komponenten können auch in anderen Ausprägungen des abstrakten Hypertexttyps *Homepage einer Person* ausgeprägt sein.

[53] Da die Privatanschrift in den meisten Fällen als Bestandteil der Dienstanschrift realisiert wird, werden *keine* separaten Hypertextsortenmodule angenommen (vgl. Abbildung 5.3, S. 286.

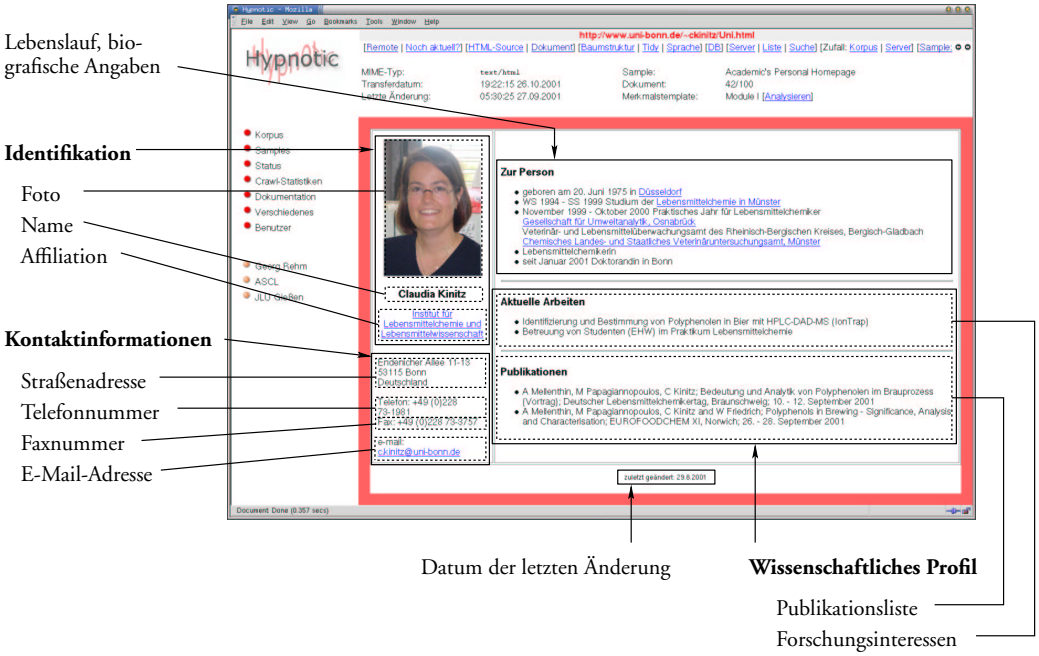


Abbildung 10.5: Hochfrequente Hypertextsortenmodule am Beispiel von HP 42

bzw. einer konkreten universitären Organisationseinheit sind. Insgesamt 54 Produzenten intendieren, mit Hilfe ihrer Homepage ein *universitäres Profil* zu etablieren, was unter anderem durch die *Angabe von Lehrveranstaltungen* und ausgeübten *Funktionen* geschieht (z. B. Gremienarbeit). Da sich der *Lebenslauf bzw. biografische Angaben* (60 Vorkommen) sowohl auf das wissenschaftliche als auch auf das universitäre Profil beziehen können, wird dieser Baustein als eigenständiges Hypertextsortenmodul aufgefasst. Es zeigt sich, dass die von Storrer (1999b, S. 6) angenommenen Funktionen der beruflichen Homepage zu kurz greifen: Den 100 Emittenten geht es nicht nur darum, „Funktionen, Kompetenzen und Zuständigkeiten abzustecken“ und sich „in institutionellen [...] Hierarchien einzuordnen“, vielmehr liegen spezifischere Funktionen vor. Bezüglich der Gestaltung vertritt Storrer die Ansicht, dass berufliche Homepages häufig auf dem Design der einbettenden Website basieren, was im Hinblick auf die 100 Homepages ebenfalls nur in Ausnahmefällen belegt werden kann.[54]

Oberflächlich betrachtet dienen die gennanten Hypertextsortenmodule zwar der Selbstdarstellung – insbesondere hinsichtlich verschiedener Facetten der beruflichen Rolle – eines Individuums, es müssen jedoch Differenzierungen vorgenommen werden, die sich unter anderem auf die intendierte Leserschaft beziehen (vgl. Abschnitt 10.4): Informationen zu Lehrveranstaltungen werden primär für die Teilnehmer der Kurse publiziert (die dem Produzenten – abhängig von Veranstaltungstyp und -größe – persönlich bekannt sind). Angaben zu Publi-

[54] Einige an der Universität Marburg angebotene Homepages basieren offenbar auf einer vom Rechenzentrum dieser Hochschule angebotenen Vorlage. Weitere Homepages ähneln einander so deutlich, dass davon ausgegangen werden muss, dass sie entweder eine gemeinsame Vorlage besitzen oder die eine Homepage auf der Grundlage des anderen Webangebots entstanden ist (dies gilt für HP 50 und HP 53 sowie HP 52 und HP 64).

Bezeichnung des Hypertextsortenmoduls	Ebene	Primärer Typ	Sekundärer Typ	Status	Verwendung	Typische Ausprägung	Frequenz
Identifikation	komplex	Inhalt/Thema		generell	obligatorisch	Einstiegsseite	100
Name des Homepage-Besitzers	atomar	Inhalt/Thema		generell	obligatorisch	Einstiegsseite	100
… begleitet vom akademischen Titel	atomar	Inhalt/Thema		spezifisch	obligatorisch	Einstiegsseite	69
… begleitet von einem Foto	atomar	Dekoration	Inhalt/Thema	generell	obligatorisch	Einstiegsseite	54
… begleitet von der Affiliation	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	34
… begleitet von einer Tätigkeitsangabe	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	27
Kontaktinformationen	komplex	Inhalt/Thema	Kommunikation	generell	obligatorisch	Einstiegsseite	100
E-Mail-Adresse	atomar	Kommunikation	Interaktion	generell	obligatorisch	Einstiegsseite	99
Straßenadresse	atomar	Inhalt/Thema		generell	obligatorisch	Einstiegsseite	90
Telefonnummer	atomar	Inhalt/Thema		generell	obligatorisch	Einstiegsseite	86
Faxnummer	atomar	Inhalt/Thema		generell	obligatorisch	Einstiegsseite	66
Zimmer- bzw. Büronummer	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	30
Sprechstunde	atomar	Inhalt/Thema		spezifisch	optional	Einstiegsseite	27
Telefonnummer (privat)	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	22
Adresse (privat)	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	18
Explizite Postadresse	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	8
Faxnummer (privat)	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	7
Telefonnummer (Sekretariat)	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	7
E-Mail-Adresse (privat)	atomar	Kommunikation	Interaktion	generell	optional	Einstiegsseite	5
URL der persönlichen Homepage	atomar	Navigation		generell	optional	Einstiegsseite	4
Mobiltelefonnummer (privat)	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	3
Informationen zur Anreise	atomar	Inhalt/Thema		generell	optional	intern	2
PGP-Key bzw. -Fingerprint	atomar	Interaktion	Kommunikation	generell	optional	intern	2
URL der privaten Homepage	atomar	Navigation		generell	optional	Einstiegsseite	2
X.500-Eintrag	atomar	Navigation	Interaktion	generell	optional	Einstiegsseite	2
SMS senden	atomar	Kommunikation	Interaktion	generell	optional	intern	1
Wissenschaftliches Profil	komplex	Inhalt/Thema		spezifisch	obligatorisch	intern	85
Publikationsliste	atomar	Inhalt/Thema	Textstrukturmuster	spezifisch	obligatorisch	intern	71
Forschungsinteressen	atomar	Inhalt/Thema		spezifisch	obligatorisch	Einstiegsseite	50
Forschungsprojekte	atomar	Inhalt/Thema		spezifisch	optional	intern	22
Prominent platzierte Bücher	atomar	Inhalt/Thema		spezifisch	optional	Einstiegsseite	7
Vorträge/Präsentationen	atomar	Inhalt/Thema	Textstrukturmuster	spezifisch	optional	intern	5
Mitgliedschaft in Fachverbänden	atomar	Inhalt/Thema		spezifisch	optional	intern	4
Technologietransfer	atomar	Inhalt/Thema		spezifisch	optional	intern	2
Eigenständige Affiliation	komplex	Inhalt/Thema		generell	obligatorisch	Einstiegsseite	79
Name der Universität (als Text)	atomar	Inhalt/Thema		spezifisch	obligatorisch	Einstiegsseite	75
Logo bzw. Siegel der Universität	atomar	Inhalt/Thema		spezifisch	optional	Einstiegsseite	16
Lebenslauf, biografische Angaben	atomar	Inhalt/Thema	Textstrukturmuster	generell	obligatorisch	intern	60
Relevante Links	komplex	Navigation	Inhalt/Thema	generell	obligatorisch	Einstiegsseite	59
… zum eigenen Institut/Arbeitsbereich	atomar	Navigation	Inhalt/Thema	spezifisch	optional	Einstiegsseite	49
… zur Einstiegsseite der Universität	atomar	Navigation	Inhalt/Thema	spezifisch	optional	Einstiegsseite	36
… zum eigenen Fachbereich	atomar	Navigation	Inhalt/Thema	spezifisch	optional	Einstiegsseite	23
Universitäres Profil	komplex	Inhalt/Thema		spezifisch	obligatorisch	intern	54
Angaben zu Lehrveranstaltungen	atomar	Inhalt/Thema	Textstrukturmuster	spezifisch	obligatorisch	intern	51
Funktionen (z. B. Gremienarbeit)	atomar	Inhalt/Thema		spezifisch	optional	intern	7
Allgemeine Studienhinweise	atomar	Inhalt/Thema		spezifisch	optional	intern	3
Angebotene Abschlussarbeiten	atomar	Inhalt/Thema		spezifisch	optional	intern	2
Datum der letzten Änderung	atomar	Metainformation		universal	optional	Einstiegsseite	42
Explizite Begrüßung	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	15
Hotlist (Liste externer Hyperlinks)	atomar	Navigation	Inhalt/Thema	generell	optional	intern	11
Alternative Version in anderer Sprache	atomar	Inhalt/Thema	Navigation	generell	optional	Einstiegsseite	10
Zugriffszähler	atomar	Metainformation		universal	optional	Einstiegsseite	9
Kontaktinformationen (Sekretariat)	komplex	Inhalt/Thema	Kommunikation	spezifisch	optional	Einstiegsseite	8
Name	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	8
E-Mail-Adresse	atomar	Kommunikation	Interaktion	generell	optional	Einstiegsseite	6
Faxnummer	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	6
Telefonnummer	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	6
Öffnungszeiten	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	5
Zimmer- bzw. Büronummer	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	4
Straßenadresse	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	3
Kontaktinformationen (Mitarbeiter)	komplex	Inhalt/Thema	Kommunikation	spezifisch	optional	Einstiegsseite	7
Name	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	7
Auflistung mehrerer Einträge	atomar	Textstrukturmuster		generell	optional	Einstiegsseite	6
Telefonnummer	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	4
E-Mail-Adresse	atomar	Kommunikation	Interaktion	generell	optional	Einstiegsseite	4
Zimmer- bzw. Büronummer	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	3
Namen von Hilfskräften	atomar	Inhalt/Thema		spezifisch	optional	Einstiegsseite	2
Straßenadresse	atomar	Inhalt/Thema		generell	optional	Einstiegsseite	2
Gästebuch	atomar	Kommunikation	Interaktion	universal	optional	intern	1

Tabelle 10.3: Die Hypertextsortenmodule der *persönlichen Homepage eines Wissenschaftlers*

kationen und Forschungsinteressen richten sich an Wissenschaftler. Kontaktinformationen werden sowohl für Angehörige der eigenen Institution als auch für interessierte Mitarbeiter anderer Hochschulen sowie für Gäste angeboten. Informationen über Aktivitäten im Bereich des Technologietransfers sind an die Privatwirtschaft gerichtet. Eine besonders ausführliche Form der Angabe lehrveranstaltungsbezogener Informationen kann als „digitaler Handapparat" bezeichnet werden: Dieser umfasst nicht nur die grundlegenden Informationen zu Kursen (Studiengang, Titel, Ort, Zeit, Zielgruppe, Kommentar etc.), sondern auch Lehrmaterialien wie z. B. Skripte, Foliensätze, Artikel, Aufgabenzettel, Lösungen und Hyperlinks auf externe Webangebote, die für die Kursteilnehmer und interessierte Dritte online zugreifbar sind.[55] Eine der Funktionen persönlicher Homepages von Wissenschaftlern ist demnach auch die Bereitstellung digitaler Lehr- und Lernmaterialien, die – im Vergleich zu einem traditionellen, in der Bibliothek aufbewahrten Handapparat – von den Dozenten schneller publiziert und gepflegt werden und auf die Studierende effizienter zugreifen können.[56]

In Anlehnung an die Definition von de Saint-Georges (1998, siehe S. 225) kann die Hypertextsorte auf Basis der Analyse – ebenso wie die *private Homepage eines Studierenden* – als subgenerische Variante der „personal home page" charakterisiert werden (vgl. Tabelle 10.3): Die *persönliche Homepage eines Wissenschaftlers* präsentiert Inhalte in hypertextueller Form, die von einem Wissenschaftler, der an einer Universität oder einer vergleichbaren Institution tätig ist, erstellt oder in seinem Auftrag angefertigt wurde und (i) den Emittenten durch einen Namen und möglicherweise ein Foto identifiziert sowie die Affiliation zur Institution zum Ausdruck bringt, (ii) Kontaktinformationen zur Verfügung stellt, (iii) die aktuellen (sowie optional auch die abgeschlossenen) Forschungsaktivitäten und -interessen in Form eines wissenschaftlichen Profils thematisiert, das unter anderem eine Publikationsliste umfasst, (iv) einen Lebenslauf enthält und (v) hochschulbezogene Aktivitäten und Aufgaben vorstellt (z. B. Gremienarbeit). Zu den Funktionen dieser Hypertextsorte gehört unter anderem die Etablierung von wissenschaftlichen und universitären Profilen und das Bereitstellen von Kontaktinformationen, digital verfügbaren Publikationen und Lehrveranstaltungsmaterialien.

In der Stichprobe studentischer Homepages sind sowohl rudimentäre als auch sehr umfangreiche Hypertextexemplare enthalten, wobei die zuerst genannten als Ausprägung des prototypischen Kerns des Hypertexttyps *Homepage einer Person* aufgefasst werden können (vgl. Abschnitt 9.7). Derartige unterschiedlich umfangreiche Ausprägungen können auch in der hier untersuchten Stichprobe beobachtet werden. Abbildung 10.6 umfasst Beispiele für drei Typen von Realisierungsformen: Die Bildschirmabzüge zeigen sechs Einstiegsseiten, die vollständig reproduziert wurden. Über die enthaltenen Hyperlinks kann auf den Umfang der Webangebote geschlossen werden (vgl. auch die Tabellen 10.1 und 10.2). Das erste Beispiel

[55] Für die weitere Entwicklung der Hypertextsorte kann prognostiziert werden, dass Materialien zu Lehrveranstaltungen immer seltener auf den persönlichen Homepages der Lehrenden bereitgestellt werden, da für diesen Zweck in jüngster Zeit an vielen Hochschulen E-Learning-Management-Systeme eingerichtet wurden, die die Pflege eines derartigen Informationsangebot in zentral administrierten Plattformen ermöglichen.

[56] Adamzik (2004, S. 93) weist darauf hin, dass Bibliografien und Rezensionsorganen eine werbende Funktion zukommt. Dies gilt ebenfalls für im WWW veröffentlichte Inhalte wie z. B. Publikationslisten, digitale Versionen von Aufsätzen und Lehrmaterialien. So sorgt etwa die bloße Existenz des Titels einer neuen Veröffentlichung in einem HTML-Dokument einer persönlichen Homepage dafür, dass bei der Aktualisierung des Datenbestandes einer Suchmaschine eben dieser Titel weltweiten Recherchen zur Verfügung steht und somit – abhängig von der Übereinstimmung zwischen Suchanfrage und Publikationstitel – als relevanter Treffer zurückgeliefert wird.

HP 33

HP 19

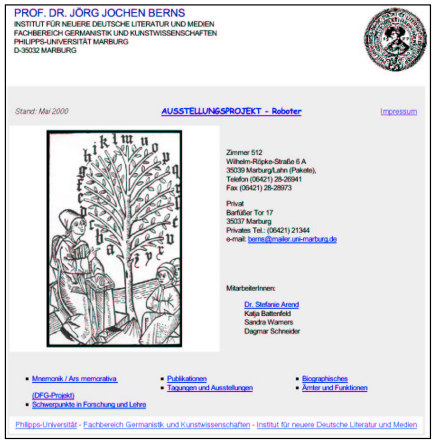
rudimentäre Ausprägung

HP 49

typische Ausprägung

HP 52

HP 53



HP 41

ausführliche Ausprägung

Abbildung 10.6: Ausprägungen *persönlicher Homepages von Wissenschaftlern* (Beispiele)

für eine rudimentäre Ausprägung (HP 33) umfasst lediglich den Namen und verschiedene Kontaktinformationen, das zweite Beispiel (HP 19) enthält darüber hinaus ein Foto und eine Publikationsliste. Die typische Ausprägung beinhaltet neben diesen Hypertextsortenmodulen einen Lebenslauf, weiterführende Hyperlinks und Angaben zu Lehrveranstaltungen (HP 52; vgl. auch Tabelle 10.3). Die ausführliche Ausprägung umfasst weitere Hypertextsortenmodule, z. B. Kontaktinformationen der Mitarbeiter oder Informationen zu Forschungsprojekten.

Auf Basis der Analysen kann eine Typologie konstruiert werden (vgl. Abbildung 10.7). Der prototypische Kern des Hypertexttyps *Homepage einer Person* entspricht dem Kern der Definition von de Saint-Georges (1998, S. 76; vgl. Abschnitt 9.7).[57] Die beiden Subtypen entspre-

[57] Die *Homepage einer Person* ist – ebenso wie z. B. *Pressetext* – als (hypothetischer) übergeordneter Hypertexttyp bzw. als eine Hypertextklasse aufzufassen, die als gemeinsame Einordnungsinstanz fungiert.

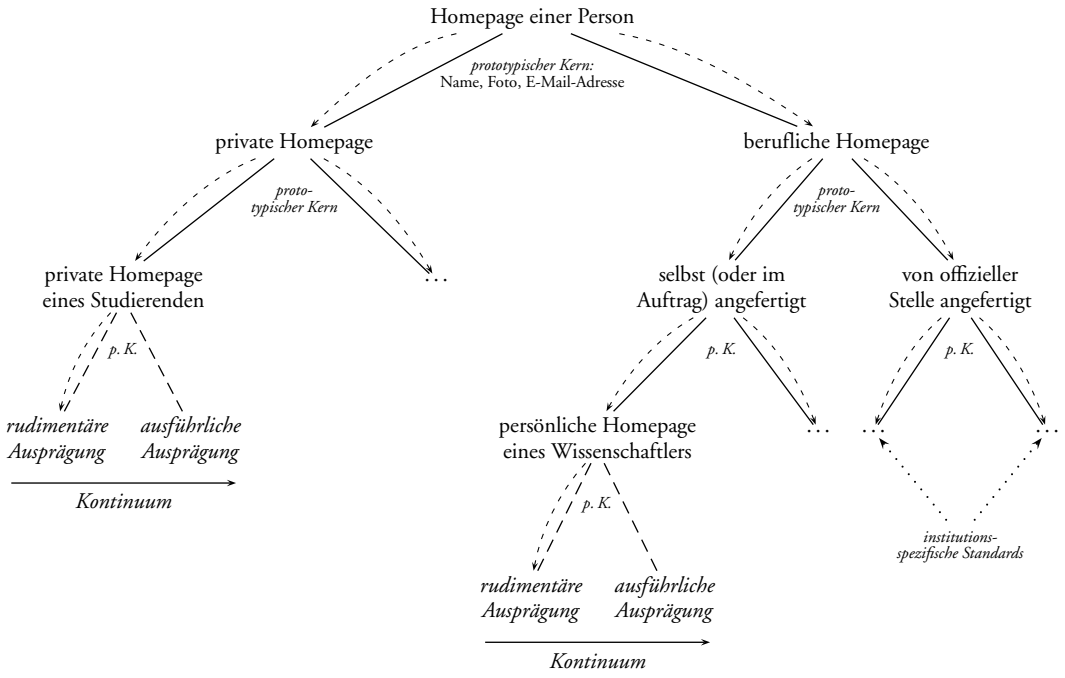


Abbildung 10.7: Typologie des Hypertexttyps *Homepage einer Person*

chen den Hypertextsorten *private Homepage* (thematisiert private Interessen des Emittenten) und *berufliche Homepage* (thematisiert die berufliche Rolle), die wiederum Hypertextsortenvarianten enthalten, deren prototypische Kerne den ermittelten Hypertextsortenprofilen entnommen werden können (vgl. die Tabellen 9.9 und 10.3), d. h. die jeweils übergeordneten prototypischen Kerne werden an die Subtypen vererbt und von diesen mit spezifischeren Belegungen angereichert. Für die Hypertextsorte *berufliche Homepage* wird eine Differenzierung angenommen in Exemplare, die von der dargestellten Person selbst angefertigt wurden und Exemplare, die von einer offiziellen Stelle (z. B. einer Abteilung für Öffentlichkeitsarbeit) realisiert wurden.[58] Letztere basieren auf institutionsspezifischen Standards, die in der Regel nicht modifiziert werden können. Hypertextsorten wie die *private Homepage eines Studierenden* oder die *persönliche Homepage eines Wissenschaftlers* besitzen unterschiedlich ausführliche Manifestationen. Diese können mit empirischen Mitteln verifiziert und auf einem Kontinuum verortet werden, das sich zwischen den Polen der rudimentären und der ausführlichen oder auch ›vollständigen‹ Ausprägung aufspannt, wobei jeweils die rudimentäre Ausprägung der Instanziierung des korrespondierenden prototypischen Kerns entspricht.

[58] Es existiert zumindest ein weiterer Subtyp (in der Abbildung nicht dargestellt), der Homepages umfasst, die im Auftrag von Personen des öffentlichen Lebens angefertigt werden (z. B. Künstler, Politiker und Schauspieler).

11 Analyse 4: Die Einstiegsseite des Webauftritts einer Universität

11.1 Einleitung

Die vierte Analyse betrifft eine zweistufige Untersuchung der Einstiegsseiten universitärer Webauftritte mit dem Ziel der Erstellung eines Profils dieser Hypertextknotensorte sowie der Identifizierung und Sammlung der Hypertextknotensorten derjenigen HTML-Dokumente, auf die von den Einstiegsseiten verwiesen wird.

Im Folgenden werden zunächst die detaillierten Ziele dieser Analyse aufgeführt, woraufhin die untersuchte Stichprobe vorgestellt wird (Abschnitt 11.3). Wie bereits angesprochen, bezieht sich die erste Phase der Untersuchung auf die Einstiegsseiten universitärer Webauftritte (Abschnitt 11.4), woraufhin die zweite Phase die Dokumente der ersten Verknüpfungsebene fokussiert (Abschnitt 11.5). Aus den Ergebnissen der beiden Untersuchungsphasen kann rekonstruiert werden, wie universitäre Webangebote entstanden sind (Abschnitt 11.6). Abschließend geht Abschnitt 11.7 auf den Einfluss des Domänenwissens auf die Analyse sowie auf die Interpretation der Instanzen von Hypertextsorten und Hypertextknotensorten ein.

11.2 Ziele und Bezüge zum Hypertextsortenmodell

Nachdem in den vorangegangenen Analysen vornehmlich unterschiedliche Hypertextsortenvarianten des Hypertexttyps *Homepage einer Person* detailliert untersucht wurden, werden im Folgenden Aspekte fokussiert, die sich auf die Einstiegsseiten universitärer Webauftritte und das Ziel der Erstellung einer Typologie von Hypertextsorten beziehen. Zunächst stehen 35 Exemplare der Hypertextknotensorte *Einstiegsseite eines universitären Webangebots* im Zentrum der Analyse, die durch sämtliche HTML-Dokumente ergänzt werden, welche durch Hyperlinks von diesen Einstiegsseiten erreichbar sind und sich innerhalb des zugehörigen Webauftritts befinden. Die Analyse der Einstiegsseiten und der ersten Verlinkungsebene basiert

auf der Hypothese, dass Erkenntnisse hinsichtlich der in diesem zentralen Bereich universitärer Webauftritte verwendeten Hypertextsorten bzw. Hypertextknotensorten maßgeblich zum Aufbau einer Typologie von Hypertextsorten für die Untersuchungsdomäne beitragen können. Eine alternative Sichtweise bezieht sich auf den Status eines generischen universitären Webauftritts, denn dieser kann als abstrakter Hypertexttyp konzeptualisiert werden, der zahlreiche Konstituenten umfasst (vgl. Kapitel 5). Die Analyse konzentriert sich demnach auf die Modellierung einer spezifischen Hypertextknotensorte sowie auf die Sammlung und Identifizierung weiterer Hypertext(knoten)sorten, die in der oberen Hierarchieebene universitärer Webauftritte verwendet werden. Diese *top-down*-Untersuchung der Untersuchungsdomäne wird durch eine *bottom-up*-Analyse komplementiert (vgl. Rehm, 2002b), die in Kapitel 12 vorgestellt wird. Eine Kombination dieser empirischen Befunde erlaubt die Konstruktion einer Typologie von Hypertextsorten, die Gegenstand von Kapitel 13 ist.

11.3 Die Stichprobe

Die Stichprobe umfasst 727 HTML-Dokumente, bei denen es sich um die Einstiegsseiten 35 zufällig ausgewählter universitärer Webauftritte sowie 692 HTML-Dokumente handelt, auf die per Hyperlink verwiesen wird. Es wurden diejenigen Dokumente in die Stichprobe aufgenommen, die in der Korpusdatenbank verfügbar waren; externe Webseiten wurden nicht berücksichtigt.[1] Unter den 35 Webauftritten befinden sich 23 allgemeine und sieben technische Universitäten, eine Wirtschaftshochschule, eine Musik- und Kunst- sowie drei sonstige Hochschulen (vgl. Tabelle 11.1 und Anhang E auf der beiliegenden CD ROM).[2]

Die Stichprobe wurde mit Hilfe eines *Perl*-Skripts generiert, in das die Datenbankkennungen der 35 zufällig ausgewählten Einstiegsseiten eingetragen wurden. Daraufhin konnte eine maschinelle Analyse der HTML-Dokumente stattfinden, die auf dem *Perl*-Modul `HTML::Parser` basiert und die enthaltenen Hyperlinks untersucht. Im Falle der Referenzierung einer externen Webseite wurde der Hyperlink verworfen, bei der Verknüpfung mit einem Dokument, das im Korpus enthalten ist, wurde die korrespondierende ID ermittelt und für die Stichprobe vorgemerkt. Abschließend wurde eine Datei mit den benötigten Informationen generiert und in die Korpusdatenbank importiert.[3]

11.4 Inhalte und makrostrukturelle Komponenten

Die Einstiegsseiten universitärer Webauftritte können zwar in erster Annäherung als institutionelle Homepages aufgefasst werden, die als offizielles Aushängeschild der jeweiligen Hochschulen dienen sollen, doch stellt sich die Frage, welche weiterführenden Gemeinsamkeiten

[1] Bei der Generierung der Stichprobe wurden insgesamt 368 Hyperlinks zurückgewiesen, da sie *JavaScript*-Anweisungen ausführen, `mailto:`-Verweise darstellen oder externe Webangebote referenzieren.

[2] Die Tabelle enthält in der zweiten Spalte eine fortlaufende Identifikationsnummer, die die HTML-Dokumente eindeutig kennzeichnet. Die Nummern werden in der nachfolgend dargestellten Analyse für ihre Referenzierung eingesetzt. Die beiliegende CD ROM enthält eine Liste aller in der Stichprobe enthaltenen Dokumente.

[3] Aus technischen Gründen wurde nicht die aktuelle, im Korpus verfügbare Einstiegsseite der Technischen Universität Darmstadt in die Stichprobe aufgenommen, sondern eine ältere Version, die sich ebenfalls in der Korpusdatenbank befindet (vgl. die Bezeichnung der URL von E 25 in Tabelle 11.1).

	ID	Hochschule	URL	Anzahl Dokumente	Download	Letzte Änderung
E 1	1	Univ. Witten/Herdecke	http://www.uni-wh.de	9	30.01.2001	k. A.
E 2	10	Katholische Univ. Eichstätt-Ingolstadt	http://www.ku-eichstaett.de	13	22.01.2001	05.12.2000
E 3	23	Univ. Trier	http://www.uni-trier.de	20	06.02.2001	09.10.2000
E 4	43	Justus-Liebig-Univ. Gießen	http://www.uni-giessen.de	12	16.01.2001	02.11.2000
E 5	55	Brandenburgische Techn. Univ. Cottbus	http://www.tu-cottbus.de	12	07.02.2001	21.11.2000
E 6	67	Techn. Univ. Chemnitz	http://www.tu-chemnitz.de	44	12.02.2001	13.09.2000
E 7	111	Univ. Leipzig	http://www.uni-leipzig.de	20	25.01.2001	25.01.2001
E 8	131	Univ. Ulm	http://www.uni-ulm.de	55	02.02.2001	01.02.2001
E 9	186	Univ. Augsburg	http://www.uni-augsburg.de	10	03.04.2001	k. A.
E 10	196	Techn. Univ. Hamburg-Harburg	http://www.tu-harburg.de	14	10.04.2001	05.04.2001
E 11	210	FernUniv. Gesamtschule Hagen	http://www.fernuni-hagen.de	7	11.04.2001	03.04.2001
E 12	217	Ruprecht-Karls-Univ. Heidelberg	http://www.uni-heidelberg.de	16	18.04.2001	30.01.2001
E 13	233	Techn. Univ. Bergakademie Freiberg	http://www.tu-freiberg.de	5	02.05.2001	k. A.
E 14	238	European Business School	http://www.ebs.de/index_de.asp	25	03.05.2001	k. A.
E 15	263	Univ. Koblenz-Landau	http://www.uni-koblenz-landau.de	33	08.05.2001	22.03.2001
E 16	296	Freie Univ. Berlin	http://www.fu-berlin.de	29	21.05.2001	k. A.
E 17	325	Friedr.-Alex.-Univ. Erlangen-Nürnberg	http://www.uni-erlangen.de	5	21.06.2001	19.06.2001
E 18	330	Univ. der Künste Berlin	http://www.hdk-berlin.de	23	22.06.2001	22.06.2001
E 19	353	Techn. Univ. Dresden	http://www.tu-dresden.de	18	16.07.2001	06.07.2001
E 20	371	Rhein.-Westf. Techn. Hochschule Aachen	http://www.rwth-aachen.de	32	25.07.2001	k. A.
E 21	403	Heinrich-Heine-Univ. Düsseldorf	http://www.uni-duesseldorf.de	23	30.07.2001	k. A.
E 22	426	Univ. Paderborn	http://www.uni-paderborn.de	9	10.08.2001	02.08.2001
E 23	435	Univ. Passau	http://www.uni-passau.de	20	13.08.2001	15.05.2001
E 24	455	Univ. Essen	http://www.uni-essen.de	31	20.08.2001	15.08.2001
E 25	486	Techn. Univ. Darmstadt	http://www.tu-darmstadt.de/homepage/alt/	22	06.09.2001	15.01.2001
E 26	508	Univ. Osnabrück	http://www.uni-osnabrueck.de	35	12.09.2001	14.08.2001
E 27	543	Univ. Kassel	http://www.uni-kassel.de	20	13.09.2001	k. A.
E 28	563	Ruhr-Univ. Bochum	http://www.ruhr-uni-bochum.de	20	24.09.2001	21.09.2001
E 29	583	Westfälische Wilhelms-Univ. Münster	http://www.uni-muenster.de	13	09.10.2001	07.05.2001
E 30	596	Univ. zu Lübeck	http://www.mu-luebeck.de	29	18.10.2001	16.11.1999
E 31	625	Univ. Dortmund	http://www.uni-dortmund.de/TOP/	15	24.10.2001	23.10.2001
E 32	640	Rheinische Friedrich-Wilhelms-Univ.	http://www.uni-bonn.de	23	26.10.2001	04.09.2001
E 33	663	Univ. Konstanz	http://www.uni-konstanz.de	26	01.11.2001	31.10.2001
E 34	689	Univ. Hohenheim	http://www.uni-hohenheim.de	13	01.11.2001	25.06.2001
E 35	702	Univ. Mannheim	http://www.uni-mannheim.de	26	05.11.2001	k. A.

Tabelle 11.1: Die untersuchten Einstiegsseiten universitärer Webauftritte

und welche spezifischen Unterschiede in Bezug auf die Einstiegsseiten kommerziell ausgerichteter Websites bestehen (vgl. ausführlich hierzu Abschnitt 4.6.2).[4] Nachfolgend werden die Ergebnisse einer Inhalts- und Makrostrukturanalyse vorgestellt, die sich ausschließlich auf die *Einstiegsseite eines universitären Webauftritts* als Hypertextknotensorte bezieht.

Die Ergebnisse der Untersuchung werden mit den Studien von Kamenz et al. (1998), Nielsen und Tahir (2002) und Schütte (2004a) kontrastiert.[5] Diesbezüglich ist zu betonen, dass es sich nicht um eine Analyse der Benutzerfreundlichkeit universitärer Einstiegsseiten handelt. Es ist also nicht das Ziel, die ermittelten Spezifika aus Sicht der Mensch-Maschine-Interaktion zu bewerten oder einen Katalog von Gestaltungshinweisen zu erarbeiten. Vielmehr sollen die Gemeinsamkeiten der hier untersuchten Dokumente auf mehreren unterschiedlichen Ebenen aufgezeigt werden, die es erlauben, die Einstiegsseite eines universitären Webauftritts als konventionalisierte Hypertextknotensorte zu charakterisieren.

[4] Initiale Ergebnisse der ersten Phase dieser Analyse wurden in Rehm (2002b, 2004c) publiziert.

[5] Sofern ein unmittelbarer Vergleich möglich war, wurden die jeweiligen Ergebnisse von Kamenz et al. (1998), Nielsen und Tahir (2002) und Schütte (2004a) in die nachfolgend diskutierten Tabellen 11.2 bis 11.7 integriert. Dabei ist jedoch zu beachten, dass sich Kamenz et al. (1998) *nicht* ausschließlich auf die Einstiegsseite eines universitären Webauftritts beziehen und darüber hinaus nicht angeben, wie viele HTML-Dokumente der Websites von Hochschulen analysiert wurden (vgl. Abschnitt 6.3.7).

11.4.1 Gestaltung und Typografie

Hinsichtlich typografischer und gestalterischer Merkmale unterscheiden sich die 35 Einstiegsseiten in vielerlei Hinsicht. Ein zeitgemäßes Webdesign besitzt für die Emittenten von beispielsweise E 2, E 23 und E 32 offenbar nur einen untergeordneten Stellenwert, wohingegen die Produzenten von z. B. E 7, E 13 und E 20 ästhetische, professionell wirkende und typografisch anspruchsvolle HTML-Dokumente erstellt haben. Trotz dieser auffälligen Unterschiede bezüglich der Gestaltung können verschiedene Konventionen ermittelt werden, die bereits bei der Wahl von Weiß als Hintergrundfarbe in 33 der 35 Einstiegsseiten (94%) ansetzen (vgl. Tabelle 11.2). Der Name der Hochschule wird ebenfalls in 94% der Dokumente in einer großen Schrifttype an einer auffälligen Position genannt, in 58% dieser Fälle wurde die obere Seitenmitte gewählt.[6] In einigen Fällen ist dieser Name Bestandteil eines universitätsspezifischen Logos, das in 91% aller Seiten verwendet wird. Während Nielsen und Tahir (2002) in 84% der von ihnen untersuchten kommerziellen Homepages eine Positionierung des Logos in der oberen linken Ecke berichten, bevorzugen die Produzenten der universitären Einstiegsseiten zu gleichen Teilen die obere Seitenmitte und die obere linke Ecke. In letzterem Fall handelt es sich meist lediglich um ein grafisches Symbol, bei einer Anordnung in der Seitenmitte wird in der Regel zusätzlich der vollständige und meist sehr lange Name der Institution innerhalb der eingebetteten Grafik aufgeführt, so dass eine Positionierung in der oberen linken Ecke aus gestalterischen Gründen nicht in Frage kommt.

In 37% der Einstiegsseiten wird – zumeist in unmittelbarer Nähe des Logos – das Siegel der Hochschule als Grafik eingebettet.[7] Trotz der hochfrequenten Benutzung von Logos kann in nur 24 Einstiegsseiten (69%) von einer Umsetzung des Corporate-Designs der jeweiligen Hochschule gesprochen werden, das z. B. spezifische Schrifttypen und Farbräume vorschreibt. Eine weitere sehr allgemeine Konvention bezieht sich auf den Einsatz einer mittels typografischer Mittel abgesetzten Fußzeile (86%), die in 28 von 30 Fällen Metainformationen und gelegentlich eine sekundäre Navigationshilfe (sechs von 30 Fällen) oder sekundäre Inhalte (fünf von 30 Fällen) enthält; in neun Dokumenten werden mehrere Fußzeilen benutzt, wobei Metainformationen in nahezu allen Fällen an finaler Position aufgeführt werden. Der Einsatz einer Kopfzeile ist mit Vorkommen in 51% der Dokumente weniger ausgeprägt, wobei sie zu nahezu gleichen Teilen Metainformationen, sekundäre Informationen und zusätzliche Navigationshilfen umfassen.

Allen 35 Einstiegsseiten ist gemein, dass sie sehr viele Informationen und Hyperlinks beinhalten, die sich z. B. auf aktuelle Neuigkeiten, Organisationseinheiten, Informationen für Ehemalige, neu eingerichtete Studiengänge oder das Studentenwerk beziehen. Diese Menge an heterogenen Informationsbausteinen und Verweisen, die sich potenziell in der Einstiegs-

[6] In den Tabellen 11.2 bis 11.7 werden unterschiedliche prozentuale Angaben aufgeführt: Die Daten für Merkmale der obersten Ebene beziehen sich in der Regel auf die Gesamtmenge von 35 Dokumenten, wohingegen sich eingerückte Submerkmale meist auf die Menge der spezifischen Vorkommen beziehen.

[7] Es ist auffällig, dass Siegel, die häufig das Gründungsjahr der jeweiligen Institution beinhalten, ausschließlich in den Einstiegsseiten der allgemeinen Universitäten verwendet werden. Dem Siegel wird offenbar die Funktion eines weiteren Identifizierungs- und Individualisierungsmerkmals zugeschrieben, das darüber hinaus in der Lage ist, die langjährige Tradition einer Hochschule zu betonen. Diese Aufgabe besitzt in vielen Fällen auch eine Abbildung der Universität (49% aller Einstiegsseiten), die in den meisten Fällen ein Foto oder eine Strichzeichnung des Hauptgebäudes zeigt.

			Kommerzielle Einstiegsseiten		*Universitäre Sites*
Merkmal	Frequenz	Prozent	Nielsen und Tahir (2002) Prozent	Schütte (2004a) Prozent	Kamenz et al. (1998) Prozent
Hintergrundfarbe					
Weiß	33	94	84	—	—
Blau	2	6	—	—	—
Prominent platzierter Name der Hochschule	33	94	—	100	—
Position: Mitte oben	19	58	—	—	—
Position: Rechts oben	8	24	—	—	—
Position: Links oben	6	18	—	—	—
Logo, Schriftzug	32	91	100	100	—
Position: Mitte oben	11	34	6	—	—
Position: Links oben	11	34	84	—	—
Position: Rechts oben	9	23	6	—	—
Position: Seitenzentrum	1	3	—	—	—
Typografisch abgesetzte Fußzeile	30	86	—	—	—
Farbe der Grundschrift					
Schwarz	25	71	72	—	—
Blau	7	20	8	—	—
Rot	1	3	—	—	—
Grau	1	3	8	—	—
(Kein ASCII-Text)	1	3	—	—	—
Unterstreichung und farbliche Gestaltung von Hyperlinks					
Unterstrichen	27	77	80	—	—
Blau	25	71	60	—	—
Rot	4	11	—	—	—
Schwarz	3	9	12	—	—
Grün	1	3	—	—	—
Grau	1	3	—	—	—
Corporate-Design erkennbar	24	69	—	—	—
Strukturierung und Gruppierung des Dokuments					
Primär zweispaltiges Layout	19	54	—	—	—
Primär einspaltiges Layout	11	31	—	—	—
Primär dreispaltiges Layout	5	14	—	—	—
Typografisch abgesetzte Kopfzeile	18	51	—	—	—
Abbildung der Universität	17	49	—	—	49
Als Foto	15	88	—	—	—
Als Strichzeichnung	2	12	—	—	—
Siegel der Universität	13	37	—	—	—
Position: Links oben	7	54	—	—	—
Position: Rechts oben	4	31	—	—	—
Position: Mitte oben	1	8	—	—	—
Position: Hintergrund	1	8	—	—	—
Primäre Navigationshilfe					
Position: Mitte	16	45	12	—	—
Position: Mitte links	7	20	—	—	—
Position: Links oben	4	11	—	—	—
Position: Mitte rechts	4	11	—	—	—
Position: Links	3	9	30	—	—
Position: Mitte unten	1	3	—	—	—
Position: Mitte oben	0	0	18	—	—
Sekundäre Navigationshilfe					
Position: Mitte oben	5	14	—	—	—
Position: Mitte unten	5	14	—	—	—
Position: Mitte rechts	3	9	—	—	—
Position: Links	2	6	—	—	—
Position: Rechts oben	1	3	—	—	—
Position: Links unten	1	3	—	—	—

Tabelle 11.2: Ergebnisse der Makrostruktur- und Inhaltsanalyse – Gestaltung und Typografie

seite befinden oder von ihr erreichbar sein sollten, stellt für den Produzenten eine Herausforderung dar, schließlich weisen sämtliche Gestaltungsratgeber darauf hin, die Einstiegsseite mit einem kompakten Layout zu versehen. Entsprechend wird in 54% der Dokumente eine zweispaltige Anordnung präferiert, um die Präsentation einer großen Menge von Informationsbausteinen zu ermöglichen. Ein einspaltiges Layout wird in 31% der Einstiegsseiten verwendet, wobei typischerweise Listen von Hyperlinks eingesetzt werden, wodurch der rechte Teil des Dokuments mehr Weißraum umfasst als die linke Seite. Ein dreispaltiges Layout wird in nur fünf Dokumenten benutzt (z. B. in E 24; vgl. auch Indikator 38 auf S. 136).

Die in den 35 Einstiegsseiten angebotenen Navigationshilfen verschaffen dem Rezipienten einen Überblick über die Inhalte und die Strukturierung der Webangebote. In 27 Fällen wird die primäre Navigationshilfe, die die zentralen Themenangebote bündelt, großflächig im mittleren Seitenbereich angeordnet. In lediglich drei Einstiegsseiten wird eine Auflistung von Hyperlinks am linken Seitenrand verwendet, in vier Fällen befindet sich diese Liste in der oberen linken Ecke. Im Vergleich mit den von Nielsen und Tahir (2002) ermittelten Daten zeigen diese Angaben einen zentralen Unterschied zu kommerziell ausgerichteten Websites auf, in denen die primäre Navigationshilfe in der Peripherie der Einstiegsseite positioniert ist, um das Zentrum des Dokuments für Inhalte wie z. B. Produkthinweise oder Neuigkeiten zu reservieren. In den Einstiegsseiten universitärer Webauftritte dominiert die Navigationshilfe, sie stellt somit ihren eigentlichen Inhalt dar. In 17 der 35 Einstiegsseiten wird neben der primären Navigationshilfe eine sekundäre Navigationshilfe angeboten, die zumeist Hyperlinks bündelt, die zu Hilfeseiten, Suchmöglichkeiten oder einem Schlagwortindex führen.[8] Die sekundäre Navigationshilfe umfasst typischerweise horizontal angeordnete Hyperlinks und befindet sich in der Regel oberhalb oder unterhalb der primären Navigationshilfe.

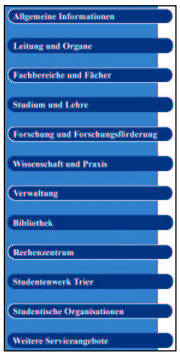
11.4.2 Navigations- und Zugriffshilfen

Jede der 35 Einstiegsseiten enthält eine primäre Navigationshilfe mit Hyperlinks zu weiterführenden Themengebieten, die sich von sekundären Navigationshilfen durch ihre Größe, Gestaltung, Positionierung und ihre inhaltlichen Kategorien unterscheidet (vgl. auch Androutsopoulos und Kraft, 2003, S. 7). Im Hinblick auf die Gestaltung der Navigationshilfen können verschiedene Konventionen ermittelt werden, wobei festzuhalten ist, dass die von Nielsen und Tahir (2002) ermittelten Typen nicht bestätigt werden können (vgl. Tabelle 11.3).[9] Abbildung 11.1 zeigt Beispiele für die ermittelte Typologie primärer Navigationshilfen, die mit Ausnahme von E 20 (vgl. Fußnote 9) sämtliche Einstiegsseiten abdeckt.

[8] Reiss (2000, S. 130 ff.) stellt verschiedene „secondardy features“ vor, die in den meisten Websites nicht als Teil der primären Navigationshilfen realisiert werden: „This lets visitors know that they are *not* subjects, but are more functional in nature.“ Bei den von Reiss empfohlenen „site tools“ handelt es sich um „Home or Main Menu“, „Contact (site owner)“, „Feedback (webmaster)“, „Site map“, „Site index“, „Disclaimer“, „What's new“, „About this site“, „First-time visitors“, „FAQ“, „Quick links“ und „Search“. Etwa die Hälfte der hier untersuchten Einstiegsseiten universitärer Webauftritte umfasst eine sekundäre Navigationshilfe, wobei jedoch eine Beschränkung auf Suchmöglichkeiten, Kontaktinformationen und die Verknüpfung zusätzlicher Versionen des Webangebots in weiteren Sprachen festzustellen ist (vgl. Abbildung 11.3, S. 488, für Beispiele).

[9] In E 20 wird am linken Seitenrand eine dynamische, *Java*-basierte Navigationshilfe verwendet, in der Hyperlinks in Listenform dargestellt werden. Diese wird als eigenständiger Dateityp aufgefasst, weshalb ihre Inhalte nicht in die nachfolgenden Analysen einfließen.

E 27

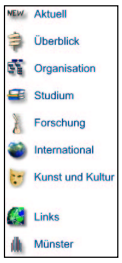


E 3

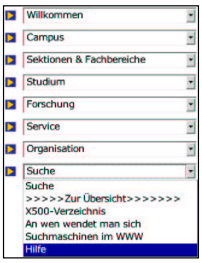
E 18

E 13

E 16



E 29



E 33

Typ 1: Einfache Liste (vertikal)

E 5

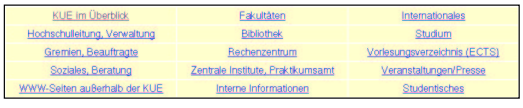
E 22

E 7

E 6

E 4

Typ 2: Komplexe Tabelle



E 2

E 19

Typ 3: Einfache Tabelle

E 30

E 24

Typ 4: Einfache Liste (horizontal)

Abbildung 11.1: Beispiele für die vier unterschiedlichen Typen *primärer Navigationshilfen* in den Einstiegsseiten der Stichprobe

Merkmal	Frequenz	Prozent	*Kommerzielle Einstiegsseiten* Nielsen und Tahir (2002) Prozent	*Universitäre Sites* Kamenz et al. (1998) Prozent
Suchmöglichkeiten insgesamt	29	83	86	52
Suche per internem HTML-Dokument	21	72	20	—
Suchbox in Einstiegsseite	8	28	81	—
Position: Mitte links	3	38	—	—
Position: Mitte unten	2	25	—	—
Position: Links unten	1	13	—	—
Position: Mitte oben	1	13	14	—
Position: Rechts oben	1	13	35	—
Position: Links oben	0	0	30	—
Bezeichnung des Hyperlinks bzw. Submit-Knopfes				
„Suche“	18	62	—	—
„Suchen“	8	28	—	—
Lupe (als Icon)	2	7	—	—
Fernglas (als Icon)	1	3	—	—
Fragezeichen (als Icon)	1	3	—	—
„Starte Suche“	1	3	—	—
Typen primärer Navigationshilfen				
Einfache Liste (vertikal)	18	51	—	—
Komplexe Tabelle (Überschriften und Schlagworte)	9	26	—	—
Schlagworte als Hyperlinks realisiert	5	56	—	—
Schlagworte nicht als Hyperlinks realisiert	3	33	—	—
Einige Schlagworte als Hyperlinks realisiert	1	11	—	—
Einfache Tabelle (ein Hyperlink pro Zelle)	5	14	—	—
Einfache Liste (horizontal)	2	6	—	—
Java-basiertes Navigationsmenü	1	3	—	—
Index, Schlagwortverzeichnis, FAQ	7	20	—	8
Zielgruppenspezifische Navigationshilfen	7	20	—	—
Reduplikation grafischer Navigationshilfen	5	14	—	—
Sitemap	2	6	48	—
„Sitemap“	1	50	—	—
„Überblick“	1	50	—	—

Tabelle 11.3: Ergebnisse der Makrostruktur- und Inhaltsanalyse – Navigation

(i) Insgesamt 18 Einstiegsseiten (51%) verwenden eine Darstellung der Hyperlinks als einfache, vertikal angeordnete Liste; eingebettete Listen werden innerhalb dieses Typs nur in sehr wenigen Fällen verwendet. Sechs der 18 Vorkommen können als Varianten der einfachen Liste aufgefasst werden und umfassen z. B. zwei nebeneinander positionierte Listen (die linke enthält deutschsprachige und die rechte englischsprachige Hyperlinks, E 28), fünf im Seitenzentrum angeordnete Listen von jeweils fünf Hyperlinks (E 26) und eine Liste von acht Aufzählungspunkten, die jedoch keine per a-Element realisierten Hyperlinks, sondern Pull-Down-Menüs umfassen (E 33). (ii) Der zweite Typ wird als komplexe Tabelle bezeichnet und enthält als Hyperlinks realisierte Überschriften, die – unterhalb oder rechts der Überschrift – mehrere untergeordnete Schlagworte umfassen, die das in der Überschrift genannte Themenfeld genauer charakterisieren. Dieser in 26% der Einstiegsseiten eingesetzte Typ nimmt zwar einen größeren Bereich ein als die einfache Hyperlinkliste, besitzt jedoch den Vorteil, dass der Rezipient durch die aufgeführten Schlagworte einen Eindruck von den Inhalten der jeweils übergeordneten inhaltlichen Kategorie erhält, weshalb dieser Typ die Aufgabe einer zentralen Kohärenzbildungshilfe potenziell besser erfüllen kann. Interessanterweise werden die charakterisierenden Schlagwörter nicht in allen Navigationshilfen dieses Typs als Hyperlinks realisiert (vgl. E 4, E 5 und E 22 in Abbildung 11.1), so dass der Rezipient gezwungen ist, zunächst das verknüpfte Überblicksdokument aufzurufen, um daraufhin das ihn interessierende Schlagwort zu lokalisieren und somit gegebenenfalls einem weiteren Hyperlink zu

	Studium	Forschung	Selbstdarstellung/Überblick	Liste dezentraler Einrichtungen	Aktuelles/Pressemitteilungen	Liste zentraler Einrichtungen	Service/Dienstleistungen	Informationen zur Region	Suche	Internationales	Nennung spez. Einrichtungen	Weiterführende Hyperlinks	Wissenstransfer/Kooperationen	Ansprechpartner/Adressen	Struktur/Organisation	Gremien und Organe	Hochschulverwaltung	Vorlesungsverzeichnis	Alumni/Förderverein	Hochschulleitung
E 1	✓		✓		✓	✓			✓	✓				✓					✓	
E 2	✓		✓	✓	✓	✓				✓	✓	✓				✓	✓	✓		✓
E 3	✓	✓	✓	✓			✓				✓		✓				✓			✓
E 4	✓	✓	✓	✓	✓	✓		✓												
E 5	✓	✓	✓					✓							✓					
E 6	✓	✓	✓	✓	✓	✓						✓				✓	✓			✓
E 7	✓	✓	✓	✓		✓	✓	✓	✓				✓	✓						
E 8	✓	✓	✓	✓	✓		✓	✓	✓		✓	✓			✓					
E 9	✓	✓	✓	✓	✓	✓		✓	✓		✓		✓							
E 10	✓	✓	✓		✓		✓													
E 11	✓	✓	✓	✓		✓			✓			✓		✓		✓	✓		✓	
E 12	✓	✓	✓	✓		✓							✓							
E 13	✓	✓	✓	✓		✓				✓										
E 14	✓		✓																	
E 15	✓	✓	✓		✓					✓				✓				✓		
E 16	✓	✓		✓	✓	✓	✓				✓							✓		
E 17	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓
E 18	✓	✓	✓		✓		✓		✓		✓	✓	✓						✓	
E 19	✓	✓			✓					✓			✓						✓	
E 21	✓	✓					✓	✓	✓	✓										
E 22	✓	✓	✓	✓		✓														
E 23			✓		✓		✓	✓												
E 24	✓	✓	✓	✓	✓	✓	✓				✓	✓		✓						
E 25	✓	✓	✓	✓		✓				✓	✓		✓		✓				✓	✓
E 26	✓	✓	✓				✓													
E 27	✓	✓	✓		✓									✓		✓			✓	
E 28	✓	✓	✓				✓	✓										✓		
E 29	✓	✓	✓		✓			✓		✓		✓			✓					
E 30				✓	✓	✓								✓				✓		
E 31	✓	✓	✓	✓	✓				✓											
E 32	✓	✓	✓	✓		✓	✓	✓		✓		✓			✓	✓	✓			
E 33	✓	✓		✓			✓		✓						✓					
E 34	✓	✓	✓				✓		✓						✓					
E 35	✓	✓					✓									✓				
Σ	32	30	29	18	18	17	15	11	11	10	9	9	8	8	7	7	6	6	6	5

Tabelle 11.4: Ergebnisse der Makrostruktur- und Inhaltsanalyse – In *primären Navigationshilfen* aufgeführte Inhalts- und Themenfelder

folgen. (iii) Neben der komplexen Tabelle existiert als dritter Typ die einfache Tabelle (14% der Einstiegsseiten), in der sich jeweils ein Hyperlink in einer Tabellenzelle befindet. Dieser Typ wird in allen Fällen im Seitenzentrum angeordnet und umfasst z. B. vier Zeilen und drei Spalten (E 9) oder neun Zeilen und drei Spalten (E 11). Die einfache tabellarische Präsentation von Themenbereichen kann naturgemäß keine Möglichkeiten der Hierarchisierung der dargebotenen Informationen verwenden, wirkt daher insgesamt unsystematisch und trägt nur partiell zur Stiftung globaler Kohärenz bei. (iv) Der letzte Typ betrifft die einfache Liste, in der die Aufzählungspunkte horizontal angeordnet sind (6%), so dass ein großer Dokumentbereich zur Präsentation von Inhalten eingesetzt werden kann (z. B. aktuelle Meldungen). Beide Vorkommen dieses Typs sind am oberen Seitenrand positioniert und enthalten fünf Hyperlinks. Sekundäre Navigationshilfen greifen auf diesen sowie den ersten Typ zurück.

Bezüglich der Inhalts- und Themenfelder, zu denen innerhalb der primären Navigationshilfe Hyperlinks aufgeführt werden, existiert eine große Varianz. Die 34 Navigationshilfen

(vgl. Fußnote 9) enthalten durchschnittlich 21,09 Hyperlinks (Min.: 5, Max.: 120, Median: 12, Modus: 6, Summe: 717) zu mehr als 40 unterschiedlichen Inhalts- und Themengebieten. Tabelle 11.4 stellt die 20 häufigsten Gebiete dar. Konventionen können sowohl im Hinblick auf die Existenz bzw. Frequenz einzelner Themenfelder, als auch bezüglich ihrer Hyperlinkanzeiger ermittelt werden (vgl. auch Androutsopoulos und Kraft, 2003).[10] Die Themenfelder „Studium" (32 Vorkommen), „Forschung" (30) und „Selbstdarstellung/Überblick" (28) werden in nahezu allen Navigationshilfen aufgeführt und können als grober inhaltlicher Kern eines universitären Webauftritts aufgefasst werden. Die Hyperlinkanzeiger der beiden ersten Kategorien weisen eine sehr ausgeprägte Konventionalisierung auf: Neben der kompakten Form „Studium" (19 Vorkommen) existieren verschiedene Varianten, z. B. „Studium und Lehre" (4 Vorkommen), „Studium und Weiterbildung" (3 Vorkommen), „Studium und Studierende", „Lehre – Studium – Weiterbildung" und „Lehre und Studium" (jeweils ein Vorkommen). Im Hinblick auf das zweite Themenfeld dominiert „Forschung" (22 Vorkommen), für das ebenfalls ähnliche Nominalphrasen eingesetzt werden, an deren initialer Position jedoch immer „Forschung" genannt wird, z. B. „Forschung und Forschungsförderung" (2 Vorkommen), „Forschung und Wissenstransfer", „Forschung und Transfer" und „Forschung und Kooperation" (je ein Vorkommen). Bezüglich des dritten Themengebiets existieren unterschiedliche Formen der Verknüpfung: Neben „Wir über uns" (4 Vorkommen; Variante: „Wir stellen uns vor"), „Universität" (4 Vorkommen; Variante: „Die Universität") werden „Überblick" (3 Vorkommen; Varianten: „Überblick/Overview", „Universität im Überblick", „Die Universität im Überblick") sowie „Hochschulführer", „Porträt", „Information" und „Allgemeine Informationen" (jeweils ein Vorkommen) verwendet.[11] Eine weitere Gruppe nennt den Namen bzw. das Akronym der Hochschule, wobei der Überblickscharakter des Zieldokuments in einigen Bezeichnungen betont wird: (i) „Universität Augsburg", „TU Hamburg-Harburg", „Universität Osnabrück", „Die Universität Passau", (ii) „KUE im Überblick", „FAU im Überblick", (iii) „HdK-Profil", „Hohenheim kompakt".

Neben der Differenzierung in studien- und forschungsbezogene Inhalte führen 18 Einstiegsseiten einen Hyperlink zu aktuellen Informationen auf, die in der Regel Veranstaltungshinweise oder Pressemitteilungen umfassen. An vierter bzw. sechster Position der hochfrequenten Hyperlinkanzeiger befinden sich Verweise auf Listen dezentraler bzw. zentraler Ein-

[10] In Tabelle 11.4 sind ausschließlich die Themenfelder der primären Navigationshilfen enthalten. Beim Typ der einfachen Liste wurden Hyperlinks in eingebetteten Listen nicht berücksichtigt, ebenso wurden beim Typ der komplexen Tabelle lediglich die Überschriften einzelner Sektionen untersucht. Falls ein Hyperlinkanzeiger mehr als einem Themengebiet zugeordnet werden konnte, wurden entsprechend viele Vorkommen notiert. Aus Darstellungsgründen konnten nicht alle mehr als einmal verwendeten Themengebiete aufgeführt werden. Jeweils drei Vorkommen enthalten „Impressum", „Campus", „Lagepläne/Anfahrt" und „Weiterbildung". Jeweils zwei Vorkommen umfassen „Listen außeruniversitärer Einrichtungen", „interne Informationen", „Information", „Sport, Soziales, Beratung" und „Multimedia". Insgesamt 24 Hyperlinkanzeiger konnten keiner Kategorie zugeordnet werden und müssen somit als untypisch bezeichnet werden, z. B. „Schwarzes Brett", „Community", „Rechtsgrundlagen", „Kunst und Kultur", „Veröffentlichungen" und „Work in Progress".

[11] Die Hypertextliteratur weist im Kontext der Kohärenzbildungshilfen darauf hin, dass Hyperlinkanzeiger möglichst deskriptiv gestaltet sein sollten, um dem Rezipienten eine Vorstellung der im Ziel präsentierten Inhalte zu vermitteln. Von dieser Perspektive aus betrachtet sind Hyperlinkanzeiger wie z. B. „Information", „Allgemeine Informationen" und „Überblick" abzulehnen, da sich hinter derartig generellen Bezeichnungen prinzipiell arbiträre Inhalte verbergen können. Mehrere dieser sehr vagen Hyperlinkanzeiger werden in E 23 aufgeführt, z. B. „Lokal vorhandene Informationen und Services" und „Externe Informationsquellen".

richtungen, die ihrerseits auf die zugehörigen Webauftritte verweisen, welche in aller Regel von Angehörigen dieser Einheiten gepflegt werden. Neben der zentral administrierten Einstiegsseite und den zugehörigen allgemeinen Dokumenten (z. B. zur Hochschulleitung und -verwaltung, Lagepläne etc.) stellen insbesondere diese dem gesamten Hypertext zugehörigen, zugleich jedoch eigenständigen Websites einen Großteil der fach- und prozessspezifischen Informationen zur Verfügung, die für Angehörige der Hochschule, Studierende und Wissenschaftler anderer Institutionen von Interesse sind (vgl. Kapitel 6). Da diese eingebetteten Hypertexte einen Großteil der eigentlichen Kerninhalte eines universitären Webauftritts im Hinblick auf die Themengebiete Studium und Forschung enthalten, ermöglichen die Produzenten der Einstiegsseiten eine Navigation über die einer Hochschule zugehörigen Einrichtungen. In den primären Navigationshilfen werden zwei grobe Muster verwendet: Entweder werden alle zentralen und dezentralen Einrichtungen unmittelbar in der Navigationshilfe aufgeführt (vgl. E 6 in Abbildung 11.1) oder ein Hyperlink führt zu einem weiteren HTML-Dokument, das eine Aufstellung dieser Organisationseinheiten enthält (vgl. z. B. E 13 in Abbildung 11.1) – bei der Verwendung des Typs der komplexen Tabelle werden die Formen in nahezu allen Fällen kombiniert. Insbesondere diese beiden Strategien sind für die eingangs genannten Unterschiede bezüglich der Anzahl von Hyperlinks verantwortlich, die in den 34 primären Navigationshilfen enthalten sind: Der Produzent entscheidet sich entweder dafür, einen Hyperlink in der Einstiegsseite zu präsentieren oder ihn erst in einem nachfolgenden Überblicksdokument zu platzieren, so dass ein strukturierendes Hyperonym diesen sowie weitere Hyperlinks zu verwandten Themengebieten subsumiert, wodurch jedoch in der Einstiegsseite der Aspekt der Kohärenzbildung beeinträchtigt sein kann.

Insgesamt 29 der 35 Einstiegsseiten (83%) geben dem Rezipienten die Möglichkeit, in der oder den zugehörigen Websites mittels Suchanfragen recherchieren zu können.[12] In acht HTML-Dokumenten wird eine kleinformatige Suchbox[13] angeboten, in die der Rezipient unmittelbar Suchbegriffe eintragen kann; in 21 Einstiegsseiten muss für diesen Zweck zunächst ein untergeordnetes HTML-Dokument aufgerufen werden. Der korrespondierende Hyperlinkanzeiger bzw. die Beschriftung des *submit*-Knopfes unterliegt einer Konvention: „Suche“ wird in 62% und „Suchen“ in 28% der Vorkommen verwendet. Darüber hinaus werden in einigen Fällen auch Icons benutzt, die z. B. eine stilisierte Lupe oder ein Fragezeichen darstellen. Eine besonders ausführliche Form der Kennzeichnung wird in E 23 eingesetzt: „Sie können im WWW-Angebot der Universität Passau auch suchen.“

Webangebote von Universitäten richten sich an unterschiedliche Zielgruppen (vgl. auch Boardman, 2005, S. 24, sowie Abschnitt 6.3.1). Insgesamt sieben Einstiegsseiten (20%) enthalten zielgruppenspezifische Navigationshilfen (vgl. z. B. Reiss, 2000, S. 72 ff.) des Typs einfache Liste, mit deren Hilfe die Produzenten versuchen, die Informationsbedarfe der einzelnen Rezipientengruppen zu antizipieren und die Leser gleichzeitig auf für sie unter Umständen interessante Angebote hinzuweisen. Zielgruppenspezifische Navigationshilfen wer-

[12] Somit erreichen die universitären Einstiegsseiten beinahe den Wert von 86%, den Nielsen und Tahir (2002) für kommerziell ausgerichtete Websites ermitteln. Auffällig ist bezüglich der Integration von Suchmöglichkeiten eine zunehmende Tendenz, da Kamenz et al. (1998) einen Wert von 52% angeben.

[13] Bezüglich der Position einer Suchbox innerhalb der Einstiegsseite kann keine Konvention ermittelt werden. Die Vorkommen innerhalb der Stichprobe befinden sich z. B. im linken, oberen oder unteren Bereich des Seitenzentrums, in der unteren linken oder der oberen rechten Ecke.

Zielgruppe	E 13	E 16	E 19	E 20	E 24	E 28	E 31
Studien-interessenten	„Abiturienten“	„Studien-bewerber“	„Studierende in spe“	„Schule“	„Studien-bewerber“	„Schülerinnen und Schüler“	„Schüler“
Studierende	„Studenten“	„Studierende“	„Studierende“	„Studierende“	„Studierende“	„Studierende“	„Studierende“
Ausländische Studierende	—	—	„Ausländische Studierende“	—	„Inter-nationales“	—	—
Ehemalige/Alumni	—	„Alumni“	„Absolventen“	—	„Alumni/Förderer“	—	—
Angehörige der Hochschule	—	„Beschäftigte“	„Hochschul-angehörige“	„RWTH-Intern“	„Mitarbeiter“	„Intranet“	„RUB-Intern“
Wissenschaftler anderer Hoch-schulen	„Wirtschaft & Forschung“	—	„Forschung“	„Forschung“	—	—	„Forschung“
Wirtschaft	„Wirtschaft & Forschung“	„Wirtschaft“	„Wirtschaft“	„Wirtschaft“	„Wirtschaft“	„Wirtschaft“	„Wirtschaft“
Besucher/Gäste	—	—	„Gäste“	—	—	—	„Besucher“
Sprecher des Englischen	—	—	„English readers“	—	—	—	—
Presse	—	„Presse“	„Presse“	„Medien“	„Presse“	—	„Presse“

Tabelle 11.5: Ergebnisse der Makrostruktur- und Inhaltsanalyse – In *zielgruppenspezifischen Navigationshilfen* verwendete Hyperlinkanzeiger

den typischerweise als eigenständige Bausteine realisiert, die in der Regel am oberen oder unteren Seitenrand positioniert werden. E 19 stellt diesbezüglich eine Ausnahme dar, weil sich in dieser Einstiegsseite *keine* nach thematischen Gesichtspunkten gegliederte Navigationshilfe befindet; stattdessen fungiert die zielgruppenspezifische Navigationshilfe, die im Zentrum des Dokuments angeordnet ist, als primäre Navigationsmöglichkeit. Tabelle 11.5 zeigt die Hyperlinkanzeiger, die in den sieben Einstiegsseiten für insgesamt zehn verschiedene Zielgruppen verwendet werden.[14] Es fallen verschiedene Konventionen auf: Rezipienten aus dem Bereich der Privatwirtschaft gelangen in allen Einstiegsseiten über den Hyperlinkanzeiger „Wirtschaft“ (bzw. „Wirtschaft & Forschung“ in E 13) zu spezifischen Inhalten. Auch Informationen für „Studierende“ werden – bis auf die Ausnahme „Studenten“ – mit einem konventionalisierten Hyperlinkanzeiger verknüpft. Ähnlich verhält es sich bei den Kategorien „Forschung“ und „Presse“, die jeweils einen abweichenden Hyperlinkanzeiger beinhalten. Besonders umfangreiche Varianzen existieren im Hinblick auf die Zielgruppen der Studieninteressenten sowie der Hochschulangehörigen. Die Informationen für die zuletzt genannte Zielgruppe werden über Synonyme („Beschäftigte“, „Hochschulangehörige“, „Mitarbeiter“) sowie indirekt über einen funktional spezifizierten, passwortgeschützten Bereich des Webauftritts markiert („RWTH-Intern“, „RUB-Intern“, „Intranet“). Auch bei der Zielgruppe der Studieninteressenten werden synonyme Ausdrücke als Hyperlinkanzeiger eingesetzt („Stu-

[14] In Tabelle 11.5 wurden ausschließlich diejenigen Hyperlinkanzeiger aufgenommen, die sich innerhalb des der zielgruppenspezifischen Navigationshilfe zugehörigen Hypertextmoduls befinden. Zwei dieser Navigationshilfen werden mit einer Nominalphrase eingeleitet: „Infos für:“ (E 16) und „Universität für“ (E 24).

dienbewerber“, „Studierende in spe“), zugleich werden jedoch auch „Abiturienten“, „Schüler“ und „Schülerinnen und Schüler“ verwendet, die den Skopus des korrespondierenden Webangebots deutlich auf die genannte Gruppe von Personen einschränken, so dass z. B. Studienabbrecher, die an dieser Hochschule einen neuen Studiengang beginnen möchten, von diesem Hyperlinkanzeiger nicht subsumiert werden.

In fünf Einstiegsseiten (14%) werden Hyperlinks redupliziert. Diese Wiederholung bezieht sich in vier Einstiegsseiten auf die Wiederholung von Hyperlinks in Textform, die in der primären Navigationshilfe als eingebettete Grafiken realisiert werden (vgl. Abbildung 11.3, S. 488, für ein Beispiel). Hierdurch werden Personen, die z. B. auf der Kommandozeile Browser wie *lynx* oder *links* oder einen *Screen Reader* einsetzen, in die Lage versetzt, den entsprechenden Hyperlinkanzeiger rezipieren zu können.[15] Eine ungewöhnliche Variante der Reduplikation wird in E 8 vorgenommen: Dieses sehr lange HTML-Dokument ist in inhaltliche Rubriken wie z. B. „Suchen und Finden“, „Studium und Lehre“, „Forschung“, „Struktur und Organisation“ und „Wir über uns“ eingeteilt. Rechts dieser Überschriften werden Hyperlinks aufgeführt, wobei jedoch keine eindeutige Zuordnung vorgenommen wird, d. h. viele der insgesamt 114 Hyperlinkanzeiger wurden in mehrere Kategorien aufgenommen.

Lediglich sieben Einstiegsseiten (20%) bieten Verweise zu einem Index, einem Schlagwortverzeichnis oder einer FAQ-Liste an. Als Hyperlinkanzeiger werden für diesen Zweck „Index“ (E 5, E 22), „FAQ“ (E 13), „Überblick“ (E 12), „Schnelleinstieg“ (E 33), „Uni von A-Z“ (E 4) und „Wer, Was und Wo“ (E 26) eingesetzt. Die verknüpften Dokumente erfüllen, wie die Hyperlinkanzeiger andeuten, heterogene Funktionen, z. B. die Bündelung von Schlagwörtern, universitätsvorstellenden Kurztexten und Ansprechpartnern. Im Vergleich mit den von Kamenz et al. angegebenen Vorkommen in 8% aller universitären Websites ist ein zunehmender Trend dieser Gruppe von Hypertextsortenmodulen festzustellen. Die Heterogenität der verknüpften Dokumente bezieht sich zusätzlich auf den Umstand, dass das in E 12 mittels „Überblick“ verknüpfte Dokument – ebenso wie E 10 – eine Sitemap enthält. Für die Tatsache, dass die Sitemap als niedrigfrequentes Hypertextsortenmodul aufgefasst werden muss, kann der mögliche Grund angeführt werden, dass die Anfertigung einer vollständigen Sitemap des (gesamten) universitären Webauftritts einen sehr aufwändigen und komplexen Prozess darstellt, den vermutlich einige Produzenten scheuen; weiterhin besteht die Möglichkeit, dass die Produzenten der Ansicht sind, dass eine Suchfunktion über das gesamte Webangebot in der Lage ist, die Funktionalität einer Sitemap vollständig zu ersetzen.

11.4.3 Primäre Inhalte und Themenbereiche

Neben einer Analyse der primären Navigationshilfen wurden hochfrequente Themenbereiche ermittelt, die in den 35 Einstiegsseiten entweder per Hyperlink referenziert werden oder unmittelbar in dem Dokument als Text oder Textfragment enthalten sind. Dabei kann grob zwischen primären und sekundären Themen differenziert werden: Erstere umfassen vornehmlich

[15] Die Usability-Literatur lehnt die Reduplikation grafischer Navigationshilfen in Textform ab, da hierdurch der kognitive Ballast potenziell erhöht wird (vgl. Abschnitt 3.4). Zudem ist es gerade in Einstiegsseiten wichtig, ein kompaktes, transparent strukturiertes und redundanzfreies Webdesign zu präsentieren. Aus technischen Gründen ist die Wiederholung von Hyperlinks unnötig, da eine textuelle Umschreibung des Hyperlinks im `alt`-Attribut des Elements `img` oder dem `title`-Attribut des `a`-Elements hinterlegt werden kann.

Inhalte, die sich auf die Kernbereiche und -aufgaben einer Hochschule beziehen (vgl. Tabelle 11.6), während unter sekundären Themen insbesondere weiterführende Informationen über eine Universität verstanden werden (vgl. Tabelle 11.7).

Die hochfrequenten primären Inhaltsgebiete „Studium", „Forschung" und „Einrichtungsübersicht" entsprechen weitestgehend den Themenfeldern, die in Abschnitt 11.4.2 diskutiert wurden. Es fällt jedoch auf, dass Unterschiede bezüglich der individuellen Frequenzen existieren (vgl. Tabelle 11.6 sowie Tabelle 11.4): Der Bereich „Forschung" wird in 30 primären Navigationshilfen (86% der Einstiegsseiten) aufgeführt, weitere Nennungen in anderen Bereichen der Einstiegsseiten konnten nicht ermittelt werden. „Studium" wird zusätzlich in einem Fall außerhalb der primären Navigationshilfe aufgeführt (94%). Dies gilt insbesondere für Auflistungen von Einrichtungen mit Vorkommen in insgesamt 29 Einstiegsseiten (in 18 Fällen in der primären Navigationshilfe). Bei den korrespondierenden Hyperlinkanzeigern bzw. Überschriften können verschiedene Konventionen ausgemacht werden: In acht Fällen wird „Fakultäten" verwendet, fünf Vorkommen sind für „Fachbereiche" zu verzeichnen, jeweils zwei Belege konnten für „Einrichtungen", „Einrichtungen der Universität", „Fakultäten und Institute" sowie „Universitäre Einrichtungen" ermittelt werden. Zusätzlich existieren verschiedene Varianten, die initial Fakultäten (z. B. „Fakultäten und Einrichtungen", „Fakultäten, Einrichtungen und Organe", „Fakultäten, Institute und Forschungseinrichtungen") oder Fachbereiche aufführen (z. B. „Fachbereiche und Fächer", „Fachbereiche und Institute", „Fachbereiche und Fakultäten", „Fachbereiche, Zentren, Kliniken").[16] Eine Sonderstellung nimmt E 14 (European Business School) ein, die eine Liste der 17 Professuren über den Hyperlinkanzeiger „Lehrstühle" zur Verfügung stellt. Weiterhin wurden die Vorkommen zentraler Organisationseinheiten ermittelt: In 54% aller Einstiegsseiten wird auf die Website der Bibliothek einer Hochschule verwiesen, ein Hyperlink zum jeweiligen Rechenzentrum wird in 29% der Dokumente angeboten. Daneben werden die Webauftritte des akademischen Auslandsamts (26%), der Pressestelle (20%) und der Mensa (9%) referenziert, zwei weitere Einstiegsseiten verweisen auf die aktuellen Speisepläne der zugehörigen Mensen.

Eines der wesentlichen Charakteristika eines Portals ist die Präsentation aktueller Informationen in der Einstiegsseite, wobei mehrere Themengebiete parallel in eigenständigen Listen aufgeführt werden. In den 35 Dokumenten der Stichprobe können derartige Listen in 23 Fällen belegt werden (66%), sie beschränken sich jedoch in der Regel auf aktuelle Veranstaltungen (öffentliche Vortragsreihen, Konferenzen etc.) oder allgemeine Neuigkeiten (z. B. ein Aufruf zur Wahl eines Hochschulgremiums).[17] Die in Abbildung 11.2 dargestellten Beispiele verdeutlichen, dass unterschiedliche Formen der Realisierung vorliegen, die verschiedene Gemeinsamkeiten aufweisen.[18] Ein geeignetes Mittel zur Beschreibung dieser Gemeinsam-

[16] Diese Beispiele zeigen, dass Konventionen der Hypertextknotensorte sowie des Hypertexttyps *Webauftritt einer Universität* von den Landeshochschulgesetzen beeinflusst werden. Während z. B. in Hessen oder Niedersachsen die Strukturierung einzelner Fachgebiete in Fachbereiche vorgesehen ist, wird in Bayern und Baden-Württemberg eine Aufteilung in Fakultäten empfohlen bzw. verbindlich vorgeschrieben (vgl. Abschnitt 13.4).

[17] In fast allen verbleibenden Einstiegsseiten, die dieses Hypertextsortenmodul nicht ausprägen, ist zumindest ein Hyperlink enthalten, der auf ein eingebettetes HTML-Dokument mit aktuellen Informationen verweist; in diesen Fällen fungiert das Hypertextsortenmodul als Hypertextknotensorte. Die Hyperlinkanzeiger heißen z. B. „Aktuelles", „News" und „Die neuesten Meldungen". In einigen Fällen werden beide Formen eingesetzt.

[18] Sämtliche Vorkommen dieses Hypertextsortenmoduls teilen sich die Eigenschaften der listenartigen Darstellung der Informationen sowie der Bereitstellung von Hyperlinks zu weiterführenden Informationen.

			Kommerzielle Einstiegsseiten	*Universitäre Sites*
Merkmal	Frequenz	Prozent	Schütte (2004a) Prozent	Kamenz et al. (1998) Prozent
„Studium“, „Studium und Lehre“ (o. ä.)	33	94	—	—
„Forschung“, „Forschung und Kooperation“ (o. ä.)	30	86	—	—
Übersicht über Einrichtungen	29	83	—	99
„Fakultäten“	8	28	—	—
„Fachbereiche“	5	17	—	—
„Einrichtungen“	2	7	—	—
„Einrichtungen an der Universität“	2	7	—	—
„Fakultäten und Institute“	2	7	—	—
„Universitäre Einrichtungen“	2	7	—	—
Aktuelle Veranstaltungen bzw. Informationen (Features)	23	66	51	—
„Aktuelles“ (als Überschrift)	3	14	—	—
„Aktuell:“ (als Überschrift)	2	9	—	—
„Schlagzeilen“ (als Überschrift)	1	5	—	—
„News“ (als Überschrift)	1	5	—	—
„Neuigkeiten“ (als Überschrift)	1	5	—	—
„Veranstaltungen – News & Events“ (als Überschrift)	1	5	—	—
„Jugend forscht 2001 an der TU“	1	5	—	—
Spezifische Einrichtungen				
„Bibliothek“ (o. ä.)	19	54	—	—
„Rechenzentrum“, „Hochschulrechenzentrum“ (o. ä.)	10	29	—	—
„Akademisches Auslandsamt“ (o. ä.)	9	26	—	59
„Pressestelle“ (o. ä.)	7	20	—	46
„Mensa“	3	9	—	—
Speiseplan der Mensa	2	6	—	27
„Mensaessen“	1	50	—	—
„Mensaplan“	1	50	—	—
Universitätsverwaltung	16	46	—	—
„Verwaltung“	7	44	—	—
„Universitätsverwaltung“	3	19	—	—
„Zentrale Universitätsverwaltung“	1	6	—	—
„Kanzler, Verwaltung“	1	6	—	—
„Organisation“	1	6	—	—
„Organisation und Struktur“	1	6	—	—
Universitätsleitung	15	43	—	58
„Universitätsleitung“	3	20	—	—
„Leitung“	1	7	—	—
„Leitung und Organe“	1	7	—	—
„Der Präsident“	1	7	—	—
„Rektor, Rektorat“	1	7	—	—
„Rektorat“	1	7	—	—
„Hochschulpolitik“	1	7	—	—
Informationen zur Region	14	40	—	—
„Region Cottbus“	1	7	—	—
„Region Freiberg“	1	7	—	—
„Münster“	1	7	—	—
„Darmstadt“	1	7	—	—
„Stadt und Region“	1	7	—	—
„Umfeld“	1	7	—	—
Alumni, Ehemalige	11	31	—	—
„Alumni“	3	27	—	—
„Alumni-Club“	1	9	—	—
„Alumni/Fördervereine“	1	9	—	—
„Alumni/Förderer“	1	9	—	—
„Absolventen“	1	9	—	—
„Absolventen-Treffpunkt“	1	9	—	—
„Weiterbildung“ (o. ä.)	8	23	—	54
„Forschungsschwerpunkte“, „Forschungsprofil“ (o. ä.)	6	17	—	56
„Forschungsbericht“ (o. ä.)	4	11	—	—
Die Institution vorstellender Kurztext	2	6	24	—

Tabelle 11.6: Ergebnisse der Makrostruktur- und Inhaltsanalyse – Primäre Inhalte und Themenfelder

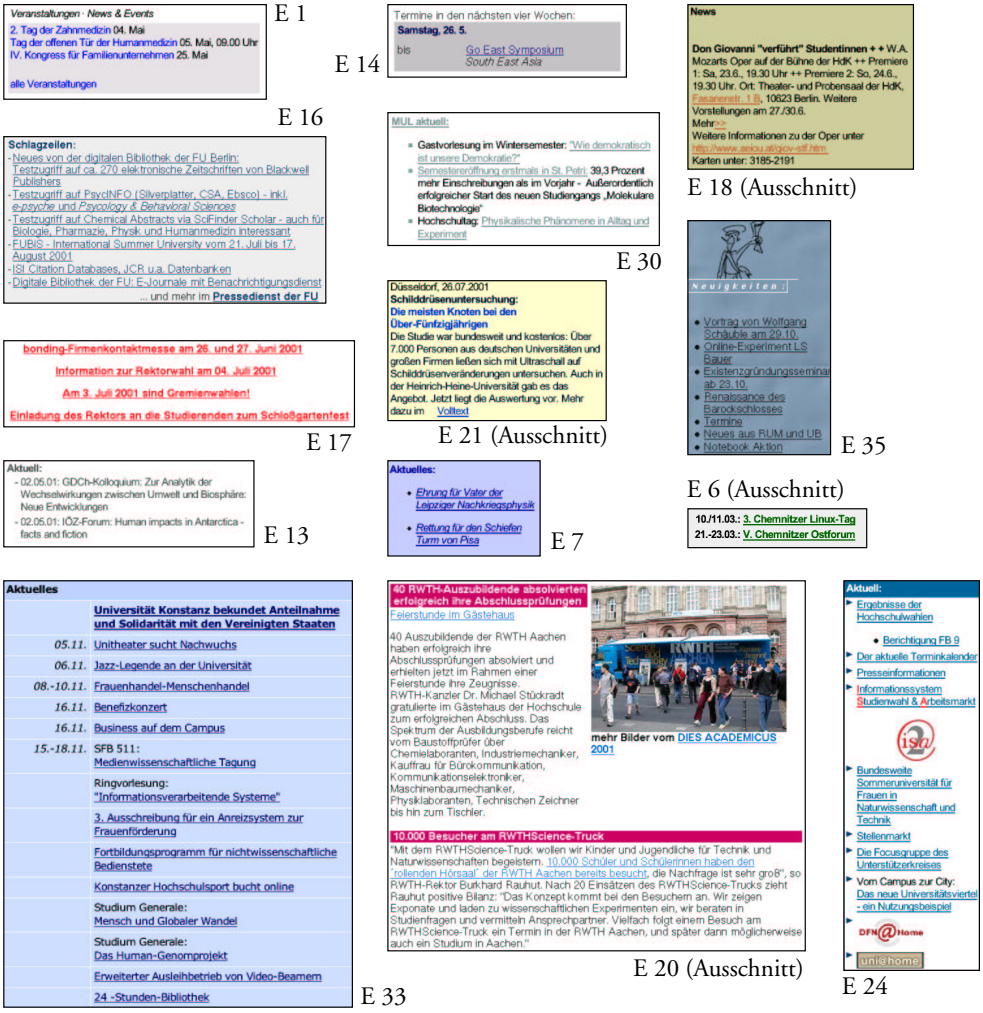


Abbildung 11.2: Beispiele für das Hypertextsortenmodul *aktuelle Neuigkeiten* in den Einstiegsseiten der Stichprobe

keiten sind nicht etwa, wie bei den primären Navigationshilfen, trennscharfe Typen, sondern (binäre) Merkmale. Die in der Abbildung aufgeführten Beispiele decken das Spektrum der Ausprägungen dieses Hypertextsortenmoduls in der Stichprobe vollständig ab: Während einige Instanzen ohne eine übergeordnete Etikettierung präsentiert werden (z. B. E 6 und E 17), sind in anderen Instanzen Überschriften enthalten; neben „Aktuelles" (insgesamt drei Vorkommen), werden diesbezüglich „Aktuell" (zwei Vorkommen) sowie „Schlagzeilen", „News", „Neuigkeiten", „Veranstaltungen – News & Events" (jeweils ein Vorkommen) verwendet. Ein weiteres Merkmal betrifft die Markierung des Datums, das in einigen Fällen nicht verwendet wird (z. B. in E 7, E 24 und E 30).[19] In anderen Dokumenten wird das Datums links

[19] Das Fehlen einer Datumsangabe macht die rezipientenseitige Einschätzung der Aktualität einer präsentierten Nachricht unmöglich, so dass damit zu rechnen ist, dass sich Datumsangaben in Zukunft zu einem konventionalisierten Bestandteil dieses Hypertextsortenmoduls entwickeln werden.

(E 6, E 13, E 33), rechts (E 1), oberhalb (E 14, E 21) oder innerhalb des Eintrags (E 17, E 18) dargestellt. Weiterhin enthalten einige Instanzen lediglich den Titel einer Veranstaltung oder einer Pressemitteilung, während in anderen Fällen zusätzlich der Anreißer oder der gesamte Text dargestellt wird. Einen Extremfall stellt E 21 dar: Diese sehr lange Einstiegsseite enthält die Anreißer bzw. vollständigen Texte von insgesamt 12 Pressemitteilungen, die etwa 90% des Inhalts dieses Dokuments ausmachen (Abbildung 11.2 stellt eine einzelne Pressemitteilung dar).[20] In ähnlicher Form ist E 20 strukturiert, wobei jedoch nur drei Pressemitteilungen präsentiert werden, weshalb das Textdesign dieses Dokuments insgesamt harmonischer und übersichtlicher wirkt. Ein weiteres Merkmal bezieht sich auf Dekorationsobjekte, insbesondere Fotos, die in einigen Fällen die präsentierten Anreißer flankieren (z. B. in E 20 und E 21). Der Umfang einer Instanz wurde bereits angesprochen und kann als weiteres Merkmal konzeptualisiert werden: Neben der Darstellung von einer (E 14) oder zwei Neuigkeiten (E 6, E 7) sind in E 21 insgesamt 12 und in E 33 sogar 15 Einträge enthalten.

Informationen zur Hochschulverwaltung bzw. -leitung werden in sechs bzw. fünf primären Navigationshilfen verknüpft. Bezüglich der vollständigen Dokumente können insgesamt 16 (46%) bzw. 15 (43%) Hyperlinkanzeiger belegt werden, d. h. für die Mehrzahl der Produzenten handelt es sich um Themengebiete, die nur von untergeordnetem Interesse sind und daher in sekundäre Navigationshilfen ausgelagert werden. Die für den ersten Bereich verwendeten Hyperlinkanzeiger können als konventionalisierter bezeichnet werden: „Verwaltung“ (7 Vorkommen), „Universitätsverwaltung“ (3) und verschiedene Varianten, z. B. „Zentrale Universitätsverwaltung“. Zur Verknüpfung des zweiten Bereichs werden „Universitätsleitung“ (3 Vorkommen), „Leitung“ (2) und insgesamt zehn Varianten eingesetzt, z. B. „Präsidialamt“, „Der Präsident“, „Rektor, Rektorat“, „Rektorat“ (vgl. Fußnote 16), „Leitung und Organe“ sowie die sehr untypische Bezeichnung „Hochschulpolitik“.

Informationen zur Region, in der eine Universität ansässig ist, werden in 11 primären Navigationshilfen und insgesamt 14 Einstiegsseiten präsentiert.[21] In fünf Einstiegsseiten besitzen die Hyperlinkanzeiger bzw. Überschriften eine identische Struktur: „Region Freiberg“, „Region Düsseldorf“, „Region Cottbus“, „Region Ulm“ und „Region Konstanz“; „Gießen und die Region“ kann als Variante aufgefasst werden. Der Name der Stadt, in der sich eine Hochschule befindet, ist in nahezu allen Fällen Bestandteil der offiziellen Bezeichnung einer Universität, die in ebenfalls nahezu allen Fällen in einer großen Schrifttype im oberen Bereich des Dokuments prominent platziert wird. Daher handelt es sich bei der Nennung der Stadt in den oben angegebenen Belegen um redundante Reduplikationen, da der Leser aus dem Kontext erschließen kann, um welche Region es sich handelt. Entsprechend sind in der Stichprobe auch Hyperlinks wie „Region“ (E 17), „Stadt und Region“ (E 28), „Unsere Region und Partner“ (E 8), „Umfeld“ (E 9) und „Um die Uni rum“ (E 35) enthalten. Einen umgekehrten Weg verfolgen die Produzenten dreier weiterer Einstiegsseiten, in denen lediglich der Name der Stadt aufgeführt wird: „Darmstadt“, „Münster“, „Chemnitz“. Den ausführlichsten Hyperlinkanzeiger enthält E 23: „Informationen zu Passau und Umgebung“.

[20] Diese Dominanz aktueller Neuigkeiten kann auch in E 18 beobachtet werden. Hierbei handelt es sich um den Webauftritt der Hochschule der Künste Berlin, die in den Pressemitteilungen Werbung für Veranstaltungen betreiben, an denen Studierende dieser Universität mitwirken.

[21] Interessanterweise werden diese Informationen nur in den Einstiegsseiten derjenigen Hochschulen aufgeführt, die in kleineren Städten ansässig sind; eine Ausnahme stellt die Universität Düsseldorf dar.

Die drei Vorkommen in den zielgruppenspezifischen Navigationshilfen eingerechnet sind in den Einstiegsseiten elf Hyperlinkanzeiger enthalten (31%), die auf Informationen für ehemalige Studierende verweisen. In drei Fällen wird „Alumni“ verwendet (E 1, E 10, E 16), wobei vier Varianten existieren: „Alumni/Fördervereine“ (E 12), „Alumni/Förderer“ (E 24), „Alumni-Club“ (E 32) und „HdK-Alumni“ (E 18). Die gebräuchlicheren Ausdrücke werden in den verbleibenden Vorkommen benutzt: „Absolventen“ (E 19), „Absolventen-Treffpunkt“ (E 27), „Verein der Ehemaligen“ (E 33) und „AbsolventUM“ (E 35). Die Existenz dieser Komponente in 31% der Einstiegsseiten deutet an, dass Themen, die sich an eine bestimmte Rezipientengruppe wenden, in Zukunft vermehrt durch zielgruppenspezifische Navigationshilfen gebündelt werden, die in lediglich 20% der Dokumente enthalten sind.

Ein Text, der eine Universität unmittelbar in der Einstiegsseite in Kurzform vorstellt, ist lediglich in zwei Webauftritten enthalten. In E 30 wird für diesen Zweck unterhalb der Begrüßung des Rezipienten ein einzelner Satz verwendet: „Die MUL bietet innovative Studien- und Forschungsmöglichkeiten an der Schnittstelle von Medizin, Naturwissenschaften und Informationstechnologie.“ In E 22 werden mehrere Texte eingesetzt, die sich unterhalb der Hyperlinks in der primären Navigationshilfe befinden (vgl. Abbildung 11.1). Diese Kurztexte erläutern die zugehörige inhaltliche Kategorie und können in ihrer Gesamtheit als universitätsvorstellender Kurztext aufgefasst werden.

11.4.4 Sekundäre Inhalte und Themenbereiche

In 26 Dokumenten (74%) werden anderssprachige Versionen der Einstiegsseiten referenziert (vgl. Tabelle 11.7). Die entsprechenden Hyperlinkanzeiger werden in nahezu allen Fällen in der sekundären Navigationshilfe positioniert, ihre Frequenzen geben Auskunft über die Verteilung der angebotenen Sprachen in den 35 Webauftritten.[22] Die Flagge Großbritanniens wird in zehn Einstiegsseiten als kleinformatiges Icon eingebettet. In acht Fällen wird der Hyperlinkanzeiger „English“ verwendet, von dem mehrere Varianten existieren, z. B. „This page in English“ (E 15), „English Version“ (E 17), „English Speakers“ (E 18), „English readers“ (E 19), „In English“ (E 35) und „In case you prefer English please refer to our international homepage“ (E 23). Neben einer englischsprachigen Version, die in allen genannten 26 Einstiegsseiten verfügbar ist, können weitere Sprachen belegt werden, die zugleich das intendierte Einzugsgebiet einer Hochschule verdeutlichen. Rechts neben der britischen Flagge in der Einstiegsseite der RWTH Aachen (E 20) verweist die Flagge der Niederlande auf eine entsprechende Sprachversion. Die Universität Lübeck (E 30) bietet ihr Webangebot zusätzlich in Französisch und Schwedisch an, wobei die Nationalflaggen die alleinigen Hyperlinkanzeiger darstellen. Der zentrale Bereich der Website der Universität Bonn (E 32) ist zusätzlich in einer französischsprachigen Variante verfügbar.

Kontaktinformationen sind in 22 Einstiegsseiten enthalten (63%), wobei in jedem Fall mindestens eine E-Mail-Adresse oder ein `mailto:`-Hyperlink mit einem Hyperlinkanzeiger wie z. B. „Webmaster“ verfügbar ist. Zusätzlich wird die Straßenadresse (8 Vorkommen), die Telefonnummer (6 Vorkommen) und die Faxnummer (4 Vorkommen) angegeben. Falls

[22] Eine untypische Präsentation der englischsprachigen Variante wird in E 28 verwendet, in der die primäre Navigationshilfe in Form zweier einfacher Listen innerhalb einer Tabelle realisiert ist. Die Liste in der linken Spalte enthält deutschsprachige und die rechte Liste englichsprachige Hyperlinkanzeiger.

Merkmal	Frequenz	Prozent	*Kommerzielle Einstiegsseiten*		*Universitäre Sites*
			Nielsen und Tahir (2002) Prozent	Schütte (2004a) Prozent	Kamenz et al. (1998) Prozent
Version in alternativer Sprache(n)	26	74	—	—	71
Britische Flagge (als Icon)	10	38	—	—	—
„English“	8	31	—	—	—
„This page in English“	1	4	—	—	—
„English version“	1	4	—	—	—
„English Speakers“	1	4	—	—	—
„English readers“	1	4	—	—	—
„In English“	1	4	—	—	—
Kontaktinformationen	22	63	90	—	—
E-Mail-Adresse	22	63	—	—	—
Straßenadresse	8	23	—	14	100
Telefonnummer	6	17	—	—	—
Faxnummer	4	11	—	—	—
Stellenanzeigen	15	43	74	—	46
„Stellenausschreibungen“	5	33	—	—	—
„Stellenmarkt“	2	13	—	—	—
„Stellenangebote“	1	7	—	—	—
„Stellenanzeigen“	1	7	—	—	—
„Stellenbörse“	1	7	—	—	—
Anfahrtsplan, Lageplan, Wegbeschreibung	13	37	—	—	60
Verzeichnisse					
Vorlesungsverzeichnis	12	34	—	—	33
„Vorlesungsverzeichnis“	7	58	—	—	—
„Lehrveranstaltungen“	1	8	—	—	—
„Vorlesungen“	1	8	—	—	—
Telefonverzeichnis	11	31	—	—	62
Telefon (als Icon)	2	18	—	—	—
„Telefon“	1	9	—	—	—
„Telefon & E-Mail“	1	9	—	—	—
„E-Mail/Tel.“	1	9	—	—	—
„Telefon- und eMail-Verzeichnis“	1	9	—	—	—
E-Mail-Verzeichnis	7	20	—	—	62
„Mitarbeiterverzeichnis (Mail, Telefon)“	1	14	—	—	—
„X500-Verzeichnis“	1	14	—	—	—
„Telefon/Email-Verzeichnisse“	1	14	—	—	—
Namensverzeichnis	6	17	—	—	—
„Personal“	1	17	—	—	—
„Personen und Fakten“	1	17	—	—	—
„Personen- und Einrichtungsverzeichnis“	1	17	—	—	—
Adressverzeichnis	5	14	—	—	—
„Adressen“	2	40	—	—	—
„Adresse“	1	20	—	—	—
„Adressen und Gremien“	1	20	—	—	—
Motto, Leitbild, Leitspruch, Slogan	5	14	—	45	—
„In omnibus veritas“	1	20	—	—	—
„Sciendo Docendo Curando“	1	20	—	—	—
„veritas iustitia libertas“	1	20	—	—	—
„Zukunft beginnt bei uns!“	1	20	—	—	—
„Zur Freiheit ermutigen!“	1	20	—	—	—
„Schwarzes Brett“ (o. ä.)	4	11	—	—	—
Ergebnisse von Hochschulrankings	2	6	—	—	—
Gästebuch	0	0	—	—	7
Benutzeranmeldung möglich	0	0	52	—	—

Tabelle 11.7: Ergebnisse der Makrostruktur- und Inhaltsanalyse – Sekundäre Inhalte und Themenfelder

keinerlei Möglichkeiten der Kontaktaufnahme in der Einstiegsseite genannt werden, liegt ein Hyperlinkanzeiger wie beispielsweise „Kontakt“, „Redaktion“ oder auch „Webmaster“ vor, über den der Rezipient zu einer Folgeseite gelangt, die die Kontaktinformationen bereitstellt. Derartige Hyperlinks liegen in insgesamt 25 Einstiegsseiten vor, in mehreren Fällen werden beide Möglichkeiten der Präsentation genutzt.

In 15 der 35 untersuchten HTML-Dokumente (43%) werden Stellenanzeigen aufgeführt, wobei zwei dieser Vorkommen nicht als Hyperlink realisiert sind, sondern als charakterisierende Schlagworte unterhalb der Kategorien „Infos“ (E 5: „Stellenausschreibungen“) und „Aktuelle Informationen“ (E 4: „Stellenangebote“) in primären Navigationshilfen vom Typ komplexe Tabelle eingesetzt werden (vgl. Abbildung 11.1). Unter dem zuletztgenannten thematischen Bereich wird auch in anderen Einstiegsseiten auf Stellenanzeigen Bezug genommen, jedoch handelt es sich dabei um Ausprägungen als Hyperlinkanzeiger (z. B. in E 6, E 8, E 10): „Stellenausschreibungen“ unterhalb von „Aktuelles“ (vgl. ebenfalls Abbildung 11.1). Alternativ verwendete Hyperlinkanzeiger lauten „Ausschreibungen“ (in E 7 unter „Informationen und Kontakte“), „Stellenanzeigen“ (in E 15 unter „News“), „Stellenmarkt“ (in E 24 unter „Aktuell“) und „Stellenausschreibungen“ (in E 35 unter „Initiativen“).[23] Einige Produzenten nehmen an, dass nur ein geringes Interesse an Stellenanzeigen besteht und platzieren diese Verknüpfung in der sekundären Navigationshilfe (E 12, E 13).

Neben den bislang genannten sekundären Inhalten werden fünf unterschiedliche Typen von Verzeichnissen referenziert. Auf ein „Vorlesungsverzeichnis“ wird – mit diesem Hyperlinkanzeiger – in sieben Knoten verwiesen (insgesamt 58%). Die Varianten geben spezifischere Informationen an, z. B. „Vorlesungsverzeichnis SS 2001 jetzt online“ (E 18) und „Vorlesungsverzeichnis (jetzt mit WS 01/02)“ (E 28). Zwei Produzenten verzichten auf den zweiten Bestandteil des Kompositums (E 35: „Vorlesungen“, E 13: „Lehrveranstaltungen“); hierdurch entfällt das Etikett der im Zieldokument zu erwartenden Hypertextsorte, wodurch die Kohärenzbildung negativ beeinflusst werden kann. Abgesehen von Verweisen auf ein Vorlesungsverzeichnis enthalten die Einstiegsseiten Hyperlinks zu Telefon- (11 Vorkommen), E-Mail- (7), Namens- (6) und Adressverzeichnissen (5). In der Mehrzahl der Fälle liegen Mischungen dieser Typen vor, z. B. „Mitarbeiterverzeichnis (Mail, Telefon)“ (E 7), „Telefon & E-Mail“ (E 13), „E-Mail/Tel.“ (E 16), „Telefon- und eMail-Verzeichnis“ (E 30) und „Personenverzeichnis der Universität, Namen, Adressen, Telefonnummern usw.“ (E 15).[24] Gelegentlich werden auch weiterführende Verzeichnisse genannt, z. B. „Personen- und Einrichtungsverzeichnis“ (E 30), „Personen und Fakten“ (E 4) und „Adressen und Gremien“ (E 7).

Das Motto bzw. der Leitspruch einer Hochschule wird in fünf Einstiegsseiten genannt und ist in drei Fällen in dem als Grafik eingebetteten Siegel der Universität enthalten, wodurch die Tradition einer Institution betont wird (E 8: „Sciendo Docendo Curando“, E 16: „veritas iustitia libertas“, E 35: „In omnibus veritas“). Die beiden anderen, jeweils als eigenständige Logos realisierten Vorkommen können als Bestandteil der Corporate-Identity aufgefasst

[23] Unter dem Aspekt der Kohärenzbildung muss die Einordnung des zuletzt genannten Hyperlinks in den Bereich „Initiativen“ als problematisch bezeichnet werden. Ein mehrdeutig etikettierter Hyperlinkanzeiger wird am unteren Seitenrand in E 18 verwendet: „Stellenausschreibung online+ + Mehr >>>“ könnte einerseits dazu dienen, eine Stellenausschreibung auf dem Webserver zu veröffentlichen oder auf die Liste verfügbarer Stellenangebote zuzugreifen. Tatsächlich führt dieser Hyperlink unmittelbar zu einer spezifischen Ausschreibung.

[24] Neben textuell realisierten Hyperlinkanzeigern werden in einigen Fällen auch Icons verwendet, die ein stilisiertes Telefon oder einen Brief zeigen (z. B. in E 3 und E 25).

werden und sollen helfen, das eigenständige Profil der Institution zu vermitteln (E 1: „Zur Freiheit ermutigen!“, E 20: „Zukunft beginnt bei uns!“). Die Profilierung einer Universität kann auch durch die Aufnahme von Informationen zu Hochschulrankings auf der Einstiegsseite hervorgehoben werden. Von dieser Möglichkeit der Werbung für Studiengänge oder Forschungsleistungen, die von einer unabhängigen Stelle als herausragend eingestuft wurden, machen jedoch nur zwei Hochschulen Gebrauch (E 13, E 33).

11.4.5 Metainformationen

In den 35 Einstiegsseiten existieren verschiedene Konventionen zur Präsentation von Metainformationen (vgl. Tabelle 11.8). In 71% der Dokumente wird das Datum der letzten Änderung genannt. Es wird in allen Fällen am unteren Seitenrand im Bereich der typografisch abgesetzten Fußzeile aufgeführt. Neben vollständigen Angaben, die das Datum und teilweise die Uhrzeit umfassen, werden in einigen Einstiegsseiten lediglich Monat und Jahr genannt. Oftmals werden diese Angaben entsprechend gekennzeichnet (z. B. in E 1: „Aktualisiert im Januar 2001“, E 2: „Letzte Änderung: 05.12.2000“), da die bloße Präsentation einer Datumsangabe als Datum der Erstellung missverstanden werden könnte.[25] In unmittelbarer Nähe der Angabe des Datums der letzten Änderung wird in 18 Einstiegsseiten (51%) der Name des Erstellers (z. B. in E 2: „Bernhard Brandel und Dr. Bernward Tewes“), eine Tätigkeitsbezeichnung (z. B. E 27: „WWW-Administration“, E 28: „Webmaster“) oder die Bezeichnung einer Gruppe von Personen genannt (z. B. E 16: „Webteam“, E 20: „Das Web-Team“); in E 12 wird der Name des Unternehmens aufgeführt, das diesen Webauftritt gestaltet hat.

Eine besonders ausgeprägte Konvention existiert bezüglich der Nennung der Person, die die juristische Verantwortung für die Website trägt. Diese Information ist in 16 Einstiegsseiten zumeist durch einen Hyperlink erreichbar, der in 57% der Vorkommen mit der aus den Printmedien stammenden Bezeichnung „Impressum“ beschriftet ist. Weitere Hyperlinkanzeiger bzw. Hinweise lauten z. B. „Inhaltliche Verantwortung“ (E 1), „V. i. S. d. P.: Martin Schulz, Rektoramt, Universität Ulm“ (E 8), „Redaktion“ (E 11), „Herausgeber: Der Rektor“ (E 21) oder „Verantwortlich für den Inhalt: Pressestelle“ (E 26). Auch diese Angabe wird in allen Fällen innerhalb bzw. unterhalb der Fußzeile aufgeführt. In diesem Bereich befindet sich darüber hinaus in 26% der Einstiegsseiten eine Bitte um Rückmeldungen oder der Aufruf zur Kontaktaufnahme, z. B. „Anregungen und Kritik bitte via Email an diese Adresse“ (E 1), „Technische Fragen bitte an den Webmaster“ (E 8), „Vermissen Sie etwas auf dieser Seite? Schreiben Sie uns!“ (E 13) und „Kommentare zu unserem Webangebot sind jederzeit willkommen.“ (E 24). Hilfeseiten werden in sieben Fällen referenziert (vgl. Abschnitt 4.6.2), wobei sich die jeweiligen Hyperlinks in der primären (drei Hyperlinks) oder der sekundären Navigationshilfe (drei Hyperlinks) oder innerhalb der Fußzeile befinden. In vier Fällen fungiert das Wort „Hilfe“ als Hyperlinkanzeiger, zusätzlich werden „Technische Informationen

[25] Das Datum der Erstellung wird lediglich in E 23 explizit genannt. Die explizite Identifizierung einer Datumsangabe als Datum der letzten Änderung erscheint somit redundant zu sein (doch siehe Abschnitt 10.5.11) und wird entsprechend in neun Dokumenten nicht verwendet. In den verbleibenden 16 Dokumenten werden unterschiedliche Formen der Markierung benutzt, „Letzte Änderung:“ wird jedoch in acht Fällen eingesetzt und kann als Konvention aufgefasst werden. Die längste Identifizierung befindet sich in E 4: „Diese Seite wurde zuletzt überarbeitet am: 2. November 2000“.

Merkmal	Frequenz	Prozent	*Kommerzielle Einstiegsseiten* Nielsen und Tahir (2002) Prozent	*Universitäre Sites* Kamenz et al. (1998) Prozent
Datum der letzten Änderung des HTML-Dokuments	25	71	—	—
Position: Mitte unten	13	52	—	—
Position: Links unten	8	32	—	—
Position: Rechts unten	4	16	—	—
Name des Autors bzw. Verantwortlichen	18	51	—	—
Position: Mitte unten	10	56	—	—
Position: Links unten	7	39	—	—
Position: Rechts unten	1	6	—	—
Typ des Eigennamens				
Bezeichnung einer Funktion bzw. Gruppe	12	67	—	—
Vor- und Nachname einer Person	5	28	—	—
Name einer Firma	1	6	—	—
Impressum, Betreuung, Redaktion	16	46	—	—
Position: Mitte unten	8	50	—	—
Position: Links unten	6	38	—	—
Position: Rechts unten	1	6	—	—
Position: In primärer Navigationshilfe	1	6	—	—
Bezeichnung des Hyperlinks				
„Impressum“	9	57	—	—
„Redaktion“	1	6	—	—
„Inhaltliche Verantwortung“	1	6	—	—
„Verantwortlich für Inhalt: Webteam“	1	6	—	—
„Verantwortlich für den Inhalt: Pressestelle“	1	6	—	—
„verantwortlich: Rektorat der Universität Bonn“	1	6	—	—
„Das Web-Team“	1	6	—	—
„WWW-Administration“	1	6	—	—
„Webmaster“	1	6	—	—
Kontaktsequenz und Begrüßung	11	31	—	—
Begrüßung	10	91	—	—
„Herzlich Willkommen“	2	20	—	—
„Willkommen“	2	20	—	—
„Willkommen / Welcome“	1	10	—	—
„Welcome – Bienvenue – Bienvenido“	1	10	—	—
„Willkommen auf unserem World Wide Web-Server!“	1	10	—	—
„Willkommen auf den WWW-Seiten der Universität Essen“	1	10	—	—
„Willkommen an der Katholischen Universität Eichstätt“	1	10	—	—
„Willkommen in der Rheinischen Friedrich-Wilhelms-Universität Bonn“	1	10	—	—
Geleitwort	1	9	—	—
Bitte um Rückmeldungen und Kontaktaufnahme	9	26	—	—
Copyright-Hinweis	9	26	—	—
Position: Mitte unten	5	56	—	—
Position: Links unten	3	33	—	—
Position: Rechts unten	1	11	—	—
Hilfeseiten, Benutzungshinweise	7	20	54	12
In primärer Navigationshilfe	3	43	—	—
In sekundärer Navigationshilfe	3	43	—	—
In Fußzeile	2	22	—	—
Zugriffszähler	2	6	—	15
Datenschutzrichtlinien, Haftungsausschluss, Disclaimer	1	3	86	—
Datum der Erstellung des HTML-Dokuments	1	3	—	—
„Optimiert für …“-Hinweis	1	3	—	—
„under construction“-Hinweis	0	0	—	8

Tabelle 11.8: Ergebnisse der Makrostruktur- und Inhaltsanalyse – Metainformationen

zum WWW-Server“ (E 8), „Über diesen Server“ (E 26) oder ein Fragezeichen als Icon verwendet (E 25). Zugriffszähler sind in nur zwei Einstiegsseiten enthalten (E 21, E 35). Bezüglich dieses Hypertextsortenmoduls scheint ein rückläufiger Trend zu existieren, denn Kamenz et al. (1998) berichten eine Frequenz von 15%.[26] Abschließend sei auf den Umstand hingewiesen, dass lediglich in einer Einstiegsseite (E 7) Informationen zur bestmöglichen Darstellung des HTML-Dokuments enthalten sind, die sich darüber hinaus innerhalb des Bereichs der primären Navigationshilfe befinden und somit eine sehr untypische Position aufweisen.

Tabelle 11.8 umfasst aus Darstellungsgründen auch Angaben zum Einsatz von Kontaktsequenzen und Begrüßungen, die streng genommen nicht als Teil der Metainformationen aufzufassen sind. Zehn Einstiegsseiten (29%) enthalten einen einfachen Willkommensgruß, der – mit Ausnahme von E 29 – in allen Fällen im oberen Seitenzentrum positioniert ist; Schütte (2004a) berichtet für die von ihr untersuchten kommerziellen Homepages einen Wert von 71%.[27] Neben der knappen und formal erscheinenden Begrüßungsfloskel „Willkommen“ (E 30, E 31) wird auch „Herzlich Willkommen“ (E 19, E 29) verwendet. Von der zuerst genannten Begrüßung existieren verschiedene Varianten, so wird dieser Gruß in zwei Einstiegsseiten in mehr als einer Sprache aufgeführt: „Willkommen / Welcome“ (E 18) sowie „Welcome – Bienvenue – Bienvenido“ (E 3), wobei auffällig ist, dass die deutschsprachige Variante fehlt. Diese Begrüßungssequenz ist als eingebettetes Bild realisiert; für den Rezipienten ist nicht unmittelbar ersichtlich, dass die drei Wörter Hyperlinkanzeiger darstellen, die zu Versionen der Einstiegsseite in den jeweiligen Sprachen führen. Die zweite Variante bezieht sich auf die Anreicherung der Willkommensfloskel mit einer Präpositionalphrase. Dabei wird in drei von vier Fällen der Name der Hochschule aufgeführt: „Willkommen an der Katholischen Universität Eichstätt“ (E 2), „Willkommen auf den WWW-Seiten der Universität Essen“ (E 24), „Willkommen in der Rheinischen Friedrich-Wilhelms-Universität Bonn“ (E 32) sowie „Willkommen auf unserem World Wide Web-Server!“ (E 23). Es existiert offenbar kein konventionalisiertes Formulierungsmuster, so werden die Rezipienten „*auf* den WWW-Seiten“, „*auf* unserem [WWW]-Server“ sowie „*in* der [...] Universität Bonn“ und „*an* der [...] Universität Eichstätt“ willkommen geheißen. Die Produzenten von E 23 und E 32 beziehen sich auf das Medium selbst, wohingegen die Autoren von E 2 und E 24 die jeweiligen Webauftritte als eine Art virtuelle Repräsentanz der Hochschule ansehen, „in“ bzw. „an“ der der Rezipient empfangen wird.[28] An den Willkommensgruß schließt sich in E 24 ein

[26] Besonders interessant ist der Zugriffszähler, der in E 21 eingesetzt wird: Bei diesem handelt es sich lediglich um eine Zahl, die sich – ohne weitere Identifizierung ihrer Funktion – in einer Sequenz von Metainformationen befindet: „30.07.01 – 13:39:38 / 227978 / 0 HSt.“. Der rückläufige Trend kann wie folgt erklärt werden: Zugriffszähler werden oftmals als eine Art Statussymbol benutzt, das indirekt die Qualität einer Website markieren soll; eine derartige Markierung scheint für die Einstiegsseite eines universitären Webangebots unangebracht zu sein, da von dort aus zahlreiche eingebettete Hypertexte erreichbar sind, so dass die Anzahl der Zugriffe auf die Einstiegsseite nur von sekundärem Interesse ist. Gerade universitäre Webangebote umfassen oftmals detaillierte statistische Auswertungen der Zugriffe auf die Website, die unter anderem auch die Anzahl der Aufrufe der Einstiegsseite beinhalten, so dass ein eigenständiger Zugriffszähler redundant erschiene.

[27] In E 29 wird unten rechts ein Foto des Hauptgebäudes der Universität Münster präsentiert, unter dem sich der Gruß „Herzlich Willkommen“ befindet. Es ist anzumerken, dass keine der Einstiegsseiten ein E-Mail-ähnliches Textstrukturmuster, eine Verabschiedung oder ein Postscriptum enthält (vgl. Abschnitt 9.6.1).

[28] Ein besonders prägnantes Beispiel für diese Metapher befindet sich in einem Dokument, das von E 11 aus erreichbar ist und eine Art Impressum mit zusätzlichen technischen Hinweisen darstellt: „*Zum Betrachten der FernUniversität* benötigen Sie einen Browser, der Frames darstellen kann“ (Hervorhebung hinzugefügt, G. R.).

Geleit- bzw. Grußwort der Rektorin an, das jedoch nicht als solches gekennzeichnet ist; der Willkommensgruß fungiert als Überschrift dieses Textes.

11.4.6 Werbung, Produkte und Dienstleistungen

Einige Elemente, die Werbezwecken dienen, wurden bereits angesprochen (z. B. die Präsentation von Hochschulrankings). Nielsen und Tahir (2002) analysieren die Frequenzen von Anzeigen für eigene (84%) oder Fremdprodukte (46%). In den universitären Einstiegsseiten befinden sich lediglich zwei Vorkommen von Werbebannern für Fremdprodukte (6%), die – entgegen der Konvention, sie im oberen Bereich des Dokuments zu präsentieren – jeweils am unteren Seitenrand dargestellt werden (E 19, E 31). Da das Konzept „kommerzielles Produkt“ im universitären Umfeld keine unmittelbare Entsprechung besitzt, können naturgemäß keine Belege für Anzeigen ermittelt werden, die sich im Sinne von Nielsen und Tahir auf eigene Produkte beziehen. Hochschulen sind aber durchaus in der Lage, Dienstleistungen unterschiedlichster Art anzubieten. Derartige Angebote sind an die Privatwirtschaft gerichtet und werden in der Regel über außeruniversitäre Gesellschaften oder der Institution zugehörige Transferstellen abgewickelt. Die insgesamt neun Vorkommen von Hyperlinks (26%) belegen, dass die Hochschulen bemüht sind, sich neuen Märkten zu öffnen, um alternative Quellen für Drittmittel zu erschließen.[29] Die verwendeten Hyperlinkanzeiger lauten z. B. „Technologietransfer“ (E 5, E 10), „Technologie und Wissenstransfer“ (E 23), „Transferzentrum Mittelhessen“ (E 4), „Technologiecentrum Chemnitz“ (E 6), „Forschung und Transfer“ (E 9) und „Angebote für Unternehmen (Forschung, Entwicklung, Beratung)“ (E 15).[30] Ein besonders interessantes Vorkommen ist in E 8 enthalten. In diesem sehr umfangreichen Dokument befindet sich innerhalb der primären Navigationshilfe (Typ komplexe Tabelle) die Kategorie „Dienstleistungen und Produkte“, die die Hyperlinkanzeiger „Materialien zu Lehre und Lernen“, „Software“, „Hardware“, „Service“, „Einrichtungen“, „Uni-Shop“, „Veröffentlichungen“ und „Ansprechpartner“ umfasst. Diese Auflistung verdeutlicht, dass der Produzent eine sehr weit gefasste und im Vergleich zu den anderen Einstiegsseiten eher untypische Lesart des Konzepts „Dienstleistung“ besitzt, die unter anderem auch an die Mitarbeiter und Studierenden der Institution selbst gerichtet ist.

11.4.7 Fazit – Das Hypertextknotensortenprofil

Aus der Inhalts- und Makrostrukturanalyse kann ein Profil der *Einstiegsseite eines universitären Webauftritts* konstruiert werden. Da jedoch nun – im Gegensatz zu den Kapiteln 9 und 10 – eine Hypertextknotensorte im Mittelpunkt des Interesses steht, müssen zusätzliche Aspekte in die aggregierende Darstellung aufgenommen werden (vgl. Kapitel 5, insbesondere Abbildung 5.4, S. 293). Tabelle 11.9 zeigt die zentralen Komponenten des Profils: Neben den ermittelten Hypertextsortenmodulen enthält die Darstellung spezifische Merkmale der Hypertextknotensorte (markiert durch „Knoten“ in der Spalte „Ebene“), die sich auf die Dekoration einer Instanz dieser Hypertextknotensorte und die Positionierung von Hypertextsortenmodulen beziehen. Gerade diese Eigenschaft stellt die Verbindung zwischen einer

[29] Kamenz et al. (1998) geben einen Wert von 40% an (vgl. Fußnote 5, S. 463).

[30] Falls eine zielgruppenspezifische Navigationshilfe Teil der Einstiegsseite ist, wird auf einen derartigen Hyperlink meist verzichtet, da er von der Kategorie „Wirtschaft“ subsumiert wird.

konventionalisierten Hypertextknotensorte und den enthaltenen Hypertextsortenmodulen dar, weshalb die in Tabelle 11.9 dargestellten komplexen und atomaren Hypertextsortenmodule bezüglich ihrer Positionierung innerhalb des instanziierenden Knotens markiert sind (Spalte „Merkmalsebene der Hypertextknotensorte").[31]

Ein zentraler Bestandteil der Hypertextknotensorte ist die primäre Navigationshilfe. Diese wird als atomares Hypertextsortenmodul aufgefasst, das Linkanker enthält, die zu untergeordneten Bereichen des Webauftritts führen. Die Analyse hat diesbezüglich drei Aspekte aufgezeigt: Die Hyperlinkanzeiger einer Navigationshilfe können sowohl hinsichtlich ihrer Benennung als auch bezüglich der allgemeinen Existenz eines thematisch-inhaltlich markierten Bereichs charakterisiert werden. Weiterhin unterliegt auch der abstrakte Typ des Hyperlinkziels einer Konvention, der bei den in Tabelle 11.9 aufgeführten Navigationshilfen in der Spalte „Merkmalsebene der Hypertextknotensorte" als „Ziel" notiert wurde: Die Hyperlinks „Studium" und „Forschung" führen z. B. typischerweise zu der Einstiegsseite (ES) eines eingebetteten Hypertextes oder zu einem Verteiler (vgl. hierzu Abschnitt 11.5), wohingegen „Selbstdarstellung" und „Informationen zur Region" unmittelbar auf einen Inhaltsknoten verweisen (vgl. Abbildung 3.8, S. 153). Das dritte Charakteristikum bezieht sich auf Aspekte der Typologisierung, denn die Analyse der primären Navigationshilfen zeigt, dass auch Hypertextsortenmodule unterschiedliche Typen aufweisen können, die in Tabelle 11.9 als individuelle Merkmale der Hypertextsortenmodule primäre und sekundäre Navigationshilfe angedeutet werden (vgl. Abschnitt 10.5).

Eine Hypertextknotensorte ist hinsichtlich ihrer kommunikativen Funktion markiert.[32] Die *Einstiegsseite* besitzt – als konventionalisierter Hypertextknotentyp – einen besonderen Stellenwert innerhalb eines Webauftritts, was auch durch den Umstand verdeutlicht wird, dass sie unter dem Lexem „Homepage" häufig mit der Gesamtheit aller Knoten eines Hypertextes gleichgesetzt wird. Diese enge Verbindung zwischen der Einstiegsseite und den verbleibenden Knoten einer Website ist dafür verantwortlich, dass eine isolierte Beschreibung der Funktion einer Einstiegsseite nur eingeschränkt möglich ist (vgl. Abschnitt 4.6.1). Die vielfältigen Funktionen eines universitären Webauftritts wurden bereits in Kapitel 6 dargestellt, er fungiert z. B. als Kommunikationsmittel zwischen den Angehörigen unterschiedlicher Organisationsbereiche einer Hochschule (beispielsweise Forschung, Lehre, Verwaltung) und bietet Möglichkeiten des Zugriffs auf verschiedene Informationsangebote. Instanzen der Hypertextknotensorte *Einstiegsseite eines universitären Webauftritts* kommt dabei eine besondere Bedeutung zu, die mit spezifischen Funktionen korreliert: Die Einstiegsseite empfängt den Rezipienten in der virtuellen Repräsentanz der Institution und stellt – aus Sicht der jeweiligen Produzenten – die verfügbaren Inhalte in möglichst übersichtlicher Form unter Zuhilfenahme einer transparenten Visualisierung der Informationsstrukturierung dar, die zumeist

[31] Die angegebenen Werte beziehen sich auf eine Einteilung des HTML-Dokuments in neun Bereiche (LO: links oben, MO: Mitte oben, RO: rechts oben, L: links, Z: Seitenzentrum usw.) und können unmittelbar den Ergebnissen der Makrostrukturanalyse entnommen werden. Falls keine Konvention festgestellt werden konnte, wurde die Position als „beliebig" markiert. Weiterhin existieren einige Hypertextsortenmodule, die zwar keine spezifische Position besitzen, jedoch typischerweise in der „Peripherie" einer zugehörigen Instanz ausgeprägt sind, z. B. in der unteren linken oder oberen rechten Ecke.

[32] Aus diesem Grund geht Tabelle 11.9 nicht auf die kommunikative Funktion der Hypertextknotensorte ein, die in ihrer Grundausprägung als globale Eigenschaft aufgefasst wird, die durch die Aufnahme spezifischer Hypertextsortenmodule in einer zugehörigen Instanz erweitert werden kann (vgl. Abschnitt 5.5).

Bezeichnung	Ebene	Primärer Typ	Sekundärer Typ	Status	Verwendung	Merkmalsebene der Hypertextknotensorte	Frequenz
Identifikation	komplex	Inhalt/Thema	Dekoration	generell	obligatorisch	Position: MO	100
Name der Institution	atomar	Inhalt/Thema	Dekoration	generell	obligatorisch	Position: MO	94
Logo bzw. Schriftzug	atomar	Dekoration	Inhalt/Thema	generell	obligatorisch	Position: MO/LO	91
Abbildung der Institution	atomar	Dekoration		generell	optional	Position: beliebig	49
Siegel	atomar	Dekoration		spezifisch	optional	Position: LO	37
Motto/Leitspruch/Slogan	atomar	Inhalt/Thema	Dekoration	generell	optional	Position: beliebig	14
Primäre Navigationshilfe	atomar	Navigation	Inhalt/Thema	generell	obligatorisch	Position: Z	100
Einfache Liste (vertikal)	Typ	—	—	—	obligatorisch	—	51
Komplexe Tabelle	Typ	—	—	—	optional	—	26
Studium	Anker	Inhalt/Thema	Navigation	spezifisch	obligatorisch	Ziel: ES/Verteiler	91
Forschung	Anker	Inhalt/Thema	Navigation	spezifisch	obligatorisch	Ziel: ES/Verteiler	86
Selbstdarstellung	Anker	Inhalt/Thema	Navigation	generell	obligatorisch	Ziel: Inhalt	83
Aktuelles	Anker	Inhalt/Thema	Navigation	generell	obligatorisch	Ziel: Einstiegsseite	51
Dezentrale Einrichtungen	Anker	Inhalt/Thema	Navigation	spezifisch	obligatorisch	Ziel: Verteiler	51
Zentrale Einrichtungen	Anker	Inhalt/Thema	Navigation	spezifisch	optional	Ziel: Verteiler	49
Service/Dienstleistungen	Anker	Inhalt/Thema	Navigation	generell	optional	Ziel: Verteiler	43
Informationen zur Region	Anker	Inhalt/Thema	Navigation	spezifisch	optional	Ziel: Inhalt	31
Suche	Anker	Inhalt/Thema	Navigation	generell	optional	Ziel: Inhalt (Maske)	31
Internationales	Anker	Inhalt/Thema	Navigation	spezifisch	optional	Ziel: Einstiegsseite	29
Spezifische Einrichtungen	Anker	Inhalt/Thema	Navigation	spezifisch	optional	Ziel: ES/Verteiler	26
Weiterführende Hyperlinks	Anker	Inhalt/Thema	Navigation	generell	optional	Ziel: Einstiegsseite	26
Wissenstransfer/Kooperation	Anker	Inhalt/Thema	Navigation	spezifisch	optional	Ziel: Einstiegsseite	23
Ansprechpartner/Adressen	Anker	Inhalt/Thema	Navigation	generell	optional	Ziel: Einstiegsseite	23
Struktur/Organisation	Anker	Inhalt/Thema	Navigation	generell	optional	Ziel: Inhalt	20
Hintergrundfarbe: Weiß	Knoten	—	—	—	obligatorisch	Dekoration	94
Fußzeile	komplex	Metainformation	Inhalt/Thema	generell	obligatorisch	Position: MU	86
Datum der letzten Änderung	atomar	Metainformation		universal	obligatorisch	Position: MU	71
Name des Autors/Produzenten	atomar	Metainformation		universal	obligatorisch	Position: MU	51
Impressum/Redaktion	atomar	Metainformation	Navigation	generell	optional	Position: MU	46
Copyright-Hinweis	atomar	Metainformation	Inhalt/Thema	universal	optional	Position: MU	26
Bitte um Kontaktaufnahme	atomar	Inhalt/Thema		generell	optional	Position: beliebig	26
Zugriffszähler	atomar	Metainformation		universal	optional	Position: MU	6
Suchfunktion	atomar	Interaktion/Funktion	Navigation	generell	obligatorisch	Position: Peripherie	83
Hyperlinks unterstrichen	Knoten	—	—	—	obligatorisch	Dekoration	77
Alt. Version in anderer Sprache	atomar	Inhalt/Thema	Navigation	generell	obligatorisch	Position: Peripherie	74
Grundschriftfarbe: Schwarz	Knoten	—	—	—	obligatorisch	Dekoration	71
Hyperlinkfarbe: Blau	Knoten	—	—	—	obligatorisch	Dekoration	71
Corporate-Design	Knoten	—	—	—	obligatorisch	Dekoration	69
Aktuelle Informationen	atomar	Inhalt/Thema	Navigation	generell	obligatorisch	Position: beliebig	66
Kontaktinformationen	komplex	Inhalt/Thema	Kommunikation	generell	obligatorisch	Position: Peripherie	63
E-Mail-Adresse	atomar	Kommunikation	Interaktion	generell	obligatorisch	Position: MU	63
Straßenadresse	atomar	Inhalt/Thema		generell	optional	Position: Peripherie	23
Telefonnummer	atomar	Inhalt/Thema		generell	optional	Position: Peripherie	17
Faxnummer	atomar	Inhalt/Thema		generell	optional	Position: Peripherie	11
Layout: Zweispaltig	Knoten	—	—	—	obligatorisch	Position./Dekoration	54
Kopfzeile	atomar	Metainformation	Inhalt/Thema	generell	obligatorisch	Position: MO	51
Sekundäre Navigationshilfe	atomar	Navigation	Inhalt/Thema	generell	optional	Position: MO/MU	49
Einfache Liste (horizontal)	Typ	—	—	—	optional	—	34
Einfache Liste (vertikal)	Typ	—	—	—	optional	—	14
Suche	Anker	Inhalt/Thema	Navigation	generell	optional	Ziel: Inhalt (Maske)	26
Englisch	Anker	Inhalt/Thema	Navigation	generell	optional	Ziel: Einstiegsseite	17
Kontakt	Anker	Inhalt/Thema	Navigation	generell	optional	Ziel: Inhalt	17
Überblick	Anker	Inhalt/Thema	Navigation	generell	optional	Ziel: Inhalt	14
Adressen/Telefon	Anker	Inhalt/Thema	Navigation	generell	optional	Ziel: Inhalt	14
Explizite Begrüßung	atomar	Inhalt/Thema		generell	optional	Position: MO	31
Zielgruppensp. Navigationshilfe	atomar	Navigation	Inhalt/Thema	—	optional	Position: MO/MU	20
Einfache Liste (horizontal)	Typ	—	—	—	optional	—	11
Einfache Liste (vertikal)	Typ	—	—	—	optional	—	9

Tabelle 11.9: Die Hypertextsortenmodule und Merkmale der Hypertextknotensorte *Einstiegsseite eines universitären Webauftritts*

durch eine primäre und eine sekundäre Navigationshilfe realisiert wird. Weiterhin können unterschiedliche Rezipientengruppen individuelle Bedarfe aufweisen, die durch zielgruppenspezifische Navigationshilfen kanalisiert werden, um den Lesern eine zusätzliche Möglichkeit des Zugriffs zur Verfügung zu stellen. Die Einstiegsseite zeichnet sich im Allgemeinen durch eine offizielle, seriöse, repräsentative, unverwechselbare und ästhetische Gestaltung aus und fungiert als virtuelles Aushängeschild der Institution im *World Wide Web*. Eine wichtige Funktion stellt die Werbung im weitesten Sinne dar: Neben Hinweisen auf aktuelle Veranstaltungen wirbt die Hochschule bereits in der Einstiegsseite für Dienstleistungen und den Technologietransfer, Studieninteressenten sollen für ein Studium an der Universität begeistert werden und Wissenschaftler, die an anderen Einrichtungen tätig sind, werden auf Stellenausschreibungen hingewiesen. Die Analysen haben gezeigt, dass nahezu alle Produzenten bestrebt sind, die zentrale Funktion eines universitären Webauftritts als permanentes, flexibles, tagesaktuelles und weltweit erreichbares Informationsangebot bereits innerhalb der Einstiegsseite zu realisieren, um eine Hochschule als innovative und gleichermaßen traditionsbewusste und progressive Institution zu positionieren (vgl. Abschnitt 6.3.2).

Diese Charakterisierung deutet an, dass funktionale Gemeinsamkeiten mit anderen Klassen von Einstiegsseiten existieren. Die Spalte „Status“ in Tabelle 11.9 bezieht sich auf die Spezifika der Hypertextknotensorte in Bezug auf den übergeordneten Hypertextknotentyp *Einstiegsseite des Webauftritts einer Institution*, wozu z. B. Unternehmen, Vereine und gemeinnützige Organisationen gehören. Die Tabelle verdeutlicht, dass sich die spezifischen Komponenten der *Einstiegsseite eines universitären Webauftritts* insbesondere auf das Hypertextsortenmodul *Identifikation*, in dem das grafisch dargestellte *Siegel* einer Hochschule als Spezifikum dieser Hypertextknotensorte aufzufassen ist, sowie auf die in der *primären Navigationshilfe* genannten inhaltlich-thematisch markierten Kategorien beziehen. In diesem Zusammenhang ist der Umstand, dass sehr markante Korrespondenzen hinsichtlich der Benennung der verwendeten Kategorienetiketten und Hyperlinkanzeiger existieren, einer der Gründe für die Annahme einer eigenständigen Hypertextknotensorte unterhalb des genannten Hypertextknotentyps. Die Gemeinsamkeiten mit anderen Klassen von Einstiegsseiten beziehen sich, wie die Vergleiche mit den Studien von Nielsen und Tahir (2002) und Schütte (2004a) gezeigt haben (vgl. auch Abschnitt 4.6.2), insbesondere auf peritextuelle Eigenschaften (Hintergrundfarbe Weiß, Hyperlinks unterstrichen) und dem Hypertextknotentyp zugehörige, konventionalisierte Hypertextsortenmodule der primären Typen Navigation, Metainformation, Kommunikation sowie Inhalt/Thema (Name und Logo der Institution, Suchfunktion, Kontaktinformationen, Selbstdarstellung etc.).[33] Diese Merkmale kommerzieller Einstiegsseiten fungieren also als Rahmenmodell (vgl. auch Androutsopoulos und Kraft, 2003, S. 9), an das sich die Produzenten nahezu aller Einstiegsseiten dieser Stichprobe anlehnen, um es durch die domänenbezogenen Spezifika des zu realisierenden Webauftritts mit Inhalten anzureichern,

[33] Ein Vergleich mit der von Askehave und Nielsen (2005) eingeführten „move structure“ der Einstiegsseite einer kommerziellen Webpräsenz (vgl. Abschnitt 4.6.2) zeigt darüber hinaus, dass in den Einstiegsseiten universitärer Websites insbesondere die Kommunikationsschritte „Greeting“ (*explizite Begrüßung*), „Identifying sender“ (*Identifikation*), „Indicating content structure“ (*primäre Navigationshilfe*), „Detailing (selected) content“ (*aktuelle Informationen*) und „Establishing contact“ (*Kontaktinformationen*) realisiert werden. Der Schritt „Attracting attention“ wird in einigen Fällen durch ein *Motto oder einen Slogan* oder eine *Abbildung* und „Establishing credentials“ durch die Integration des *Siegels* (auch: „Identifying sender“) realisiert. Auf die Schritte „Establishing a (discourse) community“ und „Promoting an external organisation“ verzichten die Produzenten.

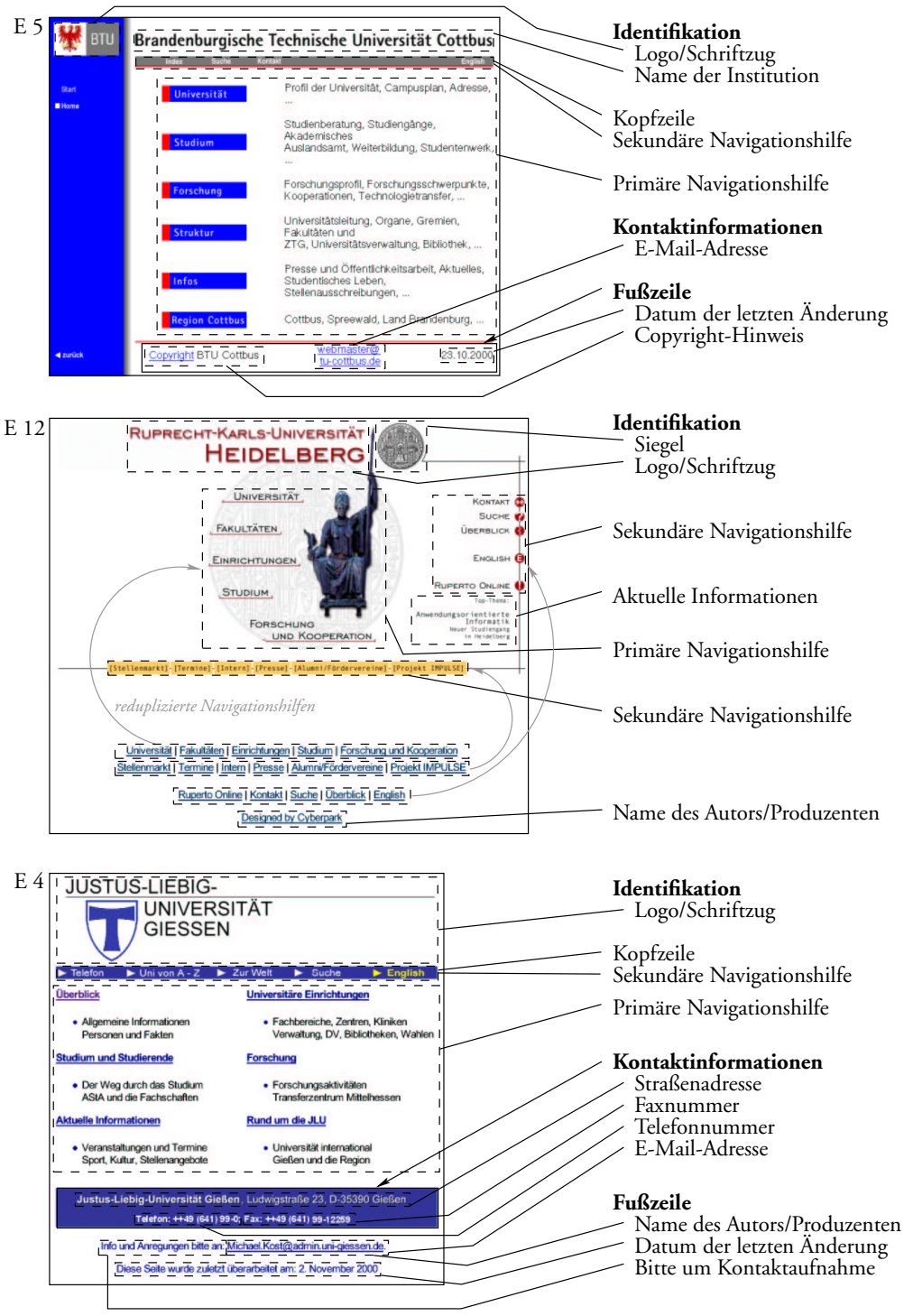


Abbildung 11.3: Die Hypertextsortenmodule der Hypertextknotensorte *Einstiegsseite eines universitären Webauftritts*

die sich wiederum insbesondere auf die Strukturierung und die thematische Markierung der primären Navigationshilfe beziehen. Nach mehr als zehn Jahren der „rhetorical anarchy“ im *World Wide Web* (Askehave und Nielsen, 2005, S. 2) existiert mit der *Einstiegsseite des Webauftritts einer Institution* also nicht nur ein durch zahlreiche Konventionen geprägter Hypertextknotentyp, es können zusätzlich subgenerische und ebenfalls konventionalisierte Varianten, d. h. Hypertextknotensorten belegt werden.

Abbildung 11.3 zeigt abschließend drei repräsentative Einstiegsseiten, in denen die wesentlichen atomaren und komplexen Hypertextsortenmodule markiert wurden (vgl. Fußnote 51, S. 455).[34] Die drei Einstiegsseiten enthalten jeweils eine *Identifikation* sowie eine *primäre* und eine *sekundäre Navigationshilfe*, wobei sich die primäre Navigationshilfe in allen Fällen im Seitenzentrum befindet und das Dokument in gestalterischer Hinsicht dominiert. Zudem veranschaulichen diese Beispiele, dass für die hier untersuchte Hypertextknotensorte die von Schütte (2004a, S. 211) für die Einstiegsseiten von Unternehmen aufgestellte Gleichung „Emittent = Thema“ Gültigkeit besitzt. E 4 und E 5 zeigen, dass die *sekundäre Navigationshilfe* als *Kopfzeile* ausgeprägt sein kann, die in anderen Fällen vornehmlich Metainformationen enthält oder nur dekorative Funktion besitzt.[35] Darüber hinaus stellt E 12 ein Beispiel für den Umstand dar, dass gelegentlich mehrere sekundären Navigationshilfen existieren; in diesem Fall besitzt die am rechten Rand angeordnete sekundäre Navigationshilfe die typische Ausprägung, wohingegen die unterhalb der primären Navigationshilfe dargestellte einfache Liste sekundäre Inhalte und Themengebiete umfasst. Auch die in dieser Einstiegsseite beobachtbare Instanz des Hypertextsortenmoduls *aktuelle Informationen* kann als untypisch bezeichnet werden, da es den Rezipienten auf lediglich ein einziges „Top-Thema“ hinweist.

11.5 Die erste Verknüpfungsebene

Die zweite Phase der Analyse bezieht sich auf die 692 HTML-Dokumente, die in den Einstiegsseiten referenziert werden (vgl. Fußnote 2, S. 462). Abschnitt 11.4 hat bereits verdeutlicht, welche Typen von Knoten zu erwarten sind (z. B. Einstiegsseiten eingebetteter Hypertexte, Überblicksdarstellungen zu spezifischen Themenbereichen, Auflistungen von Hyperlinks und Pressemitteilungen etc.). Das Ziel dieser zweiten Stufe betrifft die Identifizierung, Sammlung und Ermittlung der Frequenzen der korrespondierenden Hypertextknotensorten und darüber hinaus auch von Hypertextsorten, sofern es sich bei einem Dokument um die Einstiegsseite in einen eingebetteten Hypertext handelt. Die Ergebnisse sollen zur Konstruktion einer Typologie von Hypertextknotensorten eingesetzt werden. Dieses zweite Ziel fußt auf der Hypothese, dass insbesondere die Hypertextknotensorten, die in der – von der Einstiegsseite aus betrachtet – obersten Verknüpfungsebene Verwendung finden, zur Strukturierung der oberen Ebenen der Typologie geeignet sind.

[34] Aus Darstellungsgründen wurde in Abbildung 11.3 auf eine vollständige Markierung sämtlicher Hypertextsortenmodule verzichtet.

[35] Ein vergleichbares Beispiel enthält E 4, in dessen *Fußzeile* der *Name des Produzenten* und die *E-Mail-Adresse* des Hypertextsortenmoduls *Kontaktinformationen* von *einem* Hypertextmodul instanziiert werden, da die als `mailto:`-Hyperlink realisierte E-Mail-Adresse den Namen des Verantwortlichen im Klartext enthält.

11.5.1 Die Ergebnisse der Analyse im Überblick

Bezüglich der Methodologie wird diese Analyse mit der Problematik konfrontiert, dass eine heterogene und sehr umfangreiche Menge einzelner Knoten auf abstrakte Etiketten von Hypertextknotensorten abzubilden ist.[36] Eben dieses Problem wird von den Autoren vergleichbarer Studien als zentrale Schwierigkeit aufgeführt (vgl. auch die Abschnitte 4.4.1 bis 4.4.7), weshalb sich die vorliegende Arbeit auf eine restringierte Untersuchungsdomäne bezieht. Und in der Tat tritt diese Problematik in der vorliegenden Analyse deutlich seltener in Erscheinung (vgl. Abschnitt 4.4.8). Die nachfolgend präsentierten Ergebnisse wurden mit Hilfe einer Stichprobenanalyse erhoben, die mittels der Web-Oberfläche der Korpusdatenbank durchgeführt wurde (vgl. die Abschnitte 7.3.4 und 7.3.5). Das Inventar zugewiesener Etiketten wurde im Zuge dieser datenbankgestützten Analyse in insgesamt acht auf die gesamte Stichprobe bezogenen Restrukturierungen sukzessive verfeinert und um eine zusätzliche Analyseebene (vgl. die Abschnitte 11.7 und 11.8) ergänzt, um eine bestmögliche Homogenität und einen größtmöglichen Abdeckungsgrad gewährleisten zu konnen.[37]

Auf einer abstrakten Ebene existieren im Hypertextsystem WWW drei Typen von Knoten, die durch die Verfolgung eines Hyperlinks erreicht werden können: (i) Der Zielknoten fungiert als Einstiegsseite eines Hypertextes, (ii) er enthält eine Liste von Hyperlinks, (iii) oder er umfasst vornehmlich Inhalte (z. B. Text, Textfragmente, Tabellen, Abbildungen oder multimediale Inhalte; vgl. auch Abbildung 3.8).[38] Bei diesen drei Typen liegt naturgemäß keine trennscharfe Unterscheidung vor, d. h. ein Knoten kann eine Einstiegsseite darstellen und zugleich Text enthalten oder – neben einer primären Navigationshilfe – eine Liste von Hyperlinks präsentieren und simultan als Einstiegsseite fungieren. Obwohl die in der Stichprobe enthaltenen Dokumente durch eine hochgradige funktionale Heterogenität gekennzeichnet sind und die drei Zielknotentypen eher generischen Merkmalen eines HTML-Dokuments als trennscharfen Kategorien ähneln, konnte diese Unterscheidung erfolgreich zur initialen Strukturierung der Stichprobe eingesetzt werden.[39] Die Restriktion der Untersuchungsdo-

[36] Die Problematik wird durch den Umstand verschärft, dass viele Hypertextknotensorten keine etablierte Bezeichnung in der Alltagssprache besitzen (vgl. Abschnitt 4.3), von der textlinguistischen Forschung bislang nicht untersucht wurden und auch in die Fachsprache der Webdesigner und Informationsarchitekten keinen Einzug gehalten haben. Gleichwohl können spezifische Funktionen und Strategien der Hypertextualisierung ermittelt werden, die eine Konzeptualisierung von Hypertextknotensorten erlauben. Des Weiteren ist anzumerken, dass in der traditionellen Textlinguistik keine vergleichbare Aufgabenstellung existiert.

[37] Die Stichprobenanalyse von Crowston und Williams (2000, S. 205) basiert auf einer vergleichbaren Methodik: „We started with a list of genres and their definitions developed in a pilot study and refined it during the course of this study." Definitionen traditioneller Genres haben Crowston und Williams dem *Oxford English Dictionary* entnommen. Auf dieser Grundlage hat eine studentische Hilfskraft die zu analysierenden HTML-Dokumente Genre-Etiketten zugeordnet, wobei zusätzlich die Möglichkeit existierte, neue Etiketten anzulegen.

[38] Die Typen sind auf einer generischeren Ebene anzusiedeln als die von Haas und Grams (1998b) vorgeschlagenen „page types" (vgl. Abbildung 4.5, S. 180): Typ (i) entspricht „home page", Typ (ii) subsumiert „organizational pages", und Typ (iii) umfasst „documentation", „multimedia", „text", „tools" und „database entry".

[39] Die funktionale Heterogenität eines Knotens bezieht sich auf die Existenz unterschiedlicher Instanzen mehrerer Hypertextsortenmodule, die wiederum individuelle und spezifische Funktionen besitzen und wird im Folgenden mit Beispielen illustriert. Im Falle der Existenz mehrerer primär inhaltlich-thematisch markierter Hypertextsortenmodule in einem Knoten wurde das zugewiesene Etikett durch diejenige Instanz eines Hypertextsortenmoduls determiniert, die den Knoten dominiert (z. B. in Bezug auf ihre Abmessungen oder das Merkmal der Initialpositionierung im Inhaltsbereich des Knotens).

mäne auf universitäre Webangebote gestattete – trotz der großen Varianz – mit nur sehr wenigen Ausnahmen eine transparente Differenzierung in Kategorien, die als Hypertextknotensorten im Sinne des in Kapitel 5 dargestellten Modells verstanden werden können.

Tabelle 11.10 zeigt die Ergebnisse der Analyse im Überblick und verdeutlicht, dass die drei Knotentypen zu etwa gleichen Teilen in den 692 Dokumenten enthalten sind, wobei die Einstiegsseiten jedoch minimal überwiegen. Bezüglich dieses Typs fällt auf, dass insgesamt 100 Hyperlinks auf die *Einstiegsseiten der Webauftritte zentraler* oder *dezentraler Organisationseinheiten* verweisen.[40] Die *Einstiegsseiten zielgruppenspezifischer Webangebote* sind ebenfalls hochfrequent in der Stichprobe vertreten. Die mit Abstand umfangreichste Hypertextknotensorte bildet der *Verteiler*, der eine spezifische Ausprägung der Hyperlinkliste darstellt. Das *Suchformular*, *Kontaktinformationen*, die *Pressemitteilung* und der *Veranstaltungskalender* bilden die hochfrequenten Kategorien des dritten Typs – zwei dieser Hypertextknotensorten beinhalten aktuelle Informationen, die den Trend bestätigen, dass die Einstiegsseiten universitärer Webauftritte im Begriff sind, sich zu Informationsportalen zu entwickeln, die tagesaktuelle Neuigkeiten präsentieren. Weiterhin können zahlreiche traditionelle Textsorten belegt werden (z. B. *Studienführer*, *Stellenanzeige*, *Vorlesungsverzeichnis*, *Pressemitteilung*, *Impressum* und *Veranstaltungskalender*), deren Exemplare in unterschiedlichen Ausprägungen an die Gegebenheiten des Mediums WWW angepasst wurden. Bevor die Ergebnisse in detaillierter Form vorgestellt werden, sei der Umstand angesprochen, dass die zugewiesenen Etiketten in den meisten Fällen als trennscharf aufzufassen sind. Gleichwohl existieren einige problematische bzw. sehr generelle Hypertextknotensorten, die ebenfalls thematisiert werden.

11.5.2 Die zentralen Differenzierungskriterien

Vor der Darstellung der Charakteristika des Verteilers ist es notwendig, die Differenzierungskriterien zu erläutern, auf deren Grundlage ein Dokument als *Einstiegsseite*, *Verteiler*, *Hotlist* oder *Inhaltsknoten* kategorisiert wurde. Die Einstiegsseite eines Hypertextes muss per definitionem mindestens einen Hyperlink zu einem untergeordneten Knoten enthalten. Typischerweise umfassen Einstiegsseiten in WWW-basierten Hypertexten eine primäre und möglicherweise sekundäre Navigationshilfen. In den meisten Einstiegsseiten universitärer Webauftritte stellt die primäre Navigationshilfe den eigentlichen Inhalt dar (vgl. Abschnitt 11.4.2). In anderen Typen von Einstiegsseiten wird sie als Bestandteil des Peritextes aufgefasst, so dass eine größere Fläche des Dokuments zur Präsentation einer Begrüßung oder einer Kurzzusammenfassung der Inhalte eingesetzt werden kann. In einem Verteiler oder einer Hotlist wird dieser Bereich nahezu ausschließlich zur Präsentation von Hyperlinklisten verwendet, die mit den eigentlichen Navigationshilfen des Hypertextes somit *nicht* identisch sind. In Inhaltsknoten werden hingegen keine Listen von Hyperlinks, sondern vornehmlich Texte, Textfragmente, Listen, Abbildungen, Illustrationen und dergleichen präsentiert. In denjenigen Fällen, in denen Mischungen vorlagen, hat der Anteil des jeweiligen Hypertextsortenmoduls am Gesamtdokument darüber entschieden, ob ein Inhaltsknoten (der Inhalt überwiegt) oder eine

[40] Im Folgenden werden die Termini „Webauftritt“ und „Webangebot“ verwendet. Obwohl diese Begriffe prinzipiell als Synonyme von „Website“ und „Hypertext“ aufgefasst werden können, wird „Webauftritt“ zur Bezeichnung der Selbstdarstellung einer institutionalisierten Organisationseinheit benutzt, während sich „Webangebot“ auf arbiträre Informationsbestände bezieht, die nicht unmittelbar einer Organisationseinheit oder ihrer Selbstdarstellung zugeordnet werden können.

	Knotentyp	Hypertextknotensorte	Frequenz		Prozent	
	Einstiegsseite …			280		40,5
1.		… des Webauftritts einer zentralen Organisationseinheit	70		10,1	
2.		… eines zielgruppenspezifischen Webangebots	36		5,2	
3.		… des Webauftritts einer dezentralen Organisationseinheit	30		4,3	
4.		… einer universitären Selbstdarstellung	21		3,0	
5.		… des Webauftritts einer Universität	17		2,5	
6.		… des Webangebots eines Fachgebiets/Studiengangs/Weiterbildungsangebots	16		2,3	
7.		… des Webauftritts eines universitätsinternen Projekts	9		1,3	
8.		… eines Studienführers	9		1,3	
9.		… eines Webangebots zum Thema Forschung/Lehre	8		1,2	
10.		… des Webangebots mit einem Überblick über dezentrale Organisationseinheiten	8		1,2	
11.		… eines Webangebots mit Stellenanzeigen	7		1,0	
12.		… des Webangebots einer universitären Veranstaltung	6		0,9	
13.		… eines Webangebots mit aktuellen Informationen, Terminen, Meldungen	6		0,9	
14.		… eines Vorlesungsverzeichnisses	5		0,7	
15.		… eines Webangebots mit Pressemitteilungen	5		0,7	
16.		… einer Universitätszeitung	4		0,6	
17.		… eines Shops oder Online-Shops	4		0,6	
18.		… eines Webangebots zu angebotenen Studiengängen	3		0,4	
19.		… einer digitalen Bibliothek	3		0,4	
20.		… eines Forschungsberichts	3		0,4	
21.		… des Webangebots einer Konferenz/Tagung	2		0,3	
22.		… des Webauftritts einer Stiftung bzw. eines Fördervereins	2		0,3	
23.		… einer persönlichen Homepage	2		0,3	
24.		… eines Hypertextes mit Webserver-bezogenen Informationen (Statistiken/Layouts)	2		0,3	
25.		… des Webangebots mit studentischen Homepages	1		0,1	
26.		… des Webauftritts einer studentischen Initiative	1		0,1	
	Hyperlinkliste			214		30,9
27.		Verteiler	177		25,6	
28.		Verteiler/Hotlist (Kombination)	20		2,9	
29.		Hotlist	17		2,5	
	Inhaltsknoten			198		28,6
30.		Formular (Suchformular: 21; Posting-Formular: 1; E-Mail-Formular: 1)	23		3,3	
31.		Kontaktinformationen (Anschriften, Telefonnummern, E-Mail-Adressen)	17		2,5	
32.		Pressemitteilung	17		2,5	
33.		Veranstaltungskalender	16		2,3	
34.		Benutzungshinweise und Anleitungen (zu Netzdiensten, Geräten oder dem Webangebot)	14		2,0	
35.		Impressum	9		1,3	
36.		Kerninformationen für eine spezifische Zielgruppe	9		1,3	
37.		Lageplan, Karte, Wegbeschreibung	8		1,2	
38.		Redaktion eines Webauftritts	8		1,2	
39.		Automatische Weiterleitung (kommentiert)	7		1,0	
40.		Foto bzw. Fotogalerie	6		0,9	
41.		Kurzdarstellung einer Organisationseinheit (Profil/Porträt)	6		0,9	
42.		Ankündigung einer neuen Dienstleistung oder Ressource	5		0,7	
43.		Einladung zu einer Veranstaltung (mit Programm)	5		0,7	
44.		Kurzdarstellung einer Organisationseinheit (Funktionen und Kontaktinformationen)	5		0,7	
45.		Index (A-Z) bzw. Schlagwortverzeichnis	4		0,6	
46.		Pressemitteilung (Auflistung)	4		0,6	
47.		Copyright-Informationen	3		0,4	
48.		Grundordnung, Gesetz, juristischer Text	3		0,4	
49.		Kurzdarstellung der historischen Entwicklung einer Institution	3		0,4	
50.		Stellenanzeige	3		0,4	
51.		Textknoten mit generischer Informationsfunktion	3		0,4	
52.		Daten und Fristen eines Semesters	2		0,3	
53.		Kurzdarstellung einer Stadt bzw. Region	2		0,3	
54.		Personalia (Rufe, Ernennungen, Antrittsvorlesungen etc.)	2		0,3	
55.		Schwarzes Brett	2		0,3	
56.		Sitemap	2		0,3	
57.		Splash-Seite	2		0,3	
58.		Antragsformular	1		0,1	
59.		Ausschreibung (universitätsintern)	1		0,1	
60.		Grußwort	1		0,1	
61.		Haftungsausschluss (Disclaimer)	1		0,1	
62.		Kleinanzeigen (Auflistung von Mitfahrgelegenheiten)	1		0,1	
63.		Speisenplan der Mensa	1		0,1	
64.		„under construction"-Hinweis	1		0,1	
65.		Wahlergebnis	1		0,1	
	Σ		692	692	100	100

Tabelle 11.10: Die ermittelten Hypertextknotensorten im Überblick

Knotentyp	Hyperlinks $a \rightarrow z_{1...n}$	Autoren von $z_{1...n}$	Verhältnis von a und $z_{1...n}$
Einstiegsseite	Die Hyperlinks sind – als Navigationshilfe – ein Bestandteil des Peritextes von a.	Die Autoren von $z_{1...n}$ sind identisch mit den Autoren von a.	a und $z_{1...n}$ sind Bestandteile eines spezifischen Hypertextes.
Verteiler	Die Hyperlinks stellen den eigentlichen Inhalt von a dar.	Die Autoren von $z_{1...n}$ sind Angehörige der Institution I, der auch die Autoren von a angehören.	a und $z_{1...n}$ sind Bestandteile von maximal $n + 1$ unterschiedlichen Hypertexten, die einem übergreifenden Hypertext zugehörig sind (hier: dem Webauftritt von I).
Hotlist	Die Hyperlinks stellen den eigentlichen Inhalt von a dar.	Die Autoren von $z_{1...n}$ sind keine Angehörigen der Institution I, der die Autoren von a angehören.	$z_{1...n}$ sind Bestandteile von maximal n unterschiedlichen Hypertexten, die nicht Bestandteil des übergreifenden Hypertextes sind.
Inhaltsknoten	Die Hyperlinks sind ein optionaler Bestandteil des Inhalts (z. B. Fließtext) von a.	$z_{1...n}$ besitzen beliebige Autoren.	$z_{1...n}$ sind Knoten in beliebigen Hypertexten.

Tabelle 11.11: Die Kriterien zur Differenzierung zwischen *Einstiegsseite*, *Verteiler*, *Hotlist* und *Inhaltsknoten*

Hyperlinkliste vorliegt (die Hyperlinkliste dominiert das Dokument), wobei auch die in einem Dokument enthaltenen Überschriften und metadiskursiven Äußerungen berücksichtigt wurden, aus denen oftmals die vom Produzenten intendierte Funktion eines Knotens unmittelbar hervorgeht (z. B. in D 346: „Hotlinks – wichtige Web-Adressen“).

Es stellt sich die Frage, welche Merkmale eine Einstiegsseite in der Regel aufweist, um als solche identifiziert werden zu können. Hierzu dienen zunächst typische Überschriften, Begrüßungsfloskeln, Kurzzusammenfassungen und im Zentrum des Dokuments angeordnete Navigationshilfen. Von besonderer Bedeutung ist eine Analyse der URL des Dokuments sowie der Adressen der Knoten, auf die die Hyperlinks der primären Navigationshilfe verweisen. Da Hypertexte im WWW häufig hierarchisch organisiert sind und HTML-Dateien auf dem Webserver in hierarchisch angeordneten Verzeichnissen und Unterverzeichnissen abgelegt werden, kann ein spezifisches Dokument mit der URL `.../a/b/` schnell als Einstiegsseite erkannt werden, falls die Hyperlinks auf Dokumente verweisen, die sich in diesem Verzeichnis befinden (beispielsweise `.../a/b/k1.html` und `.../a/b/k2.html`).[41]

Die Unterscheidung zwischen einer Hotlist (vgl. Abschnitt 4.6.6) und einem Verteiler basiert auf den URLs der Zielknoten, auf die die Hyperlinks verweisen. Tabelle 11.11 stellt

[41] Wenn in der Adressierung auf die Angabe eines Dateinamens verzichtet wird (wie z. B. in `.../a/b/`), ist der Webserver dennoch in der Lage, ein Dokument auszuliefern. Für diesen Zweck existiert eine zentrale Konfigurationsdatei, in der sich eine Liste von Default-Dateinamen befindet. Der Inhalt eines Verzeichnisses wird vom Webserver sukzessive (und mit einer signifikanten Reihenfolge) auf die Existenz eines dieser Dateinamen überprüft, und falls ein Dokument wie z. B. typischerweise `index.html` oder `welcome.html` existiert, wird dieses unter einer URL wie `.../a/b/` ausgeliefert. Die Web-Oberfläche der Korpusdatenbank zeigt die URL an, unter der ein Dokument im Korpus abgelegt wurde, d. h. ein Dateiname wie `index.html` ist ebenfalls ein Indikator für das Vorliegen einer Einstiegsseite. Darüber hinaus ist auch die Benennung der beteiligten Verzeichnisse von Bedeutung, denn bei einer URL wie `http://www.uni-giessen.de/hrz/` kann unmittelbar inferiert werden, dass es sich um die Einstiegsseite des Webauftritts eines Hochschulrechenzentrums handelt.

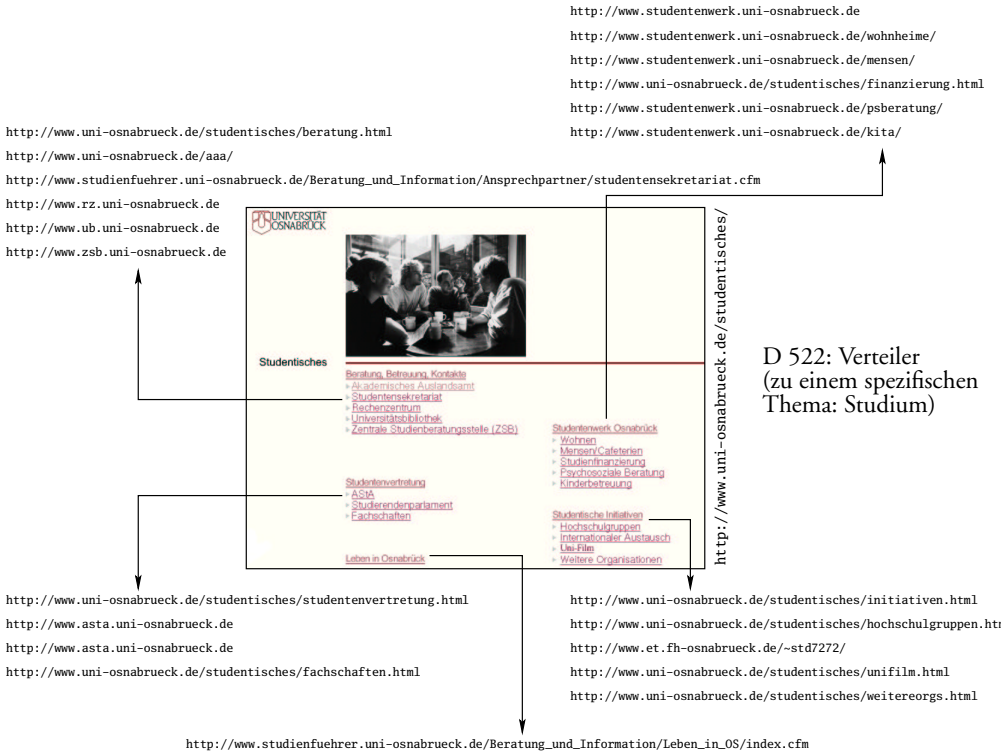


Abbildung 11.4: Beispiel einer Instanz der Hypertextknotensorte *Verteiler*

die Differenzierungskriterien dar: Ein Ausgangsknoten a wird als Hotlist verstanden, wenn die enthaltenen Hyperlinks auf die Knoten $z_{1...n}$ verweisen, die weder zum Hypertext von a noch zum übergeordneten Hypertext gehören, d. h. $z_{1...n}$ werden durch externe Hyperlinks referenziert. Ein Verteiler umfasst ebenfalls Hyperlinks, die auf Knoten verweisen, die Bestandteile anderer Hypertexte sind, diese müssen jedoch dem übergeordneten Hypertext zugehörig sein. Die traditionelle Unterscheidung zwischen externen und internen Hyperlinks wird also um eine weitere Ebene ergänzt. Abbildung 11.4 stellt einen Verteiler (D 522) und die enthaltenen Hyperlinks als Beispiel dar. Dieses Dokument wurde gewählt, weil es als sehr untypischer Vertreter des Verteilers aufzufassen ist, da einige Hyperlinks auf Knoten verweisen, die sich innerhalb des Verzeichnisses des Ausgangsknotens befinden. In gewisser Hinsicht kann dieses Dokument somit auch als Einstiegsseite konzeptualisiert werden, jedoch referenzieren deutlich mehr Hyperlinks Knoten anderer Hypertexte, wodurch die Kategorisierung als Verteiler begründet wird. Das Beispiel verdeutlicht, dass zwischen Verteiler und Einstiegsseite prinzipiell eine fließende Grenze existiert, doch weisen nur die wenigsten Exemplare des Verteilers mehr als zwei oder drei Hyperlinks zu internen Knoten auf.

11.5.3 Der Verteiler und die Hotlist

Der Hypertextknotensorte *Verteiler* ist etwa ein Viertel der 692 HTML-Dokumente zugehörig, während auf die *Hotlist* lediglich 2,5% der Webseiten entfallen. Eine Kombination

dieser unterschiedlichen Listen von Hyperlinks wird in 2,9% aller Dokumente eingesetzt (vgl. Tabelle 11.10). Das Phänomen des Verteilers erklärt sich durch den typischen Aufbau universitärer Webauftritte (vgl. Abschnitt 6.3): Diese bestehen in der Regel aus zahlreichen autarken Hypertexten, für deren Erstellung und Pflege spezifische Organisationseinheiten zuständig sind (z. B. die Bibliothek, Institute oder Dezernate). Darüber hinaus ist eine Organisationseinheit (in der Regel die Pressestelle oder das Rechenzentrum) für den übergreifenden Rahmen, also die Einstiegsseite des universitären Webauftritts und die darunter befindlichen zentralen Dokumente und Informationsangebote verantwortlich.[42] Die Produzenten des zentralen Angebots besitzen die Möglichkeit, auf Inhalte der Hypertexte anderer Organisationseinheiten zu verweisen. Gerade die prominente Hypertextknotensorte *Einstiegsseite eines universitären Webauftritts* sieht verschiedene thematischen Kategorien vor, deren Instanziierung unter anderem durch Inhalte erfolgt, die aus *anderen* Hypertexten stammen, die von der Institution innerhalb des Hypertexttyps *Webauftritt einer Universität* publiziert werden. Folglich liegt in allen 177 Verteilern eine inhaltlich-thematische Markierung vor, die sich insbesondere an den hochfrequenten Kategorien der primären Navigationshilfen orientiert, die in den Einstiegsseiten enthalten sind (vgl. Abschnitt 11.4.2).

In einem weiteren Schritt kann nun rekonstruiert werden, nach welchen Kriterien die in den Verteilern, Hotlists und Kombinationen dieser beiden Hypertextknotensorten enthaltenen Hyperlinks zusammengestellt wurden.[43] Die von anderen Organisationseinheiten angebotenen Hypertexte umfassen zahlreiche unterschiedliche Inhalte, die einen vom Produzenten des zentralen Angebots antizipierten Informationsbedarf zu decken in der Lage sind, der bereits auf der obersten Ebene des übergreifenden Hypertextes besteht. Der Produzent der hier betrachteten Verteiler dürfte in der Regel auch die Einstiegsseite angefertigt haben, die in der primären Navigationshilfe unterschiedliche inhaltliche Kategorien umfasst. Diese Informationsstrukturierung sieht Themenbereiche wie z. B. „Studium", „Forschung", „Einrichtungen" oder „Services" vor, deren jeweilige Zielknoten in vielen Fällen ausschließlich Verteiler enthalten. Dem Produzenten sind die meisten anderen Hypertexte der Hochschule bekannt (oder er wird von ihren Produzenten auf diese aufmerksam gemacht), so dass für ein spezifisches Kriterium relevante Knoten in den Verteiler aufgenommen werden. Der Verteiler kann somit als Strategie verstanden werden, bereits verfügbare Ressourcen einzusetzen, um das Erstellen eigenständiger und möglicherweise redundanter Angebote zu vermeiden. Tabelle 11.12 stellt die Subtypen der drei Hypertextknotensorten dar, die als Kriterien der Zusammenstellung von Hyperlinklisten und als distinktive Merkmale der resultierenden Hypertextknotensortenvarianten konzeptualisiert werden können: Die Verteiler wurden nach sechs Kriterien kompiliert: (i) Spezifisches Thema, (ii) Organisationseinheiten, (iii) Informa-

[42] Dieser Umstand wird in einigen Fällen explizit von den Produzenten des zentralen Angebots thematisiert. Ein *Impressum* (D 194) wird z. B. eingeleitet mit: „Das Webangebot der Universität Augsburg setzt sich aus vielen verschiedenen Quellen zusammen. Für den Inhalt der zentralen Webseiten ist die Pressestelle verantwortlich".

[43] Genau genommen handelt es sich bei dem *Verteiler* und der *Hotlist* um Hypertextsortenmodule, die als Hypertextknotensorten fungieren können. Bei der Kombination *Verteiler/Hotlist* werden Instanzen der beiden Hypertextsortenmodule in einem einzelnen Knoten aggregiert, wobei jedoch ein ausgeglichener Umfang dieser Instanzen eher untypisch ist; in der Regel umfasst eine der Instanzen deutlich mehr Hyperlinks. Während Hotlists in nahezu beliebigen Hypertextsorten eingesetzt werden können, gilt dies für Verteiler nicht. Diese setzen einen übergreifenden Hypertext voraus, der mindestens zwei eingebettete Hypertexte enthält, so dass ein minimaler Verteiler des zentralen Angebots auf mindestens zwei Knoten der eingebetteten Hypertexte verweist.

Hypertextknotensorte	Subtyp (d. h. Kriterium der Zusammenstellung)	Frequenz	
Verteiler			177
	Zu einem spezifischen Thema	70	
	• Forschung, Transfer, Kooperation	18	
	• Studium	15	
	• Beratung, Service- und Dienstleistungen	10	
	• Informationen über die Hochschule	8	
	• Freizeitaktivitäten	4	
	• Stadt, Region, Umfeld	4	
	• Suchmaschinen und Informationsrecherche	4	
	• WWW, E-Mail, Internet	4	
	• Internationale Angelegenheiten	1	
	• Publikationen und Veröffentlichungen	1	
	• Rechtliche Grundlagen	1	
	Organisationseinheiten	49	
	• Alle Einheiten (keine Beschränkung)	16	
	• Zentrale Einrichtungen	13	
	• Fakultäten/Fachbereiche	10	
	• Organe und Gremien	4	
	• Institute/Seminare	3	
	• Leitung und/oder Verwaltung	2	
	• Assoziierte Einrichtungen, An-Institute	1	
	Informationen für eine spezifische Zielgruppe	29	
	• Studierende	10	
	• Absolventen und Alumni	4	
	• Studieninteressenten	4	
	• Gäste und Nachbarn	3	
	• Mitarbeiter	3	
	• Presse	2	
	• Ausländer, ausländische Studierende	1	
	• Wirtschaft	1	
	Aktuelle Informationen	12	
	Kontaktinformationen, Adressen, Ansprechpartner, Lagepläne	10	
	Überblick/Abstract/Inhaltsverzeichnis	7	
Verteiler/Hotlist			20
	Region, regionale Webangebote	6	
	Informationen für eine spezifische Zielgruppe	6	
	Suchmaschinen, Informationsrecherche	3	
	Stellenangebote	2	
	Aktuelle Informationen	1	
	Hyperlinks zu beliebigen Angeboten	1	
	WWW und Internet	1	
Hotlist			17
	Region, regionale Webangebote	10	
	Zu einem spezifischen Thema	4	
	Hyperlinks zu beliebigen Angeboten	3	
Σ		214	214

Tabelle 11.12: Subtypen des *Verteilers*, der *Hotlist* und ihrer Kombination

tionen für eine spezifische Zielgruppe, (iv) aktuelle Informationen, (v) Kontaktinformationen, Adressen, Ansprechpartner, Lagepläne und (vi) Überblick/Abstract/Inhaltsverzeichnis. Der erste Typ bündelt Informationen, die für ein spezifisches Thema wie z. B. Forschung, Transfer und Kooperation relevant sind. Beispielsweise enthält D 156 eine sehr umfangreiche Liste, die unter anderem auf die Webauftritte der Fakultäten und Fachbereiche, interdisziplinäre Zentren, Sonderforschungsbereiche, Kompetenzzentren, Graduiertenkollegs, zentralen Einrichtungen und An-Institute der Universität Ulm verweist und zusätzlich den Forschungsbericht referenziert. Der zweite Typ beschränkt sich auf die Präsentation der Hyperlinks zu den Webauftritten spezifischer Gruppen von Organisationseinheiten (z. B. zentrale oder dezentrale Einrichtungen). Gerade die diesem Typ zugehörigen Dokumente verdeutlichen den typischen Navigationspfad, der von der *Einstiegsseite des universitären Webauftritts* zu einem *Verteiler* und von dort wiederum zu der *Einstiegsseite des Webauftritts einer zentralen* oder *dezentralen Organisationseinheit* führt (vgl. Abbildung 11.5). Die Beispiele des dritten Typs zeigen den Aspekt auf, dass in einem sehr komplexen Webauftritt verfügbare Ressourcen unter neuen Gesichtspunkten wiederverwendet werden, um den Rezipienten eine weitere Navigationsdimension zur Verfügung zu stellen: Diese Dokumente versammeln Informationen, die aus Sicht des Produzenten für eine bestimmte Zielgruppe relevant sein könnten und filtern die unwichtigen Angebote aus, so verweist z. B. D 635 (Zielgruppe: Studierende) unter anderem auf die Einstiegsseiten der Webauftritte des akademischen Auslandsamtes, des Beratungsdienstes behinderter Studierender, der einzelnen Fachbereiche, des Frauenbüros und des Prüfungsamtes.[44] Verteiler des vierten Typs werden nach dem Kriterium der Aktualität der verknüpften Informationen zusammengestellt und referenzieren typischerweise die Einstiegsseite der Pressestelle, den Speisenplan der Mensa, Stellenanzeigen, einen Veranstaltungskalender, Wahlergebnisse und universitätsinterne Projekte (z. B. neue Anwendungen im Netzwerk der Hochschulverwaltung). Der fünfte Typ versammelt Hyperlinks, die zu Kontaktinformationen, Adress- und Telefonlisten sowie Lageplänen führen. Zum Beispiel werden unter der Überschrift „Kontakt, Auskunft, Ansprechpartner/innen“ in D 298 die „Anschriften der FU Berlin“ (des Präsidenten, der Verwaltung und der Fachbereiche), Lagepläne, die Adresse der Studienberatung und weitere Kontaktinformationen aufgeführt. Verteiler des sechsten Typs stellen Hyperlinks zur Verfügung, die zu den wichtigsten Rubriken eines einzelnen Hypertextes verweisen; dieser einzelne Hypertext ist in nahezu allen Fällen der Webauftritt einer großen Organisationseinheit (z. B. die Bibliothek oder das Rechenzentrum). Diese Ausprägung gibt dem Rezipienten einen Überblick über den verknüpften Hypertext und fungiert als dessen Inhaltsverzeichnis. Da die sieben Verteiler dieses Typs Bestandteile der zentralen Angebote der jeweiligen Hochschulen sind, besitzen sie im direkten Vergleich ein entsprechend

[44] Dieses auf der Website der Universität Dortmund verfügbare Dokument weist – ebenso wie die weiteren zielgruppenspezifischen Verteiler dieser Hochschule sowie des Webauftritts der Universität Münster – ein Spezifikum auf: Die Listen von Hyperlinks zu den Einstiegsseiten interner Hypertexte werden von knappen Texten flankiert, die als Kurzzusammenfassungen und Erläuterungen des eigentlichen Inhalts (d. h. der Verteilerliste) und somit auch als Kohärenzbildungshilfe fungieren. Möglicherweise waren die Produzenten dieser Dokumente der Auffassung, dass die ohne jeglichen Kontext versehene Präsentation einer Hyperlinkliste an einer prominenten, durch einen einzigen Hyperlink von der Einstiegsseite aus erreichbaren Position innerhalb der Website die kommunikative Funktion des Knotens beeinträchtigen könnte, weshalb eine skizzenartige Erläuterung der von dieser Stelle aus verfügbaren Angebote in Textform eingefügt wurde. Die meisten Listen von Hyperlinks innerhalb der 692 Dokumente werden nicht von derartigen Texten begleitet.

konsistentes Erscheinungsbild, das Corporate-Design des zentralen Webauftritts wird konsequent beibehalten. Dennoch beinhalten sie hochgradig redundante Inhalte, denn sobald sich die Strukturierung des verknüpften Hypertextes ändert, ist der Produzent gezwungen, den Verteiler an die neuen Gegebenheiten anzupassen.[45] Tabelle 11.13 zeigt abschließend die Verteilung von Instanzen der einzelnen Hypertextknotensorten auf die 35 Hochschulen der Stichprobe.[46] Die Aufstellung verdeutlicht, dass viele Einstiegsseiten auf mehrere Verteiler verweisen, die oftmals nach unterschiedlichen Kriterien kompiliert wurden.

Auf eine detaillierte Erläuterung der unterschiedlichen Ausprägungen der Hypertextknotensorten *Hotlist* und *Verteiler/Hotlist* wird an dieser Stelle verzichtet, doch seien zwei Spezifika hervorgehoben: Regionale Informationen werden primär durch Hyperlinks auf externe Angebote verknüpft (Verteiler: 4 Vorkommen, Verteiler/Hotlist: 6, Hotlist: 10). D 75 (`http://www.tu-chemnitz.de/chemnitz/`) enthält Links zu zwei Webangeboten mit Informationen über die Stadt Chemnitz und wird durch eine metadiskursive Äußerung eingeleitet:

> Bereits seit 1994 (als die Stadt Chemnitz noch den Internet-Tiefschlaf hielt) hatte es sich die TU zur Aufgabe gemacht, auch die Stadt Chemnitz ansprechend im Internet zu präsentieren. Im Dezember 1999 ist es der Stadt Chemnitz nun endlich gelungen, eine eigene Präsentation [...] ins Internet zu stellen. Aus diesem Grund werden wir unsere Chemnitzpräsentation nicht weiter ausbauen, sonder [sic] in einem absehbaren Zeitraum aus dem Netz nehmen. [D 75]

Das von der Universität gepflegte Informationsangebot wird vom Produzenten also als redundant eingestuft, da mittlerweile ein offizieller Webauftritt der Stadt Chemnitz existiert, weshalb das Entfernen der von der Hochschule angebotenen regionalen Informationen angekündigt wird. Es handelt sich um ein Beispiel für den Umstand, dass äußere Gegebenheiten in der Lage sind, die Inhalte eines universitären Webauftritts zu beeinflussen. Der Hypertexttyp *Webauftritt einer Universität* sieht die Präsentation regionaler Informationen vor, diese werden jedoch zunehmend durch Verweise auf externe Webangebote abgedeckt. Abschließend sei angemerkt, dass Hotlists in ihrer prototypischen Form als Listen von Hyperlinks zu *beliebigen* Webangeboten (vgl. Abschnitt 4.6.6) in lediglich drei Fällen belegt werden können.

11.5.4 Einstiegsseiten eingebetteter Hypertexte

Die 692 Dokumente enthalten insgesamt 26 Hypertextknotensorten, die Einstiegsseiten eingebetteter Hypertexte darstellen und somit dem Hypertextknotentyp *Einstiegsseite* zugehörig

[45] Dieser abschließende Typ des Verteilers kann als Strategie aufgefasst werden, einen „fließenden Übergang" zu den Angeboten anderer Organisationseinheiten zu ermöglichen. Von der Einstiegsseite des universitären Webauftritts gelangt der Leser z. B. nach Aktivierung des Hyperlinks „Bibliothek" zu dem angesprochenen *Verteiler* (D 384), der zahlreiche Merkmale der *Einstiegsseite des Webauftritts einer Bibliothek* aufweist (z. B. ein Foto der Bibliothek und eine entsprechende Überschrift), der Rezipient befindet sich jedoch noch immer innerhalb des zentralen Webangebots. Erst wenn der Leser einen weiteren Hyperlink wählt, wird er auf das Angebot verwiesen, das unter der redaktionellen Kontrolle der verknüpften Organisationseinheit steht. Der Produzent des Verteilers könnte also der Auffassung sein, dass die Einstiegsseite des Webauftritts der Bibliothek zu sehr von der Gestaltung des zentralen Angebots abweicht und potenziell einen Kohärenzbruch verursachen könnte. Die Einstiegsseite könnte auch schlicht als defizitär eingeschätzt werden.

[46] In Tabelle 11.13 mussten sowohl die Namen der Hochschulen als auch einige Etiketten von Hypertextknotensorten aus Darstellungsgründen gekürzt werden. Weiterhin enthält diese Tabelle neben den bereits in Tabelle 11.10 dargestellten Hypertextknotensorten die Subtypen des Verteilers.

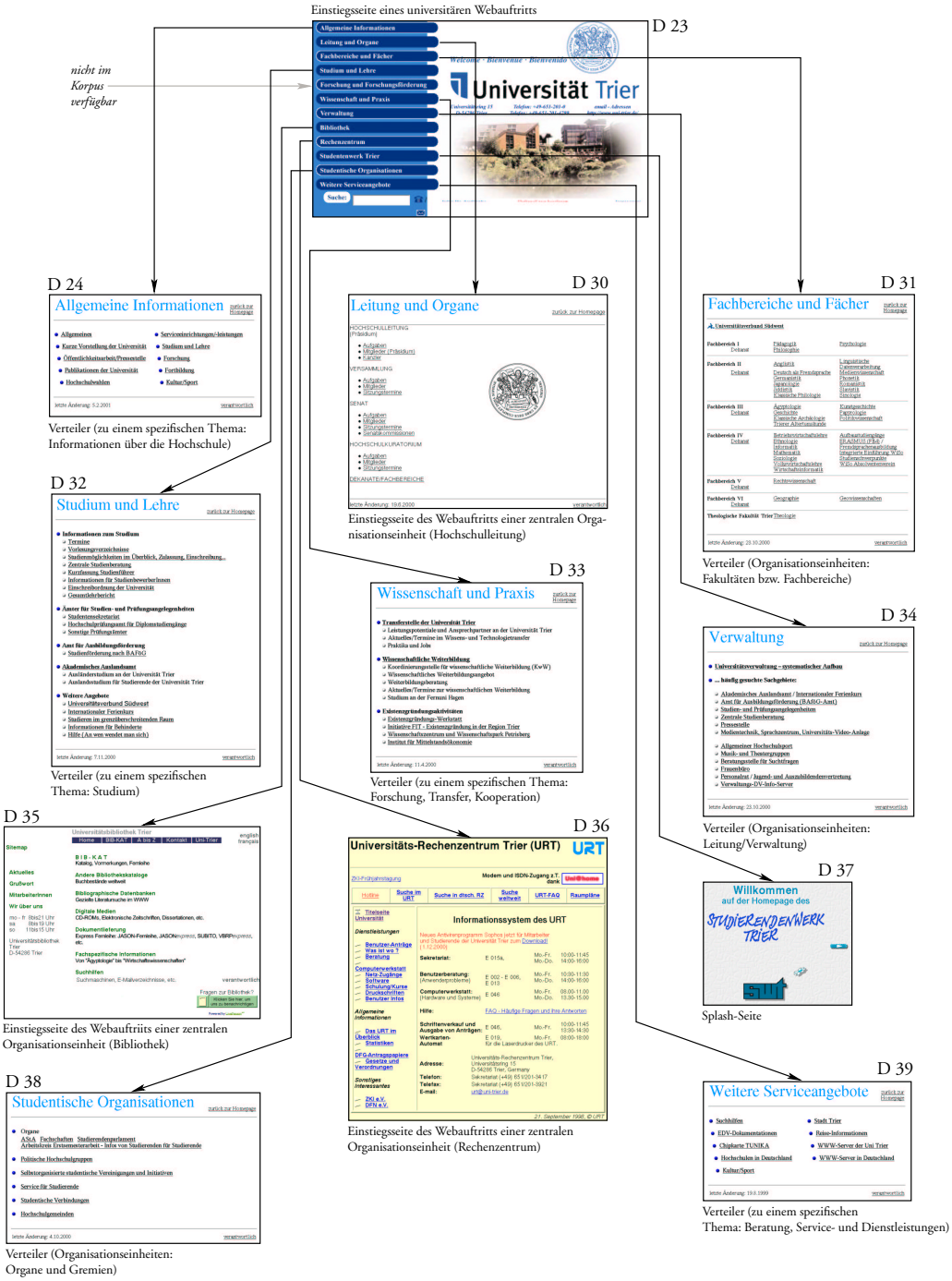


Abbildung 11.5: Einsatz von *Verteilern* zur Strukturierung universitärer Webangebote

	Ulm	Chemnitz	Osnabrück	Koblenz-Landau	Aachen	Essen	Lübeck	FU Berlin	Konstanz	Mannheim	EBS	HdK Berlin	Bonn	Düsseldorf	Darmstadt	Kassel	Leipzig	Trier	Passau	Bochum	Dresden	Heidelberg	Dortmund	Hamburg-Harb.	Eichstätt	Münster	Hohenheim	Gießen	Cottbus	Augsburg	Paderborn	Witten/Herd.	Hagen	Freiberg	Erlangen-Nürnb.	Σ
ES des Webauftritts einer zentr. OE	3	11	3	5	1	1		1		2	2	3	4	6	4	5	3	3	2	1		1		2	2			1						2	2	70
Verteiler (zu einem spezifischen Thema)	6	4	5	1	1	3	1	5	1		1	2	3		2	1	3	5	1	2	1	3			2	5	3	3	2	1	3					70
ES des Webauftritts einer Universität	2	3	2	3	2	1	2	1	2	1	1	1	1	1	2	2	1	3	1	1	1	1	1	2	1	2	2	1	1	1	2	1	1	1	1	52
Verteiler (OE)	2	1	3	2		1	4		2	5			1	2	3	1	3	2				2	1	1	5	1	1	1	1	1	2	1				49
ES eines zielgruppenspezifischen Webangebots			3	1	8	2		4	1	1	2	1	2			3	1					1	2	2			1						1			36
ES des Webauftritts einer dez. OE	9	5					1							5	2		1														6			1	30
Verteiler (für eine spezifische Zielgruppe)	1	1	1		1	6	1	1		1	1					1	1	1		2	8	1	1						1						29
Suchformular	1			1	1		1		1	1	1	1	1			1	1	2	1	1	1	1	1				1	2								21
ES einer universitären Selbstdarstellung	1		3	1	1			1	1		1		1			1	1		1	1			2	1		1	1		1		1					21
Verteiler/Hotlist	1		1	1											1		1		3	3		1			1	2		1	1	2	1					20
Hotlist		1			1		2			1		1	2	2	1				2	1					1			1	1							17
Pressemitteilung					3	1			8								2						2			1										17
Kontaktinformationen (Anschriften, etc.)	1	1		3			3			1	1		3									1	1	1				1								17
ES des Webangebots eines Fachgebiets etc.	9	1		1					2		1	1	1																							16
Veranstaltungskalender	1	1			1	1	2	1		1	1	1		2					1	1	1				1											16
Benutzungshinweise/Anleitungen		1		1	3	1	1		2				1		1	1			1														1			14
Verteiler (zu aktuellen Informationen)	1		1			1		1		1			2			1							1					1	1	1						12
Verteiler (Kontaktinformationen etc.)			1	1		1		1		1						1	1			1					1			1								10
Kerninformationen für eine spezifische Zielgruppe				6							2	1																								9
ES eines Studienführers	4		3								1																						1			9
ES des Webauftritts eines universitätsint. Projekts		1	1			2				1		1									2	1														9
Impressum	1	1					1	1				1											1	1					1	1			1			9
Lageplan/Karte/Wegbeschreibung	1					1	1				1					1			1							1								1		8
ES eines Webangeb. zum Thema Forschung/Lehre			1	1	1				1	1		1											1			1										8
ES des Webangebots mit e. Überbl. über dez. OE			1			1		1			2	2	1																							8
Redaktion eines Webauftritts					1			1		1					1			1		1							1						1			8
Automatische Weiterleitung (kommentiert)			1					1		1					2		1			1																7
Verteiler (Überblick/Abstract/Inhaltsverzeichnis)					4									2						1																7
ES eines Webangebots mit Stellenanzeigen		2				1				1							1		1		1															7
Kurzdarstellung einer OE (Profil/Porträt)	1		1								1	1									1											1			6
ES des Webangebots einer univers. Veranstaltung		2	1																		1		1	1											6
ES eines Webangebots mit aktuellen Informationen	1			1															1			1				1						1			6
Foto bzw. Fotogalerie					1			1			2	1			1																					6
Ankündigung					1			3	1																											5
ES eines Webangebots mit Pressemitteilungen		1				1	1		1		1																									5
Kurzdarst. einer OE (Funktionen und Kontaktinf.)		1								1		1			1				1																	5
Einladung zu einer Veranstaltung (mit Programm)	1					1			1	1		1																								5
ES eines Vorlesungsverzeichnisses		1		1				1		1		1																								5
ES einer Universitätszeitung	1						1							1								1														4
Index (A-Z) bzw. Schlagwortverzeichnis										1										1								1	1							4
ES eines Shops/Online-Shops	1	1								1	1																									4
Pressemitteilung (Auflistung)		1											1	1							1															4
Textknoten mit generischer Informationsfunktion						1												1		1																3
ES eines Webangebots zu angebot. Studiengängen	1			1													1																			3
Stellenanzeige												1		1									1													3
Grundordnung, Gesetz, juristischer Text						1	1								1																					3
Copyright-Informationen								1																					1	1						3
ES eines Forschungsberichts	2														1																					3
ES einer digitalen Bibliothek						1		2																												3
Kurzdarst. der hist. Entwicklung einer Institution	1			1			1																													3
Daten und Fristen eines Semesters				1																			1													2
Sitemap																						1		1												2
Schwarzes Brett			1				1																													2
ES mit Webserver-bezogenen Informationen			1			1																														2
Personalia							1							1																							2
Splash-Seite																		1											1							2
Kurzdarstellung einer Stadt bzw. Region	1		1																																	2
ES einer persönlichen Homepage																			2																	2
ES des Webauftritts e. Stiftung/e. Fördervereins		1									1																									2
ES des Webangebots einer Konferenz/Tagung		1							1																											2
Haftungsausschluss (Disclaimer)								1																												1
Grußwort											1																									1
ES des Webangebots mit student. Homepages																																	1			1
E-Mail-Formular						1																														1
Kleinanzeigen																																		1		1
Speisenplan der Mensa	1																																			1
Antragsformular							1																													1
Wahlergebnis					1																															1
Posting-Formular							1																													1
Ausschreibung (universitätsintern)									1																											1
ES des Webauftritts einer studentischen Initiative		1																																		1
„under construction"-Hinweis							1																													1
Σ	55	44	35	33	32	31	29	29	26	26	25	23	23	23	22	20	20	20	20	20	18	16	15	14	13	13	13	12	12	10	9	9	7	5	5	

Tabelle 11.13: Verteilung der Instanzen von Hypertextknotensorten

Hypertextknotensorte	Frequenz
Einstiegsseite des Webauftritts einer zentralen Organisationseinheit	70
• Akademisches Auslandsamt	10
• Bibliothek	10
• Pressestelle	8
• Dezernat für Forschungsförderung	7
• Verwaltung	7
• Rechenzentrum	5
• Transferzentrum	5
• Personalvertretung	3
• Hochschulleitung	2
• Sprachenzentrum	2
• Studentenwerk/Mensa	2
• Studierendenvertretung	2
• Studienberatung	2
• Klangwerkstatt	1
• Lehr- und Besucherbergwerk	1
• Patentinformationszentrum	1
• Sportzentrum	1
• Universitätsarchiv	1
Einstiegsseite eines zielgruppenspezifischen Webangebots	36
• Studierende	14
• Absolventen/Alumni	7
• Mitarbeiter	6
• Studieninteressenten/Schüler	4
• Wirtschaft	4
• Presse	1
Einstiegsseite des Webauftritts einer dezentralen Organisationseinheit	30
• Fachbereich/Fakultät	21
• Institut/Seminar	4
• Klinikum	4
• Professur/Arbeitsgruppe	1

Tabelle 11.14: Die Ausprägungen der drei hochfrequenten Typen von Einstiegsseiten

sind (vgl. Tabelle 11.10). Für die drei hochfrequentesten Kategorien konnten unterschiedliche Formen der Ausprägung ermittelt werden (vgl. Tabelle 11.14), die sich einerseits auf den Typ der verknüpften zentralen oder dezentralen Organisationseinheit und andererseits auf die jeweilige Zielgruppe beziehen. Die Einstiegsseiten zentraler Organisationseinheiten werden am häufigsten referenziert, wobei auf das akademische Auslandsamt und die Bibliothek jeweils 10 Vorkommen entfallen. Auffällig sind die niedrigfrequenten Organisationseinheiten, die Besonderheiten spezialisierter Hochschulen darstellen. Die „Klangwerkstatt“ ist eine Einrichtung der Hochschule der Künste Berlin, und das „Lehr- und Besucherbergwerk“ wird von der TU Bergakademie Freiberg betrieben. Bei den Einstiegsseiten dezentraler Einrichtungen überwiegen mit den Fakultäten und Fachbereichen naturgemäß diejenigen Organisationseinheiten, die spezifischere Einheiten wie z. B. Institute bzw. Seminare umfassen. Die Webauftritte dieser zuletzt genannten Einheiten sind also wiederum von den Webauftritten der Fachbereiche und Fakultäten zu erreichen. Hochfrequent sind darüber hinaus Einstiegsseiten zu Webangeboten, die zielgruppenspezifische Informationen beinhalten.

Insbesondere die zielgruppenspezifischen Webauftritte verdeutlichen unterschiedliche produzentenseitige Strategien: Falls in der *Einstiegsseite eines universitären Webauftritts* eine *zielgruppenspezifische Navigationshilfe* bereitgestellt wird, so verweisen die enthaltenen Hyperlinks entweder auf *zielgruppenspezifische Verteiler* oder auf die *Einstiegsseite eines zielgruppenspezifischen Hypertextes*, der eigens für die Informationsbedarfe einer Zielgruppe erstellt wurde. Sowohl die Erstellung als auch die fortwährende Pflege der Angebote dieses zweiten Typs stellt für den Produzenten einen größeren Aufwand dar als die einfache Verknüpfung von Angeboten, deren Pflege den Angehörigen anderer Organisationseinheiten obliegt. Anhand der Frequenzangaben wird eine weitere Strategie deutlich, die bereits in Abschnitt 11.4 angesprochen wurde: Einige Einstiegsseiten universitärer Webauftritte verweisen unmittelbar auf die Webpräsenzen einzelner Fakultäten oder Fachbereiche, während andere Einstiegsseiten zunächst auf einen Verteiler verweisen, der seinerseits die Organisationseinheiten einer Hochschule auf der Basis verschiedener Kriterien auflistet und ihre Webauftritte referenziert.

Neben den genannten Hypertextknotensorten existieren 23 weitere unterschiedliche Einstiegsseiten (vgl. Tabelle 11.10), die beispielsweise die universitäre Selbstdarstellung, das Angebot eines Fachgebiets, Studiengangs oder Weiterbildungsangebots oder die Webpräsenz eines universitätsinternen Projekts einleiten.[47] Besonders auffällig sind Vertreter mehrerer traditioneller Textsorten, die teilweise ohne signifikante Modifikationen und in anderen Fällen mit zahlreichen WWW-spezifischen Anpassungen publiziert werden. Zu den traditionellen Textsorten gehören *Studienführer*, *Stellenanzeige*, *Vorlesungsverzeichnis*, *Pressemitteilung*, *Universitätszeitung* und *Forschungsbericht*. Weitere Typen von Einstiegsseiten besitzen zumindest ein mittelbares Pendant in der Klasse gedruckter Texte, so werben z. B. studentische Initiativen häufig mit Flugzetteln oder Aushängen, universitätsinterne Projekte werden durch Rundschreiben angekündigt oder in der Universitätszeitung erläutert, und der Hochglanzprospekt einer Hochschule bündelt verschiedene Informationen, die innerhalb des Webauftritts einer Hochschule von mehreren Hypertexten instanziiert werden (z. B. die universitäre Selbstdarstellung oder Angaben zum Thema Forschung und Lehre).

Es stellt sich die Frage, ob die eingebetteten Hypertexte auf individuellen Hypertextsorten basieren. Für ihre Beantwortung wäre es notwendig, Stichproben dieser Hypertexte anzufertigen und zu analysieren, um das Ausmaß ihrer spezifischen Konventionen zu ermitteln, die die Annahme der Existenz untergeordneter Hypertextsorten rechtfertigen könnten. Da eine detaillierte Untersuchung dieses Aspekts den Rahmen der vorliegenden Arbeit sprengen würde, wurde lediglich eine unsystematische Überprüfung durchgeführt. Den Ergebnissen zufolge unterliegen z. B. die Webangebote von Studiengängen oder die Webauftritte zentraler oder dezentraler Einrichtungen zahlreichen Konventionen, die auch auf hochschulübergreifender Ebene Gültigkeit besitzen. Gerade die exakt 100 Vorkommen unterschiedlicher Webauftritte von Organisationseinheiten sind bereits innerhalb ihrer Einstiegsseiten, die auf 22 Hyper-

[47] In den 692 Dokumenten sind auch insgesamt 17 *Einstiegsseiten des Webauftritts einer Universität* enthalten. Ein Teil dieser Vorkommen ist durch den Umstand zu erklären, dass ein Dokument unterschiedliche URLs besitzen kann (vgl. Fußnote 41, S. 493), denn die Adressen `http://www.uni-ulm.de` (D 131) und `http://www.uni-ulm.de/index.html` (D 185) referenzieren das gleiche HTML-Dokument. D 185 ist Teil der Stichprobe, weil ein Hyperlink in D 131 zyklisch auf D 185 verweist. Weiterhin existieren alternative Versionen verschiedener Einstiegsseiten, so stellt z. B. D 25 die englischsprachige Variante der Einstiegsseite der Universität Trier dar, die auf Grund der Existenz verschiedener deutschsprachiger Token vom automatischen Sprachenidentifizierer als deutschsprachig klassifiziert und daher in das Korpus integriert wurde.

textknotensorten basieren, von einem hohen Grad der Konventionalisierung geprägt, der sich primär auf die interne Strukturierung der Organisationseinheiten stützt.[48]

11.5.5 Inhaltsknoten

Die 198 verbleibenden HTML-Dokumente können insgesamt 38 Hypertextknotensorten zugeordnet werden (vgl. Tabelle 11.10).[49] Die meisten Vorkommen besitzt das *Suchformular*, das insbesondere hinsichtlich des Hypertextsortenmodultyps der Interaktion markiert ist.[50] Nicht alle Suchformulare dienen der Recherche in den von einer Universität angebotenen HTML-Dokumenten. Einige der Vorkommen können z. B. auch zur Auffindung von Namen in einem Adress- oder Telefonverzeichnis eingesetzt werden.

Auf der Hypertextknotensorte *Kontaktinformationen* basieren 17 Dokumente. Diese Knoten beinhalten Informationen über Anschriften, Telefon- und Faxnummern und E-Mail-Adressen und werden in der Regel in Form einer Tabelle oder einer Liste präsentiert. Diese Darstellung ähnelt vergleichbaren Angaben, die z. B. in Studiengangsbroschüren, Prospekten einer Hochschule oder in den Anhängen von Vorlesungsverzeichnissen enthalten sind.

Ebenfalls 17 Vorkommen entfallen auf die Hypertextknotensorte *Pressemitteilung*; allein acht Pressemitteilungen werden in der Einstiegsseite des Webauftritts der Universität Konstanz referenziert. Die von einer spezifischen Hochschule angebotenen Pressemitteilungen unterliegen einer konsistenten Gestaltung und in nahezu allen Fällen weist sich die Pressestelle als Produzent dieser Dokumente aus. Hyperlinks referenzieren sowohl die Einstiegsseite des Webauftritts der Pressestelle, als auch ein Überblicksdokument, das eine Auflistung von Pressemitteilungen zur Verfügung stellt; in einigen Fällen gestattet ein Suchformular die Recherche nach spezifischen Pressemitteilungen.

Die Veröffentlichung aktueller Meldungen ist ein Kernmerkmal zahlreicher Hypertextsorten. Diesbezüglich können in den 35 universitären Webauftritten unterschiedliche Strategien beobachtet werden: Viele Hochschulen bieten ausschließlich *Pressemitteilungen* an. In der Einstiegsseite verweist ein Hyperlink auf ein Überblicksdokument, oder die Überschriften der Pressemitteilungen werden bereits in der Einstiegsseite präsentiert. Die dritte Möglichkeit betrifft die Publikation vollständiger Mitteilungen innerhalb der Einstiegsseite. Andere Hochschulen bieten derartige Informationen nicht an und verweisen stattdessen auf einen

[48] Einen besonderen Stellenwert besitzen die Einstiegsseiten sehr untypischer Organisationseinheiten wie z. B. der Klangwerkstatt oder des Lehr- und Besucherbergwerks. Für diese Typen von Einstiegsseiten sind in der Stichprobe keine Vergleichsdokumente enthalten, doch gilt auch in diesen niedrigfrequenten Fällen, dass die interne Strukturierung und Organisation der jeweiligen Einheit den Inhalt der zugehörigen Webauftritte determiniert.

[49] Dieser Abschnitt geht lediglich auf die hochfrequenten Hypertextknotensorten ein. Nachfolgend werden zwar die meisten weiteren Hypertextknotensorten diskutiert, eine vollständige Darstellung der Ergebnisse kann jedoch aus Platzgründen an dieser Stelle nicht erfolgen.

[50] Die drei *unterschiedlichen* Hypertextknotensorten *Suchformular*, *Posting-Formular* und *E-Mail-Formular* werden in Tabelle 11.10 nur aus Darstellungsgründen in einer Zeile dargestellt. Das einzige Vorkommen eines Posting-Formulars (D 622) dient der Eintragung von Nachrichten in einem interaktiven „schwarzen Brett“, das von der MU Lübeck angeboten wird. Mit Hilfe des E-Mail-Formulars (D 477) können Rezipienten eine „EMail-Verbindung zum Betreuer dieses Services“ aufnehmen. Dabei handelt es sich um ein HTML-Formular, das mit einem CGI-Skript auf dem Webserver kommuniziert, welches dem Betreuer eine E-Mail zusendet, ohne dass der Rezipient selber einen E-Mail-Client aktivieren muss. Da mittlerweile alle Browser in der Lage sind, `mailto:`-Hyperlinks zu interpretieren, werden derartige Formulare nur noch sehr selten angeboten. Dieses Exemplar wurde im Juli 1994 angelegt und enthält eine ausführliche Erläuterung des Prozederes.

Veranstaltungskalender, der die an einer Universität stattfindenden Veranstaltungen in umgekehrt chronologischer Reihenfolge als Tabelle enthält und ebenfalls von einer zentralen Organisationseinheit gepflegt wird.[51] Die Relevanz einer aktuellen Meldung ist jedoch ebenfalls zielgruppenabhängig, so dass einem generischen Rezipienten im besten Falle *alle* aktuellen Informationen präsentiert werden sollten. Einige Universitäten beschränken sich nicht auf die isolierte Referenzierung von Pressemitteilungen oder eines Veranstaltungskalenders, sondern aggregieren *sämtliche* aktuellen Meldungen unabhängig von der publizierenden Instanz (Pressestelle, Hochschulverwaltung, Fachbereiche, Institute etc.): Die *Einstiegsseite eines Webangebots mit aktuellen Informationen/Meldungen* ist mit sechs Vorkommen in der Stichprobe vertreten. Die korrespondierende Hypertextsorte ist ein seltenes Beispiel für den Umstand, dass die Organisationsstruktur einer Hochschule durchbrochen wird, um den Rezipieten ein möglichst informatives und tagesaktuelles Webangebot zur Verfügung stellen zu können.

Benutzungshinweise und Anleitungen zu Internet-Diensten, technischen Aspekten (z. B. einer Telefonanlage oder einem Internet-Wählzugang) oder dem Webangebot einer Hochschule sind mit 14 Vorkommen in der Stichprobe vertreten. D 106 präsentiert z. B. „Hinweise zur Benutzung der NetNews“, D 449 umfasst den ausführlichen Text „Das Internet – eine Einführung“, D 490 enthält „Hilfen für WWW-Leser und -Anbieter“. Auf die Funktionen, Gestaltungsmerkmale und Navigationshilfen eines Webangebots gehen ebenfalls mehrere Dokumente ein. D 373 stellt „Das Webangebot der RWTH im Überblick“ dar und enthält Rubriken wie „Wege durch das RWTH Web“ oder „Kurz vorgestellt: Die Rubriken der Navigationsleiste“. D 674 umfasst eine „Html-Dokumentation [sic] der neuen Webseiten der Universität Konstanz“ und geht unter anderem auf den „Hintergrund der Neugestaltung“ und die „Funktion der Webseiten“ ein. Sämtlichen Vertretern dieser Hypertextknotensorte ist gemein, dass primär technische Aspekte einer informationellen Ressource im weitesten Sinne erläutert werden. Ihre Instanzen richten sich vornehmlich an Rezipienten, die detaillierte Informationen über diese Ressource benötigen, z. B. weil sie Probleme mit ihrer Anwendung besitzen oder selbst als Produzent eines Webangebots tätig werden möchten.

Die Hypertextknotensorte *Impressum* besitzt in der Stichprobe neun Vorkommen und ist von der *Redaktion eines Webauftritts* zu unterscheiden (acht Vorkommen). Einige Hochschulen verwenden beide Hypertextknotensorten (z. B. die FU Berlin), während sich andere auf das Impressum beschränken, in dem zusätzlich die Namen und Kontaktinformationen der Verantwortlichen aufgeführt werden. Das Impressum ähnelt seinem Pendant in den Printmedien und fokussiert primär rechtliche Aspekte, bei der Redaktion eines Webauftritts stehen im Gegensatz dazu die oftmals mit Porträt- oder Gruppenfotos vorgestellten Mitarbeiter im Zentrum einer eher informellen Präsentation.[52] Im Impressum werden die Namen der Mitar-

[51] Die Hypertextknotensorte *Veranstaltungskalender* ist mit 16 Vorkommen in der Stichprobe vertreten. Die für diesen Zweck an der Universität Essen eingesetzte Software wurde von der HU Berlin und der Ruhr-Universität Bochum übernommen (D 478). Dieser Umstand kann als allmähliche Verhärtung eines Genres im Sinne von Yates und Sumner (1997) verstanden werden (vgl. Abschnitt 4.2.2).

[52] Beim *Impressum* und der *Redaktion eines Webauftritts* handelt es sich um Hypertextsortenmodule, die prinzipiell in arbiträrer Weise mit Instanzen anderer Hypertextsortenmodule kombiniert werden können. In der hier untersuchten Stichprobe fungieren die beiden Hypertextsortenmodule üblicherweise als Hypertextknotensorten. Die genannten rechtlichen Aspekte beziehen sich insbesondere auf einen Haftungsausschluss, der ebenfalls ein Hypertextsortenmodul darstellt und in den meisten Vorkommen ein Bestandteil des Impressums ist. Eine Hochschule bietet den *Haftungsausschluss* als eigenständigen Knoten an (vgl. Tabelle 11.10).

beiter lediglich in Form einer Liste aufgeführt. Die in der Einstiegsseite des Webauftritts der Universität Passau enthaltenen Namen der Produzenten wurden als Hyperlinks realisiert und verweisen auf die persönlichen Homepages der Autoren. Diese beiden Vorkommen erklären die in Tabelle 11.10 aufgeführte Hypertextknotensorte *Einstiegsseite einer persönlichen Homepage*, die in diesem speziellen Fall als Präsenz der Redaktion des zugehörigen Webauftritts fungieren. In den beiden sehr unterschiedlich gestalteten Homepages wird diese Funktion explizit genannt: „Unix-Systemverwaltung, WWW-Administration, Datenbanken, Modems" (D 453) und „Aufgaben (u. a.) [...] Webmaster der FMI" (D 454).

Sieben Dokumente enthalten eine *automatische Weiterleitung*. Die Weiterleitung erfolgt nach einer vorgegebenen Zeitspanne (z. B. zehn Sekunden) und wird durch ein `meta`-Element innerhalb des `head`-Bereiches realisiert, in dem die URL des eigentlichen Zielknotens hinterlegt wird. Bezüglich dieser Vorkommen wird eine eigenständige Hypertextknotensorte angenommen, da sämtliche Vorkommen von knappen Kommentaren flankiert werden, die einander sehr ähneln: „Diese Seite ist umgezogen, Sie werden automatisch umgeleitet. Die neue Seite befindet sich hier." (D 489), „Diese Seite ist umgezogen. Sie finden Sie jetzt unter index.de.tud. Nach 5 Sekunden werden Sie automatisch dorthin weitergeleitet." (D 500), „Falls diese Seite nicht automatisch ersetzt wird, bitte diesem Link folgen." (D 539), „Weiterleitung zu Adressen, Zahlen und Darstellung der Ruhr-Universität Bochum ... Falls es nicht automatisch weitergeht, bitte einmal klicken." (D 566) und „Die Adresse der Seite Forschungsberichte und Forschungskontakte hat sich geändert [*Zeilenumbruch*] Sie werden automatisch zu den von i3v generierten Seiten umgeleitet, sollte dies Ihr Browser nicht unterstuetzen [sic] klicken sie bitte hier" (D 721). Auffällig ist, dass sämtliche Erläuterungen einen Hyperlink zur neuen URL des Dokuments beinhalten, falls die Weiterleitung nicht automatisch durchgeführt werden kann. Es ist zu betonen, dass derartige Dokumente aus technischer Perspektive nicht notwendig sind, da eine automatische Weiterleitung auch in der Konfiguration des Webservers vorgenommen werden kann (vgl. auch Abschnitt 4.6.10).

Die *Kurzdarstellung einer Organisationseinheit* existiert in zwei unterschiedlichen Formen, die auf der Grundlage ihrer spezifischen Konventionen als eigenständige Hypertextknotensorten aufgefasst werden. Die erste Form fokussiert die Vorstellung einer Organisationseinheit durch die Angabe eines Profils oder eines Porträts. In vier der sechs Vorkommen steht zwar die gesamte Hochschule im Skopus des Dokuments, da aber auch z. B. Zentren und Institute derartige Texte anbieten, wurde ein abstraktes Etikett gewählt, das den Geltungsbereich der Hypertextknotensorte nicht einschränkt. Die vier Dokumente (D 211: „Virtuelle Universität", D 243: „Philosophie der ebs", D 341: „HdK-Profil" und D 354: „Kurzporträt der Technischen Universität Dresden") stellen individuelle Merkmale der Institutionen vor und gehen schlaglichtartig auf ausgewählte Aspekte ein, z. B. „Praxisorientierung", „Internationalität", „Leistungsbereitschaft", „Moderne didaktische Methoden" und „Studienbegleitende Persönlichkeitsentwicklung" (D 243). Die verbleibenden Dokumente beziehen sich auf das Forschungsprofil der Universität Osnabrück (D 521) und das Porträt einer „studentischen Unternehmensberatung", die an der Universität Ulm tätig ist (D 133). Die zweite Form beschränkt sich auf eine formale und knappe Darstellung der Zusammensetzung und Aufgaben einer Organisationseinheit und wird in fünf Dokumenten verwendet. D 96 gibt z. B. Auskunft über die Zusammensetzung des Kuratoriums der TU Chemnitz, listet die Namen der Mitglieder auf und erläutert seine Aufgaben anhand eines Auszugs des sächsischen

Hochschulgesetzes. Die weiteren vier Dokumente stellen in ähnlicher Form ein Graduiertenkolleg (D 349), das „Referat Grundsatzangelegenheiten und außeruniversitäre Kontaktaufgaben“ (D 443), die „Vereinigung von Freunden der Technischen Universität zu Darmstadt e. V.“ (D 499) und die „Gesellschaft der Freunde der Universität Mannheim e. V. (GdF)“ vor. Diese in den Dokumenten genannten Überschriften deuten bereits an, dass sich die beiden Hypertextknotensorten in einem wesentlichen Aspekt unterscheiden: Während ein Profil oder ein Porträt lediglich als Einzelknoten innerhalb eines umfangreichen Hypertextes aufzufassen ist, stellt die Kurzdarstellung der Zusammensetzung und Funktionen einer Organisationseinheit bzw. einer assoziierten Institution die *gesamte* Webpräsenz dar, sie fungiert als eine Art digitale Visitenkarte, die aus einem einzigen Knoten besteht.

Nahezu alle Hypertextknotensorten vom Typ Inhaltsknoten können als traditionelle Textsorten aufgefasst werden. In diesen Knoten sind meist vollständige Exemplare traditioneller Textsorten enthalten. Dies kann als weiteres Indiz für den Umstand aufgefasst werden, dass die Produzenten vornehmlich durch die traditionelle Schriftlichkeit geprägt sind (vgl. Eckkrammer, 2001) und eine Hypertextualisierung umfangreicherer Einzeltexte oder Informationsbestände eher als Ausnahme denn als Regelfall aufzufassen ist.[53] Spezifisch für das Medium WWW sind lediglich die Hypertextknotensorten *Suchformular*, *Posting-Formular*, *E-Mail-Formular*, *automatische Weiterleitung*, *Auflistung von Pressemitteilungen*, *Splash-Seite*, *Sitemap* und *„under construction“-Hinweis*.[54] Viele der Hypertextknotensorten mit traditionellen Pendants werden im Bereich der Printpubliktationen von Broschüren instanziiert, z. B. enthält die Hochglanzbroschüre einer Universität ein Grußwort[55] des Präsidenten oder Rektors, Kontaktinformationen, zahlreiche Fotos, ein Profil der Institution, eine Wegbeschreibung, Informationen zur Stadt oder Region, Lagepläne, die die Gebäude auf dem Campus und die historische Entwicklung[56] einer Universität erläutern. *Kerninformationen für eine spezifische Zielgruppe* finden sich in Bezug auf Studieninteressenten insbesondere

[53] Im Falle des Vorliegens umfassender Hypertexte liegen darüber hinaus, wie bereits mehrfach angesprochen, primär hierarchisch organisierte Hypertexte vor, die mit Hilfe von Konvertern aus vorhandenen Dateien (z. B. Textverarbeitungsdokumenten) erzeugt wurden.

[54] Pressemitteilungen werden einzeln veröffentlicht, indem sie z. B. per Fax an Zeitungsredaktionen verschickt werden. Daher kann die *Auflistung von Pressemitteilungen* in einem einzelnen Dokument, das darüber hinaus von derjenigen Institution veröffentlicht wird, die auch die Pressemitteilungen selbst publiziert, als WWW-spezifische Hypertextknotensorte aufgefasst werden. Die beiden Vorkommen der *Splash-Seite* sind dem Webauftritt einer zentralen Organisationseinheit (D 37, vgl. Abbildung 11.5) und dem Webauftritt der TU Cottbus (D 55) vorgeschaltet. Die Vertreter der *Sitemap* (D 209, D 220) stellen eine spezifische Form des Verteilers dar, weil sie wichtige Hyperlinks der zentralen HTML-Dokumente im Überblick darstellen (vgl. Abbildung 11.6).

[55] Das einzige in der Stichprobe enthaltene und explizit als solches identifizierte *Grußwort* (D 242) ist Teil des Webauftritts der European Business School und weist einige charakteristische Merkmale des gedruckten Grußwortes auf, es bezieht sich jedoch nicht auf eine spezifische Veranstaltung, sondern fokussiert das Würdigen der Institution als eigentlichen prototypischen Kern dieser Textsorte (vgl. Abschnitt 2.3.6).

[56] Drei HTML-Dokumente basieren auf der Hypertextknotensorte *Kurzdarstellung der historischen Entwicklung einer Institution* und besitzen eine identische Textstruktur: In D 177, D 293 und D 613 werden innerhalb einer knappen listenartigen Darstellung die wesentlichen Meilensteine der Historie einer Hochschule in chronologischer Reihenfolge aufgeführt. Der auffälligste Unterschied besteht im Umfang der Dokumente: Die Universität Koblenz-Landau hat im Jahr 2000 ihr zehnjähriges Bestehen gefeiert, weshalb D 293 lediglich zwei Einträge enthält. Die MU Lübeck wurde 1985 gegründet. D 613 umfasst acht Einträge, die die vorhergehenden Institutionen und nach 1985 eingeführte Studiengänge auflisten. D 177 stellt die Entwicklung der Universität Ulm ab 1959/60 anhand von insgesamt 52 Einträgen dar.

in Studienführern, die die Inhalte, Lehrveranstaltungen und Studienordnungen eines oder mehrerer verwandter Studiengänge in knapper Form darlegen. Oftmals enthalten Exemplare dieser Textsorte auch Kurzdarstellungen der beteiligten Organisationseinheiten (z. B. Institute bzw. Seminare oder Fachgebiete) und Angaben zu den Daten und Fristen bevorstehender Semester (vgl. Abbildung 11.6). Drei Hypertextknotensorten fungieren als traditionelle Bestandteile des Peritextes zahlreicher Textsorten, so sind *Copyright-Informationen*, ein *Schlagwortverzeichnis* und ein *Impressum* üblicherweise obligatorische Konstituenten von Fachbuchpublikationen. Andere Textsorten, die häufig als Hypertextknotensorten in universitären Webauftritten Verwendung finden, wurden vor der großflächigen Einführung digitaler Kommunikationsmedien als Rundschreiben der Universitätsleitung oder spezifischer Organisationseinheiten auf postalischem Wege verschickt, z. B. die *Ankündigungen*[57] neuer Dienstleistungen, *Einladungen zu Veranstaltungen*[58] oder *universitätsinterne Ausschreibungen*. Angaben zu Ruferteilungen, -annahmen und -ablehnungen sowie Ernennungen neuer Hochschullehrer werden traditionellerweise in der Rubrik „Personalia“[59] einer Universitätszeitung publiziert, die ebenfalls die *Wahlergebnisse* universitärer Gremien und einen *Veranstaltungskalender* enthält, der den Zeitraum bis zur Veröffentlichung der nachfolgenden Ausgabe abdeckt. Einige der Hypertextknotensorten besitzen darüber hinaus ein traditionelles Pendant in unterschiedlichen Typen von Aushängen, z. B. *Stellenanzeigen*, *Kleinanzeigen*, der *Speisenplan der Mensa* oder das generische *schwarze Brett* mit Anschlägen jeglicher Couleur.

Abbildung 11.6 stellt Beispiele dar, die unterschiedliche Formen der Anpassung an die Spezifika des Mediums WWW verdeutlichen.[60] Das *Antragsformular* macht nicht von der Möglichkeit Gebrauch, die benötigten Informationen zur Beantragung einer Benutzerkennung online eingeben zu können. Stattdessen ist der Rezipient gezwungen, dieses HTML-Dokument handschriftlich oder in einem Editor auszufüllen und in ausgedruckter Form per Brief oder Fax an das Rechenzentrum zu schicken. Die *Ausschreibung* ähnelt vergleichbaren Texten, die in der Regel auf dem postalischen Weg an die Angehörigen einer Hochschule verteilt werden. Dieses Dokument wurde vermutlich in einer Textverarbeitung erstellt und

[57] Die fünf Exemplare der Hypertextknotensorte *Ankündigung (einer neuen Dienstleistung)* beziehen sich auf drei neue Zugänge zu digitalen Bibliotheken mit Fachinformationen bzw. -publikationen an der FU Berlin (D 310–D 312), die Einführung des neuen Veranstaltungsverzeichnisses an der RWTH Aachen (D 400) und die Ausleihmöglichkeiten von Videoprojektoren an der Universität Konstanz (D 687).

[58] Die fünf in der Stichprobe enthaltenen *Einladungen zu Veranstaltungen* betreffen den Dies academicus an der Universität Ulm (D 140), ein Treffen der Ehemaligen der HdK Berlin (D 352), die „Bundesweite Sommeruniversität für Frauen in Naturwissenschaft und Technik“ (D 481), ein an der Universität Konstanz veranstaltetes Benefizkonzert (D 679) und einen vom Förderkreis des Institut für Mittelstandsforschung der Universität Mannheim organisierten Festvortrag (D 706). Auch diese Dokumente unterscheiden sich bezüglich ihres Umfangs, da sowohl Vorträge als auch mehrtätige Veranstaltungen mit einzelnen Reden und Konzerten angekündigt werden. Dennoch besitzen die Dokumente die gemeinsame kommunikative Funktion der Einladung des Rezipienten sowie der Präsentation des Programms und von Kontaktinformationen.

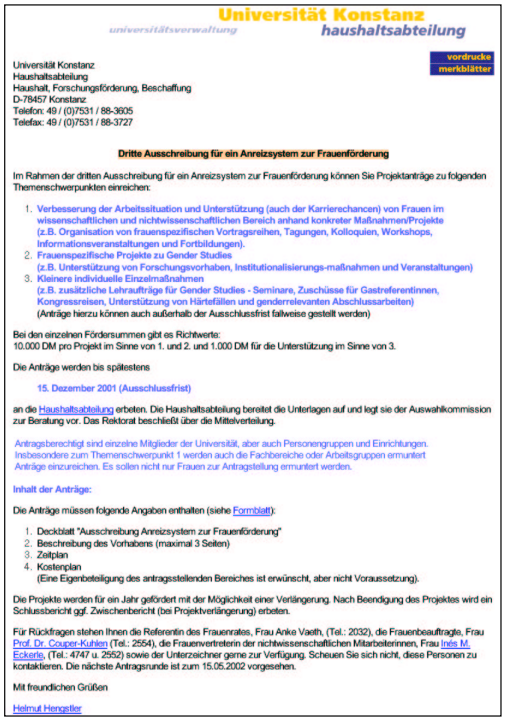
[59] Für die Hypertextknotensorte *Personalia* existieren in der Stichprobe zwei Vertreter: Bei D 425 handelt es sich um ein datenbankgestütztes Webangebot, das bevorstehende Antrittsvorlesungen an der Universität Düsseldorf präsentiert; interessanterweise werden die sechs genannten Veranstaltungen nicht in chronologischer Reihenfolge dargestellt, was auf eine fehlende Sortierklausel in der zugrunde liegenden Datenbankabfrage hindeutet. D 611 umfasst eine umfangreiche und mit „Personalia“ überschriebene Liste, die Rubriken wie z. B. „Ehrungen“, „Fachgesellschaften“, „Preise“, „Gastwissenschaftler“, „Ruf an die MUL“ und „Auswärtiger Ruf“ enthält.

[60] In diesen Screenshots wurden die meisten peritextuellen Elemente (vornehmlich Navigationshilfen) entfernt, um die Lesbarkeit der Bildschirmabzüge gewährleisten zu können.

anschließend in einen HTML-Editor importiert. Besonders auffällig sind einige in das Dokument integrierte Hyperlinks und farbliche Hervorhebungen. Die *Einstiegsseite eines Vorlesungsverzeichnisses* stellt schließlich ein Beispiel für eine aufwändig produzierte Hypertextualisierung dar, die eine intuitive Navigation erlaubt. Auch für diesen Hypertext gilt, dass er einer streng hierarchischen Organisierung unterliegt und aller Wahrscheinlichkeit nach maschinell aus dem Bestand einer Datenbank generiert wurde.

In Bezug auf die von Heinemann (2000b) diskutierten Textsorten des Kommunikationsbereichs Hochschule und Wissenschaft (vgl. Abschnitt 6.2) stellt sich nun die Frage, welche dieser Textsorten bzw. Textsortenklassen in der Stichprobe belegt werden können. Der erste von Heinemann thematisierte Teilbereich betrifft die theoriebezogenen Textsorten (vgl. Abschnitt 6.2.1), dem keine Hypertextknotensorte zugeordnet werden kann. Es könnte zwar prinzipiell angenommen werden, dass die *Einstiegsseite eines Forschungsberichts* dieser Textsortenklasse zugehörig ist, doch handelt es sich bei den drei Vorkommen um Forschungsberichte, die von einer Hochschule publiziert werden und sämtliche Forschungsprojekte einer Universität in einer sehr komprimierten und listenartigen Form vorstellen, weshalb es sich nicht um eine theoriebezogene Textsorte handelt. Während der Bereich der wissenstransmittierenden Textsorten (vgl. Abschnitt 6.2.2) ebenfalls keine Entsprechungen in der Stichprobe besitzt, können für die Textsorten der Wissenschaftsverwaltung (vgl. Abschnitt 6.2.3) mehrere korrespondierende Hypertextknotensorten angeführt werden. Heinemann unterscheidet in diesem Bereich drei Textsortenklassen, wobei für die verwaltungsinternen Dienstanweisungen und Geschäftsordnungen keine Hypertextknotensorten ermittelt werden können. In die Klasse der juristischen bzw. politischen Rahmentexte fällt die Hypertextknotensorte *Grundordnung, Gesetz, juristischer Text*, die drei Vorkommen besitzt: D 474 enthält die „Rahmenregelungen für das WWW-Informationssystem der Universität – Gesamthochschule Essen", D 506 präsentiert die „Allgemeine Benutzungsordnung für die Informationsverarbeitungs- und Kommunikations-Infrastruktur" der TU Darmstadt und D 615 stellt die „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Medizinischen Universität zu Lübeck" dar. Bei diesen drei Exemplaren handelt es sich also um sehr generelle Ordnungen im weitesten Sinne, die als normierende Rahmenvorgaben fungieren. Dem dritten Bereich der verwaltungsinternen Textsorten sind mehrere Hypertextknotensorten zugehörig: Hierzu zählen die *Ankündigung*, die *Einladung*, das *Antragsformular*, die *Ausschreibung* und das *Vorlesungsverzeichnis*.[61] Bei einer Betrachtung der zugehörigen Vertreter fällt auf, dass Heinemanns Charakterisierung der „verwaltungsinternen Textsorten" einer sehr eng gefassten Betrachtungsweise unterliegt, z. B. wurde der einzige Vertreter der Hypertextknotensorte *Antragsformular* (D 619) nicht von der Verwaltung, sondern vom „Institut für Medizinische Informatik/Rechenzentrum" der MU Lübeck veröffentlicht und stellt einen „Antrag auf Erteilung eines Modem-/Emailzugangs" dar. Die bereits diskutierten *Ankündigungen* wurden von der Universitätsbibliothek der FU Berlin, dem „Bereich Technik" der Universität Konstanz und schließlich von der Verwaltung der RWTH Aachen publiziert. Die Tatsache, dass

[61] Einige weitere Textsorten, die als Vorbilder und deren konkrete Exemplare wiederum als Quelltextlieferanten für die Instanzen der in dieser Stichprobe ermittelten Hypertextknotensorten fungieren, wurden bereits angesprochen, z. B. der Studienführer oder der Hochglanzprospekt einer Hochschule, der insbesondere der Werbung und der Imagepflege dient. Auf derartige Textsorten geht Heinemann (2000b) nicht ein, weshalb sie an dieser Stelle nicht näher betrachtet werden.



Universität Konstanz
universitätsverwaltung
haushaltsabteilung

vordrucke
merkblätter

Universität Konstanz
Haushaltsabteilung
Haushalt, Forschungsförderung, Beschaffung
D-78457 Konstanz
Telefon: 49 / (0)7531 / 88-3605
Telefax: 49 / (0)7531 / 88-3727

Dritte Ausschreibung für ein Anreizsystem zur Frauenförderung

Im Rahmen der dritten Ausschreibung für ein Anreizsystem zur Frauenförderung können Sie Projektanträge zu folgenden Themenschwerpunkten einreichen:

1. Verbesserung der Arbeitssituation und Unterstützung (auch der Karrierechancen) von Frauen im wissenschaftlichen und nichtwissenschaftlichen Bereich anhand konkreter Maßnahmen/Projekte (z.B. Organisation von frauenspezifischen Vortragsreihen, Tagungen, Kolloquien, Workshops, Informationsveranstaltungen und Fortbildungen).
2. Frauenspezifische Projekte zu Gender Studies (z.B. Unterstützung von Forschungsvorhaben, Institutionalisierungs-maßnahmen und Veranstaltungen)
3. Kleinere individuelle Einzelmaßnahmen (z.B. zusätzliche Lehraufträge für Gender Studies - Seminare, Zuschüsse für Gastreferentinnen, Kongressreisen, Unterstützung von Härtefällen und genderrelevanten Abschlussarbeiten) (Anträge hierzu können auch außerhalb der Ausschlussfrist fallweise gestellt werden)

Bei den einzelnen Fördersummen gibt es Richtwerte:
10.000 DM pro Projekt im Sinne von 1. und 2. und 1.000 DM für die Unterstützung im Sinne von 3.

Die Anträge werden bis spätestens

15. Dezember 2001 (Ausschlussfrist)

an die Haushaltsabteilung erbeten. Die Haushaltsabteilung bereitet die Unterlagen auf und legt sie der Auswahlkommission zur Beratung vor. Das Rektorat beschließt über die Mittelverteilung.

Antragsberechtigt sind einzelne Mitglieder der Universität, aber auch Personengruppen und Einrichtungen. Insbesondere zum Themenschwerpunkt 1 werden auch die Fachbereiche oder Arbeitsgruppen ermuntert Anträge einzureichen. Es sollen nicht nur Frauen zur Antragstellung ermuntert werden.

Inhalt der Anträge:

Die Anträge müssen folgende Angaben enthalten (siehe Formblatt):

1. Deckblatt "Ausschreibung Anreizsystem zur Frauenförderung"
2. Beschreibung des Vorhabens (maximal 3 Seiten)
3. Zeitplan
4. Kostenplan
   (Eine Eigenbeteiligung des antragsstellenden Bereiches ist erwünscht, aber nicht Voraussetzung).

Die Projekte werden für ein Jahr gefördert mit der Möglichkeit einer Verlängerung. Nach Beendigung des Projektes wird ein Schlussbericht ggf. Zwischenbericht (bei Projektverlängerung) erbeten.

Für Rückfragen stehen Ihnen die Referentin des Frauenrates, Frau Anke Vaeth, (Tel.: 2032), die Frauenbeauftragte, Frau Prof. Dr. Couper-Kuhlen (Tel.: 2554), die Frauenvertreterin der nichtwissenschaftlichen Mitarbeiterinnen, Frau Inés M. Eckerle, (Tel.: 4747 u. 2552) sowie der Unterzeichner gerne zur Verfügung. Scheuen Sie sich nicht, diese Personen zu kontaktieren. Die nächste Antragsrunde ist zum 15.05.2002 vorgesehen.

Mit freundlichen Grüßen

Helmut Hengstler

D 682: Ausschreibung (universitätsintern)

Freie Universität Berlin **FUBerlin**

Home | Suche | E-Mail/Tel. | Kontakt | English

SoSe 2001

Vorlesungsverzeichnis Sommersemester 2001
Lecture and Course Catalog Summer Semester 2001

- Suche im Vorlesungsverzeichnis
- Universitäts- und Ringvorlesungen, fächerübergreifende Studien
- Lehrveranstaltungen der Fachbereiche, Zentralinstitute und Zentralen Einrichtungen

Fachbereiche
- Biologie, Chemie, Pharmazie
- Erziehungswissenschaft und Psychologie
- Geowissenschaften
- Geschichts- und Kulturwissenschaften
- Humanmedizin
- Mathematik und Informatik
- Philosophie und Geisteswissenschaften
- Physik
- Politik- und Sozialwissenschaften
- Rechtswissenschaft
- Veterinärmedizin
- Wirtschaftswissenschaft

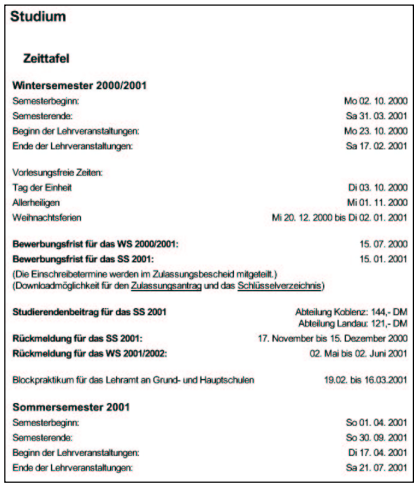
Zentralinstitute
- John-F.-Kennedy-Institut für Nordamerikastudien
- Lateinamerika-Institut
- Osteuropa-Institut

Zentraleinrichtungen
- Frauen- und Geschlechterforschung
- Hochschulsport
- Sprachlabor
- Studienberatung
- ZEDAT (Rechenzentrum)

- Kommentierte Vorlesungsverzeichnisse (Übersicht)
- Hinweise zur Benutzung des Vorlesungsverzeichnisses
- Archiv der Online-Vorlesungsverzeichnisse (ab SoSe 1998)

VV-Impressum - Redaktionsschluß: 25.1.2001

D 304: Einstiegsseite des Vorlesungsverzeichnisses



Studium

Zeittafel

Wintersemester 2000/2001

Semesterbeginn:	Mo 02. 10. 2000
Semesterende:	Sa 31. 03. 2001
Beginn der Lehrveranstaltungen:	Mo 23. 10. 2000
Ende der Lehrveranstaltungen:	Sa 17. 02. 2001
Vorlesungsfreie Zeiten:	
Tag der Einheit	Di 03. 10. 2000
Allerheiligen	Mi 01. 11. 2000
Weihnachtsferien	Mi 20. 12. 2000 bis Di 02. 01. 2001
Bewerbungsfrist für das WS 2000/2001:	15. 07. 2000
Bewerbungsfrist für das SS 2001:	15. 01. 2001
(Die Einschreibetermine werden im Zulassungsbescheid mitgeteilt.) (Downloadmöglichkeit für den Zulassungsantrag und das Schlüsselverzeichnis)	
Studierendenbeitrag für das SS 2001	Abteilung Koblenz: 144,- DM Abteilung Landau: 121,- DM
Rückmeldung für das SS 2001:	17. November bis 15. Dezember 2000
Rückmeldung für das WS 2001/2002:	02. Mai bis 02. Juni 2001
Blockpraktikum für das Lehramt an Grund- und Hauptschulen	19.02. bis 16.03.2001

Sommersemester 2001

Semesterbeginn:	So 01. 04. 2001
Semesterende:	So 30. 09. 2001
Beginn der Lehrveranstaltungen:	Di 17. 04. 2001
Ende der Lehrveranstaltungen:	Sa 21. 07. 2001

D 281: Daten und Fristen eines Semesters

Institut für Medizinische Informatik / Rechenzentrum
Antrag auf Erteilung eines Modem-/Emailzugangs

Nur eMail-Zugang, bitte hier ankreuzen ☐ Bitte in Blockschrift ausfüllen **Studiengang: Medizin/Informatik** Bitte unterstreichen

Name,Vorname: _____

Strasse, Nr: _____

Postleitzahl, Ort: _____ **Tel.:** _____

Datum: _____ **Unterschrift:** _____

Nur für Mitarbeiter: Dienststelle: _____ Telefon: _____

Über die Einwahlknoten können Sie sich analog mit 32Kbit/s oder über ISDN mit 64Kbit/s verbinden. Für beide Zugangsarten gilt der Anschluss:

0451-583210

Der Zugang ist von Wintersemester zu Wintersemester gültig und muss dann erneuert werden!
Der Zugang ist ein Jahr gültig und muss dann erneuert werden!
Bitte legen Sie dem Antrag eine Kopie des Studenten-/Mitarbeiterausweises bei !
Nach der Einwahl stehen Ihnen folgende Dienste zur Verfügung:
WWW, News, e-Mail, FTP und TELNET

Versuchter Missbrauch wie z.B. die Ausforschung fremder Passwörter oder Einbruchversuche in fremde Systeme, Datenbestände oder Rechnernetze, führt zu sofortiger Sperrung Ihres Zugangs.
Die vergebenen Passwörter können vom Benutzer nicht geändert werden. Sollten Sie den Verdacht haben, dass Ihr Passwort bekanntgeworden ist, wenden Sie sich bitte umgehend an das Rechenzentrum.
In der Regel können Sie sich Ihre Kennung und Ihr Passwort 5 Tage nach Eingang des Formulars im Rechenzentrum abholen. Mitarbeitern wird die Kennung zugesandt.
Fragen können Sie montags und freitags von 10:00-11:30 Uhr und mittwochs von 14:00-15:00 an Herrn Rosenhal (Telefon: 500-6613) richten.
Hinweis: Aufgrund von Wartungsarbeiten kann es dienstags von 8:00-12:00 Uhr zu Ausfällen des Zugangs kommen.

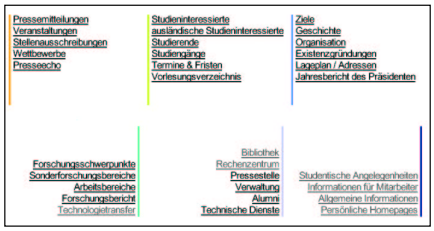
Wird vom Rechenzentrum ausgefüllt:

Antrag genehmigt

_____ Datum _____ Institut-/RZ-Leitung

Antrag bearbeitet _____ Datum Namenskürzel Mitarbeiter-/Studenten-ausweis hat vorgelegen ja ☐ nein ☐

D 619: Antragsformular



Pressemitteilungen
Veranstaltungen
Stellenausschreibungen
Wettbewerbe
Presseecho

Studieninteressierte
ausländische Studieninteressierte
Studierende
Studiengänge
Termine & Fristen
Vorlesungsverzeichnis

Ziele
Geschichte
Organisation
Existenzgründungen
Lageplan / Adressen
Jahresbericht des Präsidenten

Forschungsschwerpunkte
Sonderforschungsbereiche
Arbeitsbereiche
Forschungsbericht
Technologietransfer

Bibliothek
Rechenzentrum
Pressestelle
Verwaltung
Alumni
Technische Dienste

Studentische Angelegenheiten
Informationen für Mitarbeiter
Allgemeine Informationen
Persönliche Homepages

D 209: Sitemap

Abbildung 11.6: Beispiele für Exemplare von Hypertextknotensorten

nicht nur die Hochschulverwaltung Exemplare dieser Textsorten veröffentlicht, kann unter anderem auch für die in der Stichprobe enthaltenen *Einladungen* festgehalten werden.

11.5.6 Zum Geltungsbereich der Hypertextknotensorten

Nahezu alle in dieser Studie ermittelten Hypertextknotensorten stellen sehr trennscharfe kategoriale Einheiten mit spezifischen Geltungsbereichen dar. Es liegen jedoch zwei Kategorien mit einem sehr umfangreichen Geltungsbereich vor, die als ein direktes Resultat der Prämisse aufzufassen sind, dass die Annahme einer »Hypertextknotensorte« wie z. B. „Verschiedenes" oder „Sonstiges" unter allen Umständen vermieden werden sollte.

Die Existenz der sehr allgemeinen Hypertextknotensorte *Textknoten mit generischer Informationsfunktion* geht darauf zurück, dass drei Knoten keiner der anderen Kategorien zugeordnet werden konnten und darüber hinaus als einzige Gemeinsamkeit die Eigenschaft aufweisen, dass es sich um informierende Texte handelt.[62] Zwei dieser Knoten (D 26, D 564) besitzen ein gemeinsames Thema, sie erläutern die geschichtlichen Hintergründe der Universitätssiegel der Hochschulen in Trier und Bochum und beide Dokumente sind über einen grafisch realisierten Hyperlinkanzeiger (Abbildungen der Siegel) in der Einstiegsseite zu erreichen. Während D 26 ausschließlich die lateinischen Inschriften des Siegels erläutert, stellt D 564 Hyperlinks zu zwei Dateien bereit, die das Siegel in digitaler Form enthalten und geht auf den Bezug der abgebildeten historischen Figuren (Prometheus und Epimetheus) zur Ruhr-Universität Bochum ein: „Die moderne Wissenschaft trachtet danach, das Prometheische mit dem Epimetheischen zu verbinden. Das Emblem macht auch sichtbar, daß die [RUB] zu den Hochschulen gehört, in denen alle Wissenschaften miteinander im Gespräch sind." (D 564). Somit dient zumindest dieser Knoten auch der Profilbildung und besitzt in gewisser Hinsicht eine werbende Funktion. Darüber hinaus stellt er mit den digitalen Versionen des Siegels Ressourcen bereit, die Angehörige der Hochschule z. B. in Präsentationen oder Briefbögen einsetzen können. Die in den beiden Einstiegsseiten gegebene Verknüpfung kann dem Rezipienten jedoch erst dann deutlich werden, wenn der Mauszeiger über das Siegel bewegt wird. Die jeweiligen Zielknoten besitzen also für die Produzenten nur einen eher geringen Stellenwert und stellen keinen zentralen Bestandteil der Webauftritte dar. Der dritte Textknoten mit generischer Informationsfunktion (D 484) ist über den in der Einstiegsseite des Webauftritts der Universität Essen enthaltenen Hyperlinkanzeiger „Vom Campus zur City: Das neue Universitätsviertel – ein Nutzungsbeispiel" erreichbar (vgl. Abbildung 11.2, S. 476). Der Text erläutert Nutzungsmöglichkeiten des ehemaligen Essener Großmarktgeländes, das „eine außergewöhnliche Chance für eine zukünftige städtebauliche Gestaltung" bietet. Die Universität Essen stellt dar, dass sie als „größter Anrainer dieses Areals" bestrebt ist, an der Gestaltung dieses Gebiets mitzuwirken und schlägt „Universitätsviertel" als dessen neuen Namen vor. Weiterhin skizziert das Dokument ein Nutzungsszenario, in dessen Zentrum sich eine neu zu errichtende Bibliothek befindet. Der Knoten enthält weder kontexualisierende Hyperlinks zu weiteren Informationsangeboten noch Kontaktinformationen, weshalb eine Einschätzung seiner kommunikativen Funktion nur näherungsweise erfolgen

[62] Nach Heinemann und Viehweger (1991, S. 151) kann man der informierenden Funktion „die weitaus größte Zahl aller Textvorkommen [zuordnen], so daß diese Grundfunktion lange Zeit mit dem Kommunizieren schlechthin [...] identifiziert wurde."

kann. Der Text besitzt zunächst insbesondere eine informierende Funktion, weiterhin könnte es plausibel erscheinen, dass sich dieses Dokument vornehmlich an Vertreter der Wirtschaft (als mögliche Sponsoren) oder der Verwaltung der Stadt Essen richtet, um eventuell bereits auf anderem Wege publizierte Absichtserklärungen zu unterstreichen und das Vorhaben auch im WWW zu veröffentlichen. Somit besäße auch dieser Knoten eine werbende Funktion, die in erster Linie die Profilierung und Imagebildung der Institution betrifft.

Die zweite Hypertextknotensorte mit einem eher generellen Geltungsbereich betrifft die *Kerninformationen für eine spezifische Zielgruppe* (neun Dokumente). Sämtliche Exemplare dieser Hypertextknotensorte stellen Inhaltsknoten dar und enthalten spezifische Kerninformationen, die für eine bestimmte Zielgruppe von hoher Relevanz sind. Sechs dieser Knoten stammen aus dem Webauftritt der Universität Koblenz-Landau und erläutern den „Hochschulzugang mit Abitur“ (D 277), den „Hochschulzugang ohne Abitur“ (D 280), „Weiterbildende Studiengänge“ (D 287), „Weiterbildungsangebote“ (D 288) und „Internationale Programme“ (D 294, D 295). Während nahezu alle dieser Dokumente die wesentlichen Schritte in knapper Listenform erläutern, die der Rezipient zur Erreichung eines spezifischen Ziels durchführen muss, können D 287 und D 288 auch als eine Variante des Verteilers aufgefasst werden. Beispielsweise listet D 287 sechs weiterbildende Studiengänge auf, die mit jeweils ein bis zwei Sätzen erläutert werden, woraufhin durch die Verfolgung eines Hyperlinks „Nähere Informationen“ von dem Webauftritt derjenigen Organisationseinheit bzw. Institution eingeholt werden können, die für den jeweiligen Studiengang verantwortlich ist. D 254 stellt ebenfalls in knapper Form die Möglichkeit zur „Promotion und Habilitation“ an der European Business School vor. Für Interessenten aus der Zielgruppe der Absolventen werden der Name und die E-Mail-Adresse einer Mitarbeiterin aufgeführt, an die Anfragen gerichtet werden können. Das ebenfalls von der EBS angebotene Dokument D 257 richtet sich an Unternehmen, denen Kooperationsmöglichkeiten aufgezeigt werden, um z. B. eine Firma an der Hochschule zu präsentieren oder Praktikanten aus der Gruppe der Studierenden zu rekrutieren; auch dieser Text enthält Kontaktinformationen. Auf ein ähnliches Thema geht D 337 ein, das die unterschiedlichen Kooperationen der HdK Berlin in äußerst knapper Listenform aufzählt. Die kommunikative Funktion dieses Knotens kann nur durch seine URL erschlossen werden: `http://www.hdk-berlin.de/marketing.html`. Insbesondere die von der Universität Koblenz-Landau angebotenen Vertreter von Kerninformationen für eine spezifische Zielgruppe können zwar für die Gesamtpopulation dieser Stichprobe als sehr untypisch bezeichnet werden, sie verdeutlichen jedoch eine methodologische Problematik, für deren Lösung in nachfolgenden Analysen weitere Stichproben untersucht werden müssten. Schließlich besteht durchaus die Möglichkeit, dass diese Dokumente einen konventionalisierten Bestandteil der eingebetteten Hypertextsorte *Webangebot zum Thema Studium* darstellen.[63] Es wäre nun also notwendig, für sämtliche in Tabelle 11.10 aufgeführten Hypertextknotensorten des Typs *Einstiegsseite* korrespondierende Hypertextexemplare zu lokalisieren und diese einer Analyse zu unterziehen, um ihre konventionalisierten Konstituenten zu ermitteln.

[63] Obwohl sie in Tabelle 11.10 nicht aufgeführt wird, existieren in der Stichprobe zumindest Einstiegsseiten dieser Hypertextsorte. Da jedoch keine Unterschiede zwischen der *Einstiegsseite des Webangebots zum Thema Studium* und der *Einstiegsseite eines zielgruppenspezifischen Webangebots* ermittelt werden konnten, das sich an *Studierende* richtet, wurde die erstgenannte Hypertextknotensorte unter der zuletzt genannten subsumiert.

11.5.7 Konventionen bezüglich der Adressierung der HTML-Dokumente

Ein spezifischer Aspekt, der ebenfalls Konventionen unterliegen kann, wurde bislang noch nicht betrachtet. Dabei handelt es sich um die Dateinamen, die Produzenten einzelnen HTML-Dokumenten geben und die gleichzeitig als Teil der URL einer Webseite fungieren. Eine Betrachtung der ermittelten Hypertextknotensorten sowie der URLs zugehöriger Dokumente zeigt, dass bezüglich der Benennung von HTML-Dateien vieler Hypertextknotensorten unterschiedliche Konventionalisierungsgrade existieren; Tabelle 11.15 stellt Beispiele für 16 Hypertextknotensorten bzw. Hypertextknotensortenvarianten dar.[64]

Sämtliche Exemplare der Hypertextknotensorte *Impressum* heißen `impressum.html` oder besitzen den Namen einer Default-Datei und können somit über die Angabe des Verzeichnisses `.../impressum/` adressiert werden (vgl. Fußnote 47, S. 502).[65] Während bezüglich dieser Hypertextknotensorte aufgrund ihres Stellenwerts im Printbereich ein sehr hoher Grad der Konventionalisierung vorliegt, können in anderen Kategorien, die kein unmittelbares Pendant in den gedruckten Texten besitzen, nur partielle Übereinstimmungen ermittelt werden, die sich aber in nahezu allen Fällen auf einen gemeinsamen semantischen Kern beziehen: Die Datei- bzw. Verzeichnisnamen der Instanzen der Hypertextknotensorte *Suchformular* lauten unter anderem `suche.html`, `such.htm`, `/suche/`, `/Suchen/`, `/suchen_finden/`, `suchdienste.html`, `query.htm`, `/search/` und `search.htm`. Eine in Bezug auf die lexikalische Semantik der verwendeten Konzepte abstraktere Konvention kann hinsichtlich der *Einstiegsseite der universitären Selbstdarstellung* beobachtet werden, die Datei- bzw. Verzeichnisnamen wie z. B. `/wir_ueber_uns/`, `/allgemeines/`, `zahludat.html`, `uebersicht.html` und `ueberblick.html` umfasst. Es wurde bereits betont, dass nahezu alle hier untersuchten universitären Webauftritte primär nach der individuellen Strukturierung der jeweiligen Institution in individuelle Organisationseinheiten konzipiert werden. In einigen Fällen finden sich die spezifischen Abkürzungen zentraler Dezernate bzw. Referate auch in den URLs ihrer Hypertexte wieder (z. B. `http://www.uni-koblenz-landau.de/forschung/refa1start.html`). Derartige Adressen beeinträchtigen die Kohärenzbildung (vgl. Fußnote 88, S. 97) und können – zumindest in Bezug auf diese Stichprobe – als untypisch bezeichnet werden.

11.5.8 Zur Aggregierung verwandter Hypertextsortenmodule

Die Stichprobe enthält zahlreiche Dokumente, die die Notwendigkeit der Einführung der konzeptuellen Ebene der Hypertextsortenmodule und zugleich die Flexibilität dieses Ansatzes veranschaulichen (vgl. Kapitel 5). Die Instanzen beziehen sich insbesondere auf die Typen der Hyperlinkliste sowie Inhaltsknoten. Der vollständige Text einer *Pressemitteilung* kann z. B. – als Instanz eines Hypertextsortenmoduls – ein Bestandteil der *Einstiegsseite eines*

[64] Aus Darstellungsgründen wurden bei einigen der in Tabelle 11.15 enthaltenen URLs redundante Dateinamen wie `index.html` und `welcome.html` entfernt (vgl. Fußnote 47, S. 502).

[65] Einige der HTML-Dokumente besitzen jedoch plattformspezifische Suffixe wie z. B. `.htm` oder `.shtml`. Der erste dieser Suffixe deutet darauf hin, dass die Datei auf einer *MS Windows*-Plattform erstellt wurde. Der zweite Suffix findet ausschließlich auf *Apache*-Webservern Verwendung. Entsprechende Dokumente können Anweisungen enthalten, die zur Laufzeit, d. h. bei der Auslieferung des Dokuments berechnet werden, um z. B. das Datum der letzten Änderung, die aktuelle Uhrzeit oder eine als isoliertes HTML-Fragment gepflegte Navigationshilfe in das Dokument zu integrieren (vgl. auch Fußnote 38, S. 345).

Benutzungshinweise/Anleitungen (Gesamtvorkommen: 14): http://www.uni-koblenz-landau.de/hilfe/hilfe.html (D 266) http://www.tu-darmstadt.de/hilfe.html (D 490) http://www.uni-konstanz.de/struktur/hilfe/hilfe.html (D 674)
Copyright-Informationen (Gesamtvorkommen: 3): http://www.tu-cottbus.de/BTU/copyr.html (D 66) http://www.uni-augsburg.de/copyright.shtml (D 195) http://www.fu-berlin.de/redaktion/copyright/ (D 323)
Einstiegsseite der universitären Selbstdarstellung (Gesamtvorkommen: 21): http://www.uni-ulm.de/wir_ueber_uns/ (D 175) http://www.ebs.de/wir_ueber_uns/ (D 241) http://www.uni-augsburg.de/allgemeines/ (D 190) http://www.uni-osnabrueck.de/allgemein/zahludat.html (D 513) http://www.uni-osnabrueck.de/praesident/uebersicht.html (D 511) http://www.uni-muenster.de/de/ueberblick.html (D 585) http://www.uni-dortmund.de/TOP/ueberblick.html (D 629) http://www.verwaltung.uni-bonn.de/ueberbl/ueberblk.htm (D 647)
Einstiegsseite eines Vorlesungsverzeichnisses (Gesamtvorkommen: 5): http://www.tu-chemnitz.de/verwaltung/dez1/vlvz/ (D 88) http://www.fu-berlin.de/vv/ (D 304) http://www.hdk-berlin.de/studium/vv.html (D 350) http://www.uni-mannheim.de/uni/vlz.html (D 718)
Einstiegsseite des Webangebots mit Stellenanzeigen (Gesamtvorkommen: 7): http://www.tu-chemnitz.de/verwaltung/dez2/stellen/lststell.htm (D 87) http://www.tu-dresden.de/vd72/stellaus/stellaus.htm (D 366) http://www.uni-essen.de/stellen/ (D 482) http://www.uni-passau.de/jobs/ (D 448) http://personal.verwaltung.uni-mannheim.de/htms/jobs.htm (D 725)
Einstiegsseite des Webangebots zum Thema Forschung/Lehre (Gesamtvorkommen: 8): http://www.uni-koblenz-landau.de/forschung/refa1start.html (D 270) http://www.hdk-berlin.de/forsch/ (D 332) http://www.rwth-aachen.de/zentral/portal_forschung.html (D 392) http://www.uni-dortmund.de/TOP/forschung.html (D 627) http://www.uni-hohenheim.de/forschung/ (D 697) http://www.uni-konstanz.de/struktur/forsch/forsch.html (D 670)
Einstiegsseite des Webauftritts einer zentralen Organisationseinheit (Pressestelle; Gesamtvorkommen: 8): http://www.uni-heidelberg.de/presse/ (D 230) http://www.fu-berlin.de/presse/ (D 321) http://www.hdk-berlin.de/presse/ (D 336) http://www.uni-kassel.de/presse/ (D 555) http://www.ruhr-uni-bochum.de/pressestelle/ (D 572) http://www.uni-ulm.de/uni/leitung/pressestelle.html (D 180) http://www.verwaltung.uni-bonn.de/presse/presse.htm (D 643)
Einstiegsseite eines zielgruppenspezifischen Webangebots (Studierende; Gesamtvorkommen: 14): http://www.tu-harburg.de/studium/ (D 200) http://www.ebs.de/Studium/ (D 252) http://www.uni-kassel.de/studium/studium.ghk (D 551) http://www.verwaltung.uni-bonn.de/zsb/studium.htm (D 651) http://www.uni-konstanz.de/struktur/studium/studium.html (D 669) http://www.fu-berlin.de/studierende/ (D 318) http://www.uni-mannheim.de/uni/studierende.html (D 712) http://www.rwth-aachen.de/zentral/portal_studierende.html (D 389)
Einstiegsseite eines zielgruppenspezifischen Webangebots (Wirtschaft; Gesamtvorkommen: 4): http://www.fu-berlin.de/wirtschaft/ (D 320) http://www.rwth-aachen.de/zentral/portal_wirtschaft.html (D 391) http://www.uni-essen.de/portale/wirtschaft.html (D 465) http://www.roew.uni-dortmund.de/wirtschaft/seite-n.htm (D 634)
Suchformular (Gesamtvorkommen: 21): http://www.uni-leipzig.de/such.htm (D 129) http://www.ebs.de/Suchen/ (D 239) http://www.uni-ulm.de/suchen_finden/ (D 132) http://www.uni-heidelberg.de/suche.html (D 219) http://www.uni-hohenheim.de/suche/ (D 700) http://www.uni-koblenz.de/suche.html (D 267) http://www.uni-konstanz.de/struktur/suche/suche.html (D 672) http://www.rwth-aachen.de/zentral/uuw_suchen.htm (D 372) http://www.uni-dortmund.de/TOP/suchen.html (D 632) http://www.verwaltung.uni-bonn.de/suchen.htm (D 658) http://www.uni-mannheim.de/uni/suchdienste.html (D 710) http://www.mu-luebeck.de/suche/query.htm (D 624) http://www.hdk-berlin.de/search/ (D 340) http://www.tu-dresden.de/search.htm (D 367) http://www.rz.uni-passau.de/search.php3 (D 452) http://www.uni-kassel.de/search/ (D 559)
Impressum (Gesamtvorkommen: 9): http://www.uni-ulm.de/impressum/ (D 184) http://www.fu-berlin.de/redaktion/impressum/ (D 316) http://www.uni-paderborn.de/home/impressum/ (D 434) http://www.mu-luebeck.de/impressum.htm (D 623) http://www.tu-chemnitz.de/tu/impressum.html (D 110) http://www.tu-harburg.de/impressum.html (D 208) http://www.hdk-berlin.de/impressum.html (D 345) http://www.uni-dortmund.de/TOP/impressum.html (D 638) http://www.uni-augsburg.de/impressum.shtml (D 194)
Sitemap (Gesamtvorkommen: 2): http://www.tu-harburg.de/sitemap_d.html (D 209) http://www.uni-heidelberg.de/sitemap.html (D 220)
Kontaktinformationen (Gesamtvorkommen: 17): http://www.tu-cottbus.de/BTU/kontakt.html (D 57) http://www.uni-heidelberg.de/kontakt.html (D 218) http://www.uni-koblenz-landau.de/kontakt/kontakt.html (D 265) http://www.uni-dortmund.de/TOP/kontakt.html (D 637) http://www.uni-mannheim.de/uni/kontakt.html (D 703)
Verteiler (Kontaktinformationen, Adressen, Ansprechpartner, Lagepläne; Gesamtvorkommen: 10): http://www.fu-berlin.de/kontakt/ (D 298) http://www.uni-essen.de/service/kontakt.html (D 469) http://www.uni-muenster.de/de/kontakt.html (D 594) http://www.uni-kassel.de/presse/kontakte.ghk (D 558) http://www.uni-leipzig.de/kontakte/ (D 117)
Verteiler (zu aktuellen Informationen; Gesamtvorkommen: 12): http://www.uni-augsburg.de/aktuell/ (D 191) http://www.uni-osnabrueck.de/aktuell/ (D 515) http://www.uni-mannheim.de/uni/aktuell.html (D 705) http://www.uni-ulm.de/aktuelles/ (D 134) http://www.uni-essen.de/service/aktuelles.html (D 468) http://www.uni-dortmund.de/TOP/aktuelles.html (D 630)
Verteiler (zum Thema Studium; Gesamtvorkommen: 15): http://www.uni-heidelberg.de/studium/ (D 225) http://www.fu-berlin.de/studium/ (D 301) http://www.uni-hohenheim.de/studium/ (D 696) http://www.uni-paderborn.de/home/studium/ (D 433) http://www.uni-trier.de/uni/studium.htm (D 32) http://www.ku-eichstaett.de/studium.htm (D 16) http://www.uni-giessen.de/jlug/studierende/studium.htm (D 51) http://www.tu-cottbus.de/BTU/studium.html (D 61) http://www.uni-essen.de/studium/studium.html (D 457) http://www.uni-muenster.de/de/studium.html (D 587)

Tabelle 11.15: Konventionen hinsichtlich der Dateinamen einzelner Dokumente

universitären Webauftritts sein (vgl. E 20 in Abbildung 11.2, S. 476), das Hypertextsortenmodul kann jedoch ebenfalls als eigenständige Hypertextknotensorte fungieren, wie die 17 korrespondierenden Vorkommen zeigen.

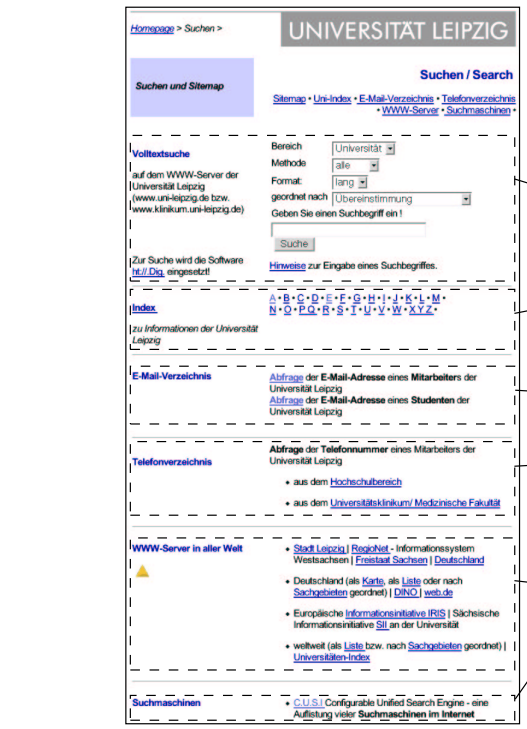
Abbildung 11.7 zeigt einige Beispiele für die Hypertextknotensorte *Suchformular*, die einerseits verdeutlichen, dass die Produzenten Instanzen von Hypertextsortenmodulen, die eine ähnlich gelagerte Typmarkierung besitzen (vgl. Abschnitt 5.6.4), oftmals in einem Einzelknoten aggregieren, andererseits wird erneut die Problematik der Zuweisung eines *eindeutigen* Hypertextknotensortenetiketts ersichtlich.[66] Bei vielen in der Stichprobe enthaltenen Knoten handelt es sich zwar um funktional heterogene Entitäten, die Kategorisierung in eine Hypertextknotensorte kann jedoch in nahezu allen Fällen eindeutig durchgeführt werden. Während z. B. im Hinblick auf D 632 und D 672 zweifelsohne die Hypertextknotensorte *Suchformular* vorliegt, könnte D 129 durchaus auch einer anderen, abstrakteren Hypertextknotensorte zugeordnet werden. In diesem Fall wurde *Suchformular* gewählt, da die Instanz des korrespondierenden Hypertextsortenmoduls die Initialposition innerhalb dieses Knotens aufweist und sowohl die Überschrift als auch die URL den funktionalen Aspekt der Suche fokussieren.[67] Die angesprochene Aggregierung der Instanzen von Hypertextsortenmodulen mit verwandten Typausprägungen verdeutlichen alle drei in Abbildung 11.7 dargestellten HTML-Dokumente: D 632 und D 672 umfassen jeweils eine Instanz der Hypertextsortenmodule *Suchformular* und *Verteiler*, wobei die Exemplare des zuletzt genannten Hypertextsortenmoduls unterschiedliche kommunikativ-funktionale Ausrichtungen der Knoten bewirken. D 632 enthält einen Verteiler, der verschiedenartige Informationen über die Universität Dortmund zur Verfügung stellt. D 672 beinhaltet hingegen einen Verteiler, der vornehmlich auf Kontaktinformationen verweist. D 129 umfasst Instanzen mehrerer Hypertextsortenmodule, die unter anderem den Zugriff auf die Hypertextbasis (*Suchformular*, *Index bzw. Schlagwortverzeichnis*) und Recherchen in E-Mail- und Telefonverzeichnissen der Hochschule erlauben und zugleich Hyperlinks zu externen Angeboten bereitstellen.

Abbildung 11.8 stellt abschließend zwei weitere Beispiele dar: D 275 ist die Einstiegsseite in einen Hypertext, der das Vorlesungsverzeichnis der Universität Koblenz-Landau enthält. Im oberen Drittel des Dokuments werden unterschiedliche Möglichkeiten des Zugriffs erläutert. In den unteren Teil hat der Produzent eine Instanz des Hypertextsortenmoduls *Daten und Fristen eines Semesters* integriert, das eine inhaltlich-thematische Verwandtschaft zu der Hypertextknotensorte *Einstiegsseite des Vorlesungsverzeichnisses* aufweist und in zwei Dokumenten der Stichprobe als eigenständige Hypertextknotensorte fungiert (D 281, D 626).[68] Im Gegensatz zu den bislang angesprochenen Beispielen liegt in D 275 das Spezifikum vor, dass diese Hypertextknotensorte aufgrund ihres Status als Einstiegsseite nicht als Hypertext-

[66] In den insgesamt fünf Bildschirmabzügen, die in den Abbildungen 11.7 und 11.8 enthalten sind, wurden einige Elemente des Peritextes entfernt, um die Lesbarkeit der Knoten gewährleisten zu können.

[67] Es fällt auf, dass D 129 die Überschrift „Suchen und Sitemap“ enthält, jedoch ist die Sitemap lediglich über einen Hyprlink (rechts der Überschrift) zu erreichen. In D 672 sind darüber hinaus drei Fehler enthalten, die sich auf die Überschrift des Verteilers („Weiter Suchmöglichkeiten“), einen Orthografiefehler in „Suchrobotor“ und den Kommentar des zweiten Hyperlinks beziehen, der einen syntaktischen Kongruenzfehler enthält.

[68] Ein Bildschirmabzug von D 281 ist in Abbildung 11.6 (S. 509) enthalten. D 281 wird ebenfalls von der Universität Koblenz-Landau angeboten und verdeutlicht, dass die Produzenten zentraler Angebote nicht immer dem Hypertextprinzip folgen, die Präsentation und Pflege redundanter informationeller Einheiten in mehreren Knoten zu vermeiden und stattdessen lediglich einen Hyperlink auf einen relevanten Knoten anzulegen.



D 129
`http://www.uni-leipzig.de/such.htm`
Hypertextknotensorte: Suchformular

Suchformular

Index bzw. Schlagwortverzeichnis

Kontaktinformationen (E-Mail)

Kontaktinformationen (Telefon)

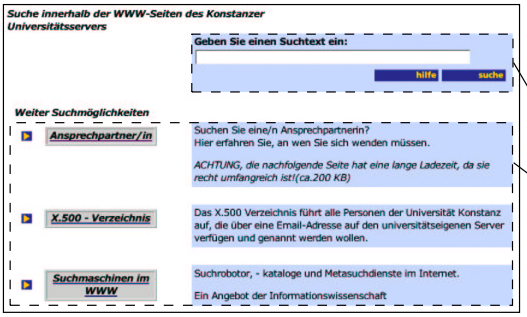
Hotlist (Region, regionale Angebote)

Hotlist (Suchmaschinen, Informationsrecherche)

D 632
`http://www.uni-dortmund.de/TOP/suchen.htm`
Hypertextknotensorte: Suchformular

Verteiler
(zu einem spezifischen Thema: Informationen über die Universität)

Suchformular

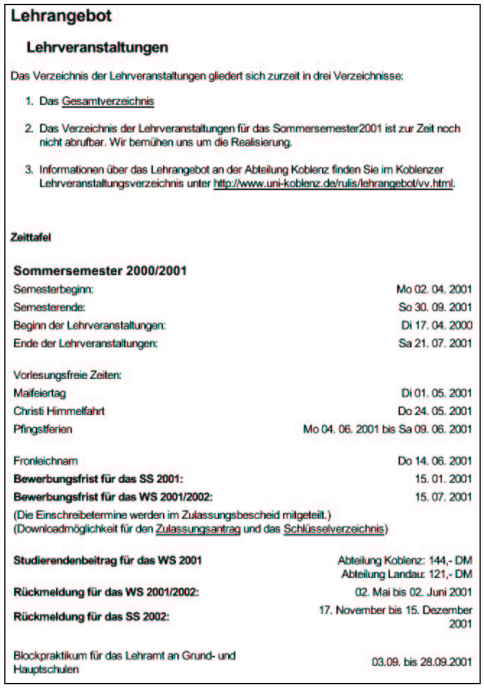


D 672
`http://www.uni-konstanz.de/struktur/suche/suche.html`
Hypertextknotensorte: Suchformular

Suchformular

Verteiler
(Kontaktinformationen, Adressen, Ansprechpartner, Lagepläne)

Abbildung 11.7: Beispiele für die Aggregierung von Hypertextsortenmodulen (Teil 1)



Lehrangebot

Lehrveranstaltungen

Das Verzeichnis der Lehrveranstaltungen gliedert sich zurzeit in drei Verzeichnisse:

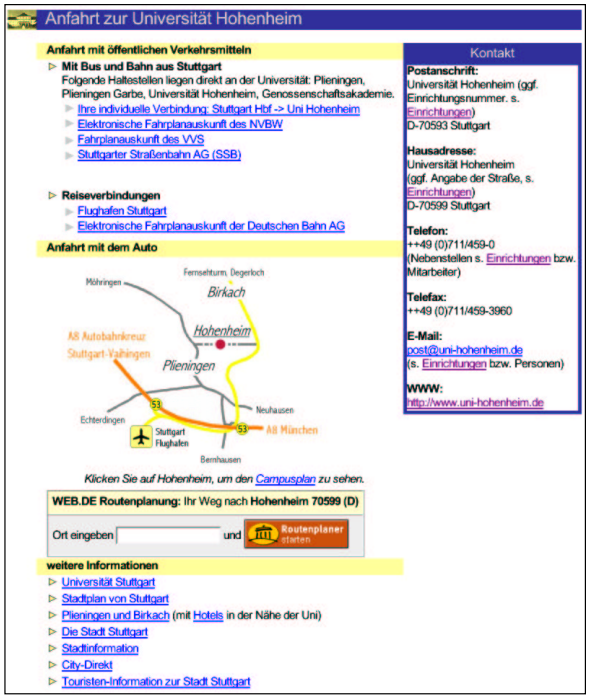
1. Das Gesamtverzeichnis
2. Das Verzeichnis der Lehrveranstaltungen für das Sommersemester2001 ist zur Zeit noch nicht abrufbar. Wir bemühen uns um die Realisierung.
3. Informationen über das Lehrangebot an der Abteilung Koblenz finden Sie im Koblenzer Lehrveranstaltungsverzeichnis unter http://www.uni-koblenz.de/rulis/lehrangebot/vv.html.

Zeittafel

Sommersemester 2000/2001

Semesterbeginn:	Mo 02. 04. 2001
Semesterende:	So 30. 09. 2001
Beginn der Lehrveranstaltungen:	Di 17. 04. 2000
Ende der Lehrveranstaltungen:	Sa 21. 07. 2001
Vorlesungsfreie Zeiten:	
Maifeiertag	Di 01. 05. 2001
Christi Himmelfahrt	Do 24. 05. 2001
Pfingstferien	Mo 04. 06. 2001 bis Sa 09. 06. 2001
Fronleichnam	Do 14. 06. 2001
Bewerbungsfrist für das SS 2001:	15. 01. 2001
Bewerbungsfrist für das WS 2001/2002:	15. 07. 2001
(Die Einschreibetermine werden im Zulassungsbescheid mitgeteilt.) (Downloadmöglichkeit für den Zulassungsantrag und das Schlüsselverzeichnis)	
Studierendenbeitrag für das WS 2001	Abteilung Koblenz: 144,- DM Abteilung Landau: 121,- DM
Rückmeldung für das WS 2001/2002:	02. Mai bis 02. Juni 2001
Rückmeldung für das SS 2002:	17. November bis 15. Dezember 2001
Blockpraktikum für das Lehramt an Grund- und Hauptschulen	03.09. bis 28.09.2001

D 275 – Hypertextknotensorte:
Einstiegsseite des Vorlesungsverzeichnisses



Anfahrt zur Universität Hohenheim

Anfahrt mit öffentlichen Verkehrsmitteln

Mit Bus und Bahn aus Stuttgart
Folgende Haltestellen liegen direkt an der Universität: Plieningen, Plieningen Garbe, Universität Hohenheim, Genossenschaftsakademie.
- Ihre individuelle Verbindung: Stuttgart Hbf -> Uni Hohenheim
- Elektronische Fahrplanauskunft des NVBW
- Fahrplanauskunft des VVS
- Stuttgarter Straßenbahn AG (SSB)

Reiseverbindungen
- Flughafen Stuttgart
- Elektronische Fahrplanauskunft der Deutschen Bahn AG

Anfahrt mit dem Auto

Möhringen, Fernsehturm, Degerloch, Birkach, A8 Autobahnkreuz Stuttgart-Vaihingen, Hohenheim, Plieningen, Echterdingen, Neuhausen, Stuttgart Flughafen, A8 München, Bernhausen

Klicken Sie auf Hohenheim, um den Campusplan zu sehen.

WEB.DE Routenplanung: Ihr Weg nach **Hohenheim 70599 (D)**

Ort eingeben und Routenplaner starten

weitere Informationen
- Universität Stuttgart
- Stadtplan von Stuttgart
- Plieningen und Birkach (mit Hotels in der Nähe der Uni)
- Die Stadt Stuttgart
- Stadtinformation
- City-Direkt
- Touristen-Information zur Stadt Stuttgart

Kontakt

Postanschrift:
Universität Hohenheim (ggf. Einrichtungsnummer. s. Einrichtungen)
D-70593 Stuttgart

Hausadresse:
Universität Hohenheim
(ggf. Angabe der Straße, s. Einrichtungen)
D-70599 Stuttgart

Telefon:
++49 (0)711/459-0
(Nebenstellen s. Einrichtungen bzw. Mitarbeiter)

Telefax:
++49 (0)711/459-3960

E-Mail:
post@uni-hohenheim.de
(s. Einrichtungen bzw. Personen)

WWW:
http://www.uni-hohenheim.de

D 694 – Hypertextknotensorte:
Lageplan, Karte, Wegbeschreibung

Abbildung 11.8: Beispiele für die Aggregierung von Hypertextsortenmodulen (Teil 2)

sortenmodul fungieren kann. D 694 kann als eine typische Instanz der Hypertextknotensorte *Lageplan, Karte, Wegbeschreibung* betrachtet werden. Die Etikettierung deutet bereits an, dass diese Hypertextknotensorte nicht auf einer spezifischen Definition basiert. Vielmehr umfassen ihre Exemplare oftmals Instanzen der Hypertextsortenmodule *Lageplan*, *Karte* und *Webbeschreibung*, die – in diesem Beispiel kombiniert mit einer Instanz des Hypertextsortenmoduls *Kontaktinformationen* – in der Form eines Einzelknotens aggregiert werden, der auf Seiten des Rezipienten unterschiedliche Informationsbedarfe im Hinblick auf die Kontaktaufnahme oder Anreise abdecken kann.[69] Da diese Hypertextknotensorte keine Entsprechung im Bereich der traditionellen Textsorten besitzt, wurde das genannte Etikett gewählt, das einen aufzählenden Charakter besitzt.

11.5.9 Begrüßungstexte in Instanzen von Hypertextknotensorten

Abschnitt 9.6.1 beschäftigt sich mit E-Mail-ähnlichen Begrüßungstexten in Exemplaren der Hypertextsorte *private Homepage eines Studierenden*, die als Instanzen eines Hypertextsortenmoduls aufgefasst wurden, das dem primären Typ Textstrukturmuster zugehörig ist. Interessanterweise enthalten 11 Dokumente der hier untersuchten Stichprobe Textstrukturmus-

[69] D 694 enthält zusätzlich am unteren Rand die Instanz einer *Hotlist*, die Hyperlinks zu externen Webangeboten zur Verfügung stellt, die spezifischere Informationen zur Region Stuttgart umfassen.

ter, die den zuvor beschriebenen Mustern bei initialer Betrachtung durchaus ähneln, jedoch unterschiedliche Ursprünge, Ausprägungen und kommunikative Funktionen besitzen.

Abbildung 11.9 stellt die 11 Dokumente im Überblick dar.[70] Ein markanter Unterschied wird in D 242 und D 455 deutlich: Diese Knoten enthalten jeweils ein Grußwort, dessen Textstrukturmuster dem eines Briefes ähnelt.[71] Während jedoch D 242 aufgrund des Umstandes, dass das *Grußwort* den einzigen Inhalt des Knotens darstellt, eben dieser Hypertextknotensorte zugeordnet wurde, ist D 455 ein größerer Bestandteil der Einstiegsseite des Webauftritts der Universität Essen und somit als Instanz des Hypertextsortenmoduls *Grußwort* aufzufassen. D 719 stellt innerhalb der 11 Dokumente das Beispiel mit dem geringsten Umfang dar: Die Rezipienten werden begrüßt, und es folgt eine prophylaktische Beschwichtigungsfloskel („[...] es wird sich demnächst noch mehr tun."). Durch die informelle Begrüßung, eine umgangssprachliche Floskel und den Einsatz eines Smileys weist D 719 darüber hinaus eine Nähe zur konzeptionellen Mündlichkeit auf (vgl. Abschnitt 2.2.7), die an den Einsatz E-Mail-ähnlicher Begrüßungstexte in studentischen Homepages erinnert (vgl. Abschnitt 9.6.1). Tatsächlich ist D 719 die Einstiegsseite des Webauftritts des AStA der Universität Mannheim.[72] Studierende beschränken den Einsatz konzeptionell mündlicher Elemente also nicht auf ihre privaten Homepages, sondern verwenden diese auch in Webangeboten mit offiziellerem Charakter, um eine authentische Atmosphäre zu erzeugen.

Die verbleibenden Knoten stellen Begrüßungstexte dar, die den Inhalt eines Webangebots zusammenfassen, den Rezipieten zum Verweilen einladen und um Rückmeldungen bitten. Alle Vorkommen sind Bestandteile von Einstiegsseiten. Sie leiten die Webauftritte zentraler Organisationseinheiten (D 102: Sportzentrum, D 123: Bibliothek, D 445: Transferzentrum, D 648: akademisches Auslandsamt), ein zielgruppenspezifisches Angebot (D 377: für Studierende), die Webpräsenz eines universitätsinternen Projekts (D 369: Thema Umweltschutz), eine universitäre Selbstdarstellung (D 629) und die Webpräsenz eines Weiterbildungsangebots ein (D 683). Tabelle 11.16 stellt unter anderem den Einsatz von Konstituenten der Textstrukturmuster Brief und E-Mail in den 11 Dokumenten dar: Bis auf D 242 und D 719 gehen sämtliche Texte in unterschiedlichen Graden der Ausführlichkeit auf den Inhalt des jeweiligen Hypertextes ein. Das Grußwort einer Rektorin (D 455) thematisiert diesen Aspekt nur in knapper Weise: „Diese vielfältigen Aufgaben darzustellen ist eine der Aufgaben unseres neugestalteten Webauftritts. Die folgenden Webseiten sollen Ihnen unter anderem einen Überblick über die wesentlichen Vorzüge der Universität Essen vermitteln." Die Positionierung des Grußworts in der Einstiegsseite des Webauftritts dieser Hochschule bestimmt den Skopus dieser Äußerung, die sich auf die *gesamte* Webpräsenz der Universität Essen bezieht.

[70] Die Bildschirmabzüge beschränken sich auf die Vorkommen der Textstrukturmuster in den jeweiligen Dokumenten, d. h. Navigationshilfen und sonstige Inhalte wurden entfernt.

[71] Abschnitt 2.3.6 geht ausführlich auf diese Textsorte ein. Sandig (2000) ermittelt das „Photo des Entbieters" als fakultative Konstituente des Grußwortes. Neben dem in D 455 enthaltenen Foto war auch in D 242 ein Porträt des Emittenten enthalten war. Da jedoch Bilder nicht in der Korpusdatenbank gespeichert werden und dieses Foto nicht mehr auf dem entfernten Webserver verfügbar ist, wurde die korrespondierende Platzhaltergrafik aus dem Bildschirmabzug entfernt. Das Foto befand sich ursprünglich zwischen Überschrift und dem Namen des Emittenten, weshalb die Überschrift nach unten verschoben wurde.

[72] D 719 wurde die Hypertextknotensorte *Einstiegsseite einer zentralen Organisationseinheit (Studierendenvertretung)* zugeordnet, um die Komplexität der Darstellung zu reduzieren. Auf der Grundlage des üblichen Sprachgebrauchs handelt es sich bei dem AStA streng genommen nicht um eine zentrale Organisationseinheit.

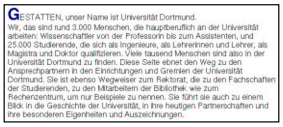
Liebe Mitglieder und Angehörige der Technischen Universität

Der Hochschulsport an unserer Universität ist als Breitensport angelegt und dient dem Ausgleich zur täglichen Studien- und Arbeitsbelastung sowie der aktiven Erholung, der Erhaltung der Gesundheit und körperlichen Fitness. Außerdem fördert er die Kommunikation zwischen den Sporttreibenden und ist Bestandteil des akademischen Lebens an unserer Universität.

Im Wintersemester 2000 / 2001 haben wir wieder viele traditionelle und neuere Sportarten im Angebot. Mit wenigen Ausnahmen (Fitnessraum, Tennis, Camps) sind die **Veranstaltungen des Hochschulsports kostenfrei.**

Mit dem vorliegenden Sportprogramm möchten wir allen Sportinteressenten den Hochschulsport näherbringen, sowohl Anfängern als auch Fortgeschrittenen und Können Informationen und Anregungen geben und die Auswahl für regelmäßige, freudvolle sportliche Betätigung erleichtern.

D 102



GESTATTEN, unser Name ist Universität Dortmund.
Wir, das sind rund 3.000 Menschen, die hauptberuflich an der Universität arbeiten: Wissenschaftler von der Professorin bis zum Assistenten, und 25.000 Studierende, die sich als Ingenieure, als Lehrerinnen und Lehrer, als Magistra und Doktor qualifizieren. Viele tausend Menschen sind also in der Universität Dortmund zu finden. Diese Seite ebnet den Weg zu den Ansprechpartnern in den Einrichtungen und Gremien der Universität Dortmund. Sie ist ebenso Wegweiser zum Rektorat, die zu den Fachschaften der Studierenden, zu den Mitarbeitern der Bibliothek wie zum Rechenzentrum, um nur Beispiele zu nennen. Sie führt sie auch zu einem Blick in die Geschichte der Universität, in ihre heutigen Partnerschaften und ihre besonderen Eigenheiten und Auszeichnungen.

D 629

Sehr geehrte Mitarbeiterinnen der TU Dresden,

im Rahmen unseres Öko-Audits sollen viele Maßnahmen zur umweltverträglichen Gestaltung aller Bereiche der TU Dresden umgesetzt werden. Für Verbesserungsvorschläge sind Sie die Expertinnen vor Ort und haben sicherlich bereits viele Ideen.

Auf dieser Seite wollen wir, die Arbeitsgruppe Öko-Audit, mit Ihrer Hilfe Ideen für den Umweltschutz in der Technischen Universität Dresden sammeln und möchten deshalb von Ihnen wissen:

Was könnte an Ihrem Arbeitsplatz, in Ihrem Arbeitsumfeld oder an der TU Dresden umweltgerechter gestaltet werden?

Dabei ist uns jeder ernst gemeinte Vorschlag wichtig. Denn kleine wie große Schritte bringen den Umweltschutz voran.
Deshalb werden unter **allen** Einsendungen kleine Preise verlost und außerdem erfolgt nach Abschluß der Ideenbörse durch eine Jury eine Prämierung der besten Ideen.

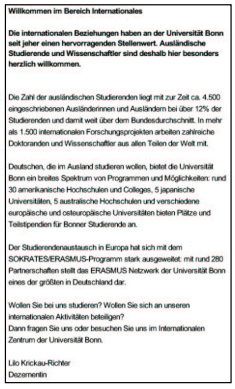
Mitmachen lohnt sich also in jedem Fall - für Sie genauso wie für die Umwelt!

Wir freuen uns auf Ihre Beiträge!

weiter ... hier erfahren Sie, wie Sie teilnehmen können.

Diese Seite wurde erstellt von J. Jagszent & J. Voß.

D 369



Willkommen im Bereich Internationales

Die internationalen Beziehungen haben an der Universität Bonn seit jeher einen hervorragenden Stellenwert. Ausländische Studierende und Wissenschaftler sind deshalb hier besonders herzlich willkommen.

Die Zahl der ausländischen Studierenden liegt mit zur Zeit ca. 4.500 eingeschriebenen Ausländerinnen und Ausländern bei über 12% der Studierenden und damit weit über dem Bundesdurchschnitt. In mehr als 1.500 internationalen Forschungsprojekten arbeiten zahlreiche Doktoranden und Wissenschaftler aus allen Teilen der Welt mit.

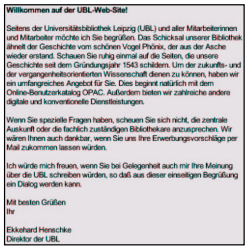
Deutschen, die im Ausland studieren wollen, bietet die Universität Bonn ein breites Spektrum von Programmen und Möglichkeiten: rund 30 amerikanische Hochschulen und Colleges, 5 japanische Universitäten, 5 australische Hochschulen und verschiedene europäische und osteuropäische Universitäten bieten Plätze und Teilstipendien für Bonner Studierende an.

Der Studierendenaustausch in Europa hat sich mit dem SOKRATES/ERASMUS-Programm stark ausgeweitet: mit rund 280 Partnerschaften stellt das ERASMUS Netzwerk der Universität Bonn eines der größten in Deutschland dar.

Wollen Sie bei uns studieren? Wollen Sie sich an unseren internationalen Aktivitäten beteiligen?
Dann fragen Sie uns oder besuchen Sie uns im Internationalen Zentrum der Universität Bonn.

Lilo Krickau-Richter
Dezernentin

D 648



Willkommen auf der UBL-Web-Site!

Seitens der Universitätsbibliothek Leipzig (UBL) und aller Mitarbeiterinnen und Mitarbeiter möchte ich Sie begrüßen. Das Schicksal unserer Bibliothek ähnelt der Geschichte vom schönen Vogel Phönix, der aus der Asche wieder erstand. Schauen Sie ruhig einmal auf die Seiten, die unsere Geschichte seit dem Gründungsjahr 1543 schildern. Um der zukunfts- und der vergangenheitsorientierten Wissenschaft dienen zu können, haben wir ein umfangreiches Angebot für Sie. Dies beginnt natürlich mit dem Online-Benutzerkatalog OPAC. Außerdem bieten wir zahlreiche andere digitale und konventionelle Dienstleistungen.

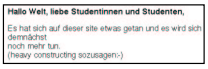
Wenn Sie spezielle Fragen haben, scheuen Sie sich nicht, die zentrale Auskunft oder die fachlich zuständigen Bibliothekare anzusprechen. Wir wären Ihnen auch dankbar, wenn Sie uns Ihre Erwerbungsvorschläge per Mail zukommen lassen würden.

Ich würde mich freuen, wenn Sie bei Gelegenheit auch mir Ihre Meinung über die UBL schreiben würden, so daß aus dieser einseitigen Begrüßung ein Dialog werden kann.

Mit besten Grüßen
Ihr

Ekkehard Henschke
Direktor der UBL

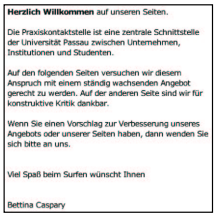
D 123



Hallo Welt, liebe Studentinnen und Studenten,

Es hat sich auf dieser site etwas getan und es wird sich demnächst noch mehr tun.
(heavy constructing sozusagen:-)

D 719



Herzlich Willkommen auf unseren Seiten.

Die Praxiskontaktstelle ist eine zentrale Schnittstelle der Universität Passau zwischen Unternehmen, Institutionen und Studenten.

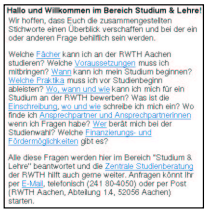
Auf den folgenden Seiten versuchen wir diesem Anspruch mit einem ständig wachsenden Angebot gerecht zu werden. Auf der anderen Seite sind wir für konstruktive Kritik dankbar.

Wenn Sie einen Vorschlag zur Verbesserung unseres Angebots oder unserer Seiten haben, dann wenden Sie sich bitte an uns.

Viel Spaß beim Surfen wünscht Ihnen

Bettina Caspary

D 445



Hallo und Willkommen im Bereich Studium & Lehre!
Wir hoffen, dass Euch die zusammengestellten Stichworte einen Überblick verschaffen und bei der ein oder anderen Frage behilflich sein werden.

Welche Fächer kann ich an der RWTH Aachen studieren? Welche Voraussetzungen muss ich mitbringen? Wann kann ich mein Studium beginnen? Welche Praktika muss ich vor Studienbeginn ableisten? Wo, wann und wie kann ich mich für ein Studium an der RWTH bewerben? Was ist die Einschreibung, wo und wie schreibe ich mich ein? Wo finde ich Ansprechpartner und Ansprechpartnerinnen wenn ich Fragen habe? Wer berät mich bei der Studienwahl? Welche Finanzierungs- und Fördermöglichkeiten gibt es?

Alle diese Fragen werden hier im Bereich "Studium & Lehre" beantwortet und die Zentrale Studienberatung der RWTH hilft auch gerne weiter. Anfragen könnt Ihr per E-Mail, telefonisch (241 80-4050) oder per Post (RWTH Aachen, Abteilung 14, 52056 Aachen) starten.

D 377

Grußwort
Prof. Dr. Dr. h.c. mult. Hans Tietmeyer
Präsident der ebs

Herzlich willkommen auf den Web-Seiten der EUROPEAN BUSINESS SCHOOL Schloß Reichartshausen!

Mit Blick auf das neue Jahrtausend hat auch die EUROPEAN BUSINESS SCHOOL (**ebs**) eine Vielzahl an Neuerungen in und um die Hochschule vorzuweisen und Planungen, die bereits seit einigen Jahren reiften, vollenden können.

An oberster Stelle ist hier die Studienstrukturreform zu nennen, die in einer neuen, durch das Hessische Ministerium für Wissenschaft und Kunst genehmigten Prüfungsordnung – und damit verbunden – in einer Verkürzung der Studiendauer von 10 auf 8 Semester mündete. Mit dieser Reform wird zukünftigen Studenten der **ebs** nicht nur ermöglicht, ein ganzes Jahr früher mit ihrer beruflichen Karriere zu beginnen, sondern auch ein positives Signal an die Wirtschaft gesendet: Ein Studium in Deutschland kann gleichzeitig theoretisch fundiert, praxisorientiert, international und zügig durchführbar sein.

Auch innerhalb des Lehrbetriebes wurden weitere richtungweisende Schritte unternommen. So konnte der im Wintersemester 99/2000 neu berufene Lehrstuhl für Investitions- und Risikomanagement unter Ulrich Hommel, Ph.D., A.M., zusammen mit dem seit Jahren in der Finanzwelt bekannten Stiftungslehrstuhl Bank- und Finanzmanagement unter Prof. Dr. Dr. Ann-Kristin Achleitner die **ebs** Finance Group ins Leben rufen und somit die traditionell starke Orientierung zum Finanzzentrum Frankfurt noch weiter ausbauen.

Der Lehrstuhl Immobilienökonomie unter Prof. Dr. Karl-Werner Schulte konnte seine internationalen Kontakte vertiefen und mit der Hanking University die erste Partneruniversität im asiatischen Raum gewinnen.

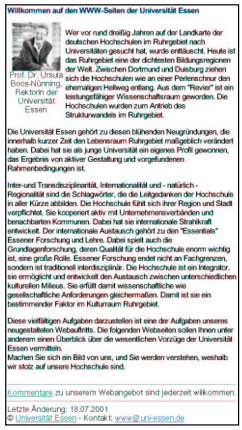
Einhergehend mit diesen inhaltlichen Entwicklungen sind auch äußerliche Veränderungen: Das neue Hörsaal- und Bibliotheksgebäude konnte bereits im Wintersemester 1999/2000 in Teilen in Betrieb genommen und die Einweihung im September 2000 gefeiert werden.

Nun stehen Dank der Unterstützung zahlreicher Freunde und Förderer mit vier neuen Hörsälen, einer sich über anderthalb Geschosse erstreckenden Bibliothek, einem für Veranstaltungen nutzbarem Forum sowie 16 Gruppenräumen unseren Studenten modernste Lehr- und Lernmethoden, mit Zugängen zum Internet und Übertragungsmöglichkeiten per Beamer in allen Räumen zur Verfügung.

Eine gute und fundierte Ausbildung erfordert diesen ständigen Wandel – was wir von unseren Studenten fordern, fordern wir auch von uns selber: Lebenslanges Lernen!

Prof. Dr. Dr. h.c. mult. Hans Tietmeyer

D 242



Willkommen auf den WWW-Seiten der Universität Essen

Prof. Dr. Ursula Boos-Nünning, Rektorin der Universität Essen

Wer vor rund dreißig Jahren auf der Landkarte der deutschen Hochschulen im Ruhrgebiet nach Universitäten gesucht hat, wurde enttäuscht. Heute ist das Ruhrgebiet eine der dichtesten Bildungsregionen der Welt. Zwischen Dortmund und Duisburg ziehen sich die Hochschulen wie an einer Perlenschnur den ehemaligen Hellweg entlang. Aus dem "Revier" ist ein leistungsfähiger Wissenschaftsraum geworden. Die Hochschulen wurden zum Antrieb des Strukturwandels im Ruhrgebiet.

Die Universität Essen gehört zu diesen blühenden Neugründungen, die innerhalb kurzer Zeit den Lebensraum Ruhrgebiet maßgeblich verändert haben. Dabei hat sie als junge Universität ein eigenes Profil gewonnen, das Ergebnis von aktiver Gestaltung und vorgefundenen Rahmenbedingungen ist.

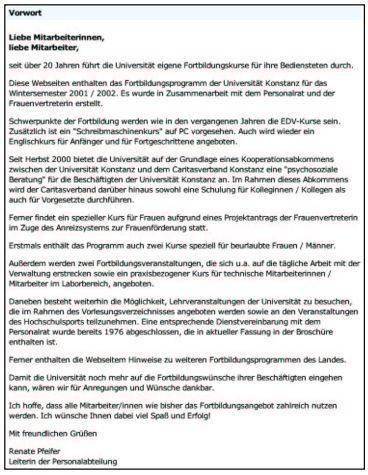
Inter-und Transdisziplinarität, Internationalität und - natürlich - Regionalität sind die Schlagwörter, die die Leitgedanken der Hochschule in aller Kürze abbilden. Die Hochschule fühlt sich ihrer Region und Stadt verpflichtet. Sie kooperiert aktiv mit Unternehmensverbänden und benachbarten Kommunen. Dabei hat sie internationale Strahlkraft entwickelt. Der internationale Austausch gehört zu den "Essentials" Essener Forschung und Lehre. Dabei spielt auch die Grundlagenforschung, deren Qualität für die Hochschule enorm wichtig ist, eine große Rolle. Essener Forschung endet nicht an Fachgrenzen, sondern ist traditionell interdisziplinär. Die Hochschule ist ein Integrator, sie ermöglicht und entwickelt den Austausch zwischen unterschiedlichen kulturellen Milieus. Sie erfüllt damit wissenschaftliche wie gesellschaftliche Anforderungen gleichermaßen. Damit ist sie ein bestimmender Faktor im Kulturraum Ruhrgebiet.

Diese vielfältigen Aufgaben darzustellen ist eine der Aufgaben unseres neugestalteten Webauftritts. Die folgenden Webseiten sollen Ihnen unter anderem einen Überblick über die wesentlichen Vorzüge der Universität Essen vermitteln.
Machen Sie sich ein Bild von uns, und Sie werden verstehen, weshalb wir stolz auf unsere Hochschule sind.

Kommentare zu unserem Webangebot sind jederzeit willkommen.

Letzte Änderung: 18.07.2001
© Universität Essen - Kontakt: www@uni-essen.de

D 455



Vorwort

Liebe Mitarbeiterinnen,
liebe Mitarbeiter,

seit über 20 Jahren führt die Universität eigene Fortbildungskurse für ihre Bediensteten durch.

Diese Webseiten enthalten das Fortbildungsprogramm der Universität Konstanz für das Wintersemester 2001 / 2002. Es wurde in Zusammenarbeit mit dem Personalrat und der Frauenvertreterin erstellt.

Schwerpunkte der Fortbildung werden wie in den vergangenen Jahren die EDV-Kurse sein. Zusätzlich ist ein "Schreibmaschinenkurs" auf PC vorgesehen. Auch wird wieder ein Englischkurs für Anfänger und für Fortgeschrittene angeboten.

Seit Herbst 2000 bietet die Universität auf der Grundlage eines Kooperationsabkommens zwischen der Universität Konstanz und dem Caritasverband Konstanz eine "psychosoziale Beratung" für die Beschäftigten der Universität Konstanz an. Im Rahmen dieses Abkommens wird der Caritasverband darüber hinaus sowohl eine Schulung für Kolleginnen / Kollegen als auch für Vorgesetzte durchführen.

Ferner findet ein spezieller Kurs für Frauen aufgrund eines Projektantrags der Frauenvertreterin im Zuge des Anreizsystems zur Frauenförderung statt.

Erstmals enthält das Programm auch zwei Kurse speziell für beurlaubte Frauen / Männer.

Außerdem werden zwei Fortbildungsveranstaltungen, die sich u.a. auf die tägliche Arbeit mit der Verwaltung erstrecken sowie ein praxisbezogener Kurs für technische Mitarbeiterinnen / Mitarbeiter im Laborbereich, angeboten.

Daneben besteht weiterhin die Möglichkeit, Lehrveranstaltungen der Universität zu besuchen, die im Rahmen des Vorlesungsverzeichnisses angeboten werden sowie an den Veranstaltungen des Hochschulsports teilzunehmen. Eine entsprechende Dienstvereinbarung mit dem Personalrat wurde bereits 1976 abgeschlossen, die in aktueller Fassung in der Broschüre enthalten ist.

Ferner enthalten die Webseiten Hinweise zu weiteren Fortbildungsprogrammen des Landes.

Damit die Universität noch mehr auf die Fortbildungswünsche ihrer Beschäftigten eingehen kann, wären wir für Anregungen und Wünsche dankbar.

Ich hoffe, dass alle Mitarbeiter/innen wie bisher das Fortbildungsangebot zahlreich nutzen werden. Ich wünsche Ihnen dabei viel Spaß und Erfolg!

Mit freundlichen Grüßen

Renate Pfeifer
Leiterin der Personalabteilung

D 683

Abbildung 11.9: Der Einsatz unterschiedlicher Textstrukturmuster in Begrüßungstexten

D 123 enthält einen Text des Direktors der Universitätsbibliothek Leipzig, der ebenfalls als Grußwort fungiert, zugleich jedoch das Informationsangebot dieser Website erläutert und den Rezipienten darüber hinaus einlädt, sich im Falle von Fragen an die Bibliothekare zu wenden und „Erwerbungsvorschläge per Mail" an die Mitarbeiter der Einrichtung zu schicken – dieser Text richtet sich somit ausschließlich an Hochschulangehörigen. Er schließt mit der Verabschiedungsfloskel „Mit besten Grüßen", die von einer Signatur begleitet wird, wodurch er wiederum in die Nähe zum formellen Brief bzw. zur formellen E-Mail rückt. D 683 stellt die Einstiegsseite des Webauftritts mit dem Weiterbildungsangebot an der Universität Konstanz dar. Der an einen Brief erinnernde Text besitzt die Überschrift „Vorwort", was jedoch nicht als Indiz dafür zu werten ist, dass er einer gedruckten Broschüre entnommen wurde. Vielmehr wurde er offenbar eigens für das WWW angefertigt, da er den Webauftritt mehrfach referenziert, z. B. „Diese Webseiten enthalten das Fortbildungsprogramm der Universität Konstanz für das Wintersemester 2001 / 2002." Es bleibt festzuhalten, dass die hier aufgeführten Texte multiple kommunikative Funktionen beinhalten und sich auf unterschiedliche Textmuster bzw. Textsorten stützen. Hierzu zählt die informelle E-Mail (D 719), das Grußwort (D 242, D 455), das Vorwort einer gedruckten Broschüre[73] (D 683) und allgemeine Rundschreiben mit Ankündigungen bzw. Handlungsaufforderungen (D 369). Nahezu alle Texte stellen die jeweilige Organisationseinheit bzw. ihren Webauftritt vor, weisen auf interessante Angebote hin und laden den Rezipienten ein, mit den Emittenten Kontakt aufzunehmen oder allgemeine Rückmeldungen einzuschicken.

Wenn die in Abbildung 11.9 dargestellten Texte generalisierend als Instanzen eines einzelnen Hypertextsortenmoduls aufgefasst werden, verdeutlicht ein Vergleich mit der Analyse von Schütte (2004a) die Existenz eines Spezifikums: Schütte ermittelt in den von ihr untersuchten Unternehmenshomepages mit der „Kontaktsequenz (primär Begrüßung)" und dem „unternehmensvorstellenden Kurztext" *zwei* makrostrukturelle Komponenten (vgl. Abschnitt 4.6.2, insbesondere Tabelle 4.9, S. 208), deren kommunikative Funktionen von dem hier diskutierten Hypertextsortenmodul, das z. B. als „institutionsvorstellender Begrüßungstext mit Bitte um Kontaktaufnahme" bezeichnet werden kann, *simultan* abgedeckt wird.

11.5.10 Rekurrent verwendete Hypertextsortenmodule

Die insgesamt 727 in dieser Stichprobe enthaltenen Dokumente umfassen einige hochfrequente Hypertextsortenmodule, die insbesondere in den Knoten der zentralen Webangebote rekurrent und von der jeweiligen Hypertextknotensorte unabhängig verwendet werden. Dabei ist wiederum zu differenzieren zwischen der Verwendung eines Hypertextsortenmoduls als Hypertextknotensorte, der Instanziierung eines Hypertextsortenmoduls innerhalb eines Knotens, der einer anderen Hypertextknotensorte zugehörig ist und der Verknüpfung eines solchen Knotens in einem HTML-Dokument.

[73] Es ist davon auszugehen, dass die Autorin von D 683 mit der Anfertigung derartiger gedruckter Broschüren vertraut ist, da sie sich explizit an diese Textsorte anlehnt. Es wäre weiterhin möglich, dass die Versendung gedruckter Broschüren mit dem Weiterbildungsangebot der Universität Konstanz aus Kostengründen eingestellt wurde, so dass die Anlehnung als Indiz für die „sehr deutliche Prägung durch die traditionelle Schriftlichkeit und ihre Produkte" aufzufassen ist, wodurch „das traditionelle Textwissen auch in den neuen medialen Umgebungen perpetuiert" wird (Eckkrammer, 2001, S. 51; vgl. auch Abschnitt 4.3.2).

	URL	Hypertextknotensorte	Begrüßung	Anrede	Text	Zussam-menfassung	Bitte um Kon-taktaufnahme	Verab-schiedung	Signatur
D 102	http://www.tu-chemnitz.de/usz/	Einstiegsseite des Webauftritts einer zentralen Organisationseinheit (Sportzentrum)	—	✓	✓	✓	—	—	—
D 123	http://www.ub.uni-leipzig.de	Einstiegsseite des Webauftritts einer zentralen Organisationseinheit (Bibliothek)	✓	—	✓	✓	✓	✓	✓
D 242	http://www.ebs.de/wir_ueber_uns/Grusswort.htm	Grußwort	✓	—	✓	—	—	—	✓
D 369	http://www.tu-dresden.de/ideenboerse/	Einstiegsseite des Webauftritts eines universitäts-internen Projekts	—	✓	✓	✓	✓	—	—
D 377	http://www.rwth-aachen.de/zentral/sul_index.html	Einstiegsseite eines zielgruppenspezifischen Webangebots (Studierende)	✓	✓	✓	✓	✓	—	—
D 445	http://www.uni-passau.de/praxiskontaktstelle/	Einstiegsseite des Webauftritts einer zentralen Organisationseinheit (Transferzentrum)	✓	—	✓	✓	✓	✓	✓
D 455	http://www.uni-essen.de	Einstiegsseite eines universitären Webauftritts	✓	—	✓	✓	—	—	—
D 629	http://www.uni-dortmund.de/TOP/ueberblick.html	Einstiegsseite der universitären Selbstdarstellung	✓	—	✓	✓	—	—	—
D 648	http://www.verwaltung.uni-bonn.de/Internationales.html	Einstiegsseite des Webauftritts einer zentra-len Organisationseinheit (akademisches Aus-landsamt)	✓	—	✓	—	✓	—	✓
D 683	http://www.uni-konstanz.de/struktur/org/personal/fortbildung/	Einstiegsseite des Webangebots eines Fach-gebiets/Studiengangs/Weiterbildungsangebots	—	✓	✓	✓	✓	✓	✓
D 719	http://www.uni-mannheim.de/studorg/asta/start.htm	Einstiegsseite des Webauftritts einer zentralen Organisationseinheit (Studierendenvertretung)	—	✓	✓	—	—	—	—

Tabelle 11.16: Einsatz von Konstituenten unterschiedlicher Textstrukturmuster

Neben den ubiquitären *Kontaktinformationen*[74] werden *Copyright-Informationen* in drei der 727 Knoten präsentiert (0,4%), Hyperlinks auf tatsächliche Instanzen dieses Hypertextsortenmoduls in einem anderen Knoten sind jedoch in 151 Dokumenten enthalten (20,8%). Ähnliche Daten können für das *Impressum* ermittelt werden: Als Hypertextknotensorte findet es in neun Dokumenten Verwendung (1,3%), 157 Webseiten referenzieren eine Instanz dieses Hypertextsortenmoduls (21,6%), das entweder als Hypertextknotensorte ausgeprägt ist oder innerhalb einer Instanz der Hypertextknotensorte *Redaktion eines Webauftritts* als Hypertextsortenmodul integriert ist. Diese beiden Hyperlinks können als rekurrente Elemente vieler Dokumente aufgefasst werden, die sich innerhalb der zentralen Webangebote befinden und sind oftmals Bestandteile einer sekundären Navigationshilfe, die in 26 Fällen (3,6%) auch einen Hyperlink auf eine *Sitemap* enthält, welche wiederum in lediglich zwei Dokumenten als Hypertextknotensorte fungiert. Eine weitere rekurrente Komponente sekundärer Navigationshilfen sind Verknüpfungen zu den acht Knoten, die einen *Lageplan*, eine *Karte*, eine *Wegbeschreibung* oder Kombinationen dieser Hypertextsortenmodule beinhalten (1,2%). Diese werden in 19 Knoten referenziert (2,6%). Sekundäre Navigationshilfen verweisen oftmals auch auf Suchformulare, mit deren Hilfe Recherchen durchgeführt werden können. Neben den 21 Instanzen (3,0%) der Hypertextknotensorte *Suchformular* in den 692 Knoten der ersten Verknüpfungsebene enthalten insgesamt 295 der 692 Dokumente dieser Stichprobe einen Verweis, der zu einem Suchformular führt (42,6%). Insgesamt 83 Knoten enthalten die Instanz eines Suchformulars als integriertes Hypertextsortenmodul (12,0%), so dass unmittelbar Recherchen ermöglicht werden, die sich vornehmlich auf das Durchsuchen von HTML-Dokumenten beziehen (74 Knoten). Suchformulare werden ebenfalls zur Recherche nach Pressemitteilungen und Veranstaltungen (jeweils zwei Vorkommen), Einrichtungen, E-Mail-Adressen, Konferenzen und innerhalb eines Bibliothekskatalogs angeboten.

11.6 Zur Entwicklung universitärer Webangebote

Im Hinblick auf das in Kapitel 4 vorgestellte Modell der allmählichen Entwicklung von Hypertexttypen bzw. Hypertextsorten stellt sich die Frage nach der initialen Erstellung von Instanzen der Hypertextsorte *Webauftritt einer Universität* – streng genommen müsste in diesem Zusammenhang von einer Protohypertextsorte die Rede sein. Diesbezüglich bestätigen die für diese Arbeit angefertigten Analysen den in Abschnitt 4.3.2 eingeführten Entwicklungszyklus. Insbesondere die von der Universität Passau und der Ruhr-Universität Bochum angebotenen Dokumente enthalten die verschiedensten Eigenschaften, die auf die initiale Anfertigung dieser Webauftritte durch Mitarbeiter der jeweiligen Rechenzentren bzw. informatischen Fakultäten hindeuten.[75] In beiden Einstiegsseiten werden Angehörige eben dieser

[74] *Kontaktinformationen* sind in 484 der 692 Dokumente (71,7%) unmittelbar enthalten, oder es steht ein `mailto:`-Hyperlink oder ein Verweis zu einem Knoten zur Verfügung, der detaillierte Kontaktinformationen umfasst (die 17 alternativen Versionen von Einstiegsseiten eines universitären Webauftritts wurden nicht berücksichtigt). Listen von Kontaktinformationen sind in 42 Knoten enthalten (6,2%; die 17 Instanzen der Hypertextknotensorte *Kontaktinformationen* wurden nicht eingerechnet).

[75] Ein detaillierter Vergleich der Entwicklungsstufen kann nur auf Basis eines Monitorkorpus erfolgen, das unterschiedliche Versionen der Webauftritte enthält. Das „Internet Archive“ umfasst auch universitäre Webauftritte, die teilweise bis zur Mitte der neunziger Jahre zurück reichen (vgl. Abschnitt 7.5.3).

Organisationseinheiten namentlich als Produzenten genannt. Hierzu zählt weiterhin der Umstand, dass innerhalb der zentralen Angebote zwar zahlreiche Hyperlinks auf externe Websites, aber nur wenige zu eigenständig erstellten Ressourcen existieren, die sich zudem auf technische Informationen (z. B. D 449: „Das Internet für Einsteiger") konzentrieren. Dabei handelt es sich vermutlich um Materialien, die den jeweiligen Produzenten bereits in digitaler Form vorlagen (beispielsweise als Skripte oder Foliensätze für Weiterbildungskurse).

Eine vergleichbare Entwicklung dürfte auch für die Mehrzahl der Webauftritte anderer Hochschulen Gültigkeit besitzen, d. h. etwa in den Jahren 1994/1995 haben Mitarbeiter zentraler oder dezentraler Organisationseinheiten, die sich mit der Informationstechnologie beschäftigen, mit den Basistechnologien des WWW experimentiert und erste Webserver aufgebaut, die als Experimentierplattform oder Forschungsgegenstand dienten. In einer nächsten Phase wurden erste öffentliche Webangebote für das Zentrum oder das Institut aufgebaut, dem die jeweiligen Mitarbeiter angehören, wobei parallel in weiteren Organisationseinheiten ebenfalls erste Websites entwickelt wurden, da das WWW Einzug in die Fachpresse hielt, einem in technologischer Hinsicht immer einfachereren Zugriff unterlag und andere Hochschulen ebenfalls mit Webpräsenzen vertreten waren, so dass ein gewisser Konkurrenzdruck entstand, die eigene Institution auch im WWW darzustellen. Die ersten Einstiegsseiten universitärer Webauftritte verknüpften somit nahezu ausschließlich die Angebote zentraler (insbesondere Rechenzentren und Bibliotheken) und dezentraler Einrichtungen (vornehmlich dezentrale Einheiten aus dem naturwissenschaftlichen und informatischen Bereich).

Mit seinem rasant ansteigenden Erfolg und der zunehmenden Präsenz in der Presse erkannten die Leitungen der Hochschulen das Potenzial des *World Wide Web* und haben in der Folgezeit individuelle Strategien entwickelt, die die Übertragung von Verantwortlichkeiten an weitere Organisationseinheiten (z. B. die Pressestelle und die Studienberatung) vorsahen, die jeweils Teile des zentralen Webangebots erstellten und – zumindest im Hinblick auf die hier untersuchte Stichprobe – nicht über den Status isolierter Hypertexte hinaus gekommen sind. Die zahlreichen Konventionen, die in den Analysen ermittelt wurden, können als ein direktes Resultat aus der Rezeption der Webauftritte anderer Hochschulen und der Integration verschiedener Merkmale in die eigene Webpräsenz aufgefasst werden (vgl. Abschnitt 4.3.2). In dieser Entwicklungsstufe wurden innerhalb des zentralen Bereichs nicht mehr externe, sondern vornehmlich lokale Angebote durch Verteiler verknüpft. Auf diese Weise entstanden sukzessive *extrem* umfangreiche Webauftritte sowie mit Informationen und Hyperlinks geradezu „überfrachtete" Einstiegsseiten und zentrale Angebote, weshalb an einigen Universitäten die Strategie der zielgruppenspezifischen Verteiler eingeführt wurde, die jedoch lediglich eine Auswahl von Hyperlinks auf bereits lokal verfügbare Ressourcen beinhalten. Den derzeitigen Abschluss der Entwicklung stellt die Anfertigung eigenständiger Hypertexte für spezifische Zielgruppen und die allmähliche Verschiebung der Einstiegsseiten zu tagesaktuellen und in Zukunft vermutlich auch personalisierbaren Informationsportalen dar.

11.7 Zum Einfluss des Domänenwissens

Die in dieser Analyse ermittelten Hypertextknotensorten für die Domäne der universitären Webauftritte abstrahieren nahezu vollständig von der Struktur einer Hochschule. Gerade diese Organisationsstruktur besitzt jedoch auf mehreren Ebenen einen zentralen Einfluss auf

die in dem Hypertextsortenmodell vorgeschlagenen Konzepte sowie auf Aspekte der Typologisierung. Hiervon ist zunächst die Interpretation eines Tupels bestehend aus Hypertextknotensorte und zugehöriger Instanz betroffen. Gerade Exemplare der Hypertextknotensorten aus der Kategorie der Inhaltsknoten wie z. B. *Kontaktinformationen*, *Suchformular*, *Impressum*, *Kurzdarstellung einer Organisationseinheit* oder *Einladung zu einer Veranstaltung* können prinzipiell in Instanzen einer *Vielzahl* eingebetteter Hypertextsorten eingesetzt werden.

Die fünf Exemplare der *Ankündigung* wurden von einer Bibliothek (D 310–D 312), einer Hochschulverwaltung (D 400) und einer zentralen Organisationseinheit publiziert, die technische Dienstleistungen erbringt (D 687). Sowohl die Zuweisung einer Hypertextknotensorte als auch die Interpretation der jeweiligen Instanz ist unmittelbar vom Typ der beteiligten Organisationseinheit betroffen, die jeweils individuelle Ausprägungen von Ankündigungen publizieren: Im Falle von D 310–D 312 werden universitätsinterne Subskriptionen verschiedener Datenbanken für den Volltextzugriff auf Zeitschriften angekündigt, D 400 erläutert die Einführung einer neuen Software-Anwendung zur Verwaltung von Lehrveranstaltungen und Raumbelegungsplänen, und D 687 kündigt einen Ausleihdienst für Videoprojektoren an. Auf Seiten der Rezipienten, die mit der internen Organisationsstruktur einer Hochschule vertraut sind, existieren spezifische Erwartungen, welche Arten von Ankündigungen von einer Organisationseinheit typischerweise veröffentlicht werden, und derartiges Wissen sollte im besten Falle auch in einer Hypertextsortentypologie repräsentiert werden.

Die Problematik der Interpretation einer Instanz wird durch die Hypertextknotensorte *Kontaktinformationen* besonders deutlich, die die tabellarische oder listenartige Präsentation der Namen und Kontaktinformationen von Hochschulangehörigen umfasst. Korrespondierende Exemplare dieser Hypertextknotensorte können ebenfalls Bestandteile zahlreicher eingebetteter Hypertexte sein, so enthält D 278 Kontaktinformationen für die Studienberatung der Universität Koblenz-Landau, D 598 umfasst die Auflistung der Namen und Kontaktmöglichkeiten des Rektorats der MU Lübeck, und D 602 enthält die Namen, Anschriften und Telefonnummern aller Lehrenden der beiden Fakultäten dieser Hochschule. D 652 präsentiert eine umfangreiche Tabelle, die Ansprechpartner in sämtlichen Organisationseinheiten der Universität Bonn beinhaltet. Der Skopus der Instanz einer Hypertextknotensorten wird also *unmittelbar* durch den abstrakten Typ der Organisationseinheit bestimmt, deren Webauftritt diesen Knoten beinhaltet. Der Typ der Organisationseinheit bildete wiederum den Ausgangspunkt der originären Entscheidung für eine spezifische Hypertextsorte, um eben diese Organisationseinheit innerhalb des *Webauftritts einer Hochschule* zu präsentieren.

Domänenwissen ist in einen weiteren Aspekt involviert: Die Hypertextknotensorte *Kurzdarstellung einer Organisationseinheit* (in der Ausprägung der Präsentation ihrer Zusammensetzung und Kontaktinformationen) kann sich auf zahlreiche Organisationseinheiten einer Hochschule beziehen. Die in der Stichprobe enthaltenen Exemplare umfassen Kurzdarstellungen eines Kuratoriums (D 96), eines Graduiertenkollegs (D 349), eines Referats innerhalb der Hochschulverwaltung (D 443) und zweier Fördervereine (D 499, D 727). Das Exemplar der abstrakten Hypertextknotensorte wird somit von dem Typ der instanziierenden Organisationseinheit determiniert. Bei dieser speziellen Hypertextknotensorte handelt es sich um eine Art digitale Visitenkarte, die den Webauftritt einer Organisationseinheit in einem Einzelknoten präsentiert. Somit kommt der Aspekt hinzu, dass der Typ der dargestellten Organisationseinheit auch die Positionierung des Knotens innerhalb der umfassenden Instanz

der Hypertextsorte *Webauftritt einer Universität* beeinflusst: D 96 ist Teil eines eingebetteten Hypertextes, der die Organe und Gremien der TU Chemnitz vorstellt. D 349 ist hingegen Teil eines übergeordneten Hypertextes zum Thema Forschung. Der eingebettete Hypertext, der Informationen über die Verwaltung der Universität Passau enthält, umfasst auch D 443, während D 499 ein Bestandteil des zentralen Webangebots der TU Darmstadt ist.

11.8 Fazit – Zur Typologisierung der Ergebnisse

Die in diesem Kapitel präsentierte zweistufige Analyse betrifft Konventionen, die sich auf die Konstituenten der Hypertextknotensorte *Einstiegsseite eines universitären Webauftritts*, die Hypertextknotensorten der ersten Verknüpfungsebene und die generelle Informationsarchitektur eines universitären Webangebots beziehen. Eines der Ziele ist die Typologisierung der Hypertextknotensorten. Die Vorgehensweise basiert auf der Hypothese, dass die Einstiegsseite eines universitären Webauftritts und die dort verknüpften Dokumente zur Konstruktion der oberen Ebenen einer derartigen Typologie gewinnbringend eingesetzt werden können. Die präsentierten Daten belegen zwar die Existenz zahlreicher Konventionen, die in der Stichprobe auf mehreren Ebenen beobachtet werden können, im Hinblick auf die Frage nach der Typologisierung sieht sich das genannte Ziel jedoch mit einer wesentlichen Schwierigkeit konfrontiert. Diese bezieht sich – neben der auch in der Textlinguistik kontrovers diskutierten Frage der Typologisierung von Textsorten (vgl. Kapitel 2) – auf die Eigenschaft der Vielschichtigkeit von Hypertextsorten. Das in Kapitel 5 eingeführte Hypertextsortenmodell kann, wie die Analysen deutlich gemacht haben, zur Beschreibung und Einordnung der Ergebnisse erfolgreich eingesetzt werden. Die zur adäquaten Repräsentation des Gegenstandes zwingend notwendige Komplexität der konzeptuellen Ebenen *Hypertextsorte*, *Hypertextknotensorte* und *Hypertextsortenmodul* sowie die ermittelten Binnenstrukturen und Subtypen sind jedoch dafür verantwortlich, dass die Konstruktion einer *einzelnen*, an die „klassischen" textlinguistischen Vorschläge angelehnten Typologie, die *sämtliche* potenziellen Konstituenten arbiträrer Hypertextsorten und sonstige beteiligte Faktoren zu repräsentieren in der Lage ist, *nicht* erfolgen kann. In Bezug auf die Hypertextknotensorten (vgl. Tabelle 11.10, S. 492) ist es z. B. nicht möglich, eine trennscharfe Typologie ohne Mehrfachzuordnung anzufertigen (vgl. die Abbildungen 11.10 bis 11.12). Lediglich der Typ *Hyperlinkliste* umfasst Hypertextknotensorten, die eindeutig voneinander abgegrenzt werden können, da sich die Kriterien, die von Produzenten zur Zusammenstellung korrespondierender Instanzen verwendet werden, hinreichend voneinander differenzieren lassen (vgl. Abschnitt 11.5.3).[76]

Für die in Abbildung 11.10 aus Darstellungs- und Argumentationsgründen fehlenden Typen *Einstiegsseite* und *Inhaltsknoten* wurden ebenfalls Subtypologien angefertigt (vgl. die Abbildungen 11.11 und 11.12), die jedoch *ausschließlich* dem Zweck dienen, die Problematik der Typologisierung einer inhaltlich und funktional hochgradig heterogenen Menge von Hypertextknotensorten zu veranschaulichen. Verantwortlich für diesen Umstand ist die genannte Vielschichtigkeit, die unter anderem die Existenz *multipler* Typologien betrifft, die in einer einzelnen Klassifikation nicht adäquat dargestellt werden können, da relationale Be-

[76] Die initiale Typologie von Hypertextknotentypen (Abbildung 3.8, S. 153) verdeutlicht, dass die Kategorisierung von Hypertextknotensorten auf zahlreichen unterschiedlichen Kriterien beruhen kann.

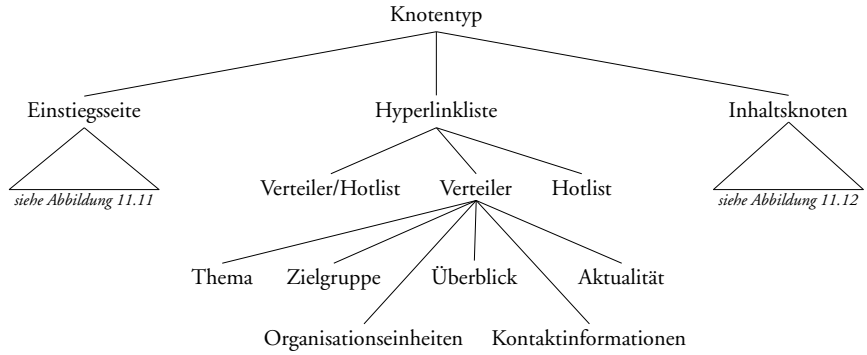


Abbildung 11.10: Eine Typologie der ermittelten Hypertextknotensorten (Teil 1)

ziehungen zwischen den Typen existieren, die für die Konstitution und Interpretation der Hypertextknotensorten essenziell sind. Diese beziehen sich auf die drei Ebenen Hypertextsorte, Hypertextknotensorte und Hypertextsortenmodul (vgl. Kapitel 5). Eine Typologie, die ausschließlich Hypertextknotensorten berücksichtigt, muss zwangsläufig sowohl von der übergeordneten als auch von der untergeordneten Ebene abstrahieren. Zwar existieren innerhalb der Einstiegsseiten und Inhaltsknoten durchaus unterschiedliche Subtypen, die Hypertextknotensorten mit spezifischen Gemeinsamkeiten bündeln; die Ermittlung eines Inventars von Merkmalen, die die 61 Hypertextknotensorten trennscharf voneinander abgrenzen und Mehrfachzuordnungen vermeiden, erscheint jedoch nicht sinnvoll.[77] Hinzu kommt, dass Hypertextsortenmodule die eigentlichen atomaren Bausteine von Hypertextsorten darstellen, da sie als Hypertextknotensorten fungieren können. Es stellt sich somit die Frage, welche weiteren Hypertextsortenmodule die Instanzen der ermittelten Hypertextknotensorten umfassen und welche obligatorischen oder optionalen Anwendungen diesbezüglich vorliegen. Zudem können auch für spezifische Hypertextsortenmodule unterschiedliche Typen ermittelt werden (z. B. für die *primäre Navigationshilfe* der Einstiegsseiten universitärer Webauftritte, vgl. Abschnitt 11.4.2, oder für den *Begrüßungstext*, vgl. Abschnitt 11.5.9). Von essenzieller Bedeutung für die Einordnung der ermittelten Hypertextknotensorten in eine möglichst homogene Typologie ist die Struktur einer generischen Hochschule, die zentrale und dezentrale Einrichtungen umfasst, welche mit jeweils eigenen Auftritten im WWW vertreten sind. Die Einbeziehung dieser Organisationsstruktur, die ebenfalls als Typologie bzw. Ontologie aufgefasst werden kann, welche das spezifische Domänenwissen umfasst, ist für die Frage nach den Konstituenten und Emittenten der jeweiligen Hypertextknotensorten unumgänglich (vgl. Abschnitt 11.7).[78] Eine adäquate Beschreibung der ermittelten Hypertextknoten-

[77] Zwar könnte das von Sandig (1972) vorgeschlagene Verfahren (vgl. Abschnitt 2.3.5) für diesen spezifischen Zweck eingesetzt werden, die Beziehungen zwischen den Hypertextknotensorten, den übergeordneten Hypertextsorten sowie den untergeordneten Hypertextsortenmodulen könnten jedoch nicht erfasst werden.

[78] In Abschnitt 11.5.1 wurde erwähnt, dass zur Aufstellung des in Tabelle 11.10 dargestellten Inventars von Hypertextknotensorten mehrere Restrukturierungsphasen notwendig waren. Diesbezüglich wurde zu einem frühen Zeitpunkt eine weitere Beschreibungsebene in das Analyseschema aufgenommen, um die jeweilige „Position“ eines Knotens innerhalb der generischen Organisationsstruktur einer Hochschule notieren zu können. Für diesen Zweck wurde in das Analyseschema ein Textfeld integriert, das eine hierarchisch gestufte Liste eben dieser Einheiten umfasst. Abbildung 7.12 (S. 356) zeigt einen Ausschnitt dieser Liste.

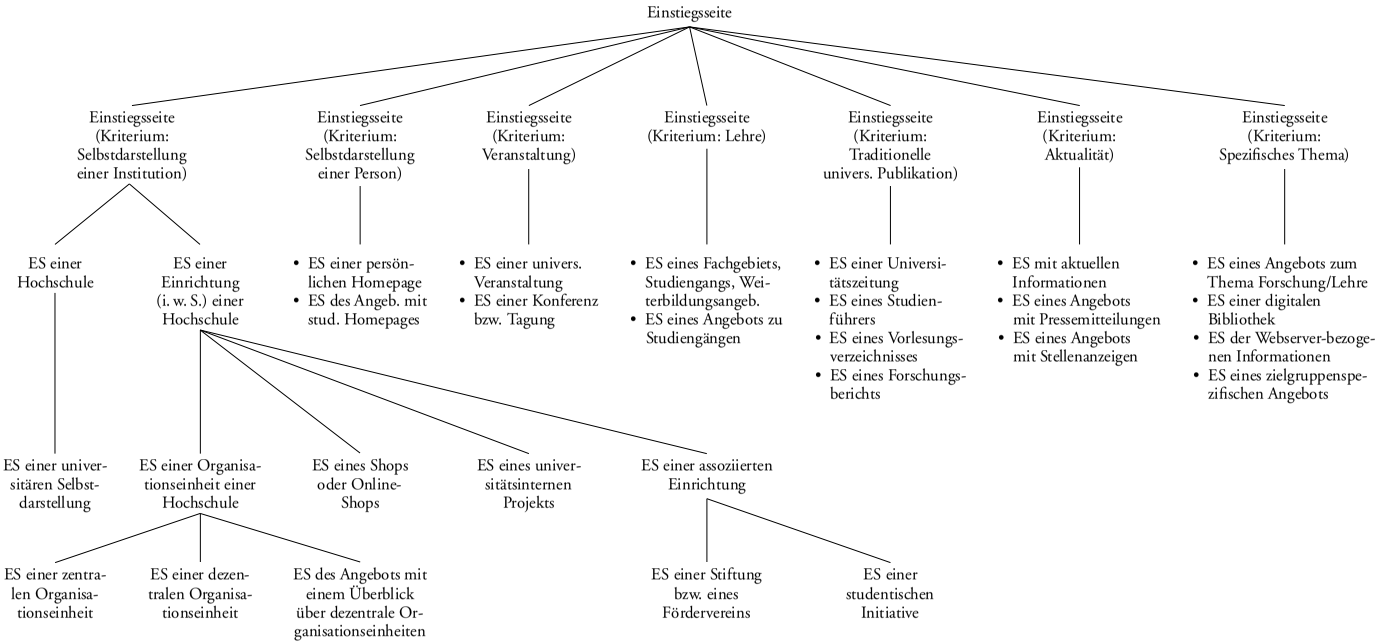


Abbildung 11.11: Eine Typologie der ermittelten Hypertextknotensorten (Teil 2)

Abbildung 11.12: Eine Typologie der ermittelten Hypertextknotensorten (Teil 3)

sorten universitärer Webangebote ist folglich nur durch den Einsatz multipler Typologien zu erzielen, die die unterschiedlichen Typen von Hypertextsorten, Hypertextknotensorten, Hypertextsortenmodulen und die generische Organisationsstruktur einer Hochschule reflektieren, da sich die einzelnen Hypertexte vornehmlich an eben dieser Strukturierung orientieren. Darüber hinaus existiert zusätzlich eine thematische Ebene, die spezifische Themengebiete wie z. B. „Forschung“ und „Studium“ umfasst, oftmals quer zu den Webauftritten einzelner Organisationseinheiten liegt und durch Verteiler oder eigens angefertigte Hypertexte realisiert wird. Zielgruppenspezifische Angebote wiederum fungieren als Filter, die die große Menge der in einem universitären Webauftritt enthaltenen Informationen nach dem Kriterium unterschiedlicher Gruppen von Rezipienten bündeln.

Die Abbildungen 11.11 und 11.12 können zur Veranschaulichung der genannten Probleme herangezogen werden: Die Hypertextknotensorten *Einstiegsseite des Webauftritts einer Hochschule* und *Einstiegsseite des Webauftritts einer Organisationseinheit einer Hochschule* besitzen das gemeinsame Merkmal, dass sie die Einstiegsseiten des Webauftritts einer Institution (im weitesten Sinne) darstellen. Zugleich sind die *Webauftritte von Organisationseinheiten* jedoch *Konstituenten* der übergeordneten Hypertextsorte *Webauftritt einer Universität*. Unklar ist weiterhin die Kategorisierung der *Einstiegsseite des Webangebots mit einem Überblick über dezentrale Organisationseinheiten* – Abbildung 11.11 fasst diese als eine Art „abstrakte Organisationseinheit“ einer Hochschule auf. Eine adäquatere Repräsentation müsste die Kernmerkmale aller dezentralen Organisationseinheiten einer generischen Hochschule einbeziehen und diese als Konstituenten der abstrakten Beschreibung einer Hypertextsorte bündeln. Die Problematik der Mehrfachzuordnung wird in Abbildung 11.12 besonders deutlich, da zahlreiche Hypertextknotensorten vom Typs Inhaltsknoten mehreren Kategorien zugeordnet werden können. Zudem existieren unterschiedliche Typen von Relationen zwischen den Konzepten, die in den Abbildungen 11.10 bis 11.12 dargestellt werden. Während das nachfolgende Kapitel 12 näher auf den Zusammenhang zwischen den Organisationseinheiten einer Hochschule und den von ihnen publizierten Instanzen spezifischer Hypertext(knoten)sorten eingeht, stellt Kapitel 13 ein Repräsentationsverfahren vor, das in der Lage ist, die hier genannten Probleme durch den Einsatz multipler Ontologien zu minimieren.

12 Analyse 5: Untersuchung 750 zufällig ausgewählter Dokumente

12.1 Einleitung

Die fünfte und abschließende Analyse betrifft die Untersuchung 750 zufällig aus dem Korpusbestand ausgewählter Dokumente mit dem Ziel der Identifizierung und Sammlung unterschiedlicher Hypertextknotensorten und Hypertextsorten in universitären Webauftritten.

Nachdem im Folgenden zunächst die Ziele und Bezüge zum Hypertextsortenmodell vorgestellt werden, geht Abschnitt 12.3 auf die zur Generierung eingesetzten Beschränkungen und den Inhalt der analysierten Stichprobe ein. Abschnitt 12.4 stellt die Ergebnisse der Untersuchung im Überblick dar. Anschließend diskutieren die Abschnitte 12.5 bis 12.7 die Analysen der drei untersuchten Ebenen in detaillierter Form, woraufhin ein Vergleich der Ergebnisse mit verwandten Arbeiten durchgeführt wird (Abschnitt 12.8).

12.2 Ziele und Bezüge zum Hypertextsortenmodell

Während sich die vorangegangenen Analysen vornehmlich auf die Untersuchung spezifischer Hypertextsorten sowie Hypertextknotensorten beschränkt haben, steht in dieser fünften und abschließenden Studie eine zufällig ausgewählte Stichprobe von HTML-Dokumenten im Zentrum des Interesses, wobei verschiedene Ziele parallel verfolgt werden. Innerhalb der Analyse werden (i) die Hypertextknotensorten der Dokumente, (ii) die übergeordneten Hypertextsorten derjenigen Hypertexte, die ein Dokument der Stichprobe beinhalten sowie (iii) die jeweiligen universitären Organisationseinheiten ermittelt, die einen spezifischen Knoten und somit auch das zugehörige Hypertextexemplar publizieren. Bei der zufälligen Zusammenstellung der Stichprobe wurden nur diejenigen Dokumente berücksichtigt, deren URLs Charakteristika aufweisen, die für das zentrale Webangebot einer Hochschule gerade *nicht* zutreffend

sind (vgl. den nachfolgenden Abschnitt 12.3). Diese Vorgehensweise basiert auf der Hypothese, dass die untersten Verknüpfungsebenen (d. h. die Blattknoten) zur Strukturierung der unteren Ebenen einer Typologie von Hypertextsorten geeignet sind. Bei dieser Studie handelt es sich somit um eine *bottom-up*-Analyse, die die *top-down* durchgeführte Untersuchung komplementiert (vgl. Kapitel 11, insbesondere die Abschnitte 11.2, 11.5 und 11.8, sowie Rehm, 2002b). Von Bedeutung ist darüber hinaus die Frage der Varianz der ermittelten Daten, d. h. welche und wie viele unterschiedliche Hypertextsorten sowie Hypertextknotensorten ermittelt werden können. Existieren Exemplare von Hypertextknotensorten, die vornehmlich innerhalb spezifischer Hypertextsorten verwendet werden? Und: Welche Abweichungen können hinsichtlich der in Analyse 4 präsentierten Ergebnisse konstatiert werden? In diesem Zusammenhang ist auch ein Vergleich mit den traditionellen Textsorten notwendig, die von Heinemann (2000b) für den Kommunikationsbereich Hochschule und Wissenschaft dargestellt werden (vgl. Abschnitt 6.2), um die spezifischen Erweiterungen bzw. Beschränkungen hinsichtlich des von den Produzenten benutzten Inventars von Hypertext(knoten)sorten zu ermitteln. Abschließend sollen die Ergebnisse dieser Analyse auch für einen Vergleich mit den in Abschnitt 4.4 diskutierten Studien eingesetzt werden, um die These zu verifizieren, dass ein homogenes Inventar von Hypertextsorten nur auf der Grundlage einer eindeutig spezifizierten Untersuchungsdomäne erstellt werden kann.

12.3 Die Stichprobe

Die im Rahmen dieser Analyse untersuchte Stichprobe besteht aus 750 zufällig aus dem Korpusbestand ausgewählten HTML-Dokumenten.[1] Zur Generierung der Stichprobe wurde die in Abschnitt 7.3.4 dargestellte Funktionalität der Web-Oberfläche der Korpusdatenbank eingesetzt, wobei verschiedene Merkmale spezifiziert wurden, die ein Dokument besitzen muss, damit es in die Liste derjenigen Dateien aufgenommen wird, aus denen abschließend die Stichprobe zufällig erzeugt wird: Zunächst wurden die Datenbestände von 35 Hochschulen angegeben, die *nicht* den Universitäten entsprechen, die in Analyse 4 untersucht wurden (Kapitel 11). Zusätzlich wurden Dokumente ausgeschlossen, deren URLs Pfade besitzen, die mit `/~` beginnen, um den typischen Adressbereich persönlicher Homepages nicht zu berücksichtigen. Ebenfalls ausgeschlossen wurden Dokumente mit URLs, die Namen von Webservern beinhalten, die nicht mit `www` beginnen. Weiterhin wurde eine minimale Einbettungstiefe von drei und eine maximale Einbettungstiefe von zehn angegeben, um möglichst viele Dokumente der dezentralen Webangebote in die Stichprobe aufzunehmen (vgl. Abschnitt 7.3.4). Insgesamt entsprechen 906 425 im Korpus enthaltene Dokumente diesen Beschränkungen, aus denen schließlich mit Hilfe der Zufallsfunktion der *MySQL*-Datenbank 750 ausgewählt wurden. Diese stammen von insgesamt 267 verschiedenen Webservern.

12.4 Die Ergebnisse der Analyse im Überblick

Die Untersuchung konzentriert sich auf die Ermittlung der Hypertextknotensorten und der übergeordneten Hypertextsorten sowie der jeweiligen universitären Organisationseinheiten,

[1] Anhang E auf der beiliegenden CD ROM enthält eine Liste der 750 in der Stichprobe enthaltenen Dokumente.

die die in der Stichprobe enthaltenen Dokumente im WWW publizieren.[2] Somit wird auch diese Studie mit der Problematik konfrontiert, dass eine große und heterogene Menge von HTML-Dokumenten auf ein Inventar abstrakter Etiketten von Hypertextknotensorten abzubilden ist (vgl. Abschnitt 11.5.1). Diese Untersuchung wurde ebenfalls mit Hilfe der in Kapitel 7 dargestellten Web-Oberfläche durchgeführt. Das Inventar von Etiketten für Hypertextsorten und Hypertextknotensorten stammt aus den bereits angefertigten Analysen. Zudem wurde der Bestand in mehreren Restrukturierungsphasen sukzessive überarbeitet, um einen bestmöglichen Abdeckungsgrad sowie eine hohe Konsistenz und Trennschärfe der Resultate gewährleisten zu können. Die in Kapitel 11 angesprochenen Aspekte bezüglich der generellen Methodologie und der Eigenschaften, die die Ergebnisse der zweiten Phase der vierten Analyse aufweisen, gelten ebenfalls für die hier diskutierte Studie, d. h. es wird aufgrund der Existenz mannigfaltiger Gemeinsamkeiten davon ausgegangen, dass es sich bei den im Folgenden vorgestellten Entitäten um Hypertextsorten und Hypertextknotensorten im Sinne des Hypertextsortenmodells handelt (vgl. auch Abschnitt 11.5.6).

Tabelle 12.1 stellt die 65 ermittelten Hypertexttypen bzw. Hypertextsorten dar. Sechs Hypertexttypen besitzen insgesmt 53 spezifische Subtypen, die als Hypertextsorten aufgefasst werden und in Tabelle 12.2 aufgeführt sind. Das Inventar weist einen wesentlichen Unterschied zu dem Bestand von Hypertextsorten auf, der in der zweiten Phase der vierten Analyse (Abschnitt 11.5) implizit[3] erhoben wurde: In den Tabellen 12.1 und 12.2 wird nicht zwischen den Webauftritten zentraler oder dezentraler Organisationseinheiten differenziert, da die Zuordnung eines Knotens zu einer spezifischen Einrichtung innerhalb der dritten Analyseebene erfolgt – der nachfolgende Abschnitt 12.5 geht genauer auf diese Thematik ein. Tabelle 12.1 verdeutlicht, dass 42,3% der in der Stichprobe enthaltenen Knoten Instanzen der beiden Hypertexttypen *Webauftritt einer Organisationseinheit* (28,4%) sowie *Webangebot einer Lehrveranstaltung* (13,9%) zuzurechnen sind. Die weiteren hochfrequenten Hypertexttypen bzw. -sorten lauten *Vorlesungsverzeichnis* (6,0%), *Software-Dokumentation* (5,3%), *Forschungsbericht, Jahresbericht, Rechenschaftsbericht* (3,7%), *Skript einer Lehrveranstaltung* (3,7%), *Fotogalerie* (3,5%) und *Pressemitteilungen* (3,2%). Den verbleibenden Hypertextsorten konnten jeweils weniger als 20 Knoten zugeordnet werden; 29 der 65 Hypertextsorten besitzen jeweils ein oder zwei zugehörige HTML-Dokumente.

Die Tabellen 12.3 und 12.4 umfassen das Inventar ermittelter Hypertextknotentypen bzw. -sorten. Zehn Hypertextknotentypen umfassen insgesamt 54 Subkategorien, die als Hypertextknotensorten betrachtet werden. Auch bezüglich dieser Analyseebene existieren verschiedene Unterschiede zur vierten Analyse. Zunächst wurde aus mehreren Gründen auf eine Differenzierung in die drei Typen Einstiegsseite, Hyperlinkliste und Inhaltsknoten verzichtet: Die Einstiegsseite wird als eigenständiger Hypertextknotentyp aufgefasst (vgl. Abschnitt 5.5), der jeweils individuellen Hypertextsorten zuzurechnen ist, die in einer separaten Analysee-

[2] Zur Bestimmung der übergeordneten Hypertextsorte sowie der publizierenden Organisationseinheit war es in vielen Fällen notwendig, die bei- und übergeordneten Dokumente zu inspizieren. Aus diesem Grund wurden nicht nur die 750 in der Stichprobe enthaltenen HTML-Dokumente, sondern insgesamt mehrere tausend Knoten untersucht. Diese Notwendigkeit kann als Indiz für den Umstand verstanden werden, dass viele Knoten Defizite in Bezug auf die Unterstützung der Kohärenzbildung aufweisen (vgl. Abschnitt 12.8.2).

[3] Da in der zweiten Phase der vierten Analyse ausschließlich Hypertextknotensorten im Zentrum des Interesses standen, reflektieren die unterschiedlichen Typen von Einstiegsseiten implizit ein Inventar zugehöriger Hypertextsorten (vgl. Abschnitt 11.5.4).

	Hypertexttyp bzw. Hypertextsorte	Subtypen	Frequenz	Prozent
1.	Webauftritt einer Organisationseinheit	24	213	28,4
2.	Webangebot einer Lehrveranstaltung	4	104	13,9
3.	Vorlesungsverzeichnis	—	45	6,0
4.	Software-Dokumentation	4	40	5,3
5.	Forschungsbericht, Jahresbericht, Rechenschaftsbericht	—	28	3,7
6.	Skript einer Lehrveranstaltung	—	28	3,7
7.	Fotogalerie	4	26	3,5
8.	Pressemitteilungen (Auflistung)	—	24	3,2
9.	Publikationsorgan einer Einrichtung	8	19	2,5
10.	Persönliche Homepage eines Wissenschaftlers	—	17	2,3
11.	Webauftritt einer Institution	9	14	1,9
12.	Private Homepage eines Studierenden	—	12	1,6
13.	Unterrichtsmaterialien für die Schule	—	11	1,5
14.	Studienführer	—	10	1,3
15.	Webangebot eines Studiengangs	—	10	1,3
16.	Studentische Präsentationen/Hausarbeiten	—	9	1,2
17.	Verzeichnis der Angehörigen einer Organisationseinheit	—	9	1,2
18.	Handbuch	—	8	1,1
19.	Virtuelles Museum	—	8	1,1
20.	Anleitungen, Benutzungshinweise, Dokumentationen	—	7	0,9
21.	Bibliothekskatalog	—	6	0,8
22.	Fachbuch/Kapitel eines Fachbuches	—	6	0,8
23.	Diplomarbeit	—	6	0,8
24.	Digitale Bibliothek	—	5	0,7
25.	Messageboard/Diskussionsforum	—	5	0,7
26.	Studentische Präsentation/Vortrag/Ausarbeitung	—	5	0,7
27.	Webangebot einer Konferenz/Tagung	—	5	0,7
28.	Klassifikation medizinischer Diagnoseprozeduren	—	4	0,5
29.	Lexikon	—	4	0,5
30.	Webangebot einer Veranstaltung/eines Wettbewerbs	—	4	0,5
31.	Zugriffsstatistik	—	4	0,5
32.	Aufgabenstellungen für Haus- oder Abschlussarbeiten	—	3	0,4
33.	Forschungsprojekte einer Organisationseinheit	—	3	0,4
34.	Medizinische Diagnosebeispiele	—	3	0,4
35.	Regelung, Ordnung, Gesetz, juristischer Text	—	3	0,4
36.	Studierendenstatistik	—	3	0,4
37.	Abschlussbericht (eines Projekts)	—	2	0,3
38.	Aktuelle Informationen, Termine, Meldungen	—	2	0,3
39.	Biografie	—	2	0,3
40.	Digitale Karte, Stadtplan	—	2	0,3
41.	Dissertation	—	2	0,3
42.	Fachinformationsportal	—	2	0,3
43.	FAQ-Dokument	—	2	0,3
44.	Grafischer Assistent zur Prozessentwicklung	—	2	0,3
45.	Internet-Zeitschrift (Rezensionsforum)	—	2	0,3
46.	Mailing-Listen-Archiv	—	2	0,3
47.	Bibliografie	—	1	0,1
48.	Bibliothekssystematik	—	1	0,1
49.	Daten historischer Bauwerke	—	1	0,1
50.	Exkursionsbericht	—	1	0,1
51.	Glossar	—	1	0,1
52.	Jahrbuch	—	1	0,1
53.	Kleinanzeigen	—	1	0,1
54.	Kochbuch	—	1	0,1
55.	Kunst- und Kulturprojekt	—	1	0,1
56.	Protokollarchiv	—	1	0,1
57.	Prüfungsordnung	—	1	0,1
58.	Richtlinien (für Studien- und Hausarbeiten)	—	1	0,1
59.	Semesterapparate	—	1	0,1
60.	Studienordnung	—	1	0,1
61.	Tageszeitung	—	1	0,1
62.	Tippspiel zu einer Sportveranstaltung	—	1	0,1
63.	Transferkatalog	—	1	0,1
64.	Virtual Library	—	1	0,1
65.	Wissenschaftlicher Artikel	—	1	0,1
	Σ		750	100

Tabelle 12.1: Die ermittelten Hypertexttypen bzw. -sorten im Überblick

	Hypertexttyp bzw. Hypertextsorte	Frequenz	Prozent
1.	*Webauftritt einer Organisationseinheit*	213	28,4
	Webauftritt einer Professur bzw. Arbeitsgruppe	84	11,2
	Webauftritt eines Instituts bzw. Seminars	28	3,7
	Webauftritt eines Projekts oder Projektverbundes	21	2,8
	Webauftritt eines spezifischen Dezernats bzw. Referats der Hochschulverwaltung	10	1,3
	Webauftritt einer Fakultät bzw. eines Fachbereiches	9	1,2
	Webauftritt eines Arbeitskreises	7	0,9
	Webauftritt eines Rechenzentrums	6	0,8
	Webauftritt einer Studienberatung	6	0,8
	Webauftritt eines Weiterbildungszentrums	5	0,7
	Webauftritt einer Bibliothek	4	0,5
	Webauftritt einer Fachschaft	4	0,5
	Webauftritt einer Studierendenvertretung	4	0,5
	Webauftritt eines Dekanats	3	0,4
	Webauftritt eines wissenschaftlichen Zentrums (generisch)	3	0,4
	Webauftritt einer Hochschulverwaltung	3	0,4
	Webauftritt eines Klinikums	3	0,4
	Webauftritt eines universitätsinternen Projekts	3	0,4
	Webauftritt einer Kommission	2	0,3
	Webauftritt einer studentischen Initiative	2	0,3
	Webauftritt eines Universitätsarchivs	2	0,3
	Webauftritt einer Frauenbeauftragen	1	0,1
	Webauftritt eines Graduiertenkollegs	1	0,1
	Webauftritt einer Pressestelle	1	0,1
	Webauftritt eines Sportzentrums	1	0,1
2.	*Webangebot einer Lehrveranstaltung*	104	13,9
	Webangebot einer regulären Lehrveranstaltung	88	11,7
	Webangebot eines Doktoranden- oder Mitarbeiterkolloquiums	6	0,8
	Webangebot eines Weiterbildungskurses	6	0,8
	Webangebot einer regulären Lehrveranstaltung (E-Learning-Angebot)	4	0,5
4.	*Software-Dokumentation*	40	5,3
	Handbuch, Manual (offizielle Dokumentation)	19	2,5
	Lehrwerk, Referenz (inoffizielle Dokumentation)	10	1,3
	Benutzungshinweise für lokal verfügbare Software	6	0,8
	Tutorial	5	0,7
7.	*Fotogalerie*	26	3,5
	Galerie mit Fotos eines Ortes bzw. einer Stadt	10	1,3
	Galerie mit Fotos einer Veranstaltung, Konferenz oder Messe	9	1,2
	Galerie mit Fotos eines Gebäudes	5	0,7
	Bildarchiv	2	0,3
9.	*Publikationsorgan einer Einrichtung*	19	2,5
	Publikationsorgan einer Hochschule (Universitätszeitung)	6	0,8
	Publikationsorgan einer Fachschaft (Fachschafts- bzw. Studierendenzeitung)	4	0,5
	Publikationsorgan eines Rechenzentrums (Mitteilungen und Informationen)	4	0,5
	Publikationsorgan einer Alumnivereinigung	1	0,1
	Publikationsorgan einer Bibliothek (Hauszeitschrift)	1	0,1
	Publikationsorgan einer hochschulpolitischen Gruppierung	1	0,1
	Publikationsorgan einer Hochschulleitung (Verkündungsblatt)	1	0,1
	Publikationsorgan eines Vereins (Vereinsmitteilungen)	1	0,1
11.	*Webauftritt einer Institution*	14	1,9
	Webauftritt eines Fachverbandes	3	0,4
	Webauftritt eines Vereins	3	0,4
	Webauftritt einer Schule	2	0,3
	Webauftritt einer Krankenpflegeschule	1	0,1
	Webauftritt eines Landesinstituts	1	0,1
	Webauftritt einer Stiftung	1	0,1
	Webauftritt einer Studentengemeinde	1	0,1
	Webauftritt einer Studentenverbindung	1	0,1
	Webauftritt einer Universität	1	0,1

Tabelle 12.2: Die Hypertextsorten der sechs ermittelten Hypertexttypen im Überblick

	Hypertextknotentyp bzw. Hypertextknotensorte	Subtypen	Frequenz	Prozent
1.	Seite/Abschnitt	20	119	15,9
2.	Folie	6	80	10,7
3.	Organisatorische Kerndaten einer Lehrveranstaltung	4	46	6,1
4.	Abstract	6	42	5,6
5.	Foto	—	29	3,9
6.	Einstiegsseite	—	24	3,2
7.	Pressemitteilung	—	24	3,2
8.	Berufliche Homepage eines Hochschulangehörigen	2	18	2,4
9.	Redaktioneller Artikel eines Publikationsorgans	6	16	2,1
10.	Primäre Navigationshilfe	—	13	1,7
11.	Kurzdarstellung eines Arbeitsgebiets (einer Organisationseinheit)	—	12	1,6
12.	Anleitung bzw. Benutzungshinweise	—	10	1,3
13.	Hotlist	—	10	1,3
14.	Persönliche Homepage eines Wissenschaftlers	—	10	1,3
15.	Übungsaufgaben (einer Lehrveranstaltung)	—	10	1,3
16.	Vorlesungsverzeichnis	3	10	1,3
17.	Ablaufplan bzw. Programm (einer Lehrveranstaltung)	—	9	1,2
18.	Publikationsliste	2	9	1,2
19.	*Zuordnung nicht möglich*	—	9	1,2
20.	Kopfzeile	—	8	1,1
21.	Studienhinweise	3	8	1,1
22.	Unterrichtsmaterialien (für die Schule)	—	8	1,1
23.	Ankündigung	—	7	0,9
24.	Fotogalerie	—	7	0,9
25.	Ausstellungsobjekt (eines virtuellen Museums)	—	6	0,8
26.	Bibliothekskatalog (Datensatz)	—	6	0,8
27.	E-Mail	—	6	0,8
28.	Kontaktinformationen	—	6	0,8
29.	Kurzdarstellung einer Organisationseinheit (Funktionen und Kontaktinformationen)	—	6	0,8
30.	Kurzdarstellung eines Dienstleistungsspektrums (im Technologietransfer-Kontext)	—	5	0,7
31.	Lexikoneintrag	—	5	0,7
32.	Lösungen von Übungsaufgaben (einer Lehrveranstaltung)	—	5	0,7
33.	Mitarbeiterverzeichnis	—	5	0,7
34.	Programmcode, Quelltext	—	5	0,7
35.	Studierendenstatistik	—	5	0,7
36.	Abgeschlossene und/oder angebotene Haus- und Abschlussarbeiten	—	4	0,5
37.	Aktuelle Meldung/Information (keine Pressemitteilung)	—	4	0,5
38.	Bibliografie	—	4	0,5
39.	Einladung	—	4	0,5
40.	Inhaltsverzeichnis	3	4	0,5
41.	Klausur- und Prüfungstermine	—	4	0,5
42.	Medizinische Diagnoseprozedur	—	4	0,5
43.	Statistische Daten (maschinell generiert)	—	4	0,5
44.	„under Construction“-Hinweis	—	4	0,5
45.	Verteiler	—	4	0,5
46.	Index bzw. Dateiliste (vom Webserver generiert)	—	4	0,5
47.	Aufgabenstellung für eine Haus- oder Abschlussarbeit	—	3	0,4
48.	Bericht zu einer Konferenz/Tagung/Veranstaltung	—	3	0,4
49.	Download-Liste (multimediale Ressourcen)	—	3	0,4
50.	Kommentar einer Lehrveranstaltung	—	3	0,4
51.	Kurzdarstellung der historischen Entwicklung einer Institution	—	3	0,4
52.	Kurzdarstellung einer Organisationseinheit/Institution (Profil/Porträt)	—	3	0,4
53.	Protokoll	—	3	0,4
54.	Q/A (Frage und Antwort)	—	3	0,4
55.	Regelung, Ordnung, Gesetz, juristischer Text	—	3	0,4
56.	Rezension	—	3	0,4
57.	Studiengangsbeschreibung	—	3	0,4
58.	Studienordnung	—	3	0,4
59.	Technische Daten/Spezifikation (Hard- oder Software)	—	3	0,4
60.	Biografie	—	2	0,3
61.	Ergebnisse einer Sportveranstaltung	—	2	0,3
62.	Expose (einer Qualifikationsarbeit)	—	2	0,3
63.	Fußzeile	—	2	0,3
64.	Glossareintrag	—	2	0,3
65.	Liste von Studiengängen	—	2	0,3
66.	Medizinisches Diagnosebeispiel	—	2	0,3
67.	Spezifikationstabelle	—	2	0,3
68.	Splash-Seite	—	2	0,3
69.	Stundenplan (eines Studiengangs)	—	2	0,3
70.	Vortrag/Rede (in Schriftform)	—	2	0,3
71.	Wissenschaftlicher Artikel	—	2	0,3
72.	Zeitungsartikel (eingescannt)	—	2	0,3
73.	Zugriffsstatistik (maschinell generiert)	—	2	0,3
74.	Anmeldeformular	—	1	0,1
75.	Anreiseinformationen, Wegbeschreibung	—	1	0,1

Tabelle 12.3: Die ermittelten Hypertextknotentypen bzw. -sorten im Überblick (Teil 1)

	Hypertextknotentyp bzw. Hypertextknotensorte	Subtypen	Frequenz	Prozent
76.	Antrag (an ein Gremium)	—	1	0,1
77.	Antragsformular	—	1	0,1
78.	Betriebsärztliche Informationen	—	1	0,1
79.	Bibliothekssystematik (Ausschnitt)	—	1	0,1
80.	Daten und Fristen (eines Semesters)	—	1	0,1
81.	Denksportaufgabe	—	1	0,1
82.	Einverständniserklärung	—	1	0,1
83.	Episodenliste einer Fernsehserie	—	1	0,1
84.	Errataliste	—	1	0,1
85.	Folien (Thumbnails; mit interaktiven Beispielen)	—	1	0,1
86.	Gästebuch	—	1	0,1
87.	Gedenkkalender	—	1	0,1
88.	Glossar	—	1	0,1
89.	Image-Map	—	1	0,1
90.	Stadtplan	—	1	0,1
91.	Kerndaten eines historischen Bauwerks	—	1	0,1
92.	Kerndaten eines Pflanzentyps	—	1	0,1
93.	Kinoprogramm	—	1	0,1
94.	Klausurergebnisse	—	1	0,1
95.	Kleinanzeige	—	1	0,1
96.	Kochrezept	—	1	0,1
97.	Kursverzeichnis (eines Vereins)	—	1	0,1
98.	Lageplan, Karte, Wegbeschreibung	—	1	0,1
99.	Lehrveranstaltungsskripte (Liste)	—	1	0,1
100.	Lexikoneinträge (Liste)	—	1	0,1
101.	Liste universitätsinterner Projekte	—	1	0,1
102.	Liste von Promotionen an einer Organisationseinheit	—	1	0,1
103.	Neue oder modifizierte Dokumente einer Website (Liste)	—	1	0,1
104.	Newsgroup (Liste der Postings)	—	1	0,1
105.	Preisliste	—	1	0,1
106.	Pressemitteilungen (Liste)	—	1	0,1
107.	Private Homepage eines Schülers/einer Schülerin	—	1	0,1
108.	Redaktion eines Webauftritts	—	1	0,1
109.	Reisetagebuch (Einzeleintrag)	—	1	0,1
110.	Rundschreiben	—	1	0,1
111.	Suchformular	—	1	0,1
112.	Teilnehmerliste (einer Lehrveranstaltung)	—	1	0,1
113.	Universitätszeitung (Übersicht über eine Ausgabe)	—	1	0,1
114.	Wahlergebnis	—	1	0,1

Tabelle 12.4: Die ermittelten Hypertextknotentypen bzw. -sorten im Überblick (Teil 2)

bene erfasst werden. Die Annahme des *Typs* Hyperlinkliste erscheint aufgrund der in der Stichprobe vertretenen Knoten dieser Kategorie nicht notwendig zu sein, da schließlich auf die *Hotlist* lediglich zehn und auf den *Verteiler* nur vier Vorkommen entfallen. Die Stichprobe wird dominiert von den Hypertextknotentypen *Seite/Abschnitt* (15,9%), *Folie* (10,7%), *organisatorische Kerndaten einer Lehrveranstaltung* (6,1%) und *Abstract* (5,3%). Obwohl Abschnitt 12.7 genauer auf dieses Inventar eingeht, erscheint eine Erläuterung des Hypertextknotentyps *Seite/Abschnitt* notwendig: Dieser subsumiert insgesamt 20 Hypertextknotensorten, deren zugehörige Instanzen vornehmlich aus Fließtext bestehen und beinahe ausnahmslos die gemeinsame Eigenschaft besitzen, gerade *nicht* für die Veröffentlichung im WWW angefertigt worden zu sein. Diese Hypertextknotensorten korrespondieren in nahezu allen Fällen unmittelbar mit traditionellen Textsorten (bzw. Teiltextsorten), die von linear sequenzierten Textexemplaren instanziiert werden, die – oftmals vollautomatisch – nach HTML konvertiert wurden. Die Autoren der in Abschnitt 4.4 diskutierten Arbeiten weisen darauf hin, dass sich die Zuordnung eines gegebenen Dokuments zu einem Genre in vielen Fällen problematisch gestaltete, da z. B. kein unmittelbares traditionelles Pendant oder keine erkennbare kommunikative Funktion vorlag (vgl. ausführlich hierzu Abschnitt 12.8.2). Diese Problematik kann von der in diesem Kapitel vorgestellten Analyse nur bedingt bestätigt werden: Zwar konnten neun HTML-Dokumente nicht zugeordnet werden (1,2%), doch lagen

bei einigen dieser Vorkommen *multiple* kommunikative Funktionen vor, die keine *eindeutige* Zuordnung erlaubten. Diese Analyse wurde vielmehr der Problematik ausgesetzt, dass mehrere der 750 Dokumente durchaus Entsprechungen im Bereich der traditionellen Textsorten besitzen, doch handelt es sich dabei um hochgradig spezialisierte Textsorten aus unterschiedlichen wissenschaftlichen Fachdisziplinen (vgl. Abschnitt 12.7.7).

12.5 Die publizierenden Institutionen und Einrichtungen

Nachfolgend werden zunächst die Institutionen und Einrichtungen thematisiert, die die 750 Dokumente publizieren. Neben zahlreichen Organisationseinheiten von Hochschulen (Abschnitt 12.5.2) handelt es sich dabei auch um einige außeruniversitäre Institutionen und Forschungseinrichtungen (Abschnitt 12.5.1). Die Abbildungen 12.1 und 12.2 stellen diese gemeinsam mit den korrespondierenden Dokumentfrequenzen im Überblick dar.[4]

12.5.1 Außeruniversitäre Einrichtungen

Die Webserver von Hochschulen werden nicht ausschließlich von universitären Organisationseinheiten und ihren Angehörigen zur Publikation von Inhalten genutzt. Insgesamt 30 HTML-Dokumente der Stichprobe (4%) können den Webauftritten außeruniversitärer Einrichtungen zugeordnet werden (vgl. Abbildung 12.1). Hierzu gehört z. B. die „Zentrale Evaluations- und Akkreditierungsagentur Hannover“ (D 560), die eine gemeinsame Einrichtung der niedersächsischen Hochschulen ist. Das sehr umfangreiche Angebot des „Bundesamts für Sicherheit in der Informationstechnik“ (D 147, D 181, D 254, D 407, D 697) befindet sich auf dem zentralen Webserver der HU Berlin. Ein Rechner an der Universität Bremen beinhaltet das Webangebot des „Landesinstitut für Schule – Bremen“ (D 58, D 125, D 558), und das Angebot der „Forschungsanstalt für Waldökologie und Forstwirtschaft Rheinland-Pfalz“ befindet sich auf dem zentralen Webserver der Universität Kaiserslautern (D 451). Daneben existieren in der Stichprobe auch Dokumente der Webauftritte von Schulen, Studentengemeinden, Studentenverbindungen, Studentenwohnheimen, eingetragenen Vereinen und wissenschaftlichen Arbeitskreisen sowie Fachverbänden.

Die Einrichtungen können unterschieden werden hinsichtlich ihrer Größe, ihres juristischen Status und ihrer jeweiligen „Nähe“ zur Universität. Diese Differenzierung erklärt, weshalb einige Dokumente, die von außeruniversitären Einrichtungen publiziert wurden, zum Zeitpunkt der Anfertigung dieser Arbeit nicht länger ein Bestandteil des jeweiligen universitären Webangebots waren: Die Website des bereits angesprochenen Bundesamtes (im Korpus: `http://www.hu-berlin.de/bsi/`) ist mittlerweile unter `http://www.bsi.de` verfügbar und auch die Webauftritte einiger Schulen besitzen inzwischen eigenständige Domains bzw. URLs, die sie nicht länger als Bestandteil eines universitären Webangebots kennzeichnen. Mitte der neunziger Jahre waren die Möglichkeiten zur Einrichtung eines Webangebots sehr begrenzt, weshalb sich außeruniversitäre Einrichtungen diesbezüglich oftmals an

[4] Die in den Abbildungen 12.1 und 12.2 in kursiver Schrift dargestellten Etiketten bezeichnen Kategorien, die in der Stichprobe nicht unmittelbar als publizierende Einrichtungen ausgeprägt sind und lediglich strukturierenden Charakter besitzen. Die bei diesen Etiketten notierten Frequenzangaben wurden über die Summe der Frequenzen der jeweiligen Subkategorien gebildet.

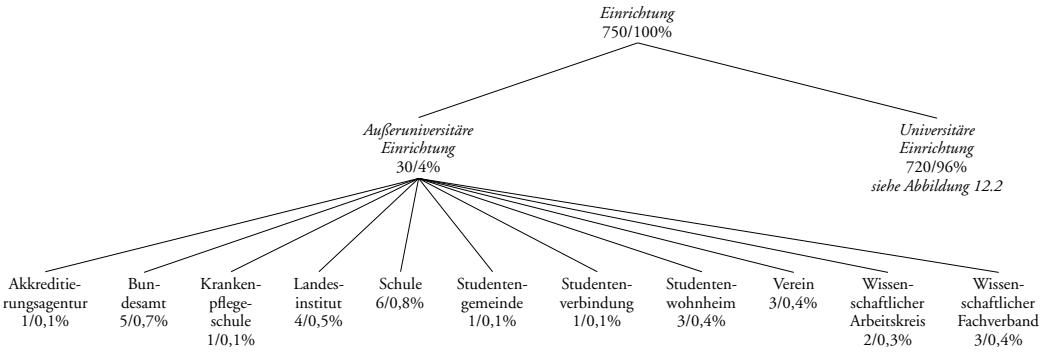


Abbildung 12.1: Die publizierenden Einrichtungen der 750 Dokumente (außeruniversitäre Einrichtungen)

die Rechenzentren der lokalen Hochschulen gewandt haben. Mittlerweile existieren zahllose Dienstleister, die Domainnamen und zugehörigen Webspace anbieten. Einige in der Stichprobe vertretenen Webangebote außeruniversitärer Einrichtungen sind auf eigenständige Webserver „umgezogen“, da eine URL wie z. B. `http://www.wilhelm-gym.de` attraktiver und auch memorierbarer ist als `http://www.tu-bs.de/schulen/Wilhelm-Gym_BS/` und möglicherweise komfortablere Bedingungen hinsichtlich der Pflege des Webangebots vorliegen (vgl. Abschnitt 4.6.10). Gerade von größeren Einrichtungen wie z. B. Bundesämtern oder Schulen wird erwartet, dass diese einen offiziellen Webauftritt mit einer in gewisser Hinsicht repräsentativen URL besitzen, und in gleicher Weise existiert dieses Interesse auch innerhalb der Leitungsspitzen der jeweiligen Einrichtungen. Bei universitätsnahen Einrichtungen wie z. B. losen wissenschaftlichen Arbeitskreisen oder sehr kleinen Fachverbänden hingegen ist die URL des Webauftritts von eher sekundärer Bedeutung, weshalb häufig auf die Etablierung eines eigenständigen Webangebots und den hierdurch hervorgerufenen finanziellen Aufwand verzichtet wird. Für die zukünftige Entwicklung ist damit zu rechnen, dass die Webserver von Hochschulen nach wie vor die Websites außeruniversitärer Einrichtungen beinhalten, diese werden sich jedoch auf kleine und universitätsnahe Institutionen beschränken.

12.5.2 Universitäre Einrichtungen

Universitäre Einrichtungen publizieren 96% der in der Stichprobe enthaltenen Dokumente (vgl. Abbildung 12.2). Es kann zwischen zentralen und dezentralen Organisationseinheiten sowie Einrichtungen aus dem Bereich Leitung und Verwaltung unterschieden werden. Für die Dokumente, die dem zentralen Angebot eines universitären Webauftritts zugehörig sind, wird eine vierte Kategorie angenommen. Die Zuordnung eines Dokuments und somit auch die iterative Anfertigung der Strukturtypologie erfolgte durch eine Analyse des jeweiligen Dokuments sowie eine Betrachtung der unmittelbar bei- und übergeordneten Knoten.[5] Ein spezifisches Vorkommen wurde notiert, sobald die einen Knoten unmittelbar publizierende

[5] Mit Abbildung 12.2 ist kein Anspruch auf Allgemeingültigkeit im Hinblick auf den generischen Aufbau deutscher Hochschulen und Universitäten verbunden. Die Kategorien dieser Typologie wurden vornehmlich aus den Organisationseinheiten abgeleitet, die in der Stichprobe vertreten sind (vgl. Kapitel 13).

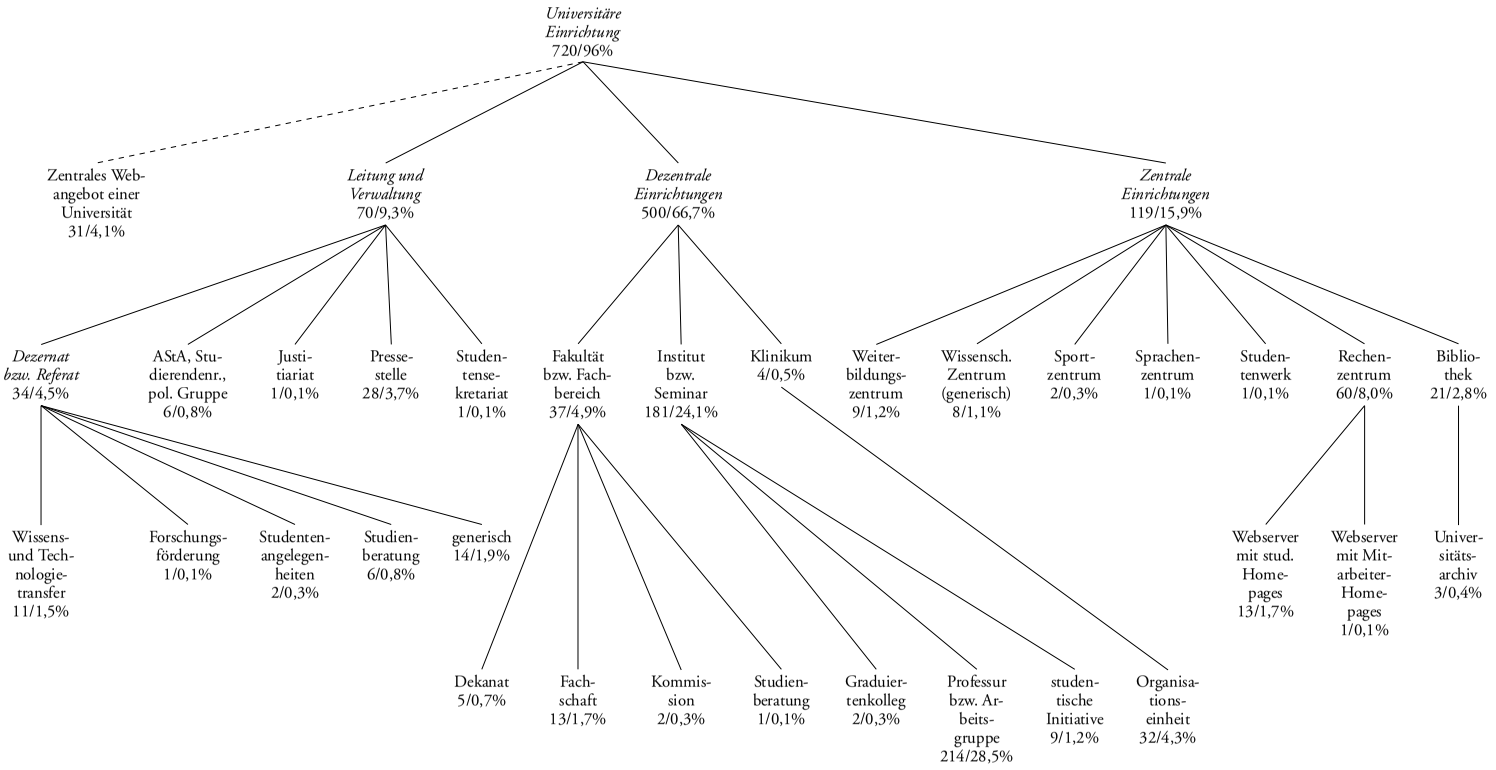


Abbildung 12.2: Die publizierenden Einrichtungen der 750 Dokumente (universitäre Einrichtungen)

Einrichtung bestimmt wurde. Einen Sonderfall stellen die Dokumente des zentralen Webangebots dar, die zwar typischerweise keine entsprechende Kennzeichnung der publizierenden Institution aufweisen, jedoch über Indikatoren wie z. B. den Aufbau der URL, ein konsistentes Webdesign und die Informationsstrukturierung zugeordnet werden konnten.

Etwa zwei Drittel der Dokumente entfallen auf dezentrale Einrichtungen; 52,6% der Knoten sind Bestandteile der Webauftritte von Instituten bzw. Seminaren (24,1%) sowie Professuren oder Arbeitsgruppen (28,5%; vgl. Abbildung 12.2 sowie Fußnote 4, S. 536). Neben diesen hochfrequenten Einrichtungen sind das Rechenzentrum mit 8,0%, die Fakultät bzw. der Fachbereich mit 4,9%, Organisationseinheiten eines Klinikums mit 4,3%, das zentrale Webangebot mit 4,1% und die Pressestelle mit 3,7% der Dokumente vertreten. Die verbleibenden Organisationseinheiten publizieren jeweils weniger als 3% der Dokumente.

12.6 Die übergeordneten Hypertexttypen und -sorten

Nachfolgend werden die Hypertexttypen und -sorten thematisiert, denen die 750 Dokumente zugehörig sind. Auf eine *vollständig* separierte Darstellung der beiden Ebenen wird jedoch verzichtet, da es bei einer Betrachtung der Hypertextsorten notwendig ist, auf die von ihnen umfassten Hypertextknotensorten einzugehen (und umgekehrt). Zunächst werden die hochfrequenten Hypertexttypen und -sorten diskutiert (Abschnitte 12.6.1 bis 12.6.5), woraufhin einige der niedrigfrequenten Hypertextsorten thematisiert werden (Abschnitt 12.6.6).[6]

12.6.1 Der Hypertexttyp Webauftritt einer Organisationseinheit

Der *Webauftritt einer Organisationseinheit* ist mit insgesamt 213 Vorkommen (28,4%) der mit Abstand hochfrequenteste Hypertexttyp, für den 24 untergeordnete Hypertextsorten ermittelt wurden (vgl. Tabelle 12.1), die jeweils den Webauftritten spezifischer Organisationseinheiten entsprechen (vgl. Tabelle 12.2). Hierbei wird eine relativ weit gefasste Lesart des Konzepts „Organisationseinheit einer Hochschule“ angewendet, so dass nicht nur z. B. das Institut, die Fakultät, das Rechenzentrum und die Bibliothek, sondern auch die Fachschaft, die Studierendenvertretung, das Dekanat, ein universitätsinternes Projekt und eine studentische Initiative unter diesem Terminus subsumiert werden.[7] Tabelle 12.2 verdeutlicht, dass die Hypertextsorte *Webauftritt einer Professur bzw. Arbeitsgruppe* die Stichprobe mit 84 Vorkommen (11,2%) dominiert, an zweiter Stelle folgt der *Webauftritt eines Instituts bzw. Seminars* (3,7%) und an dritter Stelle der *Webauftritt eines Projekts oder Projektverbundes* (2,8%).

Webauftritt einer Professur bzw. Arbeitsgruppe

Tabelle 12.5 stellt die Hypertextknotensorten dar, die Dokumenten zugeordnet wurden, die dem *Webauftritt einer Professur bzw. Arbeitsgruppe* zugehörig sind. Auffällig ist die mit 40 Kategorien sehr ausgeprägte Varianz, die darauf schließen lässt, dass individuelle Lösungen der

[6] Aus Platzgründen kann eine detaillierte Charakterisierung aller Ergebnisse an dieser Stelle nicht erfolgen.

[7] Die Hypertextsorten *Webauftritt eines spezifischen Dezernats bzw. Referats der Hochschulverwaltung* und *Webauftritt eines wissenschaftlichen Zentrums (generisch)*, die dem *Webauftritt einer Organisationseinheit* zuzuordnen sind, nehmen eine Sonderstellung ein (vgl. Tabelle 12.2), da sie sich auf spezifischere Einheiten beziehen. Zur Reduktion der Darstellungskomplexität wurde auf eine Binnendifferenzierung verzichtet (vgl. Kapitel 13).

	Hypertextknotensorte	Frequenz	Prozent
1.	Persönliche Homepage eines Wissenschaftlers	8	1,1
2.	Ablaufplan bzw. Programm (einer Lehrveranstaltung)	7	0,9
3.	Folie (Powerpoint, exportiert nach HTML)	7	0,9
4.	Kurzdarstellung eines Arbeitsgebiets (einer Organisationseinheit)	5	0,7
5.	Abstract eines Forschungsprojekts	4	0,5
6.	Berufliche Homepage eines Wissenschaftlers	4	0,5
7.	Vorlesungsverzeichnis (eines Studiengangs)	4	0,5
8.	Einstiegsseite	3	0,4
9.	Mitarbeiterverzeichnis	3	0,4
10.	Primäre Navigationshilfe	3	0,4
11.	Abgeschlossene und/oder angebotene Haus- und Abschlussarbeiten	2	0,3
12.	Berufliche Homepage eines Hochschulangehörigen	2	0,3
13.	Download-Liste (multimediale Ressourcen)	2	0,3
14.	Hotlist	2	0,3
15.	Kopfzeile	2	0,3
16.	Publikationsliste (einer Organisationseinheit)	2	0,3
17.	Abstract einer Dissertation	1	0,1
18.	Aktuelle Meldung/Information (keine Pressemitteilung)	1	0,1
19.	Bericht zu einer Konferenz/Tagung/Veranstaltung	1	0,1
20.	Einladung	1	0,1
21.	Expose (einer Qualifikationsarbeit)	1	0,1
22.	Folie (Frontpage; exportiert nach HTML)	1	0,1
23.	Foto	1	0,1
24.	Fotogalerie	1	0,1
25.	Index bzw. Dateiliste (vom Webserver generiert)	1	0,1
26.	Inhaltsverzeichnis (mit Download-Möglichkeit des Volltextes)	1	0,1
27.	*Zuordnung nicht möglich*	1	0,1
28.	Klausur- und Prüfungstermine	1	0,1
29.	Kurzdarstellung eines Dienstleistungsspektrums (im Technologietransfer-Kontext)	1	0,1
30.	Lehrveranstaltungsskripte (Liste)	1	0,1
31.	Liste von Promotionen an einer Organisationseinheit	1	0,1
32.	Organisatorische Kerndaten einer Lehrveranstaltung	1	0,1
33.	Organisatorische Kerndaten einer Lehrveranstaltung (Liste mehrerer Kurse)	1	0,1
34.	Pressemitteilung	1	0,1
35.	Programmcode, Quelltext	1	0,1
36.	Protokoll	1	0,1
37.	Publikationsliste (eines Wissenschaftlers)	1	0,1
38.	Seite/Abschnitt eines wissenschaftlichen Artikels (eingescannt)	1	0,1
39.	Studienordnung	1	0,1
40.	Vorlesungsverzeichnis (eines Dozenten)	1	0,1
	Σ	84	11,2

Tabelle 12.5: Die innerhalb der Hypertextsorte *Webauftritt einer Professur bzw. Arbeitsgruppe* verwendeten Hypertextknotensorten

Informationsstrukturierung und der Auswahl zu publizierender Inhalte existieren.[8] Die hier präsentierten Tabellen beziehen sich auf die *unmittelbar* zugeordneten Hypertextknotentypen und -sorten. Wenn ein Knoten z. B. der Hypertextsorte *Webangebot einer Lehrveranstaltung* zugewiesen wurde, kann die Instanz dieser Hypertextsorte ihrerseits einer übergeordneten Hypertextsorte zugeordnet werden, die oftmals dem *Webauftritt einer Professur bzw. Arbeitsgruppe* entspricht. Derartige Relationen, die zwischen zwei oder mehr Instanzen von Hypertextsorten beobachtet werden können, wurden in dieser Analyse nicht berücksichtigt.

Das Spektrum von Hypertextknotentypen und -sorten deckt nahezu alle Arbeitsgebiete ab, mit denen die Angehörigen einer Professur bzw. Arbeitsgruppe im Alltag beschäftigt sind: Interessanterweise steht die (auch als Hypertextknotensorte instanziierbare, vgl. Kapitel 5) *persönliche Homepage eines Wissenschaftlers* mit acht Vorkommen an erster Stelle, woraufhin lehrveranstaltungsbezogene Hypertextknotensorten folgen (*Ablaufplan bzw. Programm einer*

[8] Abschnitt 12.7 geht in detaillierter Form auf die Hypertextknotentypen und -sorten ein. Für einige Hypertextknotensorten wurden spezifische Notationsrichtlinien bezüglich der übergeordneten Hypertextsorte etabliert (z. B. im Hinblick auf die unterschiedlichen Typen von Folien), die einen Teil der Varianzen erklären.

Lehrveranstaltung und *Folie*). Die *Kurzdarstellung eines Arbeitsgebiets* umfasst typischerweise eine knappe Vorstellung der Forschungsschwerpunkte, wobei in einigen Fällen auch auf rekurrente Lehrveranstaltungsangebote hingewiesen wird. Diesem Themenkomplex widmen sich auch Hypertextknotensorten wie z. B. *Vorlesungsverzeichnis (eines Studiengangs)*, *abgeschlossene und/oder angebotene Haus- und Abschlussarbeiten* sowie *Klausur- und Prüfungstermine*. Zusätzlich sei auf die Instanz der Hypertextknotensorte *Kurzdarstellung eines Dienstleistungsspektrums* hingewiesen, die sich an die Privatwirtschaft richtet und Wissens- und Technologietransferdienstleistungen aufzeigt (vgl. Abschnitt 11.4.6).

Webauftritt eines Instituts bzw. Seminars

Professuren bzw. Arbeitsgruppen sind Bestandteile der Organisationseinheit Institut bzw. Seminar. Die 24 Hypertextknotensorten von insgesamt 28 HTML-Dokumenten, die dem *Webauftritt eines Instituts bzw. Seminars* zugeordnet wurden, sind in Tabelle 12.6 dargestellt. Die Vorkommen verdeutlichen mehrere Unterschiede zu der zuvor diskutierten Hypertextsorte, die sich z. B. auf das vollständige Fehlen wissenstransmittierender Hypertextknotensorten beziehen. Im Zentrum stehen vielmehr lehrveranstaltungsorganisierende und allgemeine studiumsbezogene Hypertextknotensorten wie z. B. *organisatorische Kerndaten einer Lehrveranstaltung*, *Studienhinweise*, *Studienordnung* und *Vorlesungsverzeichnis*. Die wissenstransmittierenden Hypertextsorten werden primär von denjenigen Organisationseinheiten instanziiert, die für die Durchführung von Lehrveranstaltungen verantwortlich zeichnen.

	Hypertextknotensorte	Frequenz	Prozent
1.	Berufliche Homepage eines Wissenschaftlers	2	0,3
2.	Einladung	2	0,3
3.	Primäre Navigationshilfe	2	0,3
4.	Publikationsliste (einer Organisationseinheit)	2	0,3
5.	Abstract eines Forschungsprojekts	1	0,1
6.	Anreiseinformationen, Wegbeschreibung	1	0,1
7.	Berufliche Homepage eines Hochschulangehörigen	1	0,1
8.	Bibliografie	1	0,1
9.	Einstiegsseite	1	0,1
10.	Gästebuch	1	0,1
11.	Hotlist	1	0,1
12.	*Zuordnung nicht möglich*	1	0,1
13.	Kurzdarstellung der historischen Entwicklung einer Institution	1	0,1
14.	Kurzdarstellung einer Organisationseinheit (Funktionen und Kontaktinformationen)	1	0,1
15.	Kurzdarstellung eines Arbeitsgebiets (einer Organisationseinheit)	1	0,1
16.	Organisatorische Kerndaten einer Lehrveranstaltung	1	0,1
17.	Persönliche Homepage eines Wissenschaftlers	1	0,1
18.	Publikationsliste (eines Wissenschaftlers)	1	0,1
19.	Studienhinweise (Grundstudium)	1	0,1
20.	Studienordnung	1	0,1
21.	Studierendenstatistik	1	0,1
22.	Verteiler	1	0,1
23.	Vorlesungsverzeichnis (einer Organisationseinheit)	1	0,1
24.	Zugriffsstatistik (maschinell generiert)	1	0,1
	Σ	28	3,7

Tabelle 12.6: Die innerhalb der Hypertextsorte *Webauftritt eines Instituts bzw. Seminars* verwendeten Hypertextknotensorten

Während der *Webauftritt einer Professur bzw. Arbeitsgruppe* Vorkommen der Hypertextknotensorte *persönliche Homepage eines Wissenschaftlers* umfasst, werden innerhalb des *Webauftritts eines Instituts bzw. Seminars* eher *berufliche Homepages von Hochschulangehörigen* ver-

öffentlicht. Von einer bestimmten Organisationseinheit publizierte Exemplare dieser Hypertextknotensorte zeichnen sich durch eine konsistente und übergreifende Gestaltung aus und basieren meist auf einer gemeinsamen Vorlage (vgl. Abschnitt 12.7.5).

Webauftritt eines Projekts oder Projektverbundes

Unter dem Etikett *Webauftritt eines Projekts oder Projektverbundes* werden sowohl die Websites einzelner Forschungsprojekte als auch die Angebote von Projektverbünden subsumiert (z. B. über DFG-Mittel finanzierte Sonderforschungsbereiche und Forschergruppen).[9] Dieser Hypertextsorte besitzt 21 Vorkommen (2,8%; vgl. Tabelle 12.7). Bei mehr als der Hälfte dieser Dokumente handelt es sich um Folien, die für eine Veröffentlichung im WWW nach HTML konvertiert und ursprünglichen im Rahmen von Vorträgen eingesetzt wurden. Das bzw. die Forschungsprojekte werden in Form eines einzelnen oder mehrerer *Abstracts* vorgestellt, die den Forschungsgegenstand und die Projektziele skizzieren und in nahezu allen Fällen Publikationslisten enthalten (vgl. Abschnitt 12.7.4). Tabelle 12.8 zeigt die Organisationseinheiten, die die 21 HTML-Dokumente publizieren. Erwartungsgemäß sind Webauftritte von Projekten vornehmlich Bestandteile der Websites von Instituten bzw. Seminaren und Fakultäten bzw. Fachbereichen. Daneben existieren in der Stichprobe auch fünf Dokumente, die aus den Webauftritten von Projekten stammen, die an Rechenzentren durchgeführt werden.

	Hypertextknotensorte	Frequenz	Prozent
1.	Folie (Powerpoint, exportiert nach HTML)	11	1,5
2.	Abstract eines Forschungsprojekts	3	0,4
3.	Abstracts mehrerer Forschungsprojekte	1	0,1
4.	Folie (HTML)	1	0,1
5.	Kopfzeile	1	0,1
6.	Kurzdarstellung einer Organisationseinheit/Institution (Profil/Porträt)	1	0,1
7.	Primäre Navigationshilfe	1	0,1
8.	Publikationsliste (einer Organisationseinheit)	1	0,1
9.	Redaktioneller Artikel eines Publikationsorgans (Universitätszeitung)	1	0,1
	Σ	21	2,8

Tabelle 12.7: Die innerhalb der Hypertextsorte *Webauftritt eines Projekts oder Projektverbundes* verwendeten Hypertextknotensorten

12.6.2 Der Hypertexttyp Webangebot einer Lehrveranstaltung

Der Hypertexttyp *Webangebot einer Lehrveranstaltung* umfasst vier Hypertextsorten, die mit unterschiedlichen Lehrveranstaltungstypen korrespondieren. Die Hypertextsorte *Webangebot einer regulären Lehrveranstaltung* (11,7%) bezieht sich auf Veranstaltungen wie z. B. Vorlesungen, Seminare, Übungen und Praktika. Das *Webangebot eines Doktoranden- oder Mitarbeiterkolloquiums* (0,8%) betrifft einen Veranstaltungstyp, in dessen Sitzungen Vorträge von Doktoranden sowie internen oder externen Wissenschaftlern gehalten werden. Oftmals wird

[9] Streng genommen müsste diesbezüglich eine Differenzierung vorgenommen werden, da der *Webauftritt eines Projektverbundes* mehrere Instanzen der Hypertextsorte *Webauftritt eines Projekts* beinhaltet, die sich auf ein gemeinsames Forschungsthema beziehen.

	Organisationseinheit	Frequenz	Prozent
1.	Dezentrale Einrichtung: Institut bzw. Seminar	8	1,1
2.	Zentrale Einrichtung: Rechenzentrum	5	0,7
3.	Dezentrale Einrichtung: Fakultät bzw. Fachbereich	4	0,5
4.	Dezentrale Einrichtung: Professur bzw. Arbeitsgruppe	1	0,1
5.	Dezentrale Einrichtung: Organisationseinheit eines Klinikums	1	0,1
6.	Leitung und Verwaltung: Dezernat bzw. Referat (generisch)	1	0,1
7.	Zentrale Einrichtung: Weiterbildungszentrum	1	0,1
	Σ	21	2,8

Tabelle 12.8: Die Organisationseinheiten, die Instanzen der Hypertextsorte *Webauftritt eines Projekts oder Projektverbundes* publizieren

auch die Öffentlichkeit eingeladen. Das *Webangebot eines Weiterbildungskurses* bezieht sich auf hochschulinterne Fortbildungsveranstaltungen, die an Mitarbeiter gerichtet sind und sehr spezifische Themen diskutieren (z. B. hochschuldidaktische Konzeptionierungen oder das Anfertigen von Präsentationen). Das *Webangebot einer regulären Lehrveranstaltung (E-Learning-Angebot)* ist eine Variante der eingangs genannten Hypertextsorte und beinhaltet umfangreiche Sammlungen von Lehrmaterialien, die von den Studierenden primär über das WWW zu rezipieren sind, d. h. es finden entweder keine oder nur sehr wenige Präsenzphasen statt. Die vier Typen besitzen jeweils eindeutig differenzierbare Inhalte und kommunikative Funktionen, wodurch die Annahme eigenständiger Hypertextsorten gerechtfertigt wird.

Tabelle 12.9 stellt die Hypertextknotensorten der vier Hypertextsorten dar. Zunächst fällt auf, dass das *Webangebot einer regulären Lehrveranstaltung* 27 verschiedene Hypertextknotensorten umfasst.[10] Hierzu zählen insbesondere unterschiedliche Typen von Folien, die meist im Rahmen von Präsentationen innerhalb einer Sitzung eingesetzt werden und den Studierenden anschließend als HTML-Dokumente zur Verfügung gestellt werden. Das WWW wird jedoch nicht nur als digitaler Semesterapparat benutzt: Hypertextknotensorten wie z. B. *Übungsaufgaben*, *Lösungen von Übungsaufgaben*, *Teilnehmerliste* und *Klausurergebnisse* zeigen, dass das Webangebot einer Lehrveranstaltung oftmals sehr flexibel eingesetzt wird und als zentrale Anlaufstelle für alle veranstaltungsrelevanten Fragen dient. Die 27 Hypertextknotensorten enthalten auch einige fachspezifische Textsorten (z. B. *Biografie*, *Lexikoneintrag* und *Programmcode, Quelltext*). Die drei verbleibenden Hypertextsorten sind mit sechs bzw. vier Vorkommen als niedrigfrequent zu bezeichnen und umfassen dementsprechend lediglich maximal drei Hypertextknotensorten. Besonders auffällig sind die beiden Hypertextknotensorten *Ablaufplan bzw. Programm* und *Abstract eines Vortrags*, die innerhalb der Hypertextsorte *Webangebot eines Doktoranden- oder Mitarbeiterkolloquiums* verwendet werden. Die beiden Ablaufpläne umfassen tabellarische Aufstellungen der Sitzungstermine und Vortragenden, die ihre Vorträge zusätzlich mit einer Zusammenfassung vorstellen und nach der Durchführung in Form von HTML-Dokumenten zur Verfügung stellen. Die Hypertextknotensorte *Seite/Abschnitt eines E-Learning-Kurses* wird in Abschnitt 12.7.1 diskutiert.

Tabelle 12.10 zeigt die Organisationseinheiten, die die vier Hypertextsorten instanziieren. Fünf Vorkommen der Hypertextsorte *Webangebot eines Weiterbildungskurses* werden inner-

[10] Die *eingebettete* Hypertextsorte *Skript einer Lehrveranstaltung* kann ebenfalls als hochfrequente Konstituente des Hypertexttyps *Webangebot einer Lehrveranstaltung* verstanden werden (vgl. hierzu auch Abschnitt 12.6.1).

	Hypertextknotensorte	Frequenz	Prozent
Webangebot einer regulären Lehrveranstaltung			
1.	Folie (Powerpoint, exportiert nach HTML)	33	4,4
2.	Übungsaufgaben (einer Lehrveranstaltung)	10	1,3
3.	Einstiegsseite	9	1,2
4.	Lösungen von Übungsaufgaben (einer Lehrveranstaltung)	5	0,7
5.	Folie (HTML)	4	0,5
6.	Primäre Navigationshilfe	3	0,4
7.	Index bzw. Dateiliste (vom Webserver generiert)	2	0,3
8.	Programmcode, Quelltext	2	0,3
9.	*Zuordnung nicht möglich*	2	0,3
10.	Anleitung bzw. Benutzungshinweise	1	0,1
11.	Biografie	1	0,1
12.	Folie (Corel Presentations, exportiert nach HTML)	1	0,1
13.	Folie (Frontpage; exportiert nach HTML)	1	0,1
14.	Folien (Thumbnails; mit interaktiven Beispielen)	1	0,1
15.	Folie (Photoshop, exportiert nach HTML)	1	0,1
16.	Foto	1	0,1
17.	Fotogalerie	1	0,1
18.	Hotlist	1	0,1
19.	Klausurergebnisse	1	0,1
20.	Kommentar einer Lehrveranstaltung	1	0,1
21.	Kontaktinformationen	1	0,1
22.	Kopfzeile	1	0,1
23.	Lexikoneintrag	1	0,1
24.	Seite/Abschnitt einer studentischen Ausarbeitung	1	0,1
25.	Splash-Seite	1	0,1
26.	Studienhinweise (Grundstudium)	1	0,1
27.	Teilnehmerliste (einer Lehrveranstaltung)	1	0,1
	Σ	88	11,7
Webangebot eines Doktoranden- oder Mitarbeiterkolloquiums			
1.	Ablaufplan bzw. Programm (einer Lehrveranstaltung)	2	0,3
2.	Abstract eines Vortrags	2	0,3
3.	Folie (Powerpoint, exportiert nach HTML)	2	0,3
Webangebot eines Weiterbildungskurses			
1.	Folie (Powerpoint, exportiert nach HTML)	4	0,5
2.	Folie (HTML)	2	0,3
Webangebot einer regulären Lehrveranstaltung (E-Learning-Angebot)			
	Seite/Abschnitt eines E-Learning-Kurses	4	0,5

Tabelle 12.9: Die innerhalb der vier Hypertextsorten des Hypertexttyps *Webangebot einer Lehrveranstaltung* verwendeten Hypertextknotensorten

	Organisationseinheit	Frequenz	Prozent
1.	Dezentrale Einrichtung: Professur bzw. Arbeitsgruppe	51	6,8
2.	Dezentrale Einrichtung: Institut bzw. Seminar	36	4,8
3.	Dezentrale Einrichtung: Organisationseinheit eines Klinikums	5	0,7
4.	Zentrale Einrichtung: Rechenzentrum	5	0,7
5.	Dezentrale Einrichtung: Fakultät bzw. Fachbereich	2	0,3
6.	Zentrale Einrichtung: Wissenschaftliches Zentrum (generisch)	2	0,3
7.	Rechenzentrum: Webserver mit Mitarbeiter-Homepages	1	0,1
8.	Rechenzentrum: Webserver mit studentischen Homepages	1	0,1
9.	Zentrale Einrichtung: Bibliothek	1	0,1
	Σ	104	13,9

Tabelle 12.10: Die Organisationseinheiten, die Instanzen der vier Hypertextsorten des Hypertexttyps *Webangebot einer Lehrveranstaltung* publizieren

halb des Webauftritts eines Rechenzentrums publiziert. Vier Instanzen des *Webangebots eines Doktoranden- oder Mitarbeiterkolloquiums* sind Bestandteile des *Webauftritts eines Instituts bzw. Seminars*, die verbleibenden zwei Exemplare gehen auf eine Professur bzw. Arbeitsgruppe zurück. Die verbleibenden in Tabelle 12.10 aufgeführten Vorkommen beziehen sich auf das *Webangebot einer regulären Lehrveranstaltung*, wobei die E-Learning-Angebote von Professuren bzw. Arbeitsgruppen, einer Organisationseinheit eines Klinikums sowie von einem Institut bzw. Seminar publiziert werden.

12.6.3 Die Hypertextsorte Vorlesungsverzeichnis

Im Allgemeinen existieren zwei Varianten des Vorlesungsverzeichnisses: Die in der Regel als Buch verkaufte Version wird typischerweise von einem Dezernat bzw. Referat der Zentralverwaltung veröffentlicht und umfasst eine nach Fakultäten bzw. Fachbereichen sowie nach Studiengängen geordnete tabellarische Aufstellung aller Lehrveranstaltungen eines Semesters. Das kommentierte Vorlesungsverzeichnis wird von dezentralen Organisationseinheiten publiziert und beinhaltet die Lehrveranstaltungen eines oder mehrerer Studiengänge, die z. B. an einem Institut angeboten werden. Die einzelnen Einträge umfassen neben den organisatorischen Kerndaten eines Kurses auch einen Kommentar, der die Inhalte und Ziele der Veranstaltung thematisiert und eine Literaturliste enthält. Typischerweise erscheinen kommentierte Vorlesungsverzeichnisse einige Monate nach der Veröffentlichung des tabellarischen Vorlesungsverzeichnisses. Exemplare dieser beiden Textsortenvarianten erlauben es dem Rezipienten, sich einen Überblick über die geplanten Lehrveranstaltungen zu machen. Das tabellarische Vorlesungsverzeichnis wird meist dazu verwendet, eine initiale Liste zu besuchender Kurse zu konzipieren, woraufhin die detaillierteren und aktuelleren Informationen des kommentierten Vorlesungsverzeichnis eingesetzt werden, um Auskunft über Änderungen zu erhalten und einen endgültigen Stundenplan aufzustellen.

	Hypertextknotensorte	Frequenz	Prozent
1.	Organisatorische Kerndaten einer Lehrveranstaltung (mit Kommentar)	33	4,4
2.	Organisatorische Kerndaten einer Lehrveranstaltung	5	0,7
3.	Kommentar einer Lehrveranstaltung	2	0,3
4.	Daten und Fristen (eines Semesters)	1	0,1
5.	Einstiegsseite	1	0,1
6.	Organisatorische Kerndaten einer Lehrveranstaltung (mit Kommentar; Liste mehrerer Kurse)	1	0,1
7.	Vorlesungsverzeichnis (einer Organisationseinheit)	1	0,1
8.	Vorlesungsverzeichnis (eines Studiengangs)	1	0,1
	Σ	45	6,0

Tabelle 12.11: Die innerhalb der Hypertextsorte *Vorlesungsverzeichnis* verwendeten Hypertextknotensorten

Im Rahmen dieser Studie wird die Existenz einer Hypertextsorte *Vorlesungsverzeichnis* angenommen, die *keine* Binnenstrukturierung aufweist, weil die im Kontext dieser spezifischen Hypertextsorte zentrale Information der publizierenden Institution (Hochschule vs. dezentrale Organisationseinheit) in einer getrennten Analyseebene erfasst wird. Der Typ der Institution, die eine Instanz des *Vorlesungsverzeichnisses* veröffentlicht, determiniert somit deren Inhalte. Tabelle 12.11 stellt die Hypertextknotensorten dar, die dieser Hypertextsorte zu-

	Organisationseinheit	Frequenz	Prozent
1.	Dezentrale Einrichtung: Institut/Seminar	27	3,6
2.	Dezentrale Einrichtung: Fakultät/Fachbereich	8	1,1
3.	Dezentrale Einrichtung: Professur/Arbeitsgruppe	5	0,7
4.	Zentrales Webangebot einer Hochschule	3	0,4
5.	Dezentrale Einrichtung: Dekanat	1	0,1
6.	Zentrale Einrichtung: Sportzentrum	1	0,1
	Σ	45	6,0

Tabelle 12.12: Die Organisationseinheiten, die Instanzen der Hypertextsorte *Vorlesungsverzeichnis* publizieren

geordnet wurden (vgl. ausführlich hierzu Abschnitt 12.7.3). Die tabellarische Aufstellung verdeutlicht, dass unterschiedliche Strukturierungsprinzipien eingesetzt werden: Zum einen werden Listen angeboten, die die Lehrveranstaltungen einer Organisationseinheit umfassen, andererseits existieren Aufstellungen der Kurse eines Studiengangs. In einer Instanz der Hypertextsorte *Webauftritt einer Professur bzw. Arbeitsgruppe* findet sich darüber hinaus eine Liste, die die Lehrveranstaltungen eines einzelnen Lehrenden beinhaltet.

Tabelle 12.12 zeigt die Organisationseinheiten, die Exemplare dieser Hypertextsorte publizieren. Die zugeordneten Knoten werden von dezentralen Einrichtungen veröffentlicht, d. h. die Mehrzahl der 45 Dokumente ist kommentierten Vorlesungsverzeichnissen zugehörig. Das *Vorlesungsverzeichnis* subsumiert auch, wie die letzte Zeile in Tabelle 12.12 verdeutlicht, das Kursangebot von Sportzentren. Dieses Dokument (D 672, Hypertextknotensorte *organisatorische Kerndaten einer Lehrveranstaltung*) ist bezüglich seiner Textstruktur mit Instanzen des tabellarischen Vorlesungsverzeichnisses vergleichbar.

12.6.4 Der Hypertexttyp Software-Dokumentation

Der Hypertexttyp *Software-Dokumentation*, deren Instanzen z. B. Applikationen, allgemeine Werkzeuge und informationstechnologische Standards erläutern, ist mit 40 Dokumenten vertreten (5,3%) und umfasst vier spezifische Ausprägungen: Die Hypertextsorte *Handbuch, Manual, Referenz* wird von denjenigen Institutionen bzw. Personengruppen instanziiert, die auch die Software, auf die sich das Textexemplar bezieht, implementiert haben. Korrespondierenden Instanzen können als „offizielle“ Dokumentation aufgefasst werden. Die Hypertextsorte *Lehrwerk, Referenz* hingegen wird von Dritten instanziiert, bei entsprechenden Exemplaren handelt es sich um „inoffizielle“ Dokumentationen, die – zumindest in Bezug auf die hier untersuchte Stichprobe – den Charakter von Lehrwerken oder Referenzen besitzen. Während sich Instanzen dieser zwei Hypertextsorten nicht auf eine spezifische Institution oder Einrichtung beziehen, sondern in vielen Fällen gespiegelte Kopien frei verfügbarer Dokumentationen darstellen, wurden Exemplare der Hypertextsorte *Benutzungshinweise für lokal verfügbare Software* eigens von Angehörigen einer Organisationseinheit angefertigt, um z. B. in einem Institut oder einem Rechenzentrum zur Verfügung stehende Anwendungen zu erläutern. Instanzen der Hypertextsorte *Tutorial* werden mit dem Zweck angefertigt, dem Rezipienten eine transparent in aufeinanderfolgende Schritte gegliederte und mit zahlreichen Beispielen versehene Anleitung zu präsentieren, deren Durcharbeitung das Erlernen einer Ap-

	Hypertextknotensorte	Frequenz	Prozent
Software-Dokumentation: Handbuch, Manual, Dokumentation			
1.	Seite/Abschnitt einer Software-Dokumentation	16	2,1
2.	Programmcode, Quelltext	2	0,3
3.	Primäre Navigationshilfe	1	0,1
Software-Dokumentation: Lehrwerk, Referenz			
	Seite/Abschnitt einer Software-Dokumentation	10	1,3
Software-Dokumentation: Benutzungshinweise für lokal verfügbare Software			
1.	Seite/Abschnitt einer Software-Dokumentation	3	0,4
2.	Einstiegsseite	2	0,3
3.	Anleitung bzw. Benutzungshinweise	1	0,1
Software-Dokumentation: Tutorial			
	Seite/Abschnitt einer Software-Dokumentation	5	0,7

Tabelle 12.13: Die innerhalb der vier Hypertextsorten des Hypertexttyps *Software-Dokumentation* verwendeten Hypertextknotensorten

	Organisationseinheit	Frequenz	Prozent
1.	Dezentrale Einrichtung: Institut bzw. Seminar	20	2,7
2.	Dezentrale Einrichtung: Professur bzw. Arbeitsgruppe	9	1,2
3.	Zentrale Einrichtung: Rechenzentrum	8	1,1
4.	Dezentrale Einrichtung: studentische Initiative bzw. Arbeitsgruppe	2	0,3
5.	Zentrales Webangebot einer Hochschule	1	0,1
	Σ	40	5,3

Tabelle 12.14: Die Organisationseinheiten, die Instanzen der vier Hypertextsorten des Hypertexttyps *Software-Dokumentation* publizieren

plikation ermöglichen soll. Während sich die Hypertextsorte *Fachbuch* auf Bücher bezieht, die ursprünglich in Papierform publiziert wurden und zusätzlich im WWW veröffentlicht werden, betrifft der Hypertexttyp *Software-Dokumentation* Hypertexte, die ursprünglich im WWW und in einigen Fällen zu einem späteren Zeitpunkt als Buch publiziert werden.

Tabelle 12.14 zeigt die Organisationseinheiten, die Instanzen der vier Hypertextsorten des Hypertexttyps *Software-Dokumentation* publizieren. Institute bzw. Seminare sowie Professuren bzw. Arbeitsgruppen bieten in der Regel Dokumentationen für sehr spezifische Software-Pakete an, die im Lehr- und Forschungsbetrieb eingesetzt werden. Rechenzentren dokumentieren vornehmlich ihren sehr umfangreichen Bestand lokal verfügbarer Applikationen (z. B. Compiler für verschiedene Programmiersprachen).

12.6.5 Die Hypertextsorte Forschungs-, Jahres-, Rechenschaftsbericht

Abhängig vom Bundesland, in dem eine Hochschule ansässig ist, kann das jeweilige Hochschulgesetz die Institution zur Publikation eines Forschungs-, Jahres- oder Rechenschaftsberichts verpflichten, der der Dokumentation von Forschungs- und gegebenenfalls Lehraktivitäten dient. Da sich die 28 Dokumente (3,7%), die dieser Klasse zugeordnet wurden, in vielerlei Hinsicht ähneln, wird die Existenz einer einzelnen Hypertextsorte angenommen, die

die drei Textsortenvarianten in ihrem Etikett subsumiert. Neben Hochschulen publizieren jedoch auch zentrale und dezentrale Organisationseinheiten Forschungs- und Jahresberichte, die ebenfalls im WWW veröffentlicht werden.[11] Wie bereits bei der Hypertextsorte *Vorlesungsverzeichnis* gezeigt wurde (vgl. Abschnitt 12.6.3), liegen auch hier einrichtungsspezifische Unterschiede vor, die sich jedoch nicht auf die verwendeten Hypertextknotensorten, sondern lediglich auf den durchschnittlichen Umfang einer Instanz auswirken, da eine zentrale oder dezentrale Organisationseinheit naturgemäß weniger Forschungsvorhaben darstellt als der Forschungsbericht einer Hochschule. Die zu präsentierenden Texte und Informationen werden typischerweise aus einer zentralen Forschungsdatenbank bezogen, die von den Angehörigen der relevanten Einrichtungen gepflegt werden.

	Hypertextknotensorte	Frequenz	Prozent
1.	Abstract eines Forschungsprojekts	18	2,4
2.	Seite/Abschnitt eines Forschungs-, Jahres-, Rechenschaftsberichts	4	0,5
3.	Kurzdarstellung einer Organisationseinheit (Funktionen und Kontaktinformationen)	2	0,3
4.	Abgeschlossene und/oder angebotene Haus- und Abschlussarbeiten	1	0,1
5.	Abstracts mehrerer Forschungsprojekte	1	0,1
6.	Kontaktinformationen	1	0,1
7.	Q/A (Frage und Antwort)	1	0,1
	Σ	28	3,7

Tabelle 12.15: Die innerhalb der Hypertextsorte *Forschungsbericht, Jahresbericht, Rechenschaftsbericht* verwendeten Hypertextknotensorten

Tabelle 12.15 zeigt die Hypertextknotensorten, deren Instanzen der Hypertextsorte *Forschungs-, Jahres-, Rechenschaftsbericht* zugeordnet wurden. Der Großteil dieser Dokumente besteht aus Abstracts, die den Forschungsgegenstand und die Ziele eines Projekts in sehr knapper Form thematisieren und die beteiligten Mitarbeiter samt Kontaktinformationen auflisten. Oftmals enthalten derartige Abstracts auch eine Publikationsliste sowie Informationen über den Drittmittelgeber (z. B. die DFG). Alle Dokumente teilen sich die Eigenschaft, einen sehr knappen Umfang zu besitzen, da lediglich Kerninformationen aufgeführt werden. Tabelle 12.16 stellt die publizierenden Organisationseinheiten dar: Die Mehrzahl der Knoten stammt aus dem Bereich des zentralen Webangebots einer Hochschule, in einigen Fällen werden die Instanzen, die sich auf eine gesamte Universität beziehen, von einem Dezernat bzw. Referat der Zentralverwaltung publiziert. Daneben existieren einige Forschungsberichte dezentraler Einrichtungen, wozu z. B. ein Institut und ein Graduiertenkolleg zählen.

12.6.6 Weitere Hypertexttypen und -sorten

Neben den in den vorangegangenen Abschnitten dargestellten Hypertexttypen und Hypertextsorten führt Tabelle 12.1 insgesamt 60 weitere Einträge auf, von denen die meisten Entsprechungen in der Klasse traditioneller Textsorten besitzen (z. B. *Skript einer Lehrveranstaltung*, *Studienführer*, *Handbuch*, *Diplomarbeit*, *Lexikon*, *Dissertation*, *Bibliografie*, *Glossar*, *Jahrbuch*, *Kochbuch*, *Prüfungsordnung*, *Studienordnung* und *Tageszeitung*). Da diese Hypertextsorten weitestgehend ihren papierbasierten Pendants entsprechen, weil sie primär durch

[11] Diese Hypertextsorte ist von dem Forschungsbericht zu unterscheiden, der die Ergebnisse eines einzelnen Forschungsvorhabens in detaillierter Form erläutert.

	Organisationseinheit	Frequenz	Prozent
1.	Zentrales Webangebot einer Hochschule	14	1,9
2.	Leitung und Verwaltung: Dezernat bzw. Referat für Wissens- und Technologietransfer	6	0,8
3.	Dezentrale Einrichtung: Institut bzw. Seminar	3	0,4
4.	Leitung und Verwaltung: Dezernat bzw. Referat (generisch)	2	0,3
5.	Dezentrale Einrichtung: Graduiertenkolleg	1	0,1
6.	Dezentrale Einrichtung: Organisationseinheit eines Klinikums	1	0,1
7.	Leitung und Verwaltung: Pressestelle	1	0,1
	Σ	28	3,7

Tabelle 12.16: Die Organisationseinheiten, die Instanzen der Hypertextsorte *Forschungsbericht, Jahresbericht, Rechenschaftsbericht* publizieren

mit Hilfe einer maschinellen Konvertierung generierte HTML-Dokumente instanziiert werden (vgl. die Abschnitte 3.3.6 und 4.3.2), werden im Folgenden einige Aspekte derjenigen Hypertextsorten hervorgehoben, die ausschließlich im *World Wide Web* verwendet werden bzw. sich in diesem Informations- und Kommunikationsmedium erst gebildet haben.

Der Hypertexttyp Virtuelles Museum

Die Hintergründe und Vorlagen der Webauftritte von Institutionen und Organisationseinheiten wurden bereits thematisiert: Diese können zwar in ihrer Gesamtheit als genuine Hypertexttypen bzw. -sorten aufgefasst werden, sie vereinen jedoch Inhalte und Funktionen, die in den papierbasierten Medien durch mehrere unterschiedliche Textsorten realisiert werden (vgl. Abschnitt 11.5.5).[12] Einen besonders interessanten Hypertexttyp stellt das *virtuelle Museum* dar, dem acht HTML-Dokumente zugeordnet wurden (1,1%). Hierbei handelt es sich um eine der wenigen Hypertexttypen bzw. -sorten, die ein spezifisches und konventionalisiertes Etikett besitzen. Aufgrund der wenigen Gemeinsamkeiten, die zwischen den acht Knoten beobachtet werden können, wird jedoch von der Existenz eines Hypertexttyps ausgegangen, der mehrere Subkategorien umfasst. Ein wesentlicher Unterschied betrifft den Umstand, dass einige der virtuellen Museen ein in der realen Welt existentes Museum als Gegenstück besitzen, während sich andere als rein virtuelle Ausstellungen im *World Wide Web* verstehen.[13]

[12] Besonders deutlich wird dieser Umstand bei einer Betrachtung des Hypertexttyps *Webauftritt einer Hochschule bzw. Universität*. In diesem Zusammenhang ist der Hypertexttyp *digitale Bibliothek* von dem *Webauftritt einer Bibliothek* zu differenzieren: Letzterer umfasst die Selbstdarstellung und Dienstleistungen einer Bibliothek und beinhaltet eine oder mehrere Instanzen des Hypertexttyps *digitale Bibliothek* (vgl. Fußnote 12, S. 306).

[13] „Das virtuelle Nahverkehrsmuseum“ (D 290) stellt „Die Museumsfahrzeuge des Hamburger Omnibus Vereins e. V.“ vor, und D 170 präsentiert ein spezifisches Ausstellungsobjekt in einer „virtuellen Museumsführung“ durch die Räumlichkeiten des Instituts für Geschichte der Arabisch-Islamischen Wissenschaften (Universität Frankfurt). Dagegen ist D 48 ein Bestandteil des ausschließlich im WWW existierenden „Chernikow Virtual Museum“, das von zwei Architekturstudenten im Rahmen einer Hausarbeit entworfen wurde, um das Werk des russischen Architekten und Künstlers im WWW zu präsentieren. Eine metadiskursive Äußerung im Impressum dieses Hypertextes erläutert die Ziele: „Das Chernikov Virtual Museum ist der Entwurf eines Museums im WorldWideWeb. In diesem Entwurfsstadium vermittelt es keinen umfassenden Überblick über Jakov Chernikovs Werk. Es werden vielmehr Möglichkeiten gezeigt, eine Sammlung im Internet auszustellen.“

Der Hypertexttyp Messageboard/Diskussionsforum und die Hypertextsorte Mailing-Listen-Archiv

Das *Messageboard/Diskussionsforum* stellt einen weiteren genuinen Hypertexttyp dar (0,7%). Unter diesem Etikett werden im WWW angebotene Diskussionsforen und WWW-basierte Oberflächen zum Zugriff auf Newsgroups subsumiert: An der Universität Bamberg wird ein System betrieben, das die Nutzung lokaler Newsgroups ermöglicht (D 330). Die verbleibenden vier Dokumente stammen aus proprietären Diskussionsforen, deren Einträge eine Textstruktur besitzen, die prinzipiell einer generischen E-Mail entspricht (Betreffzeile, Autor, Datum, Text), weshalb diese vier Knoten vereinfachend der Hypertextknotensorte *E-Mail* zugeordnet wurden – die Differenzierung wird durch den übergeordneten Hypertexttyp expliziert. Die Knoten umfassen z. B. Beiträge zu den Themen „Lineare Algebra" (D 698) und „Re: bin krank" (D 350). Die Hypertextsorte *Mailing-Listen-Archiv* (0,3%) umfasst ebenfalls Dokumente der Hypertextknotensorte *E-Mail*, die ursprünglich als E-Mail entstanden und zu einem späteren Zeitpunkt von einem Konvertierungswerkzeug in ein HTML-basiertes Archiv integriert worden sind (vgl. die Abschnitte 3.3.6 und 4.3.2).[14] Der Hypertexttyp *Messageboard/Diskussionforum* und die Hypertextsorte *Mailing-Listen-Archiv* sind demnach Beispiele für die Konvergenz unterschiedlicher Internet-Dienste (WWW und E-Mail sowie WWW und Usenet) sowie die Emulation der Mailing-Liste im WWW (Diskussionsforum).

Die Hypertextsorte Virtual Library

Die Hypertextsorte *Virtual Library* ist in der Stichprobe mit lediglich einem HTML-Dokument vertreten (0,1%). Die meisten Dokumente einer *Virtual Library* sind – wie auch D 194 – der Hypertextknotensorte *Hotlist* zugehörig, weshalb diese Hypertextsorte prinzipiell auch als *Sammlung von Hotlists* bezeichnet werden kann. Der Terminus *Virtual Library* geht auf Tim Berners-Lee zurück (vgl. Abschnitt 3.2), der diese Initiative kurze Zeit nach der Vorstellung des WWW ins Leben gerufen hat (vgl. `http://vlib.org`). Die *Virtual Library* sollte Hotlists zu den verschiedensten Themenbereichen umfassen, die von Experten des jeweiligen Feldes unter freiwilliger Mitarbeit zusammengestellt werden, um auf diese Weise einen virtuellen Bibliothekskatalog zu erstellen, der auf ausgesuchte und geprüfte Angebote verweist. Es existiert jedoch (bis heute) kein zentraler Datenbestand, so dass sich die *Virtual Library* über mehrere hundert Webserver erstreckt, die jeweils strukturierende Linklisten und Hotlists umfassen, die identische Textstrukturmuster verwenden. War diese Initiative anfangs eine sehr wertvolle Ressource bei der Informationsrecherche im *World Wide Web*, ist sie mit der steigenden Popularität und dem deutlich umfangreicheren Datenbestand von Webkatalogen wie z. B. *Yahoo!* relativ schnell in Vergessenheit geraten.

12.7 Die Hypertextknotentypen und -sorten

Die Tabellen 12.3 und 12.4 stellen die ermittelten Hypertextknotentypen bzw. -sorten dar (vgl. Abschnitt 12.4, insbesondere S. 534 f.). Nachfolgend werden zunächst die hochfrequen-

[14] Diese Zuordnung ist lediglich als Approximation zu verstehen, da es sich bei E-Mail um einen Internet-basierten Kommunikationsdienst und nicht um eine Text- bzw. Hypertextknotensorte handelt. Streng genommen müssten eigenständige Textsorten differenziert werden (vgl. die Abschnitte 4.2.1 und 4.2.4).

	Hypertextknotensorte	Frequenz	Prozent
1.	Seite/Abschnitt einer Software-Dokumentation	34	4,5
2.	Seite/Abschnitt eines Skripts	28	3,7
3.	Seite/Abschnitt einer studentischen Ausarbeitung	9	1,2
4.	Seite/Abschnitt eines Handbuchs	8	1,1
5.	Seite/Abschnitt eines Fachbuches	6	0,8
6.	Seite/Abschnitt einer Diplomarbeit	6	0,8
7.	Seite/Abschnitt eines Studienführers	5	0,7
8.	Seite/Abschnitt eines E-Learning-Kurses	4	0,5
9.	Seite/Abschnitt eines Forschungsberichts, Jahresberichts, Rechenschaftsberichts	4	0,5
10.	Seite/Abschnitt einer Regelung, einer Ordnung, eines Gesetzes, eines juristischen Textes	3	0,4
11.	Seite/Abschnitt eines Abschlussberichts	2	0,3
12.	Seite/Abschnitt einer Dissertation	2	0,3
13.	Seite/Abschnitt einer Biografie	1	0,1
14.	Seite/Abschnitt eines Exkursionsberichts	1	0,1
15.	Seite/Abschnitt mit Fachinformationen	1	0,1
16.	Seite/Abschnitt eines Jahrbuches	1	0,1
17.	Seite/Abschnitt einer Prüfungsordnung	1	0,1
18.	Seite/Abschnitt einer Studienordnung	1	0,1
19.	Seite/Abschnitt eines wissenschaftlichen Artikels	1	0,1
20.	Seite/Abschnitt eines wissenschaftlichen Artikels (eingescannt)	1	0,1

Tabelle 12.17: Der Hypertextknotentyp *Seite/Abschnintt*

ten Hypertextknotentypen vorgestellt (Abschnitte 12.7.1 bis 12.7.6). Abschnitt 12.7.7 geht auf einige weitere Hypertextknotensorten ein und diskutiert verschiedene Problemfälle.

12.7.1 Der Hypertextknotentyp Seite/Abschnitt

Insgesamt 119 Dokumente (15,9%) wurden dem Hypertextknotentyp *Seite/Abschnitt* zugeordnet, der 20 Hypertextknotensorten umfasst (vgl. Tabelle 12.17). Nahezu alle Instanzen dieser 20 Hypertextknotensorten bestehen vornehmlich aus Fließtextabschnitten und wurden *nicht* eigens für die Veröffentlichung im WWW angefertigt. Bei den korrespondierenden Instanzen handelt es sich bis auf wenige Ausnahmen um Dokumente, die ursprünglich Bestandteile linear organisierter Texte waren, die entweder mittels *Copy & Paste* oder durch eine maschinelle Konvertierung in multiple HTML-Dokumente strukturiert wurden. Die Hypertextknotensorten korrespondieren somit in nahezu allen Fällen mit traditionellen Textsorten bzw. spezifischen Teiltextsorten als Konstituenten übergeordneter Textsorten. Durch den Prozess der automatischen Konvertierung – z. B. mit Hilfe des innerhalb der Stichprobe häufig verwendeten Werkzeugs *LaTeX2HTML* – entstehen bei kürzeren Ursprungstexten oftmals linear sequenzierte Texte, die durch „vor"- und „zurück"-Hyperlinks navigiert werden können. Bei längeren Texten wird zusätzlich eine hierarchische Ebene integriert, so dass auch die Navigation von einem Unterabschnitt zu einem Abschnitt oder von einem Kapitel zu einem Abschnitt ermöglicht wird (vgl. Abschnitt 3.3.6).[15]

Die 20 Hypertextknotensorten des Typs *Seite/Abschnitt* besitzen jeweils korrespondierende Hypertextsorten, so ist z. B. die *Seite/Abschnitt einer Diplomarbeit* ein Bestandteil der Hypertextsorte *Diplomarbeit*, bei der übergeordneten Hypertextsorte von *Seite/Abschnitt eines Studienführers* handelt es sich um den *Studienführer*. Da diese unmittelbare Zuordnung nahezu alle der 119 HTML-Dokumente betrifft, wird auf eine tabellarische Darstellung der

[15] Aus diesem Grund wird dieser Hypertextknotentyp auch als „Seite/Abschnitt" bezeichnet, da die Korrespondenzen zum vornehmlich linear sequenzierten Text sehr viel deutlicher ausgeprägt sind als die Tendenz zur prototypischen Realisierung des Konzepts Hypertext (vgl. Fußnote 20, S. 73, sowie Abschnitt 3.7), so dass eine Etikettierung dieser Hypertextknotensorten als „Text*knoten*" nicht gerechtfertigt erscheint.

übergeordneten Hypertextsorten verzichtet. Dieser Umstand verdeutlicht ein übergeordnetes Problem bezüglich der Etikettierung der Hypertextknotensorten, denn bei ihnen handelt es sich um Sammelbezeichnungen für *mehrere* konventionalisierte Konstituenten übergeordneter Textsorten: Eine *Diplomarbeit* umfasst z. B. eine *Titelseite*, ein *Inhaltsverzeichnis*, eine *Zusammenfassung*, mehrere *Kapitel*, *Abbildungen*, ein *Literaturverzeichnis* und eine *Erklärung*, die ihrerseits als standardisierte Teiltextsorten konzeptualisiert werden können, die wiederum in *mehreren* übergeordneten Textsorten Verwendung finden.[16] Eine *Dissertation* enthält ebenfalls die genannten Konstituenten, besitzt jedoch eine unterschiedliche Funktion, andere Realisierungsvoraussetzungen und typischerweise einen größeren Durchschnittsumfang.[17]

Die Hypertextknotensorte Seite/Abschnitt eines E-Learning-Kurses

Da die meisten der 20 Hypertextknotensorten Bestandteile traditioneller und ausführlich untersuchter Textsorten sind, werden zwei untypische Varianten des Hypertextknotentyps *Seite/Abschnitt* genauer betrachtet: Die Hypertextknotensorte *Seite/Abschnitt eines E-Learning-Kurses* wurde vier HTML-Dokumenten zugeordnet (D 65, D 158, D 331, D 648), die Bestandteile der Lehrmaterialien synchron oder asynchron durchgeführter E-Learning-Veranstaltungen sind (vgl. Abschnitt 12.6.2). Abgesehen von ihrem Verwendungszweck weisen die korrespondierenden Hypertexte nur wenige Gemeinsamkeiten auf und unterscheiden sich zudem von traditionellen Lehrmaterialien, die in Präsenzveranstaltungen Verwendung finden und deren Veröffentlichung im WWW nicht zu ihren Kernfunktionen gehört. D 65 erinnert bezüglich der Textstruktur an die Folie eines traditionellen oder digitalen Foliensatzes, andere Dokumente dieses Hypertextes enthalten Tabellen, Übungsaufgaben, Lösungen, Sitzungsprotokolle und eine „Mini-Vorlesung in [sic] Audio-Format“. Im Gegensatz dazu besitzt D 158 eine stringentere Strukturierung, in der zunächst das Lernziel explizit diskutiert wird, woraufhin monosequenzierte Knoten die geometrisch korrekte Entzerrung eines Fotos erläutern. D 331 wird im „Workshop Anatomie fürs Internet“ eingesetzt und basiert auf dem „Visible Human Project“. Dieser Hypertext besitzt eine äußerst verwirrende Strukturierung und qualitativ minderwertige, nicht hochauflösende Abbildungen. D 648 ist ein Bestandteil des Webangebots eines Kurses zum Thema „Schreibkompetenz“ und gliedert einzelne Sitzungen in jeweils ein HTML-Dokument, das eine Kurzzusammenfassung der Inhalte, als PDF-Dokumente verfügbare Foliensätze, eine Videodatei und Übungsaufgaben enthält.

Es ist zwar damit zu rechnen, dass diese Hypertextknotensorte sowie die übergeordnete Hypertextsorte in einen Prozess der Standardisierung eintreten werden. Dieser wird sich jedoch auf den Einsatz spezieller E-Learning-Plattformen beziehen, die unterschiedlich restriktive Vorlagen zur Anfertigung von Lerninhalten bereitstellen.[18] Derzeit werden didak-

[16] Es handelt sich also um „partial documents“ (Crowston und Williams, 2000, S. 213; vgl. Abschnitt 4.5.4).

[17] Zum Zwecke der transparenteren Veranschaulichung der Ergebnisse dieser Untersuchung wurde von der ursprünglich durchgeführten Etikettierung der spezifischen Teiltextsorten Abstand genommen.

[18] Der Idee, Internet-gestützte Lehrveranstaltungen ohne Präsenzphasen durchzuführen, wurde anfangs mit sehr viel Skepsis begegnet. Meines Erachtens ist diese Skepsis auch auf die Textsortenproblematik zurückzuführen, denn es war – wie die genannten Beispiele zeigen – in vielen Fällen unklar, welche Textsorte für diese spezifische Kommunikationssituation eingesetzt werden sollte. In der Konsequenz weisen viele der in der Anfangsphase entwickelten und oftmals kontrovers diskutierten E-Learning-Materialien zahlreiche Merkmale traditioneller wissenstransmittierender Textsorten auf (z. B. Lehrveranstaltungsskript, Foliensatz und Lehrbuch).

tische Konzeptionen für E-Learning-Szenarien (z. B. synchron vs. asynchron) diskutiert. Es ist abzusehen, dass sich die Vorlagen für Lehr- und Lernmaterialien der in der jüngsten Vergangenheit auf nationaler und internationaler Ebene eingeführten Plattformen im Laufe der kommenden Jahre zu mehreren neuen Hypertextsorten entwickeln werden. Da mittlerweile nahezu alle deutschen Hochschulen E-Learning-Plattformen betreiben, werden E-Learning-Materialien in Zukunft immer seltener Bestandteile der Webauftritte von z. B. Instituten, Seminaren oder Professuren sein, die sich lediglich auf die Verknüpfung der jeweiligen E-Learning-Plattform beschränken werden.

Die Hypertextknotensorte Seite/Abschnitt mit Fachinformationen

Unter dem Schlagwort „Fachinformationen“ subsumieren insbesondere die Webauftritte von Bibliotheken Sammlungen unterschiedlicher Texte und Referenzen, die primär eine Informationsfunktion besitzen und sich an Vertreter oder Studierende spezifischer Fachdisziplinen richten. Hierzu zählen z. B. spezialisierte Datenbanken, Nachschlagewerke, Lexika und Verknüpfungen zu digitalen Bibliotheken. In der Stichprobe befindet sich ein Dokument, dem die Hypertextknotensorte *Seite/Abschnitt mit Fachinformationen* zugeordnet wurde (D 222). Es enthält die Überschriften „Pakistan (Sind)“ und „(Bewaffneter Konflikt)“, den Namen des Verfassers und diskutiert in sechs Absätzen die Ursachen der gewalttätigen Auseinandersetzungen in dem genannten Land. Das Dokument ist Bestandteil einer sehr umfangreichen, hierarchisch sequenzierten Gruppe von Hypertexten, die über die „Kriege und bewaffnete[n] Konflikte“ jeweils eines Jahres informieren (publiziert von der „Forschungsstelle Kriege, Rüstung und Entwicklung und Arbeitsgemeinschaft Kriegsursachenforschung“ am Institut für Politische Wissenschaft der Universität Hamburg). Da in diesen Hypertexten keine explizite Textsortenbezeichnung enthalten ist, wurde D 222 die eingangs genannte Hypertextknotensorte zugewiesen, weil es sich im Kontext der Politikwissenschaft um Fachinformationen handelt, für die keine spezifischere Funktion ermittelt werden konnte.[19]

12.7.2 Der Hypertextknotentyp Folie

Dem Hypertextknotentyp *Folie* wurden 80 HTML-Dokumente zugeordnet, die sich auf sechs Hypertextknotensorten verteilen (vgl. Tabelle 12.18). Fast alle Dokumente wurden maschinell generiert, indem mit einer entsprechenden Applikation erstellte Foliensätze nach HTML konvertiert wurden.[20] Nur sieben Dokumente wurden – ihrer Gestaltung sowie den `meta`-Elementen zufolge (vgl. Abschnitt A.4.7) – manuell angefertigt. Die Differenzierung zwischen den sechs Hypertextknotensorten gründet sich in der Unterscheidung zwischen maschinell und manuell hergestellten Dokumenten und in den verwendeten Anwendungen.

[19] Da die Hypertexte jährlich produziert werden, könnten sie alternativ als eine Variante des Jahresberichts der Forschungsstelle aufgefasst werden. Darüber hinaus besteht die Möglichkeit, dass die Text- bzw. Hypertextsorte innerhalb der Politikwissenschaft ein spezifisches, jedoch in den Hypertexten nicht aufgeführtes Etikett besitzt.

[20] Die Folie einer Präsentation ist nicht als Textsorte zu verstehen, da sie ein Trägermedium darstellt. Die Stichprobe enthält nahezu ausschließlich Folien, die im Bereich Hochschule als typische Bestandteile von Präsentationen aufgefasst werden können, die in Kommunikationssituationen wie z. B. Lehrveranstaltung, Weiterbildungskurs und wissenschaftlicher Vortrag verwendet werden. Sie bestehen in der Regel aus einer Titelzeile, in Listenform angeordneten Phrasen und Sätzen sowie unterschiedlichen Typen von Visualisierungen.

	Hypertextknotensorte	Frequenz	Prozent
1.	Microsoft Powerpoint-Folie (exportiert nach HTML)	67	8,9
2.	HTML-Folie	7	0,9
3.	Microsoft Frontpage-Folie (exportiert nach HTML)	2	0,3
4.	Adobe Photoshop-Folie (exportiert nach HTML)	2	0,3
5.	Corel-Presentations-Folie (exportiert nach HTML)	1	0,1
6.	MagicPoint-Folie (exportiert nach HTML)	1	0,1

Tabelle 12.18: Der Hypertextknotentyp *Folie*

Die in Tabelle 12.18 dargestellte Differenzierung ist prinzipiell nicht notwendig, da alle korrespondierenden Knoten einerseits die jeweilige Folie – als eingebettete Grafik oder als Text – und andererseits eine Navigationshilfe enthalten. Darüber hinaus besitzen nahezu alle maschinell erzeugten Dokumente sowohl ein identisches Textstrukturmuster als auch ein sehr ähnliches Webdesign: Die Folie selbst wird im Seitenzentrum angeordnet, und unmittelbar unterhalb der Folie wird die Navigationshilfe präsentiert, die zur ersten, vorherigen, nächsten und letzten Folie sowie zum Inhaltsverzeichnis führt. Die Verteilung der Frequenzen zeigt, dass das Produkt der Firma *Microsoft* den Marktführer darstellt. An das Ergebnis dessen Exportfilters zur Erzeugung der HTML-Version eines Foliensatzes lehnen sich nahezu alle Konkurrenzprodukte an.[21] Mittlerweile generieren aktuelle Versionen von *Powerpoint* sehr komplex strukturierte HTML-Dokumente, die – in mehreren Frames eines Framesets – die Folie selbst, eine Navigationshilfe und gegebenenfalls enthaltene Notizen darstellen. Es ist damit zu rechnen, dass die Mitbewerber in Zukunft ebenfalls in der Lage sein werden, HTML-Dokumente dieser Strukturierungsform zu erstellen.

Weil die übergeordnete Hypertextsorte des Hypertextknotentyps *Folie* in jedem einzelnen Fall *Präsentation eines Foliensatzes* lautet, wurde stattdessen die übergeordnete Hypertextsorte eben dieser Instanz notiert. Auf das *Webangebot einer Lehrveranstaltung* entfallen 49 HTML-Dokumente (6,5%), der *Webauftritt eines Projekts bzw. Projektverbundes* umfasst 12 (1,6%) und der *Webauftritt einer Professur bzw. Arbeitsgruppe* acht Exemplare (1,1%).

12.7.3 Die Hypertextknotentypen Organisatorische Kerndaten einer Lehrveranstaltung und Vorlesungsverzeichnis

Der Hypertextknotentyp *organisatorische Kerndaten einer Lehrveranstaltung* umfasst die vier in Tabelle 12.19 dargestellten Hypertextknotensorten, denen insgesamt 46 HTML-Dokumente zugeordnet wurden (6,1%). Als übergeordnete Hypertextsorte fungiert in 39 Fällen das *Vorlesungsverzeichnis* (5,2%; vgl. Abschnitt 12.6.3), die verbleibenden Vorkommen beziehen sich auf die Webauftritte dezentraler (Professur bzw. Arbeitsgruppe sowie Institut bzw. Seminar) und zentraler Organisationseinheiten (Weiterbildungs- sowie Sportzentrum).

Die Vorkommen der vier Ausprägungen werden dominiert von der Hypertextknotensorte *organisatorische Kerndaten einer Lehrveranstaltung (mit Kommentar)*, der 34 der 46 HTML-Dokumente zugewiesen wurden. Korrespondierende Instanzen umfassen sowohl die Kerndaten einer Lehrveranstaltung (Zeit, Gebäude, Raum, Veranstaltungstyp, Zielgruppe bzw. Stu-

[21] Diese Beobachtung kann als indirekte Bestätigung des Modells der Entwicklung von Hypertextsorten im Hinblick auf die maschinelle Erstellung von HTML-Dokumenten interpretiert werden (vgl. Abschnitt 4.3.2).

	Hypertextknotensorte	Frequenz	Prozent
1.	Organisatorische Kerndaten einer Lehrveranstaltung (mit Kommentar)	34	4,5
2.	Organisatorische Kerndaten einer Lehrveranstaltung	9	1,2
3.	Organisatorische Kerndaten einer Lehrveranstaltung (Liste mehrerer Kurse)	2	0,3
4.	Organisatorische Kerndaten einer Lehrveranstaltung (mit Kommentar; Liste mehrerer Kurse)	1	0,1

Tabelle 12.19: Der Hypertextknotentyp *Organisatorische Kerndaten einer Lehrveranstaltung*

diengänge, Anzahl der Semesterwochenstunden) als auch einen Kommentar, der den Gegenstand des Kurses erläutert, unter Umständen auf die Teilnahmevoraussetzungen hinweist und mit einer Literaturliste schließt. In Instanzen der Hypertextknotensorte *organisatorische Kerndaten einer Lehrveranstaltung* fehlt der Kommentar, d. h. es werden ausschließlich die Kerndaten präsentiert (1,2%). Von diesen beiden Hypertextknotensorten existieren zwei Varianten, in denen jeweils *mehrere* derartige Einträge innerhalb eines Knotens dargestellt werden. Dieses Phänomen kann durch einen Rückgriff auf das Verhältnis zwischen Hypertextknotensorten und Hypertextsortenmodulen erklärt werden (vgl. Kapitel 5): Letztere stellen die eigentlichen atomaren Bausteine einer Hypertextknotensorte dar, die wiederum hinsichtlich der Existenz einer oder mehrerer Instanzen eines Hypertextsortenmoduls markiert sein kann. Wenn der einzelne Eintrag (d. h. die organisatorischen Kerndaten mit bzw. ohne Kommentar) als Hypertextsortenmodul aufgefasst wird, existiert eine deutliche Markierung hinsichtlich der einzelnen Ausprägung (vgl. Tabelle 12.19).

Traditionelle Textsorten enthalten oftmals konventionalisierte Teiltextsorten (vgl. hierzu Abschnitt 12.7.1). Das Etikett des Hypertexttyps *organisatorische Kerndaten einer Lehrveranstaltung* verdeutlicht, dass auch im Bereich der *traditionellen* Textsorten Konstituenten existieren, die *kein* etabliertes Etikett aufweisen (vgl. Abschnitt 4.3.1), denn der einzelne Eintrag eines tabellarischen oder kommentierten Vorlesungsverzeichnisses besitzt – von abstrakten Begriffen wie „Datensatz“, „Kurseintrag“ oder „Lehrveranstaltung“ abgesehen – keine standardisierte Bezeichnung. Aus eben diesem Grund wurde die Phrase „organisatorische Kerndaten einer Lehrveranstaltung“ gewählt, die durch Zusätze, die den Merkmalen ± Kommentar und ± Liste entsprechen, spezifiziert werden kann. Das Phänomen, dass in der Alltagssprache typischerweise unetikettierte Teiltextsorten, die Konstituenten etablierter Textsorten darstellen, im WWW als *einzelne* HTML-Dokumente instanziiert und somit als einzelner Text betrachtet werden, ist einer der wesentlichen Gründe für den Umstand, dass nahezu alle Verfasser verwandter Studien (vgl. Abschnitt 4.4) von Problemen bei der Zuordnung eines Dokuments zu einem Genre berichten und somit gezwungen sind, Etiketten von Genres bzw. Textsortenbezeichnungen zu konstruieren (vgl. auch Abschnitt 2.3.2).

Bei dem *Vorlesungsverzeichnis* handelt es sich um einen eng verwandten Hypertextknotentyp, der drei Hypertextknotensorten umfasst, denen 10 Dokumente (1,3%) zugeordnet wurden. Während die Instanzen des Hypertextknotentyps *organisatorische Kerndaten einer Lehrveranstaltung* einen oder mehrere Kurse innerhalb eines Einzelknotens beinahe ohne jeglichen Kontext präsentieren, handelt es sich bei den Instanzen des Hypertextknotentyps *Vorlesungsverzeichnis* um Knoten, die eine Art kontextuelle Geschlossenheit aufweisen, da sie das *vollständige* Vorlesungsverzeichnis eines Studiengangs, einer Organisationseinheit oder eines Dozenten enthalten und *kein* Bestandteil einer Instanz der Hypertextsorte *Vorlesungsverzeich-*

	Hypertextknotensorte	Frequenz	Prozent
1.	Vorlesungsverzeichnis eines Studiengangs	6	0,8
2.	Vorlesungsverzeichnis einer Organisationseinheit	3	0,4
3.	Vorlesungsverzeichnis eines Dozenten	1	0,1

Tabelle 12.20: Der Hypertextknotentyp *Vorlesungsverzeichnis*

nis sind (vgl. Tabelle 12.20).[22] Neben diesem Differenzierungskriterium besitzen Instanzen oftmals eine Überschrift wie z. B. „Veranstaltungen Sommersemester 2001“ (D 611) oder „Lehrveranstaltungen im Wintersemester 2001/2002“ (D 626) und weisen darüber hinaus eine Textstrukturierung auf, die sich in nahezu allen Fällen deutlich von gedruckten Vorlesungsverzeichnissen entfernt, indem z. B. von Verknüpfungsstrategien und innovativen Gestaltungen Gebrauch gemacht wird. Die Instanzen des Hypertextknotentyps *organisatorische Kerndaten einer Lernveranstaltung* hingegen imitieren entsprechende Einträge in gedruckten Verzeichnissen. Eine weitere Unterscheidung bezieht sich auf die Organisationseinheiten, die Instanzen des Hypertextknotentyps *Vorlesungsverzeichnis* publizieren: Bei ihnen handelt es sich in den meisten Fällen um Professuren bzw. Arbeitsgruppen, die bestrebt sind, ihr spezifisches Lehrangebot in *einem* HTML-Dokument anzubieten.

Abschließend ist generalisierend und mit Bezug auf das Hypertextsortenmodell (vgl. Kapitel 5) festzuhalten, dass sich die beiden hier diskutierten Hypertextknotentypen vorwiegend auf das Hypertextsortenmodul *organisatorische Kerndaten einer Lehrveranstaltung* stützen, das die Eigenschaften besitzt, mit oder ohne Kommentar sowie einzeln oder in Listenform ausgeprägt zu werden. Die Hypertextknotensorten des Typs *organisatorische Kerndaten einer Lehrveranstaltung* sind – insbesondere in Gestalt der Einzelausprägung – primär konventionalisierte Bestandteile der übergeordneten Hypertextsorte *Vorlesungsverzeichnis*, wohingegen der Hypertextknotentyp *Vorlesungsverzeichnis* eine Art minimale und nur selten realisierte Variante der identisch benannten Hypertextsorte darstellt und ausschließlich mehrere Instanzen des Hypertextsortenmoduls *organisatorische Kerndaten einer Lehrveranstaltung* umfasst.

12.7.4 Der Hypertextknotentyp Abstract

Der Hypertextknotentyp *Abstract* umfasst sechs Hypertextknotensorten, denen 42 Dokumente zugeordnet wurden (5,6%; vgl. Tabelle 12.21). Vier der sechs Hypertextknotensorten sind traditionelle Textsorten bzw. Teiltextsorten aus dem Kommunikationsbereich Hochschule und Wissenschaft, die verbleibenden Sorten *Abstract eines Forschungsprojekts* sowie *Abstracts mehrerer Forschungsprojekte* (vgl. Tabelle 12.3) bedürfen einer Erläuterung: Bei einem

[22] Die Hypertextknotensorte *Vorlesungsverzeichnis eines Dozenten* unterscheidet sich von dem Hypertextsortenmodul *Angaben zu Lehrveranstaltungen* insbesondere durch die übergeordnete Hypertextsorte: *Angaben zu Lehrveranstaltungen* sind ein Bestandteil der *persönlichen Homepage eines Wissenschaftlers* (vgl. Abschnitt 10.5.7), wohingegen das *Vorlesungsverzeichnis eines Dozenten* Teil des Webauftritts einer Professur bzw. einer Arbeitsgruppe ist. Dennoch existieren naturgemäß zahlreiche Korrespondenzen. Es wurden zwei unterschiedliche Etiketten gewählt, da in den persönlichen Homepages eine größere Varianz bezüglich der Aggregierung dieses Hypertextsortenmoduls mit anderen Hypertextsortenmodulen vorliegt, wohingegen die in diesem Kapitel untersuchte Stichprobe die Annahme eines Hypertextknotentyps nahe legt, der sich wiederum auf eine Sequenz von Instanzen des korrespondierenden Hypertextsortenmoduls stützt.

	Hypertextknotensorte	Frequenz	Prozent
1.	Abstract eines Forschungsprojekts	30	4,0
2.	Abstract eines Vortrags	3	0,4
3.	Abstract einer wissenschaftlichen Publikation	3	0,4
4.	Abstract einer Dissertation	3	0,4
5.	Abstracts mehrerer Forschungsprojekte	2	0,3
6.	Abstract eines Posters	1	0,1

Tabelle 12.21: Der Hypertextknotentyp *Abstract*

Abstract handelt es sich typischerweise um die Zusammenfassung eines Textes, der einer spezifischen Textsorte zugehörig ist (vgl. Endres-Niggemeyer, 1998, 2004). In einigen Fällen ist das Abstract ein konventionalisierter Bestandteil eines Textsortenexemplars. Die innerhalb dieses Hypertextknotentyps hochfrequente Hypertextknotensorte *Abstract eines Forschungsprojekts* und ihre Variante *Abstracts mehrerer Forschungsprojekte*[23] beziehen sich jedoch nicht auf den Text einer spezifischen Textsorte, sondern auf ein wissenschaftliches Forschungsprojekt, das entweder durch Drittmittel oder durch die Grundausstattung einer universitären Organisationseinheit finanziert wird. Die Notwendigkeit der knappen Darstellung der Ziele und des Untersuchungsgegenstandes eines Projekts kann auf verschiedene Rahmenbedingungen zurückgeführt werden: Zunächst sind Hochschulen in der Regel gesetzlich verpflichtet, einen Forschungsbericht zu publizieren (vgl. Abschnitt 12.6.5), der sämtliche Forschungsprojekte einer Universität mitsamt der zentralen Kerndaten (beteiligte Mitarbeiter, Kurzdarstellung, Publikationsliste) aufführt. Weiterhin werden von den meisten Drittmittelgebern jährlich kurze Zwischenberichte angefordert, die über den Fortschritt eines Projekts informieren. Schließlich sind natürlich auch die Mitarbeiter eines Projekts aus verschiedenen Gründen bestrebt, ihr Vorhaben und die bereits erzielten Ergebnisse im WWW darzustellen. Zur Realisierung eben dieses Webangebots werden in der Regel Passagen aus den bereits vorhandenen Texten der genannten Klassen oder des initial bei einem Drittmittelgeber eingereichten Projektantrags verwendet, die manuell (z. B. per *Copy & Paste*) oder maschinell mit einem Konverter wie *LaTeX2HTML* nach HTML überführt werden.

Tabelle 12.22 stellt die übergeordneten Hypertextsorten der Instanzen des Hypertextknotentyps *Abstract* dar. Es wird deutlich, dass das *Abstract eines Forschungsprojekts* vornehmlich innerhalb der Hypertextsorte *Forschungs-, Jahres-, Rechenschaftsbericht* (vgl. Abschnitt 12.6.5), aber auch im *Webauftritt einer Professur bzw. Arbeitsgruppe* verwendet wird. Die Instanzen von *Abstract einer Dissertation* werden innerhalb zweier Exemplare der Hypertextsorte *digitale Bibliothek* publiziert, die digitale Versionen von Dissertationen veröffentlichen.

12.7.5 Der Hypertextknotentyp Berufliche Homepage eines Hochschulangehörigen

Der Hypertextknotentyp *berufliche Homepage eines Hochschulangehörigen* umfasst die Hypertextknotensorten *berufliche Homepage eines Wissenschaftlers* (11 Vorkommen; 1,5%) und *berufliche Homepage eines Hochschulangehörigen* (sieben Vorkommen; 0,9%). Während sich

[23] Bei dem *Abstract eines Forschungsprojekts* handelt es sich demnach um ein Hypertextsortenmodul, das, wie die Frequenzen zeigen, in der Regel einzeln in einem Knoten ausgeprägt wird (vgl. Kapitel 5).

	übergeordnete Hypertextsorte	Hypertextknotensorte	Frequenz	Prozent
1.	Forschungsbericht, Jahresbericht, Rechenschaftsbericht	Abstract eines Forschungsprojekts	18	2,4
2.	Webauftritt einer Professur bzw. Arbeitsgruppe	Abstract eines Forschungsprojekts	4	0,5
3.	Forschungsprojekte einer Organisationseinheit	Abstract eines Forschungsprojekts	3	0,4
4.	Webauftritt eines Projekts oder Projektverbundes	Abstract eines Forschungsprojekts	3	0,4
5.	Digitale Bibliothek	Abstract einer Dissertation	2	0,3
6.	Webangebot einer Lehrveranstaltung (Doktoranden- oder Mitarbeiterkolloquium)	Abstract eines Vortrags	2	0,3
7.	Forschungsbericht, Jahresbericht, Rechenschaftsbericht	Abstracts mehrerer Forschungsprojekte	1	0,1
8.	Persönliche Homepage eines Wissenschaftlers	Abstract einer wissenschaftlichen Publikation	1	0,1
9.	Unterrichtsmaterialien für die Schule	Abstract einer wissenschaftlichen Publikation	1	0,1
10.	Webangebot einer Konferenz/Tagung	Abstract einer wissenschaftlichen Publikation	1	0,1
11.	Webangebot einer Konferenz/Tagung	Abstract eines Posters	1	0,1
12.	Webangebot einer Konferenz/Tagung	Abstract eines Vortrags	1	0,1
13.	Webauftritt eines Arbeitskreises	Abstract eines Forschungsprojekts	1	0,1
14.	Webauftritt eines Instituts bzw. Seminar	Abstract eines Forschungsprojekts	1	0,1
15.	Webauftritt einer Professur bzw. Arbeitsgruppe	Abstract einer Dissertation	1	0,1
16.	Webauftritt eines Projekts oder Projektverbundes	Abstracts mehrerer Forschungsprojekte	1	0,1
	Σ		42	5,6

Tabelle 12.22: Die übergeordneten Hypertextsorten des Hypertextknotentyps *Abstract*

die erstgenannte Hypertextknotensorte ausschließlich auf Wissenschaftler bezieht, handelt es sich bei der zweiten Kategorie um eine Sammelbezeichnung, die unter anderem die Homepages von studentischen Hilfskräften (D 184, D 437) und administrativ-technischen Mitarbeitern (D 284, D 341) beinhaltet.

Zusätzlich existiert die Hypertextknotensorte *persönliche Homepage eines Wissenschaftlers* (1,3%) sowie die gleichnamige Hypertextsorte (2,3%; vgl. Tabelle 12.1).[24] Da es sich bei der *persönlichen Homepage eines Wissenschaftlers* ebenfalls um eine berufliche Homepage handelt, müsste die Bezeichnung des Hypertextknotentyps *berufliche Homepage eines Hochschulangehörigen* streng genommen mit dem Zusatz „von offizieller Stelle angefertigt" versehen werden (vgl. Abschnitt 10.6, insbesondere Abbildung 10.7, S. 460).

Es existieren drei Unterschiede zwischen der *beruflichen Homepage eines Hochschulangehörigen* und der *persönlichen Homepage eines Wissenschaftlers*: Letztere werden in der Regel von der in einer Homepage dargestellten Person selbst angefertigt und besitzen eine individuelle Gestaltung. Die beruflichen Homepages, von denen oftmals mehrere Instanzen z. B. innerhalb des Webauftritts einer Professur publiziert werden, weisen hingegen eine konsistente Gestaltung sowie identische Hypertextsortenmodule auf und werden typischerweise von einer studentischen Hilfskraft oder einem Techniker angefertigt und gepflegt. Der dritte Unterschied betrifft den Inhalt der Hypertextknotensorte *beruflichen Homepage eines Hochschulangehörigen*, denn dieser bezieht sich *ausschließlich* auf die Rolle der dargestellten Person im Kontext einer spezifischen Organisationseinheit: D 477 ist Teil des *Webauftritts eines Dekanats* und präsentiert die dargestellte Person in der Rolle als „Studiendekan" mit seinem Namen, einem Porträtfoto und verschiedenen Kontaktinformationen. D 640 ist Teil des *Webauftritts einer*

[24] Weil die Hypertextsortenmodule der Hypertextsorte *persönliche Homepage eines Wissenschaftlers* (vgl. Tabelle 10.3, S. 457) auch in einem einzelnen Knoten realisiert werden können, muss die Existenz einer identisch etikettierten Hypertextknotensorte angenommen werden, da 17 derartige HTML-Dokumente in der Stichprobe enthalten sind. Da Kapitel 10 ausführlich auf die genannte Hypertextsorte eingeht, kann an dieser Stelle auf eine detaillierte Charakterisierung der einzelnen Hypertextknotensorten verzichtet werden. Es ist jedoch anzumerken, dass zahlreiche Korrespondenzen zwischen den Hypertextsortenmodulen der in dieser Stichprobe enthaltenen persönlichen Homepages und den in Analyse 3 untersuchten Dokumenten existieren.

Fachschaft, besitzt ebenfalls das genannte Inventar von Hypertextsortenmodulen und weist die dargestellte Person als „Referent für Studienangelegenheiten“ aus.

12.7.6 Der Hypertextknotentyp Redaktioneller Artikel eines Publikationsorgans

Gerade die größeren Organisationseinheiten einer Hochschule betreiben oftmals Publikationsorgane, so veröffentlicht eine Universität selbst eine Universitätszeitung und eventuell parallel ein Hochglanzmagazin. Ein Rechenzentrum publiziert ein Mitteilungsblatt, dessen Beiträge eine vornehmlich technische Ausrichtung aufweisen. Derartige Zeitungen und Zeitschriften besitzen in vielen Fällen korrespondierende Webangebote, die die Artikel der bislang erschienenen Ausgaben online zur Verfügung stellen. In der Stichprobe sind 19 Vorkommen des Hypertexttyps *Publikationsorgan einer Einrichtung* enthalten (2,5%), der acht Hypertextsorten umfasst, die wiederum den unterschiedlichen Typen von Publikationen entsprechen. Dem Hypertextknotentyp *redaktioneller Artikel eines Publikationsorgans* wurden 16 Dokumente zugeordnet (2,1%; vgl. Tabelle 12.23), die jedoch interessanterweise nicht alle als Bestandteile des Hypertexttyps *Publikationsorgan einer Einrichtung* aufzufassen sind.[25] Beispielsweise ist D 274 ein Bestandteil des *Webauftritts eines Projekts oder Projektverbundes* und enthält den Artikel „Konsultationssystem für die gastroenterologische Endoskopie“, der in der Kopfzeile explizit als „Beitrag in den Mitteilungen der Technischen Universität München, Ausgabe 4-00/01, S. 14/15“ identifiziert wird – der Hyperlink führt zur Einstiegsseite des Webauftritts dieses Organs, in dem der Artikel ebenfalls online zugänglich ist. Anstatt das grundlegende Hypertext-Konzept der Verknüpfung des Artikels zu realisieren, wird von dieser dezentralen Organisationseinheit eine *Kopie* des Textes angeboten. In D 200 kann dieses Phänomen ebenfalls beobachtet werden: Es handelt sich um einen kurzen Beitrag zu genetischen Algorithmen, der ursprünglich in der Universitätszeitschrift der Universität Kiel veröffentlicht wurde. Das Dokument stellt jedoch eine Kopie des Artikels dar, die ein Bestandteil der *persönlichen Homepage eines Wissenschaftlers* ist.[26]

12.7.7 Weitere Hypertextknotensorten

Da eine ausführliche Darstellung sämtlicher Hypertextknotentypen und -sorten den Rahmen der vorliegenden Arbeit sprengte, werden nachfolgend zunächst die verbleibenden hochfrequenten Hypertextknotensorten erläutert. Anschließend werden spezialisierte Hypertextknotensorten diskutiert, woraufhin diejenigen Dokumente vorgestellt werden, die aus mehreren Gründen nicht zugeordnet werden konnten.

[25] Die HTML-Dokumente, bei denen es sich nicht um redaktionelle Artikel handelt, denen aber dennoch der übergeordnete Hypertexttyp *Publikationsorgan einer Einrichtung* zugewiesen wurde, besitzen Hypertextknotensorten wie z. B. *Suchformular* und *Errataliste*.

[26] Meiner Erfahrung nach handelt es sich hierbei keinesfalls um ein seltenes Phänomen. Es ist zu vermuten, dass die Autoren der Artikel davon ausgehen, dass die innerhalb des Webangebots des jeweiligen Publikationsorgans angebotenen Versionen in der Zukunft nicht mehr zur Verfügung stehen oder dass sich ihre URLs ändern könnten, weshalb eine Kopie innerhalb des von den Autoren kontrollierten Webangebots präsentiert wird, um eine alternative Zugangsmöglichkeit zu gewährleisten (vgl. auch Abschnitt 7.2.5). Mit anderen Typen von Publikationen (z. B. Artikel in wissenschaftlichen Zeitschriften) wird oftmals in gleicher Weise verfahren.

	Hypertextknotensorte	Frequenz	Prozent
1.	Redaktioneller Artikel einer Universitätszeitung	5	0,7
2.	Redaktioneller Artikel einer Fachschafts- bzw. Studierendenzeitung	4	0,5
3.	Redaktioneller Artikel von Rechenzentrumsinformationen bzw. -mitteilungen	3	0,4
4.	Redaktioneller Artikel einer Universitätszeitung (Liste von Artikeln)	2	0,3
5.	Redaktioneller Artikel einer Alumnizeitung	1	0,1
6.	Redaktioneller Artikel einer Bibliothekszeitschrift	1	0,1

Tabelle 12.23: Der Hypertextknotentyp *Redaktioneller Artikel eines Publikationsorgans*

Die verbleibenden hochfrequenten Hypertextknotensorten

Der hochfrequenten Hypertextknotensorte *Foto* wurden 29 Dokumente zugewiesen (3,9%; vgl. Tabelle 12.3), die ein Digitalfoto und eine Navigationshilfe beinhalten. Als übergeordneter Hypertexttyp fungiert in nahezu allen Fällen die *Fotogalerie*. Instanzen dieses Typs besitzen oftmals einen identischen Aufbau: Eine oder mehrere Indexseiten, die den traditionellen Kontaktabzügen ähneln, zeigen kleinformatige *thumbnails* der Fotos, die als Hyperlinkanzeiger zu einer Darstellung in voller Größe führen.[27] Dennoch werden unterschiedliche Hypertextsorten angenommen, die sich auf die Motive sowie die wiederum übergeordnete Hypertextsorteninstanz beziehen: Die Hypertextsorte *Fotogalerie – Rundgang durch ein Gebäude* wird nur innerhalb der Webauftritte größerer Organisationseinheiten instanziiert, die einen virtuellen Rundgang durch ihre Räumlichkeiten präsentieren. Die *Fotogalerie – Veranstaltung, Messe, Konferenz* wird in korrespondierenden Webangeboten eingesetzt und beinhaltet z. B. Impressionen einer Tagung, Wissenschaftler, die einen Vortrag halten oder Schnappschüsse der Diplomfeier eines Studiengangs. Daneben existieren zwei Vorkommen der Hypertextsorte *Fotogalerie – Bildarchiv*, in denen Fotos von Gemälden präsentiert und aus kunsthistorischer Perspektive interpretiert werden. Die vierte Hypertextsorte zeigt Fotos, die in einer bestimmten Stadt oder an einem spezifischen Ort entstanden sind.

Einen besonderen Stellenwert nimmt die Hypertextknotensorte *Einstiegsseite* ein, die in der Stichprobe 24 Vorkommen besitzt (3,2%). In diesen Fällen wurde in der zweiten Analyseebene nicht die übergeordnete Hypertextsorte, sondern die Hypertextsorte notiert, deren Instanz durch die Einstiegsseite eingeleitet wird. Dabei handelt es sich unter anderem um das *Webangebot einer Lehrveranstaltung* (9 Vorkommen), den *Webauftritt einer Professur bzw. Arbeitsgruppe* (3), die *Software-Dokumentation (Benutzungshinweise für lokal verfügbare Software)* (2) und verschiedene weitere Webauftritte von Organisationseinheiten.

[27] Der in der Stichprobe nicht enthaltene Hypertextknotentyp *Einstiegsseite einer Fotogalerie*, der die enthaltenen Fotos, wie geschildert, in Form von *thumbnails* präsentiert, ist ein gutes Beispiel zur Erläuterung des Einflusses spezifischer Software-Pakete auf die Entwicklung konventionalisierter Hypertextknotensorten (vgl. Abschnitt 4.3.2): Die manuelle Anfertigung einer derartigen Einstiegsseite ist mit sehr viel Aufwand verbunden. Zunächst müssen Kopien aller digitalen Fotos angefertigt werden, deren Größe daraufhin so skaliert wird, dass sie gemeinsam in der Einstiegsseite präsentiert werden können. Anschließend muss der HTML-Code der Einstiegsseite erstellt werden, der aufgrund der vielen eingebetteten Fotos und der jeweiligen Hyperlinkanzeiger eine recht große Komplexität aufweist. Seit für die Erstellung solcher Fotogalerien spezialisierte, diesen Prozess automatisierende Werkzeuge existieren, wird die „traditionelle" Realisierung einer derartigen Einstiegsseite als Liste von Hyperlinks, die zu den Fotos führen, kaum noch realisiert, d. h. die Möglichkeiten der Software zur Erstellung einer *Fotogalerie* haben diese Hypertextsorte und die von ihr umfassten Hypertextknotensorten (d. h. *Einstiegsseite* und *Foto*) maßgeblich beeinflusst.

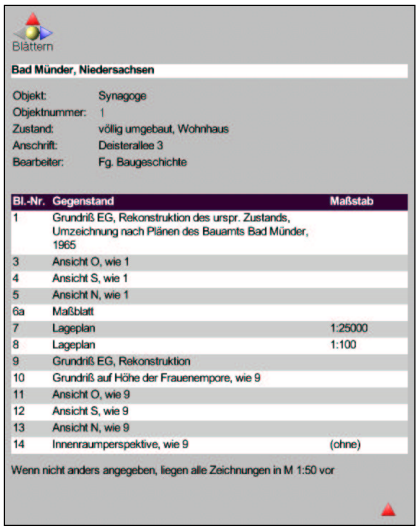
Ebenfalls 24 Dokumenten wurde die Hypertextknotensorte *Pressemitteilung* zugewiesen. Bei den Instanzen handelt es sich um einzelne Pressemitteilungen, die in nahezu allen Fällen eingebettete Bestandteile des *Webauftritts einer Pressestelle* sind. Die meisten Pressestellen organisieren die online zugänglichen Pressemitteilungen in hierarchischer Form, so dass als übergeordnete Hypertextknotensorte zunächst *Pressemitteilung (Liste)* angenommen wird, die typischerweise alle Pressemitteilungen eines Monats oder eines Jahres in Form einer Hyperlinkliste umfasst. Diese Instanzen werden wiederum von Exemplaren der Hypertextsorte *Webangebot mit Pressemitteilungen* subsumiert, die ihrerseits Konstituenten des *Webauftritts einer Pressestelle* sind (vgl. auch Kapitel 11). Die *Einstiegsseite eines universitären Webauftritts* verweist typischerweise nicht auf die *Einstiegsseite einer Pressestelle*, sondern unmittelbar auf das *Webangebot mit Pressemitteilungen* (vgl. Kapitel 11).

In der Aufstellung der Hypertextknotensorten sind auch die Einträge *primäre Navigationshilfe*, *Kopfzeile* und *Fußzeile* enthalten (vgl. die Tabellen 12.3 und 12.4). Bei den insgesamt 23 zugehörigen Dokumenten (3,1%) wurden diese Hypertextknotensorten notiert, weil es sich um *eigenständige* HTML-Dokumente handelt, die jedoch bei einer Betrachtung der jeweiligen Website mit einem Browser nicht in dieser Form in Erscheinung treten, da es sich um die Inhalte von Frames einzelner Framesets handelt. Die Aufnahme derartiger Dokumente in eine Stichprobe könnte zwar prinzipiell vermieden werden, die hierfür notwendigen Verfahren sind jedoch außerordentlich komplex, denn es müsste der gesamte Korpusbestand untersucht werden, um diejenigen HTML-Dokumente zu ermitteln, die eine Frameset-Definition besitzen. Anschließend müssten rekursiv sämtliche referenzierten Dokumente analysiert werden, um festzustellen, ob sie innerhalb oder außerhalb eines Framesets präsentiert werden. Diese Information müsste anschließend in einer Datenbanktabelle hinterlegt werden, wobei jedoch das Problem besteht, dass mehrere Verweise auf ein Dokument existieren, die es einerseits innerhalb und andererseits außerhalb eines Framesets präsentieren.

Spezialisierte Hypertextknotensorten

Die Überblicksdarstellung der Ergebnisse (Abschnitt 12.4) ist auf den Umstand eingegangen, dass in der Stichprobe mehrere Hypertextsorten und Hypertextknotensorten enthalten sind, die auf hochgradig spezialisierten Textsorten aus mehreren Fachgebieten basieren. Diese Studie wurde in vielen Fällen mit dem Umstand konfrontiert, dass bei einem gegebenen HTML-Dokument zwar eindeutig eine spezifische Textsorte bzw. eine aus ihr abgeleitete Hypertextknotensorte vorliegt, aufgrund des auf Seiten des Verfassers fehlenden fachsprachlichen Wissens jedoch nicht das etablierte Etikett dieser Text- bzw. Teiltextsorte benannt werden konnte (falls es denn existiert, vgl. Fußnote 19, S. 553). In einigen Fällen konnten in den bei- und übergeordneten Dokumenten entsprechende Bezeichnungen ermittelt werden, die wiederum als Etikett verwendet wurden.

Abbildung 12.3 stellt drei *vollständig* reproduzierte HTML-Dokumente als Beispiele dar. D 236 (und D 337, D 585 sowie D 663) wurden mit der Hypertextknotensorte *medizinische Diagnoseprozedur* etikettiert, da der übergeordnete Hypertext die Überschrift „Internationale Klassifikation der Prozeduren in der Medizin Version 1.1 vom 21.09.1995 [...]“ besitzt (publiziert innerhalb des Webauftritts der Klinik für Herz- und Thoraxchirurgie, Universität zu Köln). Die zugehörige Hypertextsorte wurde mit der Bezeichnung *Klassifikation medizi-*



Blättern

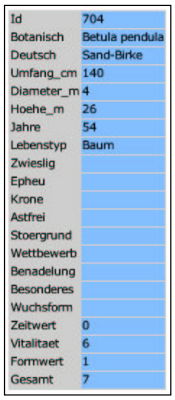
Bad Münder, Niedersachsen

Objekt:	Synagoge
Objektnummer:	1
Zustand:	völlig umgebaut, Wohnhaus
Anschrift:	Deisterallee 3
Bearbeiter:	Fg. Baugeschichte

Bl.-Nr.	Gegenstand	Maßstab
1	Grundriß EG, Rekonstruktion des urspr. Zustands, Umzeichnung nach Plänen des Bauamts Bad Münder, 1965	
3	Ansicht O, wie 1	
4	Ansicht S, wie 1	
5	Ansicht N, wie 1	
6a	Maßblatt	
7	Lageplan	1:25000
8	Lageplan	1:100
9	Grundriß EG, Rekonstruktion	
10	Grundriß auf Höhe der Frauenempore, wie 9	
11	Ansicht O, wie 9	
12	Ansicht S, wie 9	
13	Ansicht N, wie 9	
14	Innenraumperspektive, wie 9	(ohne)

Wenn nicht anders angegeben, liegen alle Zeichnungen in M 1:50 vor

D 116



Id	704
Botanisch	Betula pendula
Deutsch	Sand-Birke
Umfang_cm	140
Diameter_m	4
Hoehe_m	26
Jahre	54
Lebenstyp	Baum
Zwieslig	
Epheu	
Krone	
Astfrei	
Stoergrund	
Wettbewerb	
Benadelung	
Besonderes	
Wuchsform	
Zeitwert	0
Vitalitaet	6
Formwert	1
Gesamt	7

D 720

D 236

Übersicht - Vorwort, Ausblick, Anschriften

Systematik für 5-369: und parvi

- 1: DIAGNOSTISCHE MASSNAHMEN
- 5: OPERATIONEN
  - 5-35...5-37: Operationen am Herzen
    Exkl.:
    Perkutan-transluminale Gefäßintervention (8-836)
    Perkutan-transluminaleKoronarangioplastie (8-837.1)
    Shuntoperationenzwischen großem und kleinem Kreislauf (5-390)
    Hinweis:
    DieAnwendung mikrochirurgischer Technik ist, sofern nicht als eigener Kode angegeben, zusätzlich zu kodieren (5-980)
    DieDurchführung der Operation im Rahmen der Versorgung einer Mehrfachverletzung ist zusätzlich zu kodieren (5-981)
    DieDurchführung der Operation im Rahmen der Versorgung eines Polytraumas ist zusätzlich zu kodieren (5-982)
    DieDurchführung einer Reoperation ist, sofern nicht als eigener Kode angegeben, zusätzlich zu kodieren (5-983)
    Dervorzeitige Abbruch einer Operation ist zusätzlich zu kodieren (5-995)
    - 5-36: Operationen an den Koronargefäßen
      Exkl.:
      Operationen bei kongenitalen Gefäßanomalien (5-357)
      Hinweis:
      DieAnwendung der Herz-Lungen-Maschine ist, sofern nicht im Kode enthalten, zusätzlich zu kodieren (8-851)
      - 5-369: Andere Operationen an den Koronargefäßen
        - 5-369.0: Naht (nach Verletzung)
        - 5-369.1: Korrektur eines Aneurysmas
        - 5-369.2: Verschluß einer erworbenen koronaren Fistel
        - 5-369.3: Rekonstruktion des Koronarostiums
        - 5-369.4: Sympathektomie der Koronararterien
        - 5-369.5: Reoperation
        - 5-369.x: Sonstige
        - 5-369.y: N.n.bez.
- 8: NICHTOPERATIVE THERAPEUTISCHE MASSNAHMEN
- 9: ERGÄNZENDE MASSNAHMEN

Copyright by: Hagen LOTZE, Obermainanlage 24, 60314 Frankfurt am Main, Tel. und Fax: +49 (69) 441831
EMail me via CompuServe: 100520,3457 or Internet: Hagen.Lotze@t-online.de ; Infos und Solutions

für den Webmaster der Universitäts-Kinderklinik, Frankfurt am Main

Abbildung 12.3: Beispiele für spezialisierte Text- bzw. Teiltextsorten

nischer Diagnoseprozeduren versehen. Für die beiden weiteren in Abbildung 12.3 enthaltenen Teiltextsorten konnten keine Etiketten ermittelt werden, weshalb sie als *Kerndaten eines Pflanzentyps* (D 720) sowie *Kerndaten eines historischen Bauwerks* bezeichnet wurden (D 116). Als übergeordnete Hypertextsorte von D 116 wurde *Daten historischer Bauwerke* gewählt, da es sich um einen sehr umfangreichen, hierarchisch sequenzierten Hypertext handelt, der – neben einigen Übersichtsseiten – ausschließlich HTML-Dokumente dieser Hypertextknotensorte umfasst („Synagogenprojekt – Die Dokumentation ritueller Bauwerke jüdischer Gemeinden", publiziert vom Institut für Bau- und Stadtbaugeschichte der Technischen Universität Braunschweig). Die übergeordnete Hypertextsorte von D 720 wurde als *digitale Karte, Stadtplan* notiert, da es sich um eine Karte handelt, in der einzelne Bereiche mit Hyperlinkankern versehen wurden, die zu HTML-Dokumenten wie z. B. D 720 führen. Diese Karte stammt aus einer Diplomarbeit mit dem Titel „Der dendrologische Park Blücherhof", die am Institut für Management in der Umweltplanung (TU Berlin) angefertigt wurde.

Problemfälle – Zuordnung nicht möglich

Bis auf neun Dokumente konnten alle 750 in der Stichprobe enthaltenen Knoten einer Hypertextknotensorte zugewiesen werden. Bei den verbleibenden Dokumenten liegt entweder das Problem vor, dass der jeweilige Knoten auf keiner erkennbaren Textsorte beruht und somit keine unmittelbar ersichtliche kommunikative Funktion besitzt oder aber, dass zahlreiche Instanzen der unterschiedlichsten Hypertextsortenmodule enthalten sind, so dass keine eindeutige Zuordnung vorgenommen werden konnte.[28]

Abbildung 12.4 stellt – mit Ausnahme von D 749 – *vollständig* reproduzierte HTML-Dokumente als Beispiele dar: D 26 (übergeordnete Hypertextsorte: *virtuelles Museum*) umfasst offenbar, wie D 652 (übergeordnete Hypertextsorte: *Webauftritt einer Professur bzw. Arbeitsgruppe*), lediglich eine Überschrift oder eine Bildunterschrift.[29] D 106 beinhaltet den Namen und die Berufsbezeichnungen einer Person (übergeordnete Hypertextsorte: *Webangebot einer Lehrveranstaltung*). Diese Datei wurde im Zuge der Konvertierung eines Dokuments in die *Hypertext Markup Language* von der verwendeten Textverarbeitung automatisch generiert.

Im Fall von D 319 konnte keine Funktion des Dokuments bestimmt werden. Die übergeordnete Hypertextsorte wurde als *Kunst- und Kulturprojekt* bestimmt: Es handelt sich um den Hypertext „Ein Friedhof senkrecht in den Himmel", der auf dem 1973 entwickelten, gleichnamigen Entwurf einer neuen Form der Totenkultur von Bazon Brock basiert (vgl. `http://www.brock.uni-wuppertal.de/Projekte/friedhof/welcome.htm`). Dieser Hypertext umfasst biografische Angaben mehrerer Personen, die Selbstbeschreibungen auf mehreren subjektiven und objekten Ebenen darstellen. Zusätzlich kann eine virtuelle „Kiste" angelegt werden, die persönliche Gegenstände aufnimmt. Bei D 319 handelt es sich um die Darstellung eines solchen Gegenstandes. D 319 ist das einzige in der Stichprobe enthaltene

[28] Im Falle von D 96 liegt zudem das Problem vor, dass sich das korrespondierende Dokument nicht im Korpus befindet. Bei der Erzeugung der in die Datenbank zu importierenden Tabelle wurde diesbezüglich keine Fehlermeldung geliefert (vgl. Abschnitt 7.2.4). Auch eine detaillierte Analyse der Protokolldateien konnte nicht die Frage klären, weshalb das Dokument nicht verfügbar ist. Möglicherweise liegt die Ursache in einem Fehler des verwendeten *Crawlers Pavuk* oder in einem vom Webserver fehlerhaft generierten HTTP-Response-Header.

[29] Den im Korpus enthaltenen bei- und übergeordneten Dokumenten zufolge wurden D 26 und D 652 *nicht* als Bestandteile von Framesets konzipiert.

D 319

Zum Verhältnis von Emblemliteratur zur Malerei

D 26

Aktuelles

D 652

Prof. Dr. Joachim Vogel, Tübingen
Richter am Oberlandesgericht

D 106

Bitte nutzen Sie auch die neuen Seiten des D.I.R.: **www.dir-info.de**

Interkultureller Rat in Deutschland

[Zurück zur Hauptseite Literatur] [Zurück zum Inhaltsverzeichnis dieser Broschüre]

Gesetzestexte

Art. 3 Grundgesetz

1. Alle Menschen sind vor dem Gesetz gleich.
2. Männer und Frauen sind gleichberechtigt.
3. Niemand darf wegen seines Geschlechts, seiner Abstammung, seiner Rasse, seiner Sprache, seiner Heimat und Herkunft, seines Glaubens, seiner religiösen oder politischen Anschauungen benachteiligt oder bevorzugt werden.

§ 130 StGB Volksverhetzung

Wer in einer Weise, die geeignet ist, den öffentlichen Frieden zu stören, die Menschenwürde anderer dadurch angreift, daß er

1. zum Haß gegen Teile der Bevölkerung aufstachelt,
2. zu Gewalt- oder Willkürmaßnahmen gegen sie auffordert,
3. sie beschimpft, böswillig verächtlich macht oder verleumdet, wird mit einer Freiheitsstrafe von drei Monaten bis zu fünf Jahren bestraft.

§ 131 StGB Gewaltdarstellung. Aufstachelung zum Rassenhaß

1. Wer Schriften, die zum Rassenhaß aufstacheln oder die grausame oder sonst unmenschliche Gewalttätigkeiten gegen Menschen in einer Art schildern, die eine Verherrlichung oder Verharmlosung solcher Gewalttätigkeiten ausdrückt oder die das Grausame oder Unmenschliche des Vorgangs in einer die Menschenwürde verletzenden Weise darstellt,
   1. verbreitet,
   2. öffentlich ausstellt, anschlägt, vorführt oder sonst zugänglich macht,
   3. einer Person unter 18 Jahren anbietet, überläßt oder zugänglich macht oder
   4. herstellt, bezieht, liefert, vorrätig hält, ankündigt, anpreist, in den räumlichen Geltungsbereich dieses Gesetzes einzuführen oder daraus auszuführen unternimmt, um sie oder aus ihnen gewonnene Stücke im Sinne der Nr. 1 bis 3 zu verwenden oder einem anderen eine solche Verwendung zu ermöglichen,

   wird mit einer Freiheitsstrafe bis zu einem Jahr oder mit Geldstrafen bestraft.
2. Rundfunk.
3. Ausnahme für Berichterstattung über Vorgänge des Zeitgeschehens oder der Geschichte.
4. Ausnahme für Sorgeberechtigte hinsichtlich Nr. 3.

§ 185 StGB Beleidigung

Die Beleidigung wird mit Freiheitsstrafe bis zu einem Jahr oder mit Geldstrafe und, wenn die Beleidigung mittels einer Tätigkeit begangen wird, mit Freiheitsstrafe bis zu zwei Jahren oder mit Geldstrafe bestraft.

Was ist der Interkulturelle Rat?

Im Interkulturellen Rat haben sich Deutsche, Ausländerinnen und Ausländer zusammengefunden, die bei Gewerkschaften, Arbeitgeberverbänden, Kirchen, Migrantenorganisationen oder staatlichen Stellen mitarbeiten. Er setzt sich für ein friedliches und gleichberechtigtes Zusammenleben von Einheimischen, Zuwandernden und ethnischen, kulturellen und religiösen Minderheiten ein. Ziel ist die Überwindung von Fremdenfeindlichkeit und Rassismus.

- Der Interkulturelle Rat wendet sich gegen jede Form staatlicher oder gesellschaftlicher Diskriminierung. Er bemüht sich um interkulturelle Dialoge.
- Der Interkulturelle Rat fordert von den politisch Verantwortlichen, durch sachliche Aufklärung zum Abbau von Ängsten, Fremdenfeindlichkeit und Rassismus beizutragen. Einwanderung ist politisch zu gestalten, eine interkulturelle Politik ist zu entwickeln.
- Der Interkulturelle Rat setzt sich im einzelnen ein für
  - die Erweiterung des kommunalen Wahlrechts für Ausländerinnen und Ausländer über die Bürger der Europäischen Union hinaus
  - die doppelte Staatsbürgerschaft
  - die Einführung von Antidiskriminierungsregelungen, die geeignet sind, fremdenfeindliche, rassistische und antisemitische Äußerungen und Handlungen wirksam zu verhindern.
- Der Interkulturelle Rat besteht zu gleichen Teilen aus Deutschen und Menschen anderer Nationalität. Er wurde am 31. August 1994 in Frankfurt/Main gegründet. Gleichberechtigte Vorsitzende sind Dr. Jürgen Micksch, Darmstadt, Interkultureller Beauftragter der Evangelischen Kirche in Hessen und Nassau sowie Saskia Santoso, Nürnberg, Vorsitzende der Arbeitsgemeinschaft der Ausländerbeiräte Bayerns. Schatzmeister ist Leo Monz, Düsseldorf, Referatsleiter beim Bundesvorstand des Deutschen Gewerkschaftsbundes (DGB). Die laufenden Geschäfte werden von Regina Zöllmann wahrgenommen.
- Der Interkulturelle Rat in Deutschland e.V. ist vom Finanzamt Frankfurt Süd unter der Nr. 452 508 1109 - K37- als gemeinnützig anerkannt. Für seine Arbeit ist er auf Spenden angewiesen.

Spendenkonto Nr. 647150 – 604 bei der Postbank Frankfurt/Main, BLZ 500 100 60. Spendenbescheinigungen werden ab DM 100,- zugesandt.

Adressenverzeichnis zur Thematik (Auswahl)

Telefone (Auswahl)

- Aachen: 0241/3 84 68 Antifaschistisches/Antirassistisches Telefon
- Berlin: 030/7 85 72 81 Antirassistische Initiative e.V./Antirassistisches Telefon
- Bonn: 0228/24 11 00 Bonn Telefon
- Braunschweig: 02531/34 11 34 Antirassistisches Telefon
- Bremen: 0421/70 64 44 Antirassismus - Büro
- Darmstadt: 06151/2 64 80 Darmstädter Bündnis gegen Rassismus
- Duisburg: 0203/28 46 73 Anti-Rassismus-Informationszentrum in Nordrhein-Westfalen (ARIC-NRW)
- Düsseldorf: 0211/9 92 00 00 Nottelefon gegen Rassismus in Düsseldorf
- Essen: 0201/23 20 60 Antirassismus Telefon
- Frankfurt: 069/70 33 37 Antifaschistisches/Antirassistisches Notruftelefon
- Hamburg: 040/43 15 87 Antirassistisches Telefon Hamburg
- Mannheim: 0621/1 56 41 41 Antirassistisches Telefon
- München: 089/5 43 96 12 Antirassistisches Telefon
- Nürnberg: 0911/26 20 88 Antirassistisches Telefon
- Passau 0851/3 61 06 Infotelefon Antifa

Impressum:

Herausgeber: Interkultureller Rat in Deutschland e.V., mit freundlicher Unterstützung der Europäischen Kommission, des Büros für Einwanderer und Flüchtlinge des Landes Hessen, des Bundesvorstandes des Deutschen Gewerkschaftsbundes (DGB), der Evangelischen Kirche in Hessen und Nassau und des Fördervereins PRO ASYL e.V.

Redaktion: Regina Zöllmann mit Unterstützung von Manuel Campos, Sabine Kriechhammer-Yagerur, Bernd Mesovic, Jürgen Micksch, Leo Monz, Saskia Santoso, Mechthild Schirmer, Robin Schneider, Georgis Tsapanos und Cornelia Wilss-Hasenkamp.

Redaktionsschluß: 20. Januar 1996

Umschlaggestaltung: Rambow + van de Sand Graphik und Design, Frankfurt/Main unter Verwendung eines Photos der Photoagentur Helga Lade/Keres, Frankfurt/Main
Layout: Wolfgang Scheffler, Mainz
Druck: Linea Plus GmbH, Frankfurt/Main
Bestellungen:
Bestelladresse: Interkultureller Rat in Deutschland e.V.,
Postfach 10 11 43, 60011 Frankfurt/Main,
Telefon: 069/23 40 54, Fax: 069/23 06 50
Einzelbestellung der Broschüre "Gleichbehandlung statt Diskriminierung": 1,50 DM zuzüglich Versandkosten

Außerdem können beim Interkulturellen Rat in Deutschland e.V. folgende Materialien bestellt werden:

Plakatserie "Rassismus macht einsam" (4 Motive):

(Einzelbestellung oder Bestellung als komplette Serie solange Vorrat reicht, jeweils zuzüglich Versandkosten)

Plakat "**Frankfurt den Frankfurtern**"
(Format DIN A0; Preis 2,50 DM /Expl.; Format DIN A2; Preis 0,50 DM /Expl.)

Plakat "**Hamburg den Hamburgern**"
(Format DIN A0; Preis 2,50 DM /Expl.; Format DIN A2; Preis 0,50 DM /Expl.)

Plakat "**Sachsen den Sachsen**"
(Format DIN A0; Preis 2,50 DM /Expl.; Format DIN A2; Preis 0,50 DM /Expl.)

Plakat "**Bayern den Bayern**"
(Format DIN A0; Preis 2,50 DM /Expl.; Format DIN A2; Preis 0,50 DM /Expl.)

[Zurück zur Hauptseite Literatur] [Zurück zum Inhaltsverzeichnis dieser Broschüre]

D.I.R. Page maintained by Info-Zentrum für Rassismusforschung, dir@mailer.uni-marburg.de.
Copyright© D.I.R. e.V. Created: 5.4.1996 Updated: 5.4.1996

D 749

Abbildung 12.4: Beispiele für Dokumente, bei denen keine eindeutige Zuordnung zu einer Hypertextknotensorte durchgeführt werden konnte

Dokument, das kein Bestandteil einer Gebrauchshypertextsorte ist. Bei der Erstellung eines Hypertextes zur Realisierung eines kulturtheoretischen und künstlerisch wirkenden Projekts ist naturgemäß zu erwarten, dass von etablierten Textsorten kein Gebrauch gemacht wird.

Ein besonders prägnantes Beispiel ist D 749: Hierbei handelt es sich um ein extrem langes Dokument, das auf einer gedruckten Broschüre basiert, die den Verein „Interkultureller Rat in Deutschland e. V." vorstellt.[30] Der in Abbildung 12.4 enthaltene Bildschirmabzug stellt lediglich einen Auszug dieses Dokuments dar und wurde aus vier verschiedenen Screenshots zusammengesetzt; die Hypertextversion der „Broschüre" besteht aus insgesamt sechs HTML-Dokumenten ähnlicher Länge. Dokument D 749 umfasst Instanzen unterschiedlichster Hypertextsortenmodule, z. B. Auszüge aus Gesetzestexten, einen Aufruf, einen selbstdarstellenden Vorstellungstext, mehrere Adresslisten, ein Telefonverzeichnis, ein Impressum und einen Produktkatalog. Aufgrund dieser Vielzahl heterogener Hypertextsortenmodule, die in einer sehr untypischen und keinesfalls als konventionalisiert geltenden Kombination vorliegen, konnte für D 749 keine eindeutige Zuordnung vergenommen werden; auf die Zuweisung einer Hypertextknotensorte wie z. B. *Seite/Abschnitt einer konvertierten gedruckten Broschüre* wurde verzichtet, da es sich um einen sehr ungewöhnlichen Knoten handelt, dessen zugehöriger Hypertext ebenfalls eine untypische Struktur aufweist.

Dass in D 26, D 106 und D 652 keine Hypertextknotensorten vorliegen, kann zumindest teilweise durch das Hypertextsortenmodell erklärt werden: Die Inhalte von D 26 und D 652 können lediglich auf der Ebene der Textoberfläche als isolierte Textfragmente, d. h. als Hypertextmodule interpretiert werden (vgl. Abschnitt 5.6.2). Eine funktionale Differenzierung kann nicht vorgenommen werden, da keine weiteren Hypertextmodule in diesen Dokumenten enthalten sind und somit der Kontext fehlt, um z. B. das in D 652 enthaltene Hypertextmodul als Überschrift auffassen zu können. Der Inhalt von D 106 kann zwar als untypische Instanz des komplexen Hypertextsortenmoduls *Identifikation* interpretiert werden (vgl. Tabelle 10.3, S. 457), jedoch rekurriert die in D 106 verwendete, sehr rudimentäre Ausprägung des Hypertextsortenmoduls nicht auf ein etabliertes Textstrukturmuster, weshalb es von Autoren in der Regel nicht als *einziges* Hypertextsortenmodul eines Knotens eingesetzt wird, so dass es auch nicht als konventionalisierte Hypertextknotensorte aufgefasst werden kann (vgl. die Abschnitte 5.5 und 5.6).

12.8 Die Ergebnisse im Kontext verwandter Arbeiten

Im Folgenden werden die Ergebnisse der in diesem Kapitel vorgestellten Studie mit den traditionellen Textsorten des Kommunikationsbereiches Hochschule und Wissenschaft (Abschnitt 12.8.1) sowie denjenigen Arbeiten verglichen, in denen Identifizierungen und Sammlungen von Web-Genres angefertigt wurden (Abschnitt 12.8.2).

[30] Im Kopf des Dokuments befindet sich der Hyperlinkanzeiger „Zurück zum Inhaltsverzeichnis dieser Broschüre". Die URL des Dokuments – `http://www.uni-marburg.de/dir/MATERIAL/BUCH/KOMPLETT/INTERKU5.HTML` – enthält interessanterweise den Bestandteil „Buch": Unter `http/www.uni-marburg.de/dir/MATERIAL/BUCH/` stehen verschiedene Publikationslisten und auch „Komplette Broschüren, Bücher ..." bereit. Der Hypertext ist Teil des Webangebots des Vereins „Dokumentations- und Informationszentrum für Rassismusforschung D.I.R. e. V." (vgl. `http://www.uni-marburg.de/dir/` und `http://www.dir-info.de`).

12.8.1 Traditionelle Textsorten in der Untersuchungsdomäne

Tabelle 12.24 stellt die Verteilung der ermittelten Hypertextknotensorten sowie der ihnen zugeordneten Dokumente auf mehrere Textsortenklassen dar. Hierzu gehören zunächst die von Heinemann (2000b) vorgeschlagenen „Textsorten innerhalb des Kommunikationsbereiches Hochschule und Wissenschaft".[31] Zusätzlich werden drei weitere Textsortenklassen angenommen, die ebenfalls den Textsorten der Hochschule zuzurechnen sind: Die „studiumsbezogenen Textsorten" beziehen sich auf die von Heinemann nicht diskutierten Hypertextknotensorten wie z. B. *Studienordnung* und *Stundenplan*. Die „weiteren Textsorten aus dem universitären Bereich" umfassen diejenigen Hypertextknotensorten, die zwar dem genannten Bereich zugehörig sind, von der in Abbildung 6.1 dargestellten Typologie jedoch nicht abgedeckt werden (z. B. *Pressemitteilung*, *Bibliothekskatalog*, *Bibliothekssystematik*, *Mitarbeiterverzeichnis* und *Publikationsliste*).[32] Dies gilt auch für die „spezialisierten Fachtextsorten", die zusätzlich die Eigenschaft der Fachspezifik aufweisen (vgl. Abschnitt 12.7.7). Die Kategorie „allgemeine Textsorten" bezieht sich auf etablierte Textsorten, die nicht unmittelbar dem universitären Bereich zugeordnet werden können (z. B. *Anleitung bzw. Benutzungshinweise*, *Fotogalerie*, *Preisliste*, *Kochrezept*, *Kleinanzeige*, *Denksportaufgabe* und *Kinoprogramm*), und die Kategorie „neue Hypertextsorten bzw. Hypertextknotensorten" beinhaltet diejenigen Hypertext(knoten)sorten, die kein unmittelbares Pendant in den traditionellen Textsorten aufweisen (z. B. *Einstiegsseite*, die Varianten der *persönlichen Homepage*, *Hotlist*, *Verteiler* und *Suchformular*). Die letzte Textsortenklasse umfasst diejenigen Vorkommen, die entweder nicht als Hypertextknotensorte aufgefasst werden können (*primäre Navigationshilfe*, *Kopfzeile*, *Fußzeile*) oder einer solchen nicht zugeordnet werden konnten (vgl. Abschnitt 12.7.7).

Tabelle 12.24 verdeutlicht, dass lediglich 66,5% der untersuchten Instanzen Hypertextknotensorten zuzuordnen sind, die zu den universitären Textsorten gehören. Etwa ein Drittel der Knoten basiert entweder auf den unterschiedlichsten allgemeinen Textsorten (vornehmlich aus dem Bereich der Gebrauchstextsorten) oder auf neuen Hypertext(knoten)sorten. Es kann somit festgehalten werden, dass sich universitäre Webangebote nicht nur auf Textsorten aus den Kommunikationsbereich Hochschule und Wissenschaft stützen, sondern auch zahlreiche weitere Textsorten sowie Hypertext(knoten)sorten verwenden, die sich seit der Etablierung des Mediums *World Wide Web* gebildet haben.

12.8.2 Zur Restriktion der Untersuchungsdomäne

Die von Crowston und Williams (1997, 2000), Roussinov et al. (2001), Shepherd und Watters (1999) sowie Haas und Grams (1998a,b, 2000) angefertigten Studien zur Identifizierung und Sammlung von Web-Genres werden in Abschnitt 4.4 diskutiert. Da die jeweiligen Listen

[31] Die in Tabelle 12.24 dargestellten Kategorien beziehen sich auf Abbildung 6.1 (S. 305), die verschiedene Erweiterungen an Heinemanns Typologie umfasst (vgl. Abschnitt 6.2.4). Tabelle 12.24 basiert auf den Hypertextknotentypen und -sorten, die in den Tabellen 12.3 und 12.4 präsentiert werden. Statt des Hypertextknotentyps *Seite/Abschnitt* wurden jedoch die 20 Hypertextknotensorten dieses Typs verwendet, so dass insgesamt 134 Etiketten zugeordnet wurden. Da in Abbildung 6.1 Mehrfachzuordnung erlaubt ist, wurde diese auch in Tabelle 12.24 bei 22 Instanzen von Hypertextknotensorten angewendet.

[32] Die Hypertextknotensorten dieser Kategorie befinden sich in einem Grenzbereich, d. h. die genannten Hypertextknotensorten werden zusätzlich in mehreren anderen Kommunikationsbereichen instanziiert, können jedoch im Arbeitsalltag eines Hochschulangehörigen durchaus als typische Textsorten verstanden werden.

Textsortenklasse	Hypertextknotensorten		Instanzen der Hypertextknotensorten	
	Frequenz	Prozent	Frequenz	Prozent
Textsorten in Hochschule und Wissenschaft (vgl. Abbildung 6.1, S. 305)				
• Theoriebezogene Textsorten				
∗ Primäre Textsorten	11	8,2	24	3,1
∗ Sekundäre Textsorten	3	2,2	45	5,8
• Wissenstransmittierende Textsorten				
∗ Lehrveranstaltungsorganisierende Textsorten	8	6,0	82	10,6
∗ Primäre Textsorten (Lehrmaterialien)	9	6,7	135	17,5
∗ Sekundäre Textsorten (Wissensverarbeitung)	—	0,0	—	0,0
∗ Studentische Schrifttexte	6	4,5	26	3,4
∗ Textsorten der Überprüfung	5	3,7	22	2,8
• Textsorten der Wissenschaftsverwaltung				
∗ Verwaltungsexterne Textsorten	2	1,5	6	0,8
∗ Verwaltungsinterne Dienstanweisungen/Geschäftsordnungen	—	0,0	—	0,0
∗ Verwaltungsinterne Textsorten	12	9,0	37	4,8
Studiumsbezogene Textsorten	9	6,7	26	3,4
Weitere Textsorten aus dem universitären Bereich	19	14,2	100	13,0
Spezialisierte Fachtextsorten	5	3,7	10	1,3
Allgemeine Textsorten	26	19,4	121	15,7
Neue Hypertextsorten bzw. Hypertextknotensorten	20	14,9	106	13,7
Zuordnung nicht möglich	4	3,0	32	4,1

Tabelle 12.24: Verteilung der ermittelten Hypertextknotensorten auf Textsortenklassen

von Genres bzw. Web-Genres nur partiell deckungsgleich sind und die prozentualen Vorkommen spezifischer Kategorien nicht in allen Fällen aufgeführt werden, wird an dieser Stelle auf eine direkte Gegenüberstellung der Ergebnisse verzichtet.

Es können viele der Schwierigkeiten bestätigt werden, von denen die Verfasser der genannten Studien berichten. Haas und Grams (1998b, S. 102) merken z. B. an, dass sich in ihrer Stichprobe zahlreiche interne Knoten befinden, deren Funktionen nur durch eine Betrachtung der bei- und übergeordneten Dokumente genauer ermittelt werden können. Haas und Grams *verzichten* jedoch auf die Analyse des Kontextes, da sie sich auf „page types“ konzentrieren. Das in Kapitel 5 eingeführte Hypertextsortenmodell wurde unter anderem vor dem Hintergrund der in der Literatur berichteten Problemkomplexe sowie auf der Basis mehrerer Vorstudien entwickelt und stellt einen geeigneten Ansatz für ihre Lösung dar. Die von Haas und Grams angesprochene Schwierigkeit der Zuordnung sowie die in den vorangegangenen Kapiteln präsentierten Analysen zeigen, dass HTML-Dokumente keinesfalls isoliert betrachtet werden können. Vielmehr ist es notwendig, eine Differenzierung zwischen ihrer Hypertextknotensorte und der Hypertextsorte des übergeordneten funktionalen Ganzen vorzunehmen. Zusätzlich existiert die Schwierigkeit der Einhaltung adäquater und einheitlicher Abstraktionsebenen. Die von Crowston und Williams und Roussinov et al. präsentierten Listen von Web-Genres können als Beleg für diese Notwendigkeit aufgefasst werden, denn ihre Genre-Bezeichnungen beziehen sich in gleicher Weise auf Hypertexttypen, Hypertextsorten, Hypertextknotentypen und Hypertextknotensorten (vgl. Tabelle 4.1, S. 176).

Eine weitere Gemeinsamkeit der verwandten Arbeiten besteht in der sehr ausgeprägten Heterogenität ihrer Ergebnisse, weil Stichproben untersucht wurden, die nahezu ohne jegliche Einschränkungen aus den Beständen von Suchmaschinen zusammengestellt wurden (vgl. Abschnitt 4.4.8). So berichten z. B. (Crowston und Williams, 1997, S. 34), dass bei einigen Dokumenten zwar durchaus Instanzen standardisierter Genres vorliegen, die jedoch von den Verfassern nicht benannt werden konnten, weil sie keine Mitglieder der Diskursgemeinschaften sind, die diese Genres verwenden (ähnlich bei Crowston und Williams, 2000, S. 205). Aus eben diesem Grund wurde die Untersuchungsdomäne der vorliegenden Arbeit auf die Webangebote von Universitäten eingeschränkt. Diese Restriktion hat sich als geeignetes Mittel der weitgehenden Vermeidung des von Crowston und Williams angesprochenen Problems erwiesen. Da universitäre Webauftritte jedoch zwangsläufig auch die Inhalte zahlreicher Fachdisziplinen umfassen, konnten spezialisierte Textsorten aus diesen Fachdisziplinen (zehn der 750 Dokumente) nicht vollständig umgangen werden.

12.9 Fazit

Im Rahmen dieser Analyse wurden 750 zufällig ausgewählten HTML-Dokumenten 65 übergeordnete Hypertexttypen bzw. -sorten zugewiesen. Sechs Hypertexttypen weisen Subkategorien auf, die insgesamt 53 Hypertextsorten darstellen. Bezüglich der zweiten Analyseebene konnten 114 Hypertextknotentypen bzw. -sorten ermittelt werden, wobei in zehn Fällen Subkategorien in Form von insgesamt 54 Hypertextknotensorten vorliegen. Im Hinblick auf die publizierenden Institutionen und Organisationseinheiten ist festzuhalten, dass neben universitären (96% der Dokumente) auch außeruniversitäre Einrichtungen (4%) Informationsangebote auf den Webservern von Hochschulen veröffentlichen. Zusätzlich ist anzumerken, dass *keines* der 750 Dokumente als Knoten bzw. informationelle Einheit eines prototypischen Hypertextes aufgefasst werden kann (vgl. Abschnitt 3.7).[33]

Sowohl die Restriktion der Untersuchungsdomäne als auch das Hypertextsortenmodell haben sich als geeignete Instrumente zur Vermeidung derjenigen Problemfälle erwiesen, die von nahezu allen verwandten Arbeiten berichtet werden. Dennoch war gerade die Zuweisung der übergeordneten Hypertextsorte eines Dokuments in vielen Fällen nur durch eine Analyse bei- und übergeordneter Knoten möglich. Zudem wurde in dieser Studie vollständig von der Ebene der Hypertextsortenmodule abstrahiert, woduch jedoch nur in sehr wenigen Fällen potenzielle Ambiguitäten durch mehrere Instanzen unterschiedlicher Hypertextsortenmodule hervorgerufen wurden. Das Inventar der ermittelten Hypertextknotentypen und -sorten besitzt zwar eine ausreichende Trennschärfe (z. B. war in keinem Fall eine Mehrfachzuordnung notwendig), sie ist jedoch im Vergleich zu der in Kapitel 11 untersuchten Stichprobe insgesamt etwas schwächer ausgeprägt.

Die Bandbreite der ermittelten Hypertextknotentypen und -sorten kann als sehr umfassend bezeichnet werden: Etwa 42% sind den traditionellen Textsorten des Kommunikationsbereiches Hochschule und Wissenschaft zugehörig, wobei die wissenstransmittierenden

[33] Die ausgeprägteste Tendenz zum unsequenzierten, d. h. nicht-linear organisierten Hypertext liegt in D 236, D 337, D 586 und D 663 vor (vgl. auch Abbildung 12.3, S. 562). Bei diesen Dokumenten handelt es sich um Instanzen der Hypertextknotensorte *medizinische Diagnoseprozedur*, deren zugehöriger Hypertext jedoch primär hierarchisch sequenziert ist und darüber hinaus maschinell erzeugt wurde.

Textsorten mit 21% überwiegen. Weiterhin können die verwendeten Hypertextknotentypen und -sorten als studiumsbezogene Textsorten (7%), weitere Textsorten aus dem universitären Bereich (14%) und spezialisierte Textsorten aus unterschiedlichen Fachdisziplinen aufgefasst werden (3%). Aus dem sehr umfangreichen Bereich der allgemeinen Textsorten werden vornehmlich Gebrauchstextsorten eingesetzt (19%). Etwa 14% der Hypertextknotentypen und -sorten besitzen kein unmittelbares Pendant in der Gruppe der traditionellen Textsorten.

Die Umsetzung traditioneller Textsorten im WWW erfolgt auf der Grundlage unterschiedlicher Prinzipien. So bieten kleinere universitäre Organisationseinheiten z. B. ein Vorlesungsverzeichnis an, das als einzelnes HTML-Dokument realisiert ist, andere bieten eine Instanz der Hypertextsorte *Vorlesungsverzeichnis* an, zudem wird oftmals alternativ oder auch zusätzlich auf das Gesamtverzeichnis einer Hochschule verwiesen. Die Analyse bestätigt zwei zentrale Aspekte: Erstens umfassen die unterschiedlichen Ausprägungen des Hypertexttyps *Webauftritt einer Organisationseinheit* jeweils spezifische Hypertextknotensorten sowie eingebettete Hypertextsorten. Zweitens wird eben dieses Repertoire vom Status einer Einrichtung determiniert, d. h. eine Vielzahl der Textsorten, mit denen Angehörige einer bestimmten Organisationseinheit im Arbeitsalltag umgehen, wird auch in ihrem Webauftritt eingesetzt. Dabei handelt es sich primär um Textsorten mit allgemeiner Informationsfunktion, deren Textexemplare sich an unterschiedliche Zielgruppen richten. Der generische Aufbau einer Organisationseinheit beeinflusst nicht nur das Repertoire der im korrespondierenden Webauftritt verwendeten Textsorten, sondern auch die Strukturierung der Website, so wird z. B. der *Webauftritt einer Professur bzw. Arbeitsgruppe* oftmals in Rubriken wie „Lehrveranstaltungen“, „Projekte“, „Publikationen“, „Mitarbeiter“, „Forschungsschwerpunkte“ und „Kontakt“ eingeteilt, denen wiederum Instanzen von Hypertext(knoten)sorten wie z. B. *Webangebot einer Lehrveranstaltung*, *persönliche Homepage eines Wissenschaftlers*, *Kurzdarstellung des Arbeitsgebiets (einer Organisationseinheit)* und *Abstract eines Forschungsprojekts* zugeordnet werden können. Diese Instanzen basieren wiederum überwiegend auf bereits in digitaler Form vorhandenen Dokumenten (z. B. Präsentationen, Vorlesungsverzeichnisse, Veröffentlichungen jeglicher Art, Studien- und Prüfungsordnungen, Projektanträge, Abschlussberichte, Publikationslisten etc.), die entweder nach HTML konvertiert oder mittels *Copy & Paste* in entsprechend vorbereitete Webseitengerüste eingefügt werden. Insbesondere für die Einstiegsseite und die strukturierenden Dokumente liegen jedoch keine präfabrizierten Texte vor, so dass sich das Modell der Entwicklung von Hypertextsorten (vgl. Abschnitt 4.3.2) bei der manuellen Anfertigung von HTML-Dokumenten insbesondere auf diejenigen Hypertextknotensorten bezieht, die in den traditionellen Textsorten kein unmittelbares Pendant besitzen.

Die Resultate der Studie weisen im Vergleich zu den Ergebnissen, die in der zweiten Phase der vierten Analyse ermittelt wurden, zahlreiche Unterschiede auf, die bereits in der Verteilung der jeweils untersuchten Knoten in grobe Typen deutlich werden (Einstiegsseite: 40,5% vs. 3,2% aller Knoten; Hyperlinkliste: 30,9% vs. 1,8%; Inhaltsknoten: 28,6% vs. 95%). Ein Unterschied besteht auch in der Varianz der erhobenen Hypertextknotensorten, die in dieser abschließenden Analyse deutlich umfangreicher ausfällt. Weiterhin zeigt diese Analyse zahlreiche weitere Beispiele für den Umstand auf, dass Instanzen einer Vielzahl traditioneller Textsorten, die sowohl aus dem Kommunikationsbereich Hochschule und Wissenschaft, als auch aus der sehr weitläufigen Klasse der Gebrauchstextsorten stammen, in universitären Webauftritten eingesetzt werden. Es kann somit insgesamt festgehalten werden, dass die

Webangebote von Universitäten und Hochschulen von einer sehr heterogenen Menge von Hypertextsorten und Hypertextknotensorten gekennzeichnet sind, wodurch die bereits in Abschnitt 11.8 ausführlich diskutierte Problematik der Typologisierung von Hypertextsorten sowie beteiligter Hypertextknotensorten und Hypertextsortenmodule zusätzlich verschärft wird. Das nachfolgende Kapitel 13 zeigt eine Vorgehensweise auf, mit der dieser Problematik in textlinguistischer und texttechnologischer Hinsicht adäquat begegnet werden kann.

Teil IV

Technologische Umsetzung

Überblick

Dieser vierte und abschließende Teil der Arbeit diskutiert technologische Aspekte. In Kapitel 13 wird zunächst ein Repräsentationsformat fur Hypertextsorten eingeführt. Es basiert auf dem texttechnologischen Standardformalismus *Web Ontology Language*, der in der vom *World Wide Web Consortium* ins Leben gerufenen *Semantic Web*-Initiative zur Repräsentation von Ontologien entwickelt wurde. Ontologien können als spezifische Form semantischer Netze aufgefasst werden, die wiederum eine konzeptuelle Verwandtschaft zu Hypertexten aufweisen. Die Strukturierung der Hypertextsortenontologie basiert auf den im Hypertextsortenmodell (Kapitel 5) enthaltenen Ebenen. Die in Teil III erarbeiteten Ergebnisse wurden in diese Ontologie integriert, die darüber hinaus von zwei weiteren Ontologien flankiert wird, bei denen es sich um eine Ontologie wissenschaftlicher Themen und Fachgebiete und eine Domänenontologie handelt. Letztere modelliert unter anderem den Aufbau einer generischen Hochschule und weist sehr enge Beziehungen zu den ermittelten Hypertextsorten auf. Kapitel 14 geht von den bislang vorliegenden Ansätzen zur maschinellen Erkennung von Genres und Web-Genres aus und stellt Diskrepanzen bezüglich der im *World Wide Web* existenten realen Gegebenheiten dar. Ausgehend von diesen Beobachtungen wird eine Systemarchitektur vorgestellt, die zur Berücksichtigung dieser realen Gegebenheiten und somit für den Einsatz derartiger Verfahren in Produktionssystemen benötigt werden. Daraufhin wird die grundlegende Komponente der Architektur anhand des Prototyps eines Textparsers für HTML-Dokumente exemplifiziert, der auf die Identifizierung von Hypertextsortenmodulen abzielt. Anschließend werden die Funktionen weiterer Komponenten erläutert, die in einer vollständigen Implementierung der Systemarchitektur zwingend benötigt werden. Das Kapitel schließt mit einer Überblick über mögliche Einsatzgebiete von Hypertextsorten in sprach- und informationstechnologischen Anwendungen.

13

Repräsentation von Hypertextsorten auf der Basis von Ontologien

13.1 Einleitung

Dieses Kapitel stellt einen textlinguistisch motivierten und texttechnologisch realisierten Ansatz zur Modellierung und Repräsentation von Hypertextsorten vor, der auf einen gewinnbringenden Einsatz im Rahmen der computerlinguistischen Erkennung und Verarbeitung von Hypertextsorten abzielt. Es ist bereits jetzt darauf hinzuweisen, dass sich das Repräsentationsformat nicht an den derzeit realisierbaren texttechnologischen und computerlinguistischen Verfahren orientiert. Vielmehr basiert es auf dem in Kapitel 5 eingeführten Hypertextsortenmodell, um die im *World Wide Web* existenten realen Gegebenheiten erfassen zu können und umfasst die Ergebnisse der in Teil III vorgestellten Stichprobenanalysen. Für eine detaillierte und maschinell durchgeführte Identifizierung von Hypertextsorten, Hypertextknotensorten und Hypertextsortenmodulen ist die vollständige Unterstützung des Repräsentationsformats bzw. sein Einsatz als Ressource in einem sprachverarbeitenden System notwendig. Diesbezüglich existieren jedoch noch zahlreiche Schwierigkeiten, die einen unmittelbaren Einsatz verhindern. Kapitel 14 geht ausführlich auf die Grenzen des derzeit technisch Machbaren ein und zeigt verschiedene Lösungsansätze auf.

Das Repräsentationsformat basiert auf Ontologien, die mit Hilfe der *Web Ontology Language* (OWL) modelliert werden. OWL wiederum lehnt sich an semantische Netze und die Wissensrepräsentation an und ist ein integraler Bestandteil der vom *World Wide Web Consortium* ins Leben gerufenen *Semantic Web*-Initiative (Abschnitt 13.2). Eine maschinell erzeugte Ontologie von wissenschaftlichen Disziplinen und Fachgebieten bildet die erste von drei Wissensbasen (Abschnitt 13.3). Diese wird komplementiert von einem formalen Modell der Untersuchungsdomäne (Abschnitt 13.4). Die dritte Ontologie beschreibt Hypertexttypen und -sorten sowie ihre Konstituenten (Abschnitt 13.5).

13.2 Ontologien und das Semantic Web

Das in den nachfolgenden Abschnitten vorgestellte Repräsentationsformat basiert auf Ontologien. Diese werden im Bereich der Wissensrepräsentation eingesetzt, um das in einer zu beschreibenden Domäne enthaltene Wissen formal modellieren zu können, so dass es weiterführenden Verarbeitungsprozessen zur Verfügung gestellt werden kann, die im Kontext der *Semantic Web*-Initiative oftmals als „intelligente Agenten" bezeichnet werden.

Abschnitt 13.2.1 geht zunächst auf die historischen Ursprünge von Ontologien ein. Daraufhin stellt Abschnitt 13.2.2 die vom *World Wide Web Consortium* initiierte Idee des *Semantic Web* vor. Ein zentraler Bestandteil der Schichtenarchitektur des *Semantic Web* sind Ontologien. Abschnitt 13.2.3 gibt einen knappen Überblick über die vom W3C standardisierte *Web Ontology Language* (OWL), woraufhin Abschnitt 13.2.4 einerseits auf die Parallelen, die zwischen Hypertexten und semantischen Netzen existieren und andererseits auf Anwendungen von Ontologien im Kontext von Hypertexten eingeht.

13.2.1 Semantische Netze und Wissensrepräsentation

Der Terminus *semantic network* geht auf Quillian (1967) zurück, der ein psychologisches Modell zum Aufbau des menschlichen Langzeitgedächtnisses im Hinblick auf die Speicherung und den Vergleich konzeptuellen, d. h. semantischen Wissens vorstellt. Quillian bildet die Bedeutungen von Lexemen bzw. die unterschiedlichen Lesarten polysemer oder homonymer Begriffe auf semantische Konzepte ab (vgl. Abschnitt 2.2.6), die spezifische Eigenschaften aufweisen, welche über unterschiedliche Typen von Relationen realisiert werden und einem traditionellen Wörterbuch entnommen wurden. Auf diese Weise entsteht „a mass of nodes, interconnected by different kinds of associative links." (Quillian, 1967, S. 411).[1]

Die Repräsentation von Wissen besitzt für die KI-Forschung einen zentralen Stellenwert (vgl. Luger, 2001): Wenn Systeme zur Simulation kognitiver Prozesse der Sprachverarbeitung implementiert werden sollen, ist es unumgänglich, diesen Anwendungen Sprach- und Allgemeinwissen zur Verfügung zu stellen, so dass z. B. Verfahren des automatischen Schließens auf diese Wissensbestände angewendet werden können. Die Arbeit von Quillian (1967) hat die Ausrichtung dieses als Wissensrepräsentation bezeichneten Teilgebiets der KI maßgeblich beeinflusst (vgl. Brachman und Levesque, 1985, S. 97): Enzyklopädisches Wissen wird oftmals in Form von Graphen repräsentiert, in denen Konzeptknoten durch zwei unterschiedliche Relationen (vgl. Woods, 1975) mit Knoten verbunden oder durch Eigenschaften spezifiziert werden: Die Relation *is-a* gilt zwischen zwei Knoten und markiert Super-

[1] Diese Erläuterung könnte auch als abstrakte Definition des Konzepts Hypertext verwendet werden. Abschnitt 13.2.4 geht genauer auf den Zusammenhang zwischen semantischen Netzen und Hypertext ein. Darüber hinaus ist auf die enge Verbindung zwischen semantischen Netzen und computer- bzw. textlinguistischen Ansätzen zur Modellierung von Textbedeutungen hinzuweisen (vgl. auch die Abschnitte 2.2.2 sowie 2.2.6). Diesbezüglich sei exemplarisch das hochgradig komplexe Modell zur Repräsentation der Textkohärenz von de Beaugrande und Dressler (1981, S. 100–117) angesprochen, das auf Primär- und Sekundärkonzepten sowie verschiedenen Operatoren basiert, die Verbindungen zwischen Konzepten genauer charakterisieren. Die Anwendung dieses Modells resultiert in einem Graph, der die in einem Text enthaltenen Konzepte gemäß des Textsinns zueinander in Beziehung setzt (de Beaugrande und Dressler, 1981, S. 117, fassen Kohärenz als „Ergebnis der Bedeutungsaktualisierung" auf, „die den Zweck der »Sinn-Erzeugung« verfolgt.").

und Subkonzepte („Kanarienvogel“ *is-a* „Vogel“), wohingegen eine generische Relation wie z. B. *has-property* die Eigenschaft eines Konzepts repräsentiert (vgl. auch Reimer, 1991). In der Folgezeit entstanden zahlreiche Wissensrepräsentationsformalismen, die sich jeweils in ihren Grundprinzipien an semantische Netze anlehnen und spezifische Aspekte des menschlichen Gedächtnisses, der menschlichen Sprachverarbeitung oder der lexikalischen Semantik betonen (vgl. z. B. den Überblick in Brachman, 1979, sowie Fellbaum, 1998).

Über die transitive Relation *is-a* ist es möglich, Vererbungshierarchien zu realisieren, so dass beispielsweise die Eigenschaften des Konzepts „Vogel“ an sämtliche Subkonzepte vererbt werden können. Einige Formalismen erlauben auch die Mehrfachvererbung (Konzepte erben die Eigenschaften mehrerer Superkonzepte) sowie die Überschreibung geerbter Eigenschaften, so dass z. B. bei dem Konzept „Pinguin“ – als Subkategorie von „Vogel“ – die Eigenschaft *kann-fliegen* entfernt werden kann. Daneben werden oftmals die partitiven Relationen *has-part* und *part-of* eingesetzt, um Teil-Ganzes-Beziehungen zwischen Konzepten wie z. B. „Auto“ und „Lenkrad“ zu markieren. Neben der Meronymie werden auch Synonymie und Antonymie in vielen Formalismen benutzt, um zwischen Konzepten geltende lexikalisch-semantische Relationen für ihre weiterführende Spezifikation zu verwenden. Die Instanzrelation wird schließlich dazu eingesetzt, Vertreter von Konzepten in ein semantisches Netz zu integrieren.

13.2.2 Die Semantic Web-Initiative

Die ursprüngliche technologische Realisierung des WWW wurde – primär aus dem Blickwinkel der Hypertexttheorie – in Kapitel 3 vorgestellt. Mit dem zunehmenden Erfolg des *World Wide Web* traten die konzeptionellen Grenzen seiner Basistechnologien HTML und HTTP immer deutlicher zu Tage, weshalb das *World Wide Web Consortium* Ende der neunziger Jahre die *Semantic Web*-Initiative ins Leben gerufen hat (vgl. Berners-Lee et al., 2001).

HTML ist lediglich eine Auszeichnungssprache zur Strukturierung und Verknüpfung beliebiger Dokumente. Das *Semantic Web* basiert hingegen auf der Idee, die maschinelle Verarbeitung von Daten und Informationen zu ermöglichen, so dass „a web of data that can be processed directly or indirectly by machines“ entsteht (Berners-Lee, 1999, S. 177). Zur Realisierung dieses Vorhabens ist die Entwicklung mehrerer neuer Technologien notwendig, die in Form einer Schichtenarchitektur angeordnet sind und auch in dieser Reihenfolge vom W3C entwickelt und standardisiert werden, so dass die Standards sukzessive etabliert und von Anwendungen flankiert werden können. Abbildung 13.1 stellt die auch als „Semantic Web pyramid of languages“ (Fensel et al., 2003a, S. 12) bezeichnete Schichtenarchitektur dar.[2]

Die Grundlage des Schichtenmodells bilden der Unicode-Standard zur Kodierung sämtlicher international gebräuchlicher Zeichen und Zeichensätze sowie der URI-Standard zur Addressierung arbiträrer Ressourcen im WWW (RFC 2396). Die vom W3C standardisierte *Extensible Markup Language* wird zur Strukturierung von Dokumenten und Informationen eingesetzt, um einen flexiblen Datenaustausch zu gewährleisten. XML-Instanzen können auf einer Dokumentgrammatik basieren, die entweder als DTD (Bray et al., 2004b)

[2] Berners-Lee (1999, S. 177–198) erläutert die Architektur anhand verschiedener Anwendungsszenarien (vgl. zusammenfassend Malik, 2003, sowie die zahlreichen Fallbeispiele bei Fensel et al., 2003b, und Davies et al., 2003b). Abbildung 13.1 basiert auf einer Grafik, die Berners-Lee im Rahmen eines Vortrags bei der Konferenz „XML 2000“ präsentiert hat (vgl. `http://www.w3.org/2000/talks/1206-xml2k-tbl/slide10-0.html`).

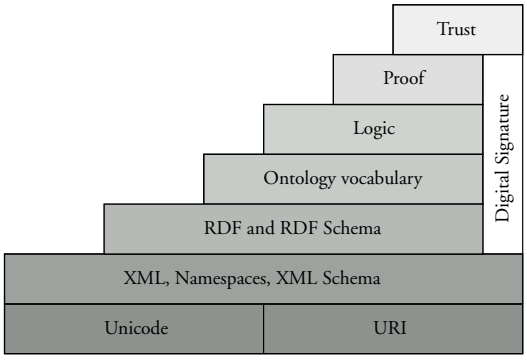


Abbildung 13.1: Die Schichtenarchitektur des Semantic Web

oder als Schema (Fallside et al., 2001) formalisiert wird (vgl. auch Lobin, 2004, sowie Abschnitt 5.2.1). Mit Hilfe von XML Namespaces (Bray et al., 2004a) können Elemente mehrerer Auszeichnungssprachen in einer Instanz verwendet werden. Das *Resource Description Framework* (RDF) wurde zur Kodierung von Metadaten entworfen, in der dritten Schicht können somit Informationen *über* Informationen modelliert werden (vgl. Schmidt, 2004). Ein RDF-Statement bezieht sich auf ein per URI adressierbares *Objekt*, das ein *Attribut* mit einem bestimmten *Wert* besitzt. Dabei können auch weitere Objekte als Werte fungieren, weshalb komplexe RDF-Statements oftmals als gerichtete Graphen repräsentiert werden. Die Notation von RDF erfolgt in einer XML-Syntax. Da RDF kein Inventar von Attributbezeichnungen umfasst, erlaubt *RDF Schema* (RDFS) die Erstellung derartiger Vokabularien, die wiederum mit spezifischen Objekttypen assoziiert werden können. Auf der Basis von Schlüsselwörtern wie z. B. `Class` und `subClassOf` sowie `Property` und `subPropertyOf` kann für die erstellten Typen eine Hierarchie erzeugt werden. Die in RDFS vorgesehenen Mechanismen können bereits zum Aufbau einfacher Ontologien eingesetzt werden (vgl. Fensel et al., 2003a, S. 15), jedoch existieren nur wenige Möglichkeiten zur Spezifizierung von Beschränkungen (z. B. des Wertebereichs). Dieser Ausdrucksschwäche wurde mit der Verabschiedung der *Web Ontology Language*[3] begegnet: OWL besitzt eine wohldefinierte Syntax, eine formale Semantik, eine große Ausdrucksstärke und kann mit Hilfe von Inferenzmaschinen maschinell verarbeitet werden (vgl. Antoniou und van Harmelen, 2004). OWL-Ontologien machen Gebrauch von RDF und den in RDFS definierten Schlüsselwörtern und nutzen zur Notation ebenfalls eine XML-basierte Syntax (vgl. den nachfolgenden Abschnitt 13.2.3).

Für die bislang dargestellten Schichten der Architektur liegen bereits W3C-Standards und zahlreiche korrespondierende Software-Anwendungen vor. Die verbleibenden Schichten befinden sich derzeit noch in der Phase der Konzeptionierung und Entwicklung: Die Ebenen „Logic“ und „Proof“ werden der weiterführenden Verarbeitung von OWL-Ontologien dienen: Zu diesem Zweck sollen Sprachen zur Spezifizierung von Inferenzregeln konzipiert

[3] Dem *Resource Description Framework* und *RDF Schema* liegen insgesamt sechs W3C-Standards zugrunde: Beckett (2004), Brickley und Guha (2004), Manola und Miller (2004), Klyne und Carroll (2004), Hayes (2004) sowie Grant und Beckett (2004). Der Standard OWL besteht ebenfalls aus sechs Teilen: McGuinness und van Harmelen (2004), Smith et al. (2004), Bechhofer et al. (2004), Patel-Schneider et al. (2004), Carroll und Roo (2004) und Heflin (2004).

werden, die wiederum maschinellen Schließverfahren zur Verfügung gestellt werden, so dass komplexe Prozesslogiken implementiert werden können, die sich nicht länger in monolithischen und nicht manipulierbaren Software-Applikationen befinden, sondern in offenen und allgemein zugänglichen Standardformaten. Die Ebene „Trust“ schließlich bezieht sich auf das Erschließen der Vertrauenswürdigkeit einer bestimmten Ressource (z. B. einer spezifischen RDF-Beschreibung). Für diese Schicht sind digitale Signaturen, die sich quer das Schichtenmodell ziehen, von besonderer Bedeutung: Ähnlich dem für E-Mail entwickelten PGP-Mechanismus werden diese zum Signieren von Ressourcen im *Semantic Web* eingesetzt, so dass Dritte eindeutig verifizieren können, dass eine bestimmte Person oder Institution eine Ressource mit dem jeweiligen Schlüssel signiert und publiziert hat. Das „Web of Trust“ (Berners-Lee, 1999, S. 193) soll nun durch die großflächige Repräsentation von Aussagen entstehen, die Informationen über die Vertrauenswürdigkeit[4] anderer Gruppen von Aussagen auf der Grundlage spezifischer Schlüssel besitzen, was Berners-Lee als „the real mirroring of society in technology“ bezeichnet (ebd.) und wie folgt exemplifiziert: Eine Person möchte sich an einem Webserver anmelden, der über eine „trust“-gesteuerte Zugangskontrolle verfügt, die ausschließlich für Mitarbeiter gilt. Der Benutzer besitzt keinen unmittelbaren Zugriff, kann sich aber dennoch als berechtigt ausweisen, indem auf Ressourcen verwiesen wird, die von anderen Personen angeboten und vom Rechner als vertrauenswürdig eingeschätzt werden. Diese somit auch vertrauenswürdigen Ressourcen wiederum bestehen aus mehreren Regeln, aus denen mittels einer Inferenzmaschine gefolgert werden kann, dass es sich bei der Person, die den Zugang nachfragt, tatsächlich um einen Mitarbeiter der Institution handelt. Dieses vereinfacht dargestellte Beispiel verdeutlicht, weshalb es sich bei dieser letzten Schicht der *Semantic Web*-Architektur um „the most complex of cases“ handelt (ebd.), der sich entweder als „a technical dream“ oder als „a legal nightmare“ herausstellen wird (ebd., S. 198).

13.2.3 Die Web Ontology Language

Der Definition von Gruber (1993, S. 199) zufolge handelt es sich bei einer Ontologie im Kontext der Wissensrepräsentation um „an explicit specification of a conceptualization“. Dieser Begriff bezieht sich wiederum auf eine Menge von Wissen und Annahmen, die die Klassen, Konzepte und Instanzen einer Beschreibungsdomäne und die zwischen ihnen existenten Relationen beinhalten. Eine Ontologie umfasst eine explizite Spezifizierung und Repräsentation dieses Wissens in Form von Termen, mit deren Hilfe Entitäten definiert werden können, welche wiederum mit Axiomen assoziiert werden, die die Interpretation der Terme einschränken (ebd.). Eine Ontologie kann somit als Wissensbasis aufgefasst werden, die in gewisser Hinsicht eine spezielle Form eines semantischen Netzes darstellt.[5]

[4] Der zentrale Aspekt der letzten Schicht betrifft den Umstand, dass nicht jede im *Semantic Web* hinterlegte Information oder Ressource pauschal als „wahr“ angenommen werden soll, sondern dass gute Gründe existieren müssen, damit ein Rechner eine Ressource als vertrauenswürdig einschätzen kann.

[5] Das in einer Ontologie enthaltene Wissen korrespondiert immer mit einem spezifischen Weltausschnitt bzw. einer Beschreibungsdomäne, die vor dem Hintergrund einer spezifischen Anwendung modelliert wird. Aus diesem Grund besitzt jede Ontologie naturgemäß Defizite in ihren konzeptuellen Randbereichen, die für die eigentliche Fragestellung oder Modellierungsaufgabe nur von sekundärem Interesse sind: „While human knowledge is often tacit and hard to describe in formal models, there is also no single correct mapping of knowledge into discrete structures.“ (Knublauch et al., 2004a, S. 2).

Ontologien in wissens- und sprachverarbeitenden Systemen

Neben wissensbasierten Anwendungen werden Ontologien auch in komplexen computerlinguistischen und sprachverarbeitenden Systemen, die einen großen Abdeckungsgrad besitzen, eingesetzt, da die Disambiguierung multipler Lesarten als Resultat der Syntaxanalyse und die Konstruktion der Semantik eines Eingabesatzes auf die formale Repräsentation von Weltwissen angewiesen sind, das entweder als Teil des Lexikons oder als separate Ontologie gepflegt wird (vgl. Vossen, 2003). Obwohl sich viele Ansätze auf die eingangs genannte Definition von Gruber stützen, kann keine eindeutige Grenze zwischen Lexika und Ontologien gezogen werden (vgl. Hirst, 2004). Dies zeigt das lexikalisch-semantische Netzwerk WordNet (Fellbaum, 1998), in dem synonyme Termini einer Wortart zu Konzepten („synsets“) gebündelt und mit anderen Konzepten über semantische Relationen in Beziehung gesetzt werden.[6] Derartige semantische Ressourcen werden unter anderem in IR-Anwendungen (zur Erweiterung der Suchanfrage durch eine automatische Expansion der vom Benutzer eingegebenen Termini), in Textklassifikationssystemen (zur Organisierung eines manuell oder maschinell erstellten Kategoriensystems) und im Bereich *Information Extraction* eingesetzt (zur Verknüpfung der von einem Template spezifizierten Informationen mit einer Ontologie).[7]

Ein Überblick über die Web Ontology Language

Die *Web Ontology Language* stellt einen XML- und RDF/RDFS-basierten Formalismus zur Repräsentation von Ontologien im *Semantic Web*-Paradigma dar.[8] Nachfolgend werden die wesentlichen Eigenschaften und Konstrukte von OWL vorgestellt (nach Antoniou und van Harmelen, 2004, und Bechhofer et al., 2004). Eine OWL-Ontologie besteht aus Klassen, Individuen (Instanzen von Klassen) und Propertys (Eigenschaften, die zwischen Individuen gelten). Klassen sind Spezifikationen von Konzepten, die als Mengen von Individuen aufgefasst werden und besitzen formale Beschreibungen, die Restriktionen über Eigenschaften ausdrücken. Erfüllt ein Individuum die Restriktionen einer Klasse, handelt es sich um einen Vertreter dieser Klasse. Klassen (und auch ihre Eigenschaften) werden in einer Subsumptionshierarchie (Taxonomie) angeordnet, die die Klasse `owl:Thing` als Wurzelknoten enthält.

Von zentraler Bedeutung ist die Tatsache, dass OWL eine formale Semantik besitzt, auf deren Basis die Bedeutung des in einer Ontologie enthaltenen Wissens eindeutig beschrieben werden kann. Darüber hinaus ermöglicht sie Inferenzprozesse: Wenn x die Instanz einer

[6] Vossen (2003) geht auf die unterschiedlichen Disziplinen ein, in denen der Begriff „Ontologie“ verwendet wird und stellt die jeweiligen Spezifika heraus.

[7] Franconi (2003) bietet einen Überblick über wissensbasierte Systeme im Bereich der Sprachtechnologie. Vossen (2003) nennt weitere Anwendungen von Ontologien und lexikalisch-semantischen Netzwerken. Erdmann (2001, S. 72) weist darauf hin, dass in Applikationen „ein ganzes Spektrum von Ontologien zum Einsatz [kommt], d. h. unter dem Begriff der Ontologie werden die unterschiedlichsten Wissensstrukturen verstanden. Sie reichen von kaum strukturierten Glossaren oder Thesauri über Begriffstaxonomien oder objekt-orientierte Modelle bis hin zu vollständig formalisierten, logischen Theorien.“

[8] OWL basiert auf den drei Formalismen DAML (*DARPA Agent Markup Language*), OIL (*Ontology Inference Layer*) und DAML+OIL (vgl. etwa McGuinness et al., 2003, und Klein et al., 2003). Mit *XML Topic Maps* (Pepper und Moore, 2001) existiert ein ähnliches Format, das auf einem ursprünglich für SGML entwickelten Standard basiert (ISO/IEC 13250). XTM ist jedoch keine W3C-Entwicklung, so dass bereits jetzt abzusehen ist, dass *XML Topic Maps* von der *Web Ontology Language* unter anderem aufgrund der umfangreichen Unterstützung durch verfügbare Werkzeuge verdrängt werden wird.

Klasse C ist, und C eine Subklasse von D ist, dann kann geschlossen werden, dass x auch eine Instanz von D ist. Neben der Klassenzugehörigkeit kann auch die Äquivalenz von Klassen gefolgert werden: Wenn eine Klasse A äquivalent zu B ist und B ist äquivalent zu C, dann ist A ebenfalls äquivalent zu C. Zusätzlich können Klassifikationsprozesse durchgeführt werden: Wenn bestimmte Attribut-Wert-Paare hinreichende Bedingungen für die Mitgliedschaft einer Klasse A sind und x diese Bedingungen erfüllt, kann geschlossen werden, dass x eine Instanz von A sein muss. Derartige Inferenzprozesse werden durchgeführt, indem die Ontologie auf einen Logikformalismus abgebildet wird, der von einer Inferenzmaschine verarbeitet werden kann. OWL verwendet für diesen Zweck die *Description Logic* (Beschreibungslogik) als Untermenge der Prädikatenlogik erster Stufe (vgl. Baader et al., 2003, 2004).

Der OWL-Standard (vgl. Fußnote 3, S. 576) sieht drei Ausprägungen der Sprache vor, die bezüglich ihrer Ausdrucksstärke abnehmen und eine Aufwärtskompatibilität besitzen: In OWL Full können sämtliche Konstrukte verwendet werden, jedoch sind derartige Ontologien aufgrund ihrer Mächtigkeit nicht berechenbar. Dieser Umstand gilt nicht in OWL DL, das verschiedenen Einschränkungen unterliegt, die Entscheidbarkeit und Kompatibilität mit der Beschreibungslogik und somit den Einsatz maschineller Verfahren gewährleisten. Die dritte Ausprägung, OWL Lite, weist im Vergleich zu OWL DL zusätzliche Restriktionen auf.

Da OWL auf XML basiert, sind OWL-Ontologien zugleich XML-Instanzen. Darüber hinaus basiert OWL auf den Konstrukten, die in RDF und RDFS definiert werden, so dass eine OWL-Ontologie ebenfalls ein RDF-Dokument darstellt und das Wurzelelement `rdf:RDF` besitzt, das Namespace-Deklarationen enthält (z. B. für den Präfix `owl`). Ähnlich dem in HTML verwendeten `head` umfasst das Element `owl:Ontology` den Kopf einer Ontologie, der z. B. Kommentare und Versionsangaben enthalten kann und darüber hinaus mittels `owl:imports` zusätzliche Ontologien importieren kann. Konzepte werden in OWL als Klassen bezeichnet, die mit Hilfe von `owl:Class` definiert werden. Über das Attribut `rdf:ID` wird einer Klasse ein Etikett zugewiesen und als Elementinhalt enthält `owl:Class` das Element `rdfs:subClassOf`, das auf die Superklasse verweist – Klassen der obersten Ebene referenzieren `owl:Thing`. Mittels `owl:disjointWith` und `owl:equivalentClass` können darüber hinaus Klassen angegeben werden, die zur definierten Klasse disjunkt oder äquivalent sind.

Ein Merkmal wird in OWL als Property bezeichnet und definiert eine binäre Relation, die entweder zwischen zwei Objekten gilt oder ein Objekt mit einem Datentyp in Beziehung setzt.[9] Die erlaubten Datentypen wurden zwar dem Inventar von XML Schema entnommen (vgl. Abbildung 13.1), die Definition weiterer Datentypen ist jedoch nicht möglich. Mittels `owl:DatatypeProperty` kann ausgedrückt werden, dass eine Eigenschaft wie z. B. `age` nur durch eine positive ganze Zahl instanziiert werden kann. Mit Hilfe von `owl:ObjectProperty` kann eine Eigenschaft wie z. B. `isTaughtBy` definiert werden, die Objekte der Klasse `Course` (die Menge der Ausgangsobjekte wird auch als *domain* bezeichnet) mit Objekten der Klasse `AcademicStaffMember` in Relation setzt (die Zielobjekte werden *range* genannt). Mit `owl:inverseOf` kann für eine Eigenschaft wie `teaches` spezifiziert werden, dass es sich um die inverse Relation von `isTaughtBy` handelt; Äquivalenz wird durch `owl:equivalentProperty` ausgedrückt. Darüber hinaus können nicht nur für Klassen, sondern auch für Eigenschaften Hierarchien gepflegt werden (mittels `rdfs:subPropertyOf`).

[9] Darüber hinaus existiert ein dritter Typ (*annotation properties*), mit dessen Hilfe Metadaten über Klassen, Individuen oder Eigenschaften notiert werden können.

Mit Hilfe von owl:Restriction werden Beschränkungen realisiert. Auf diese Weise kann für eine Klasse wie z. B. FirstYearCourse eine Restriktion spezifiziert werden, die auf die Eigenschaft isTaughtBy Bezug nimmt (mittels owl:onProperty), deren Werte *ausschließlich* der Klasse Professor entstammen dürfen. Diese Beschränkung wird mittels owl:allValuesFrom angegeben, wobei alternativ owl:hasValue (muss exakt *einen* spezifischen Wert besitzen) und owl:someValuesFrom eingesetzt werden können. In Bezug auf die zugrunde liegende *Description Logic* entspricht owl:allValuesFrom dem Allquantor (∀) und owl:someValuesFrom dem Existenzquantor (∃). Zusätzlich kann auch die minimale (owl:minCardinality) oder maximale Kardinalität (owl:maxCardinality) einer Relation spezifiziert werden, um z. B. zu repräsentieren, dass ein Course von mindestens einem Professor unterrichtet wird. Eine Eigenschaft kann darüber hinaus vier unterschiedlichen Typen zugewiesen werden: Das Schlüsselwort owl:TransitiveProperty markiert eine Relation als transitiv (z. B. isTallerThan), und owl:SymmetricProperty spezifiziert eine symmetrische Relation (hasSameGradeAs). Mit owl:FunctionalProperty wird eingeschränkt, dass eine Relation wie z. B. age oder height maximal einen eindeutigen Wert besitzt. Das Gegenstück owl:InverseFunctionalProperty definiert eine Eigenschaft, für die zwei unterschiedliche Objekte keinen identischen Wert besitzen dürfen (z. B. isPassportNumberOf).[10] Das Element owl:Restriction kann ebenfalls auf Klassen angewendet werden, um z. B. Course und StaffMember als disjunkte Klassen zu markieren (owl:isComplementOf oder owl:disjointWith). Mittels owl:intersectionOf und owl:unionOf können in Klassendefinitionen Schnittmengen (FacultyInCS setzt sich aus den Angehörigen von Faculty zusammen, die die Eigenschaft belongsTo CSDepartment besitzen) und Vereinigungsmengen (StaffMember und Student bilden PeopleAtUniversity) erzeugt werden, die mit der definierten Klasse äquivalent sind. Mit owl:oneOf können alle Instanzen einer Klasse aufgezählt werden (z. B. die Wochentage als DaysOfTheWeek).

Die Definition einer OWL-Ontologie befindet sich in der Regel in einer Datei, die auch die Instanzen enthält, die entweder als RDF-Ausdrücke oder als XML-Elemente notiert werden, deren Namen den Namen der jeweiligen Klassen entsprechen und innerhalb des Attributs rdf:ID Namen besitzen. Die *Unique Name Assumption* gilt in OWL nicht, d. h. bei unterschiedlichen Namen von Objekten wird nicht pauschal angenommen, dass auch unterschiedliche Individuen vorliegen; die Ungleichheit zweier Instanzen muss explizit notiert werden (owl:distinctMembers, owl:AllDifferent).

Zur Erstellung und Verarbeitung von OWL-Ontologien

OWL-Ontologien können zwar prinzipiell mit beliebigen Text- oder XML-Editoren bearbeitet werden, aufgrund ihrer Komplexität werden sie jedoch üblicherweise mit hierfür vorgesehenen Werkzeugen erstellt und gepflegt, die ebenfalls die Prozesse der Visualisierung, interaktiven Manipulation und Verarbeitung mit einer Inferenzmaschine erlauben.[11]

[10] In OWL DL gelten im Vergleich zu OWL Full verschiedene Einschränkungen: Inverse, funktionale, inversfunktionale und symmetrische Relationen dürfen nicht auf die Eigenschaften von Datentypen angewendet werden. Außerdem ist es nicht erlaubt, für transitive Relationen Kardinalitätsrestriktionen zu spezifizieren.

[11] Dieser Abschnitt beschränkt sich auf eine Darstellung der Werkzeuge, die im Rahmen der vorliegenden Arbeit eingesetzt wurden. Mizoguchi (2004) geht ausführlich auf „Ontology Engineering Environments“ ein und Fluit et al. (2004) widmen sich unterschiedlichen Möglichkeiten der Visualisierung von Ontologien.

Die Erstellung und Pflege der in den nachfolgenden Abschnitten dargestellten Ontologien erfolgte mit Hilfe der Plattform *protégé* (`http://protege.stanford.edu`), die der Modellierung von Ontologien und der Wissensakquisition dient, als Open-Source-Software frei verfügbar ist und über zahlreiche Erweiterungen an spezifische Formalismen und Arbeitskontexte angepasst werden kann (vgl. Noy et al., 2001). Das *OWL Plug-in* stellt eine dieser Erweiterungen dar und dient der Manipulation von OWL-Ontologien (vgl. Knublauch et al., 2004a,b, sowie Abbildung 13.2). Den Entwicklern geht es vornehmlich um den Zugriff, die Visualisierung, das Editieren und den Einsatz von OWL-Ontologien, wobei insbesondere der oftmals hochgradig komplexe Prozess der Konstruktion und sukzessiven Erweiterung einer Ontologie durch integrierte Werkzeuge unterstützt wird. Zu diesem Zweck existieren z. B. ein Versionierungssystem, eine Unterstützung der kollaborativen Wissensmodellierung und verschiedene, bei Bedarf parametrisierbare Testfunktionen; diese weisen auf typische Modellierungsfehler, in der verwendeten OWL-Sprache nicht erlaubte Konstrukte und *best practice*-Verfahren hin. Die grafische Benutzeroberfläche des Editors erlaubt den Zugriff auf Klassen, Eigenschaften und Individuen, die jeweils in hierarchischer Form dargestellt und editiert werden können. Dabei abstrahiert die Oberfläche von der kryptischen OWL- bzw. RDF/RDFS-Syntax, deren Darstellung sehr viel Bildschirmfläche in Anspruch nähme und visualisiert z. B. Restriktionen über Eigenschaften mit Hilfe von Symbolen, die an die Prädikatenlogik angelehnt sind. Zusätzlich erlaubt das *OWL Plug-in* den Zugriff auf Inferenzmaschinen, die für eine gegebene Ontologie einen Konsistenztest und eine Subsumptionsklassifikation durchführen können; das Schließen über Instanzen ist derzeit noch nicht möglich.[12] Die grafische Visualisierung einer Ontologie (vgl. Geroimenko und Chen, 2003) kann mit verschiedenen Erweiterungen erfolgen, die in das *OWL Plug-in* integriert werden und z. B. Ansichten als Baum oder Graph erzeugen. Einige dieser Werkzeuge erlauben eine interaktive Navigation, indem z. B. Kontextmenüs angeboten werden, die das Ein- oder Ausblenden spezifischer Super- oder Subklassen ermöglichen. In anderen Erweiterungen können die in eine Visualisierung aufzunehmenden Klassen und die zwischen ihnen geltenden Relationen in detaillierter Form spezifiziert werden.

Die Bedeutung der Web Ontology Language für das Semantic Web

OWL ist ein Formalismus zur Wissensrepräsentation und es stellt sich die Frage, welche spezifischen Eigenschaften OWL von den zahlreichen anderen Vorschlägen (vgl. z. B. Reimer, 1991) unterscheiden: Frühere Formalismen wurden primär für spezifische Anwendungen entwickelt, die im Rahmen von Forschungsprojekten implementiert wurden. Sobald derartige Systeme die Marktreife erlangen und als Produkte vertrieben werden, wird die Veröffentlichung weiterführender Informationen eingestellt. Mit der *Semantic Web*-Initiative ist nun jedoch ein Bedarf entstanden, derartige Formalismen zu standardisieren, da zahlreiche Unternehmen, die das WWW nicht nur zur Werbung, Kundenkommunikation und zum Vertrieb nutzen, bestrebt sind, Wissen über die Strukturen und Produkte zu repräsentieren, um hierdurch effizientere interne und externe Prozesse realisieren zu können. Gleichzeitig

[12] In der vorliegenden Arbeit wurde für diesen Zweck die Inferenzmaschine RACER (*Renamed ABox and Concept Expression Reasoner*) eingesetzt, die für Forschungs- und Lehrzwecke kostenlos erhältlich ist (vgl. Möller und Haarslev, 2003, sowie `http://www.sts.tu-harburg.de/~r.f.moeller/racer/`).

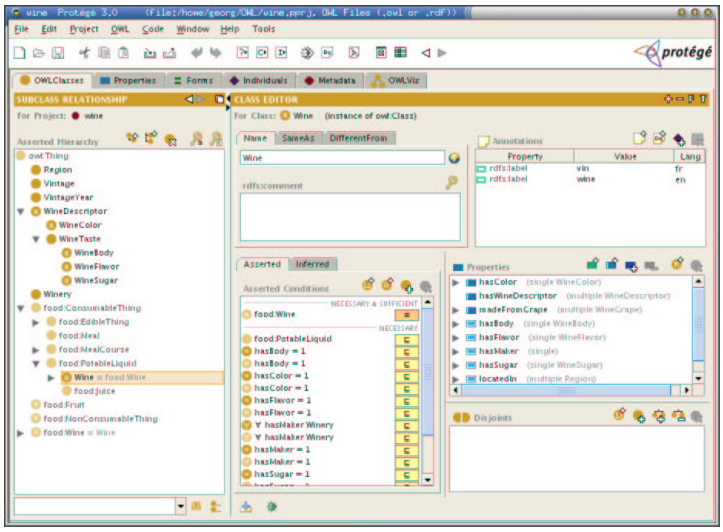


Abbildung 13.2: Die Ontologie-Entwicklungsumgebung *protégé* mit dem *OWL Plug-in*

wird erkannt, dass die von HTML, HTTP und XML vorgegebenen Möglichkeiten für diesen Zweck nur eingeschränkt nutzbar sind und proprietäre Verfahren die Interoperabilität, d. h. den Datenaustausch mit Kunden und anderen Unternehmen beeinträchtigen.

Da das *Semantic Web* auf etablierten Standards aufbaut (vgl. Abbildung 13.1), können in den öffentlich diskutierten Standardisierungsverfahren des W3C neue Formalismen entwickelt und verabschiedet werden, die einerseits in der Lage sind, den Bedarf der Industrie zu decken (Davies et al., 2003a, S. 1: „knowledge is now the key battleground for competition“) und andererseits an den wissenschaftlichen Forschungsstand anknüpfen, ohne jedoch primär von wirtschaftlichen Interessen geleitet zu werden. Das sehr große Interesse an OWL und die Diskussionen zur Konzeptionierung der verbleibenden Schichten geben Anlass zur Vermutung, dass das *Semantic Web* in einigen Jahren Realität werden und den Umgang im und mit dem WWW revolutionieren wird: „We will solve large analytical problems by turning computer power loose on the hard data of the Semantic Web.“ (Berners-Lee, 1999, S. 201). OWL stellt einen zentralen Baustein dieses Vorhabens dar, denn mit ihrer Hilfe sollen große Ontologien und Netze von Ontologien entstehen, die einen Großteil des „knowledge of the world“ (Quillian, 1967, S. 429) repräsentieren, so dass sich auch für computerlinguistische und sprachtechnologische Anwendungen völlig neue Perspektiven eröffnen werden.[13]

[13] Ebenso wie das WWW besitzt das *Semantic Web* eine dezentrale Struktur (vgl. Kapitel 3), d. h. es wird nicht angestrebt, sämtliche Ontologien auf eine Meta-Ontologie abzubilden oder ein einzelnes Vokabular für Konzepte und ihre Merkmale einzuführen. Die eigentliche Mächtigkeit des *Semantic Web* soll darin bestehen, spezialisierte Ontologien aufeinander abbilden zu können. Mit Bezug auf das Open-Source-Paradigma und die kollaborative Entwicklung von Inhalten könnte auf diese Weise eine Ontologie für Allgemeinwissen entstehen: Erfolgreiche Projekte wie die Wikipedia (vgl. Abschnitt 4.6.5) oder das *Open Directory Project* (`http://dmoz.org`) zeigen, dass engagierte Freiwillige in der Lage sind, hochwertige Inhalte zu erstellen. Es ist davon auszugehen, dass ein Projekt zur kollaborativen Erstellung einer Datenbasis von Allgemeinwissen (z. B. nach dem Vorbild von CYC, Lenat, 1995) initiiert werden wird, sobald die technologischen Grundlagen des *Semantic Web* ausgereift sind und die ersten großflächig eingesetzten Anwendungen existieren.

13.2.4 Hypertext, semantische Netze und Ontologien

Hypertexte und semantische Netze weisen verschiedene Gemeinsamkeiten auf (vgl. Fußnote 1, S. 574): Ein Hypertext besteht aus Knoten mit textuellen, tabellarischen oder multimedialen Komponenten, die durch Hyperlinks miteinander verbunden werden, so dass ein gerichteter Graph entsteht, der aufgrund seiner multilinearen Organisation auf individuellen Pfaden rezipiert werden kann (vgl. Abschnitt 3.3.4). Semantische Netze können ebenfalls als Graphen aufgefasst werden: Auch sie bestehen aus Knoten, die über Kanten miteinander verbunden werden, jedoch repräsentieren die Knoten abstrakte Konzepte (d. h. sie enthalten keine Texte), die durch unterschiedliche Typen von Relationen miteinander verknüpft werden, um einen Ausschnitt des Wissens über eine Beschreibungsdomäne zu modellieren. Im Gegensatz zu Hypertexten werden semantische Netze nicht unmittelbar rezipiert, sondern sie fungieren in sprachverstehenden und wissensbasierten Systemen als Ressourcen und dienen innerhalb der Kognitionspsychologie als abstrakte Modelle des Langzeitgedächtnisses.

Diese enge Verbindung zwischen Hypertexten und semantischen Netzen wurde früh erkannt, so dass die in ihren Grundzügen bereits von Bush (1945a) vorgetragene These der kognitiven Plausibilität des Konzepts Hypertext von dem gerade in den siebziger und achtziger Jahren populären Ansatz der semantischen Netze zusätzlichen Auftrieb erfuhr (vgl. auch Abschnitt 3.4.1). Jonassen (1989, S. 11) führt eine oftmals geäußerte Verbindung an: „This network of hypertext ideas may be constructed by the designer or the user to resemble the subject matter structure or the semantic network of the user." Und Rada (1991, S. 146) vertritt die Ansicht „[t]he nodes and links of hypertext may be viewed as a semantic net." Kuhlen (1991, S. 102 ff.) nähert sich den Korrespondenzen zwischen Hypertexten und semantischen Netzen ausgehend von den Relationen als zentrale Gemeinsamkeit: In Hypertexten werden Verknüpfungen zur Relationierung und Kontextualisierung von Knoten eingesetzt und in semantischen Netzen werden Konzepte über unterschiedliche Typen von Relationen verbunden. Eine unmittelbare Übertragung ist jedoch nicht möglich, so stellt sich z. B. die Frage, wie die in semantischen Netzen häufig verwendete *is-a*-Relation in einem Hypertext umgesetzt werden könnte: Kuhlen (1997, S. 104) führt einen exemplarischen Hypertext an, der unter anderem Knoten zu den Themen „Drucker" und „Laserdrucker" enthält, gibt jedoch zu bedenken, dass dann „nicht Einheiten verknüpft [sind], sondern Begriffe als spezielle Elemente in ihnen." Aus einer Diskussion unterschiedlicher Typen von Relationen leitet Kuhlen (1997, S. 106) schließlich eine Typologie von Hyperlinks ab (vgl. Abbildung 3.3, S. 108).[14]

Neben den konzeptionellen und terminologischen Parallelen zwischen semantischen Netzen und Hypertexten existieren Ansätze, in denen die Anreicherung von Hypertextsystemen

[14] Kuhlen (1991, S. 227–236) geht im Kontext von IR-Verfahren auf die Verbindungen zwischen Hypertexten und semantischen Netzen ein und plädiert dafür, „eine wissensbasierte Sicht für den Aufbau von Hypertextsystemen" einzunehmen (ebd., S. 228). Mit diesem Thema beschäftigt sich auch Rada (1991, S. 146), der die geplante Funktionalität des *Semantic Web* im Rahmen eines Rechercheszenarios sehr präzise antizipiert: Ein Anwender sucht in den digitalen Gelben Seiten nach einem Restaurant. Bei einer solchen Suchanfrage könnten entweder hunderte oder auch gar keine Treffer ermittelt werden, aber „the semantic network behind the *Electronic Yellow Pages* can be used as a guide. For example, grocery stores and restaurants are related in that they share the common ancestor 'business' and both sell 'food'. If a consumer was looking for a restaurant in a specific locale and none existed, the system might present the consumer with names of nearby grocery stores." (ebd.; Guha et al., 2003, gehen aus Sicht des *Semantic Web* auf ähnliche Szenarien ein).

mit Domänenwissen angestrebt wird. In den meisten Fällen geht es dabei um den wissensgestützten Aufbau von Hypertexten: Eine Ontologie repräsentiert Wissen über eine spezifische Domäne, und eben dieses Wissen wird eingesetzt, um die manuelle Hypertextualisierung eines gegebenen Textes durch Werkzeuge zu unterstützen, diesen Prozess automatisch durchzuführen oder in einem Text oder einer Hypertextbasis neue, d. h. nicht explizit kodierte Hyperlinks zu ermitteln (vgl. z. B. Beißwenger et al., 2004). Carr et al. (2004, S. 518) sprechen konsequenterweise von „ontological hypertexts“ (vgl. auch Miles-Board et al., 2001): Basierend auf einer Sammlung von Konzepten, die als Vokabular einer natürlichen Sprache ausgedrückt werden, durch Relationen verknüpft und für eine bestimmte Domäne spezifisch sind, sollen einerseits Hyperlinkanker in einzelnen Knoten erkannt werden, andererseits soll dieses ontologische Wissen eingesetzt werden, um die maschinell erstellten Links ausschließlich auf die für eine spezifische Anwendung oder Benutzungssituation relevanten Konzepte zu beschränken (z. B. Schlagwörter oder Eigennamen). Für diesen Zweck müssen mehrere Ressourcen vorgesehen werden: Von zentraler Bedeutung ist eine Ontologie, die die Domäne des Hypertextthemas (auch: Text- oder Diskurswelt) modelliert, d. h. für jedes neu zu verarbeitende Thema ist es notwendig, eine derartige Ontologie zu konstruieren oder eine verfügbare Ressource zu modifizieren. Während diese Ontologie zur Ermittlung möglicher Konzepte eingesetzt wird, die potenziell als Hyperlinkanker fungieren können, wird weiterhin ein – wiederum themenspezifischer – Datenbestand benötigt, der mögliche Hyperlinkziele umfasst (z. B. eine *Linkbase*, d. h. eine Datenbank von Links). Weiterhin müssen Strategien implementiert werden, um ein „over-linking“ zu vermeiden, die Menge der potenziellen Linkanker also auf die tatsächlich relevanten zu reduzieren. Hierzu kann die Ontologie eingesetzt werden, so wird z. B. zunächst bestimmt, welches Konzept ein in der Verarbeitung befindliches Dokument beschreibt, um potenzielle Hyperlinkanker, die auf eben dieses Konzept verwiesen, aus der maschinellen Verlinkung auszuschließen, aber verwandte und somit für den Rezipienten relevante Konzepte beizubehalten. Der Zweck von „ontological hypertexts“ ist es, dem Rezipienten effizientere Zugriffsmöglichkeiten auf eine Hypertextbasis zu bieten, die maschinell durch Einbeziehung eines strukturierten Wissensmodells berechnet werden. Hierfür ist es notwendig, entweder die Dokumente durch Hinzufügung von Hyperlinks oder die Browseroberfläche durch Integration zusätzlicher Navigationshilfen zu modifizieren (Carr et al., 2004, S. 528).

13.3 Die Ontologie wissenschaftlicher Themen und Fachgebiete

Thematische Hierarchien werden häufig im Bereich der Textklassifikation eingesetzt, die die maschinelle Bestimmung des Textthemas betrifft (vgl. Rehm, 2004d). Beim Einsatz überwachter maschineller Lernverfahren konstituiert sich die Hierarchie aus Trainingstexten, denen von einem Experten die Etiketten des verwendeten Kategoriensystems zugewiesen werden, das entweder als Liste oder als Baum organisiert ist, so dass ein Knoten wie z. B. „Sport“ Kategorien wie „Rasensport“, „Wintersport“ und „Hallensport“ subsumiert, die ihrerseits Themen wie „Fußball“, „Hockey“, „Langlauf“, „Handball“ und „Basketball“ beinhalten. In der ersten Phase werden derartige Algorithmen mit manuell zugewiesenen Texten trainiert,

so dass Wörter, die für eine spezifische Kategorie – dem Algorithmus zufolge – signifikant sind, unmittelbar in dem Knoten der Themenhierarchie hinterlegt werden. In der zweiten Phase werden unbekannte Texte klassifiziert, und das Lernverfahren bezieht die notwendigen Regeln, Merkmale oder statistischen Informationen direkt aus der Hierarchie.

Die in Abschnitt 13.5 dargestellte Ontologie von Hypertextsorten umfasst nicht unmittelbar eine Themenhierarchie. Da jedoch spezifische Hypertextsorten, Hypertextknotensorten oder Hypertextsortenmodule durchaus mit bestimmten Themen korrelieren, wird eine separate Themenontologie angenommen. Darüber hinaus dient diese strikte Separierung der Ebenen *Thema* und *Hypertextsorte* der Veranschaulichung des in Kapitel 1 genannten Ziels der maschinellen Klassifikation von Hypertextsorten: Die thematische Klassifikation benötigt keine Informationen über Hypertextsorten. Ein Verfahren zur Klassifikation von Hypertextsorten setzt jedoch Wissen über den Aufbau einer spezifischen Hypertextsorte und die beteiligten, empirisch ermittelten Konstituenten voraus und sollte *zusätzlich* von einer Themenhierarchie flankiert werden, um das Thema maschinell ermitteln und als beschränkenden Parameter in den Klassifikationsprozess einfließen zu lassen.

13.3.1 Verwendete Datenquelle

Da sich die vorliegende Arbeit auf die Untersuchungsdomäne der universitären Webangebote stützt, wurde eine aus dem Bibliothekswesen stammende wissenschaftliche Systematik als Grundlage der Ontologie eingesetzt. Die entsprechenden Daten wurden vom Projekt GERHARD (*German Harvest Automated Retrieval and Directory*) zur Verfügung gestellt (vgl. Wätjen, 1998, sowie Wätjen et al., 1998). Dieses Projekt hat eine Suchmaschine für wissenschaftliche Publikationen entwickelt, die den in einem Themenkatalog enthaltenen Kategorien zugewiesen werden (vgl. `http://www.gerhard.de`). Zu diesem Zweck werden die Dokumente nahezu aller wissenschaftlichen Webserver Deutschlands – ähnlich dem in Kapitel 7 dargestellten Verfahren – mit dem *Crawler* einer Suchmaschine traversiert, in einen Index aufgenommen und klassifiziert. Neben der Suche nach Schlagwörtern im Datenbestand kann der Katalog auch als Hypertext navigiert werden.

Die in GERHARD verwendete Klassifikation basiert nicht auf maschinellen Lernverfahren, sondern auf einer computerlinguistischen Auswertung, deren Ergebnisse auf eine speziell angepasste Variante der UDK (*Universale Dezimalklassifikation*) abgebildet werden. Die UDK ist eine vorwiegend hierarchisch organisierte Klassifikation, die in einer maschinenlesbaren Form vorliegt. Die Kategorien sind nach dem Dezimalzahlprinzip organisiert, so bezeichnet „5“ den Bereich „Mathematik/Naturwissenschaft“, „53“ bezieht sich auf „Physik“ und „536.11“ repräsentiert „Grundbegriffe der Wärmetheorie“ (vgl. Wätjen et al., 1998, S. 14 f.). Neben dieser implizit notierten Subsumption existieren weitere Relationen wie z. B. „Beiordnung“, „Erstreckung“ und „Ortsanhang“, die über Sonderzeichen repräsentiert werden (vgl. Möller, 1997). Die eingesetzte Version der UDK umfasst etwa 60 000 Einträge, die insbesondere die naturwissenschaftlichen Wissensgebiete in umfassender Weise abdecken. Für die Aufgabe der Klassifikation wurde die UDK in ein Lexikon konvertiert, so dass die in den zu klassifizierenden Dokumenten enthaltenen Wörter und Phrasen morphologisch analysiert und auf das UDK-Lexikon abgebildet werden können. Die ermittelten Kategorien dienen der statistischen Bestimmung der Themenkategorien eines Dokuments. Wätjen et al. (1998,

S. 31) geben an, dass die Klassifikation mit einer Präzision von 80% operiert und dass ein Dokument auf durchschnittlich sechs bis sieben UDK-Einträge abgebildet wird (ebd., S. 18).

13.3.2 Konvertierung der Daten in die Web Ontology Language

Die Ontologie wissenschaftlicher Themen und Fachgebiete basiert nicht auf dem UDK-Lexikon, sondern auf einem Abzug der Datenbanktabellen, die innerhalb des GERHARD-Systems verwendet werden. Mit Hilfe eines in *Perl* implementierten Skripts wurden diese Tabellen eingelesen, zu einer einheitlichen Datenstruktur aggregiert und anschließend als etwa 25 000 Klassen umfassende OWL-Ontologie exportiert (vgl. Abbildung 13.3). Mittels `owl:imports` kann diese Ontologie in andere OWL-Repräsentationen eingebunden werden.

Da die Quelldaten nicht nur Tupel der Form *Kategorienummer, Kategoriebezeichnung* umfassen, sondern zusätzlich mehrere Relationen, die zwischen zwei Kategorien gelten, in unterschiedlichen Kodierungsvarianten enthalten, beschränkt sich die Konvertierung auf die Abbildung der von den Dezimalzahlen ausgedrückten Subsumptionshierarchie.[15] Weil viele Kategorien keine übergeordneten Einträge besitzen, mussten innerhalb der OWL-Ontologie ca. 5 000 abstrakte Klassen eingeführt werden, die als Superkonzepte fungieren.

13.4 Die Domänenontologie

Die vorliegende Arbeit stützt sich auf die Untersuchungsdomäne der universitären Webangebote. Es wurde wiederholt thematisiert, dass eben diese Untersuchungsdomäne für die Bestimmung der Hypertextsorten einen zentralen Stellenwert besitzt (vgl. die Kapitel 6, 11 und 12). Aus diesem Grund wurde ein Domänenmodell konstruiert, das den typischen Aufbau einer Universität beschreibt, so dass dieses Modell – ebenso wie die im Ontologie wissenschaftlicher Themen und Fachgebiete – innerhalb der Ontologie von Hypertextsorten für ihre Beschreibung und maschinelle Repräsentation Verwendung finden kann.

13.4.1 Verwendete Quellen

Die Aufgabe der Erstellung eines Domänenmodells zur Repräsentation des Aufbaus einer „generischen“ deutschen Hochschule sieht sich mit zwei grundlegenden Schwierigkeiten konfrontiert: Erstens liegen zu diesem Themenkomplex nahezu keine Publikationen vor, auf die sich die Modellierung primär hätte beziehen können, zweitens besitzen einzelne Universitäten zahlreiche individuelle Charakteristika (vgl. auch Kapitel 12), so dass die entstandene Ontologie zwangsläufig einem sehr hohen Abstraktionsgrad unterliegt (vgl. Abschnitt 13.4.2).[16]

[15] Die Verarbeitung dieser Relationen hätte die Implementierung eines Parsers für die Notationstypen und ihre Varianten erfordert, worauf aus Komplexitätsgründen verzichtet wurde. Zur Repräsentation dieser Relationen in OWL böten sich Propertys an, die innerhalb der beteiligten Klassen als Restriktionen definiert werden.

[16] Aufgrund der zahlreichen individuellen Charakteristika liegt die „Frage nahe, ob es überhaupt möglich ist, Aussagen über die Organisationsstrukturen der Hochschulen im allgemeinen zu treffen, ob nicht nur Aussagen über einzelne Hochschulen möglich sind. Eine nähere Betrachtung zeigt nun allerdings, daß es im deutschen Hochschulwesen einige Grundfragen gibt, die die Organisationsstrukturen bedingen [...]. Es sind das (i) die fachliche Gliederung, (ii) die zentrale und dezentrale Organisation, (iii) die Gliederung nach Gruppen und (iv) die Universitätsspitze.“ (Thieme, 1996, S. 814).

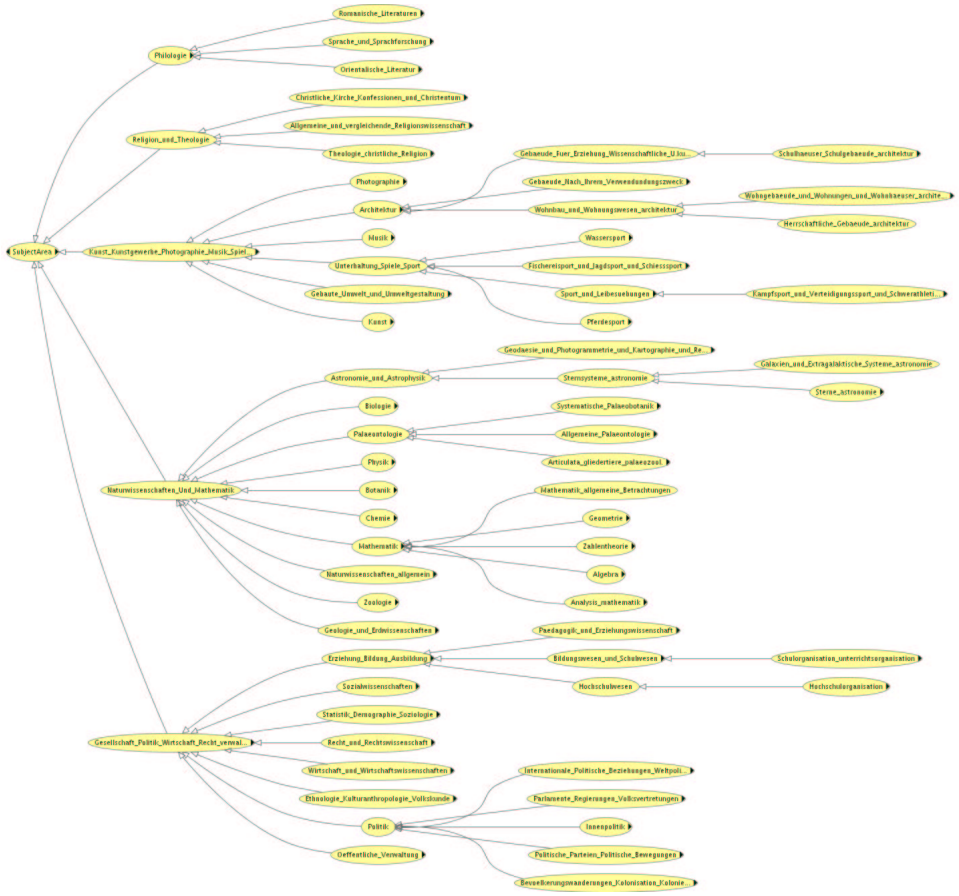


Abbildung 13.3: Ein Ausschnitt der Ontologie wissenschaftlicher Themen und Fachgebiete

Das als OWL-Ontologie entwickelte Domänenmodell basiert – neben dem Wissen des Verfassers über den Aufbau von Universitäten – auf mehreren unterschiedlichen Quellen. Zunächst wurde eine initiale Ontologie aufgebaut, die sukzessive um weiterführende Aspekte ergänzt wurde. Während frühere Fassungen des bundesweit geltenden Hochschulrahmengesetzes in einigen Bereichen strikte Vorgaben über universitäre Organisationsstrukturen enthielten, werden in der aktuellen Version des Gesetzes diesbezüglich nahezu keine obligatorischen Aussagen mehr gemacht.[17] Die ehemaligen Vorgaben wurden verschiedenen Kommentaren des Gesetzes entnommen, die sich im *Handbuch des Wissensschaftsrechts* (Flämig et al., 1996) befinden.[18] Weiterhin wurden die Strukturen und Relationen in die Ontologie integriert, die sich in insgesamt 14 Organigrammen befinden, die innerhalb der Webauftritte von Hochschulen publiziert wurden. Zusätzlich wurden einige Aspekte der „Science"-Ontologie

[17] Auf einen detaillierten Vergleich der Hochschulgesetze sämtlicher Bundesländer und die Ermittlung der Gemeinsamkeiten und Unterschiede wurde aus Komplexitätsgründen verzichtet.

[18] Ausgewertet und in die Ontologie integriert wurden Sachverhalte, Bezeichnungen und Informationen, die von Thieme (1996), Schuster (1996), Leuze (1996), Janson (1996), Gattermann (1996) und Tettinger (1996) dargestellt werden und als gemeinsame Eigenschaften deutscher Hochschulen aufgefasst werden können.

hinzugefügt, die auf der Website des Editors *protégé* als Beispiel zur Verfügung steht.[19] Einige Relationen und *datatype properties* wurden einem Band über die Problematik der Einrichtung eines *data warehouse* an einer Hochschule entnommen (Nusselein, 2003). Abschließend wurden verschiedene Aspekte aus der von Adamzik (2004, S. 87 ff.) diskutierten Betrachtung unterschiedlicher Rollen im Hochschulbetrieb in die Ontologie integriert.

Die angesprochenen individuellen Charakteristika einzelner Hochschulen und die allgemein zwischen ihnen herrschenden Unterschiede beziehen sich auf eine Vielzahl von Aspekten (z. B. bezüglich der Benennung von Einrichtungen und Organisationseinheiten, vgl. Janson, 1996, sowie Fußnote 16, S. 474), von denen an dieser Stelle nur einige stichwortartig aufgeführt werden sollen, um die generelle Problematik der Konstruktion einer Ontologie der Untersuchungsdomäne zu veranschaulichen: Einige Hochschulen besitzen eine zweistufige, andere eine dreistufige Gliederung in Organisationseinheiten. Die in den siebziger Jahren notwendig gewordene Aufteilung der philosophischen Fakultäten in einzelne Fachbereiche erfolgte uneinheitlich, die Universitäten haben „vielmehr – jede für sich – eine individuelle Lösung ihrer Probleme gefunden“ (Thieme, 1996, S. 820), wodurch sich „im Bereich der ehemaligen Philosophischen Fakultät heute [...] ein buntgeschecktes organisatorisches Bild [zeigt], das sich nicht unter ein einheitliches Schema bringen läßt.“ (ebd., S. 817). Es existieren mehrere Grenzfälle, in denen eine spezifische Organisationseinheit an einer Universität als Zentrum und an einer anderen als dezentrales Institut organisiert ist: Falls z. B. Breitensport für alle Universitätsangehörigen angeboten wird, handelt es sich typischerweise um ein Sportzentrum, falls aber primär die Sportwissenschaft, d. h. Forschung und Lehre im Vordergrund stehen, wird diese Einrichtung oftmals als Sportinstitut geführt, das einem Fachbereich zugeordnet ist (ebd., S. 825). Einige Landeshochschulgesetze sehen im Leitungsgremium einer Universität einen Rektor vor, der das Amt für einen Zeitraum von z. B. zwei Jahren übernimmt, an anderen Hochschulen existiert eine Präsidialverfassung, die für das Amt typischerweise einen längeren Zeitraum und andere Kompetenzen sowie Aufgabenbereiche vorsieht; sämtliche Landeshochschulgesetze benennen jedoch den Kanzler als leitenden Verwaltungsbeamten (vgl. Schuster, 1996). Auch im Bereich der Selbstverwaltung existieren zahllose individuelle Unterscheide und nur wenige Gemeinsamkeiten, die jedoch, wenngleich in sehr abstrakter Form, in der Ontologie enthalten sind.

13.4.2 Inhalt und Umfang der Ontologie

Die Domänenontologie besteht aus insgesamt etwa 400 Klassen und ca. 200 Relationen. Der modellierte Ausschnitt der Domäne beschränkt sich auf diejenigen Konzepte, die eine unmittelbare Relevanz für die Inhalte und die abstrakte Struktur eines universitären Webauftritts besitzen, d. h. auf sehr detaillierte Binnendifferenzierungen – z. B. von Rechenzentren oder Bibliotheken – wurde aufgrund der Komplexität eines derartigen Vorhabens verzichtet. Auf der obersten Ebene existieren die Klassen `CommunicationDevice`, `Event`, `Location`, `Result`,

[19] Darüber hinaus existieren im WWW verschiedene weitere Sammlungen von Ontologien. Die Website `http://www.daml.org/ontologies/ontologies.html` umfasst z. B. mehr als 300 DAML+OIL-Ontologien. Einige von ihnen modellieren den Aufbau eines Instituts oder eines Forschungsprojekts und stellen somit jeweils nur einen Ausschnitt des Wissens dar, das für die vorliegende Fragestellung benötigt wird. Zudem orientieren sich diese Ontologie an den Organisationseinheiten US-amerikanischer Hochschulen, die sich von deutschen Hochschulen in verschiedenen Merkmalen unterscheiden.

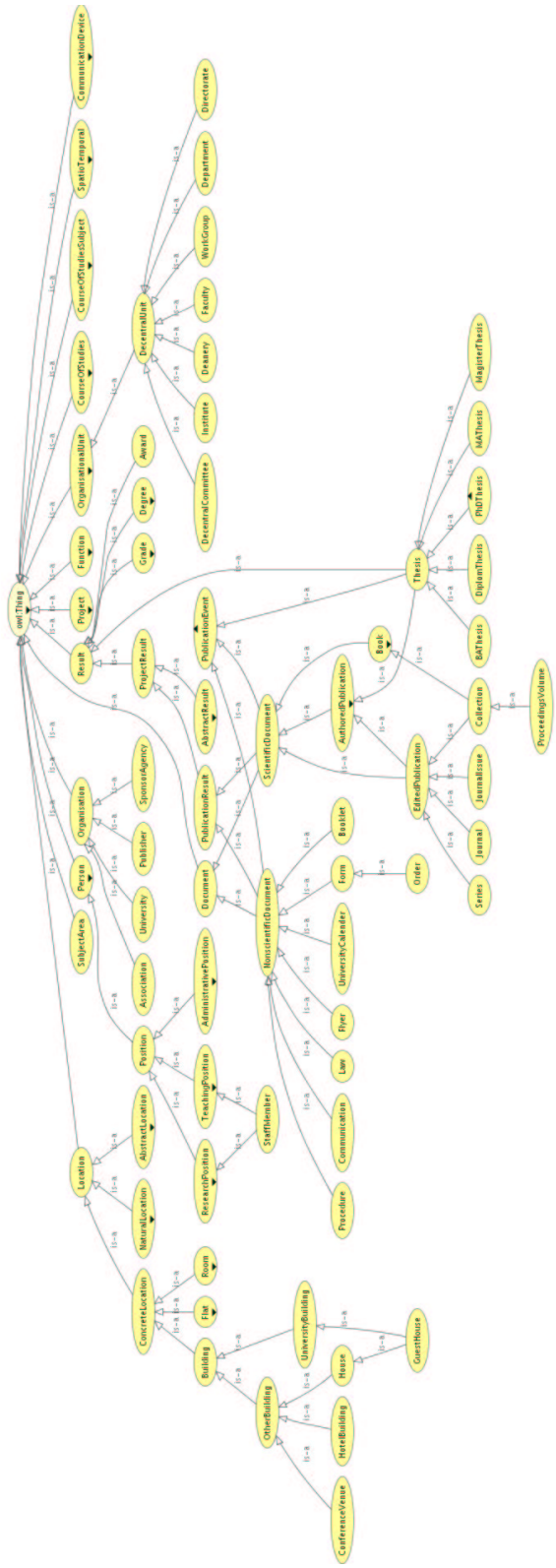


Abbildung 13.4: Ein Ausschnitt der oberen Hierarchieebene der Domänenontologie

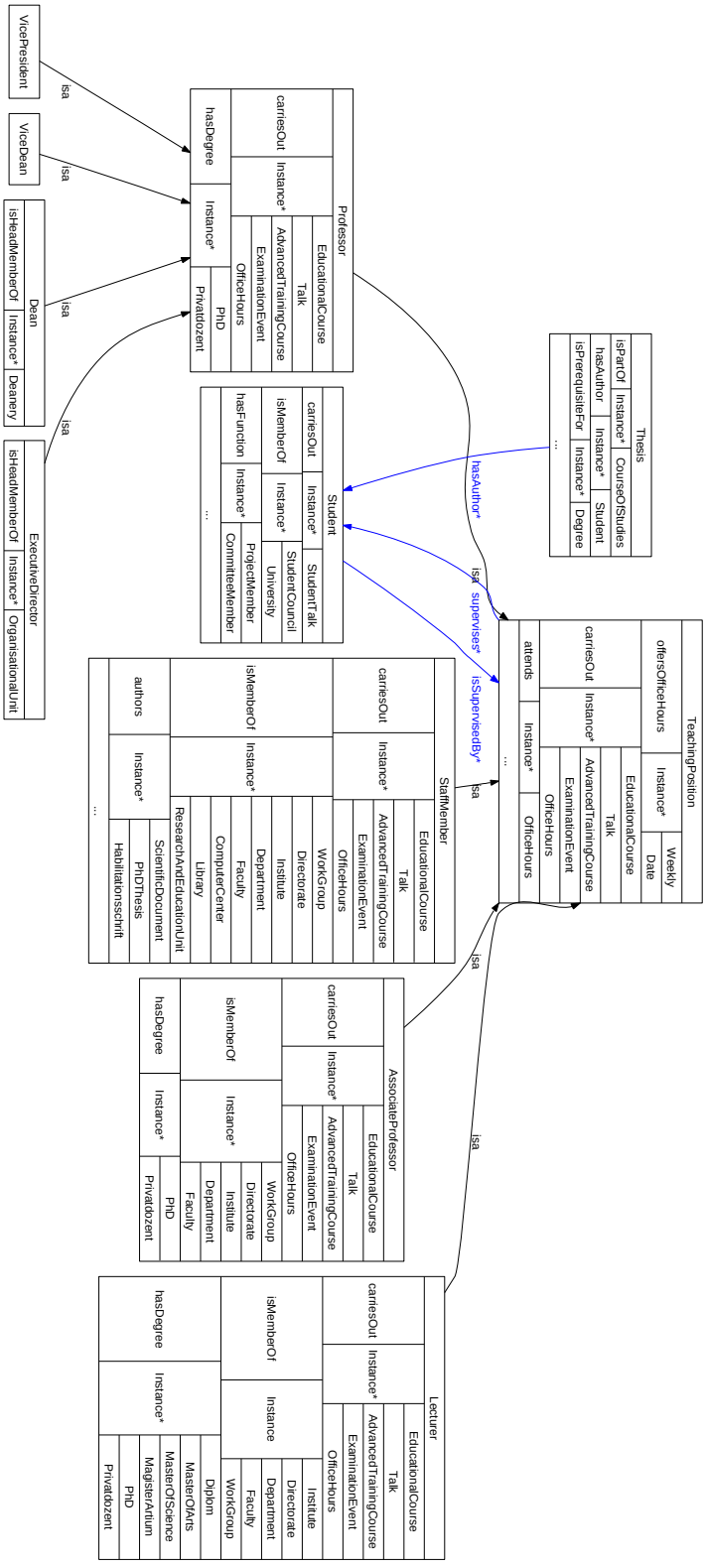


Abbildung 13.5: Ein Ausschnitt der Domänenontologie (Klassen und Relationen)

SpatioTemporal und SubjectArea, die domänenunabhängige Teile der Ontologie beinhalten (z. B. Ereignisse, Lokationen, Ergebnisse, räumliche und zeitliche Konzepte etc.).[20] Die Klasse SubjectArea stellt die Verbindung zu der in Abschnitt 13.3 dargestellten Ontologie wissenschaftlicher Themen und Fachgebiete dar, die an dieser Stelle mittels owl:imports integriert wird. Die Klassen CourseOfStudies, CourseOfStudiesSubject, Document, Function, Organisation, OrganisationalUnit, Person und Project beschreiben auf Grundlage der erwähnten konzeptuellen Basis den generischen Aufbau einer Universität, wozu z. B. Studierende, Mitarbeiter, Hochschullehrer, unterschiedliche Typen von Studiengängen, Organisationseinheiten und die Verwaltung gehören. Da eine natürlichsprachliche Paraphrasierung des vollständigen Inhalts des Domänenmodells an dieser Stelle nicht erfolgen kann, können Visualisierungen einige ausgewählte Bereiche hervorheben: Abbildung 13.4 stellt die oberen Klassen der Ontologie und einige ausgewählte Subklassen dar. Die in dem Graphen enthaltenen Kanten beziehen sich ausschließlich auf die *is-a*-Relation, die zwischen einer Klasse und ihren Subklassen besteht. Abbildung 13.5 stellt einige Subklassen von TeachingPosition dar und verdeutlicht zwei Relationen, die zwischen Lehrenden und Studierenden, sowie zwischen Studierenden und Abschlussarbeiten existieren. Zusätzlich zeigt dieser Graph für jede der dargestellten Klassen einen Ausschnitt ihrer Merkmale auf.

13.5 Die Hypertextsortenontologie

Der in diesem Kapitel eingeführte Ansatz zum Aufbau und zur Notation einer Hypertextsortenontologie basiert auf einer Verwendungsweise von OWL, die sich *nicht* – wie in Abschnitt 13.2.4 skizziert – auf die Unterstützung von Hypertextualisierungsprozessen oder die Ermittlung zusätzlicher Hyperlinks, sondern auf die Repräsentation von Informationen über Hypertextsorten, ihre Konstituenten und zusätzlich beteiligte Ressourcen bezieht. Das Repräsentationsformat zielt insbesondere auf einen Einsatz in text- und computerlinguistischen sowie texttechnologischen Szenarien und Anwendungen ab.

Zunächst wird die allgemeine Ausrichtung des Repräsentationsformats dargestellt, die sowohl auf den drei Ebenen des Hypertextsortenmodells (vgl. Kapitel 5) als auch auf den Ergebnisse der Stichprobenanalysen basiert (vgl. die Kapitel 8 bis 12). Anschließend wird die Strukturierung der Ontologie eingeführt (Abschnitt 13.5.2), woraufhin die Klassen, die den drei Ebenen des Modells entsprechen, erläutert werden (Abschnitte 13.5.3 bis 13.5.5). Abschnitt 13.5.6 thematisiert weiterführende Beispiele, woraufhin das Spannungsverhältnis zwischen der *Web Ontology Language* und XML-basierten Dokumentgrammatiken diskutiert wird (Abschnitt 13.5.7). Abschließend wird die Anreicherung der Hypertextsortenontologie mit Dokumentgrammatiken besprochen (Abschnitt 13.5.8).

[20] Die in der Ontologie verwendeten Konzepte und Eigenschaften besitzen aus drei Gründen Bezeichnungen in englischer Sprache: Zum einen lehnen sich zahlreiche Etiketten an etablierte Entwurfsmuster und Bezeichnungen von Relationen innerhalb der Wissensrepräsentation an, die traditionellerweise in dieser Sprache benannt werden (z. B. *is-a* und *has-part*), zum anderen soll das Domänenmodell nach Beendigung der Arbeit in eine der im WWW existenten Ontologie-Sammlungen eingebracht werden. Die Kategorien umfassen zusätzlich deutschsprachige Bezeichnungen, die als rdfs:label hinterlegt wurden. Der dritte Grund betrifft die einfachere Unterscheidbarkeit zwischen Klassen und Relationen, die aus dem Domänenmodell stammen und Bestandteilen der Hypertextsortenontologie, die deutschsprachig etikettiert wurde.

13.5.1 Zur Ausrichtung des Repräsentationsformats

Die Repräsentation von Hypertextsorten erfolgt durch ein Format, das die Ebenen der Hypertextsorten, Hypertextknotensorten und Hypertextsortenmodule integriert und die Modellierung multipler Typologien erlaubt. Zusätzlich werden übergreifende Ebenen vorgesehen, um Beziehungen zwischen den Konstituenten erfassen zu können (vgl. Abschnitt 5.2.1). Als verkapselndes Rahmenformat fungiert der texttechnologische Standard *Web Ontology Language*. Mit der Fundierung des Repräsentationsformats auf das Paradigma der Texttechnologie – also den Einsatz XML-basierter Auszeichnungssprachen – sind mehrere Vorteile verbunden, die z. B. den Datenaustausch, die Verwendung multipler Annotationsebenen und die Benutzung einer Vielzahl von Werkzeugen betreffen (doch siehe Abschnitt 13.5.8).

OWL wird innerhalb der Hypertextsortenontologie somit nicht zur Repräsentation von Wissen in der traditionellen Auffassung der Wissensrepräsentation (vgl. Abschnitt 13.2.1), sondern als flexibles Format zur Modellierung textlinguistischer und texttechnologischer Eigenschaften von Hypertextsorten verwendet, so dass diese Informationen als Bestandteil des *Semantic Web* aufgefasst werden.[21] Die computergestützte Repräsentation derartiger Konventionen erlaubt verschiedene neuartige Perspektiven und bezieht sich in Verbindung mit dem Hypertextsortenmodell als theoretische Fundierung auf die Forderung von Adamzik (2000b, S. 109), die Textsortenforschung um ein „Kriterium der Einbettung von Textsorten in umfassendere kommunikative Strukturen und ihre Vernetztheit miteinander“ zu erweitern.[22]

Ontologien können als semantische Netze aufgefasst werden, die Konzepte, Merkmale von Konzepten, typisierte Relationen und Individuen beinhalten. Die *Web Ontology Language* stellt einen relationalen Formalismus dar, der die Kriterien zur Repräsentation von Wissen über Hypertextsorten erfüllt, die in Abschnitt 6.5 dargestellt wurden. Eine im Kontext der Typologisierung von Hypertextsorten wichtige Prämisse betrifft die Vererbung von Eigenschaften, so dass z. B. die Hypertextsorten *persönliche Homepage eines Wissenschaftlers* und *private Homepage eines Studierenden* als subgenerische Varietäten des Hypertexttyps *Homepage einer Person* repräsentiert werden können (vgl. Abschnitt 10.6, insbesondere Abbildung 10.7, S. 460). Die Erfassung von Hypertextsorten auf der Grundlage von OWL ermöglicht eine solche Vorgehensweise, da OWL-Ontologien aus Vererbungshierarchien bestehen, in denen Subklassen die Eigenschaften ihrer Superklasse erben, wobei auch die Mehrfachvererbung möglich ist. Hypertextsorten können somit als hierarchisch gestufte Klassen einer Ontologie aufgefasst werden, die über Merkmale, die individuell restringiert werden können, als generische Eigenschaften verfügen. Hypertextsorten, Hypertextknotensorten und Hypertext-

[21] Ontologien repräsentieren in der Regel Wissen über eine spezifische Domäne, so dass z. B. das in Abschnitt 13.4 vorgestellte Domänenmodell als typische Ontologie aufgefasst werden kann. Die Hypertextsortenontologie enthält ebenfalls Wissen, dieses bezieht sich jedoch nicht auf die Objekte, Eigenschaften, Relationen und Individuen eines bestimmten Weltausschnitts, sondern auf die Konstituenten des Hypertextsortenmodells (Kapitel 5) sowie die in Teil III ermittelten Analyseergebnisse.

[22] Die Vernetzung könnte z. B. Genre-Systeme betreffen (Bazerman, 1994), die die temporale Sequenz der wechselseitigen Realisierung von Exemplaren spezifischer Text- oder die Rezeption präzise sequenzierter Knoten spezifischer Hypertextsorten reflektieren (vgl. Abschnitt 2.3.7). Die Realisierungssequenz bezieht sich letzten Endes auf Metainformationen über Text- bzw. Hypertextsorten und die jeweiligen Situationsspezifika, z. B. das Verhältnis der Produzenten und Rezipienten und die Bezugnahme auf Teile der Textexemplare. DTDs der beteiligten Textsorten könnten formal repräsentiert und durch eine abstraktere und in OWL modellierte Ebene angereichert werden, um derartige Relationen und Sequenzspezifikationen auszudrücken.

sortenmodule werden also als Klassen konzeptualisiert, die Subklassen umfassen, zwischen denen Relationen herrschen. Diese Vorgehensweise ermöglicht es, das in Abschnitt 4.5.4 thematisierte Problem der Relationierung einzelner Komponenten von Hypertextsorten zu lösen: Crowston und Williams (1997, S. 31) führen ein Beispiel an, das auf eine *is-a* Relation reduziert werden kann, denn das „social science paper“ *is-a* „research paper“ *is-a* „paper“, d. h. Subklassen erben generische Eigenschaften von ihren Superklassen und besitzen zusätzlich individuelle Merkmale, die die Annahme einer Subklasse erst rechtfertigen. Toms und Campbell (1999) hingegen beschreiben ein Beispiel, das auf einer partitiven Relation basiert, denn „wissenschaftlicher Artikel“ *has-part* „Literaturliste“ *has-part* „Literatureintrag“. Diese Relation bezieht sich im Kontext traditioneller Textsorten auf Teiltextsorten bzw. Instanzen von Teiltextsorten in Textexemplaren. Auf ähnliche Weise können auch generische Hypertextsequenzierungen und Verknüpfungen einzelner Knoten repräsentiert werden.

Propertys können neben Datentypen auch Relationen zu anderen Klassen und arbiträre RDF-Annotationen umfassen, die zur Referenzierung externer Dokumentgrammatiken eingesetzt werden können. Auf diese Weise ist es z. B. möglich, natürlichsprachliche Beschreibungen von Hypertextsorten und ihren spezifischen Aspekten in der Ontologie zu hinterlegen. Durch diesen Mechanismus werden auch Referenzierungen von Hypertextknotensorten und Hypertextsortenmodulen realisiert, deren Repräsentationen in weiteren Schichten der Ontologie gepflegt werden. Somit können universale Hypertextsortenmodule, durch eine Relation in der obersten konzeptuellen Stufe der Hypertextsortenontologie (d. h. in `Hypertexttyp`) verankert und an sämtliche Hypertexttypen und -sorten vererbt werden. Weiterhin können RDF-Annotationen auf allen Ebenen zur Aufnahme von Merkmalsinformationen eingesetzt werden, die zur maschinellen Detektion und weiterführenden Verarbeitung von Hypertextexemplaren benötigt werden (vgl. Potok et al., 2002). Von besonderer Bedeutung sind derartige Merkmale in Bezug auf Hypertextsortenmodule, da diese die atomaren makrostrukturellen Bestandteile von Hypertextsorten darstellen und somit die Schnittstelle zwischen der Erkennung von Hypertextmodulen und den Konstituenten abstrakt definierter Hypertextsorten bilden (vgl. Kapitel 14).

Die in Teil III präsentierten Analysen stellen eine exemplarische Anwendung des Hypertextsortenmodells dar. Aus den Ergebnissen dieser Analysen kann eine Ontologie von Hypertextsorten konstruiert werden. Es wurde bereits betont, dass der Gleichung „Emittent = Thema“ (Schütte, 2004a, S. 211) für die Untersuchungsdomäne ein besonderes Gewicht zukommt. Aus diesem Grund wurde das in Abschnitt 13.4 diskutierte Domänenmodell entwickelt, das die generische Struktur deutscher Hochschulen als OWL-Ontologie formalisiert und somit in gewisser Weise als Themenhierarchie im Sinne von Jakobs (2003, S. 241) fungiert (vgl. Abschnitt 5.2.2), schließlich besteht ein unmittelbarer Zusammenhang zwischen den innerhalb einer Hochschule existierenden Strukturen und Organisationseinheiten sowie den Hypertexten, die von diesen Einheiten publiziert werden (vgl. die Abschnitte 6.3.4 und 6.5). Daher ist es notwendig, die generische Struktur einer Hochschule in die Beschreibung von Hypertextsorten einzubeziehen, um Korrespondenzen zu den jeweiligen Hypertextsorten und somit ihre Geltungsbereiche erfassen zu können. Die dritte Ontologie, die eine Hierarchie wissenschaftlicher Themen und Fachgebiete auf Basis der universalen Dezimalklassifikation (UDK) umfasst, erlaubt die Referenzierung spezifischer Themen (vgl. Abschnitt 13.3). Abbildung 13.6 stellt die beteiligten Ontologien im Überblick dar.

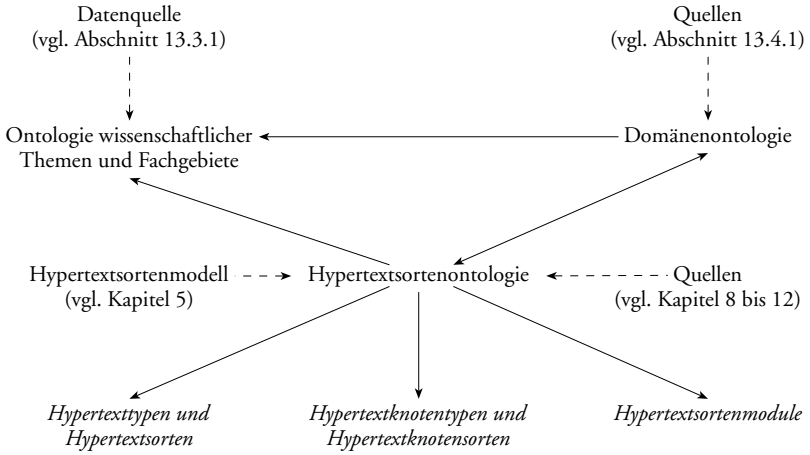


Abbildung 13.6: Die Bestandteile der Hypertextsortenontologie im Überblick

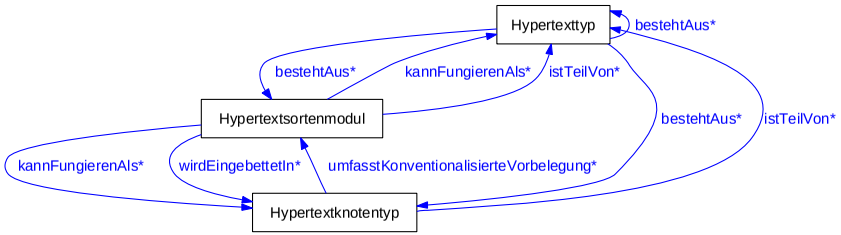
13.5.2 Die drei Ebenen des Hypertextsortenmodells als Konzepte

Die Hypertextsortenontologie besteht aus den übergeordneten Konzepten `Hypertexttyp`, `Hypertextknotentyp` und `Hypertextsortenmodul`, die den drei Ebenen des in Kapitel 5 vorgestellten Hypertextsortenmodells unmittelbar entsprechen (vgl. Abbildung 13.6, sowie Abbildung 5.4, S. 293). Es werden die Etiketten „Hypertexttyp“ (statt „Hypertextsorte“) und „Hypertextknotentyp“ (statt „Hypertextknotensorte“) verwendet, da es sich auf dieser obersten typologischen Ebene um Klassen von Hypertexten bzw. Hypertextknoten handelt, die nur eine geringe Anzahl distinktiver Merkmale und einen entsprechend großen Abdeckungsgrad besitzen (vgl. Abschnitt 2.3.2 sowie Kapitel 5). Da die Differenzierung zwischen Hypertexttypen und Hypertextsorten – analog zu Texttypen und Textsorten – von der spezifischen Betrachtungsweise und den Text- bzw. Hypertextklassen abhängig ist, die verglichen werden, wird in der Hypertextsortenontologie keine präzise Unterscheidung angenommen, ab welcher ontologischen Hierarchiestufe ein Konzept z. B. als Hypertexttyp, Hypertextsorte oder Hypertextsortenvariante aufzufassen ist (vgl. Abbildung 2.2, S. 40).

Die grundlegenden, zwischen diesen drei Klassen existierenden Relationen basieren unmittelbar auf dem Hypertextsortenmodell: Zwischen `Hypertexttyp` und `Hypertextknotentyp` sowie `Hypertextsortenmodul` fungiert die Relation `bestehtAus`, die darüber hinaus zyklisch auf `Hypertexttyp` selbst verweist, da Hypertexttypen bzw. -sorten Instanzen zusätzlicher Hypertextsorten einbetten können. Die Relationen `wirdEingebettetIn` und `kannFungierenAls` setzen `Hypertextsortenmodul` in Verbindung zur Klasse `Hypertextknotentyp`, wobei die Relation `kannFungierenAls` ebenfalls auf `Hypertexttyp` verweist; in umgekehrter Richtung existiert die Relation `umfasstKonventionalisierteVorbelegung`.[23] Darüber hinaus wird sowohl in `Hypertextsortenmodul` als auch in `Hypertextknotentyp` markiert, dass es sich um

[23] Die nachfolgenden Beispiele werden zeigen, wie diese generischen Relationen innerhalb der Definitionen spezifischer Hypertextsorten, Hypertextknotensorten und Hypertextsortenmodule durch Restriktionen spezifiziert werden, indem z. B. ihr Geltungsbereich eingeschränkt wird.

(a) Die Klassen *Hypertexttyp*, *Hypertextknotentyp* und *Hypertextsortenmodul* und ihre Relationen (vgl. Abbildung 5.4, S. 293):



(b) Die Klassen im Kontext der Domänenontologie:

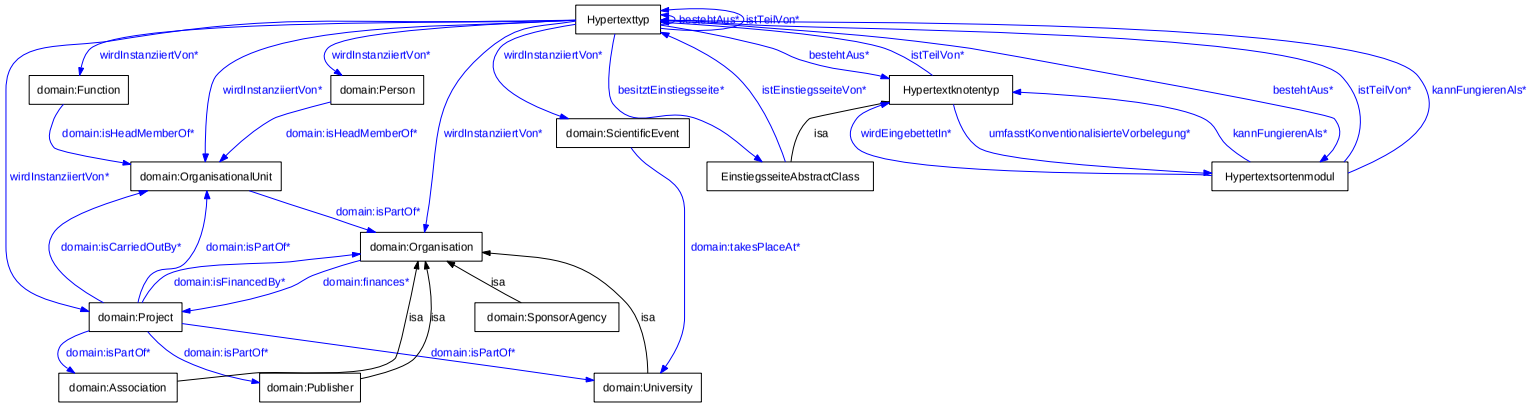


Abbildung 13.7: Die drei zentralen Klassen der Hypertextsortenontologie

Bestandteile von `Hypertexttyp` handelt (mittels `istTeilVon` als Inverse von `bestehtAus`). Abbildung 13.7 stellt die Klassen und Relationen in grafischer Form dar.[24]

Die in Abschnitt 13.3 vorgestellte Ontologie wissenschaftlicher Themen und Fachgebiete wird mittels `owl:imports` in das Domänenmodell (Abschnitt 13.4) integriert, das wiederum in die Hypertextsortenontologie importiert wird. Eine zentrale Verbindung zwischen der Hypertextsortenontologie und dem Domänenmodell wird im unteren Teil von Abbildung 13.7 dargestellt: Die Relation `wirdInstanziiertVon` herrscht zwischen `Hypertexttyp` und den potenziellen Emittenten innerhalb der Untersuchungsdomäne (z. B. Universitäten und ihre Organisationseinheiten, Forschungsprojekte, Verbände, Tagungen, Konferenzen und Personen).[25] Auf diese Weise kann auf einer generischen Ebene die Menge sämtlicher in Frage kommender Emittenten angegeben werden, die in spezifischeren Hypertextsorten wie z. B. *persönliche Homepage eines Wissenschaftlers* über Restriktionen eingeschränkt werden kann. In dem Beispiel könnte für die Relation `wirdInstanziiertVon` die Beschränkung spezifiziert werden, dass nur Vertreter der Klassen `TeachingPosition` oder `ResearchPosition` (Angestellte einer Hochschule, die in Forschung oder Lehre tätig sind) als Produzenten fungieren können, wohingegen die Hypertextsorte *private Homepage eines Studierenden* auf Individuen der Klasse *Student* eingeschränkt wird. Die abstrakte Klasse `HomepageEinerPerson` kann von beliebigen Instanzen der übergeordneten Klasse `Person` instanziiert werden.

13.5.3 Ebene 1: Hypertexttypen und Hypertextsorten

Innerhalb von OWL herrscht zwischen einer Superklasse und ihren Subklassen eine *is-a*-Relation, d. h. Subklassen stellen spezifischere Klassen dar, erben die Merkmale der übergeordneten Klasse(n) und können zusätzliche distinktive Merkmale ausprägen. Auf dieser grundlegenden Relation basieren auch die in der Hypertextsortenontologie repräsentierten Konzepte. Die in der vierten und fünften Analyse (vgl. die Kapitel 11 und 12) ermittelten Hypertexttypen bzw. -sorten sowie Hypertextknotentypen bzw. -sorten wurden – basierend auf den tabellarisch angeordneten Analyseergebnissen, die wiederum der Korpusdatenbank entnommen wurden – vollständig in die Ontologie integriert. Für die Binnenstrukturierung wurden zunächst die gegebenenfalls ermittelten Subtypen in die Ontologie aufgenommen, z. B. hinsichtlich des Hypertextknotentyps *Seite/Abschnitt* (vgl. Abschnitt 12.7.1). Da in OWL kein Mechanismus zur Definition einer abstrakten Klasse[26] existiert, wurden die übergeordneten Konzepte mit dem Suffix `AbstractClass` versehen, so dass der Hypertextknotentyp *Seite/Abschnitt* das Etikett `SeiteAbschnittAbstractClass` besitzt.

[24] Abbildung 13.7 und die weiteren Darstellungen wurden mit Hilfe von *OntoViz* angefertigt, das als *Plug-in* in *protégé* integriert werden kann. *OntoViz* erlaubt die Spezifizierung der in eine Visualisierung aufzunehmenden Klassen, wobei für jede Klasse Parameter aktiviert werden können, die die Aufnahme zusätzlicher Informationen in die Visualisierung bewirken. Das Werkzeug basiert auf dem Tool *dot* aus dem *GraphViz*-Paket, das die Visualisierung von Graphen – basierend auf einer Eingabedatei – vollautomatisch durchführt (vgl. auch Fußnote 30, S. 656). Die Abbildungen 13.3 und 13.4 wurden mit Hilfe des *Plug-ins OWLViz* angefertigt, das primär der grafischen Navigation und Exploration der Konzepte einer Ontologie dient.

[25] Da die Domänen- in die Hypertextsortenontologie importiert wird, besitzen ihre Klassen einen eigenen Namensraum, der über den Präfix `domain` markiert wird. Die Klassen der Ontologie wissenschaftlicher Themen und Fachgebiete, die in das Domänenmodell importiert wird, besitzen ebenfalls diesen Präfix.

[26] Abstrakte Klassen oder auch Gruppenklassen (Reimer, 1991, S. 21) besitzen die Eigenschaft, dass sie *nicht* unmittelbar instanziiert werden dürfen, sie dienen also lediglich der Strukturierung einer Wissensbasis.

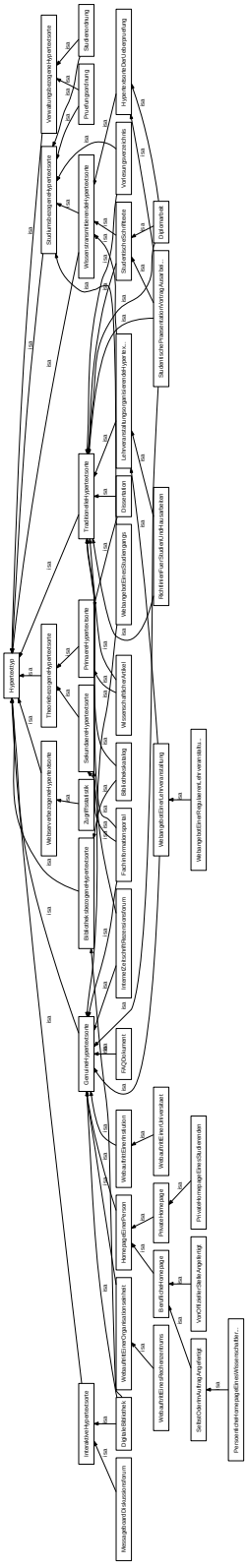


Abbildung 13.8: Typologisierung von Hypertextsorten (Ausschnitt)

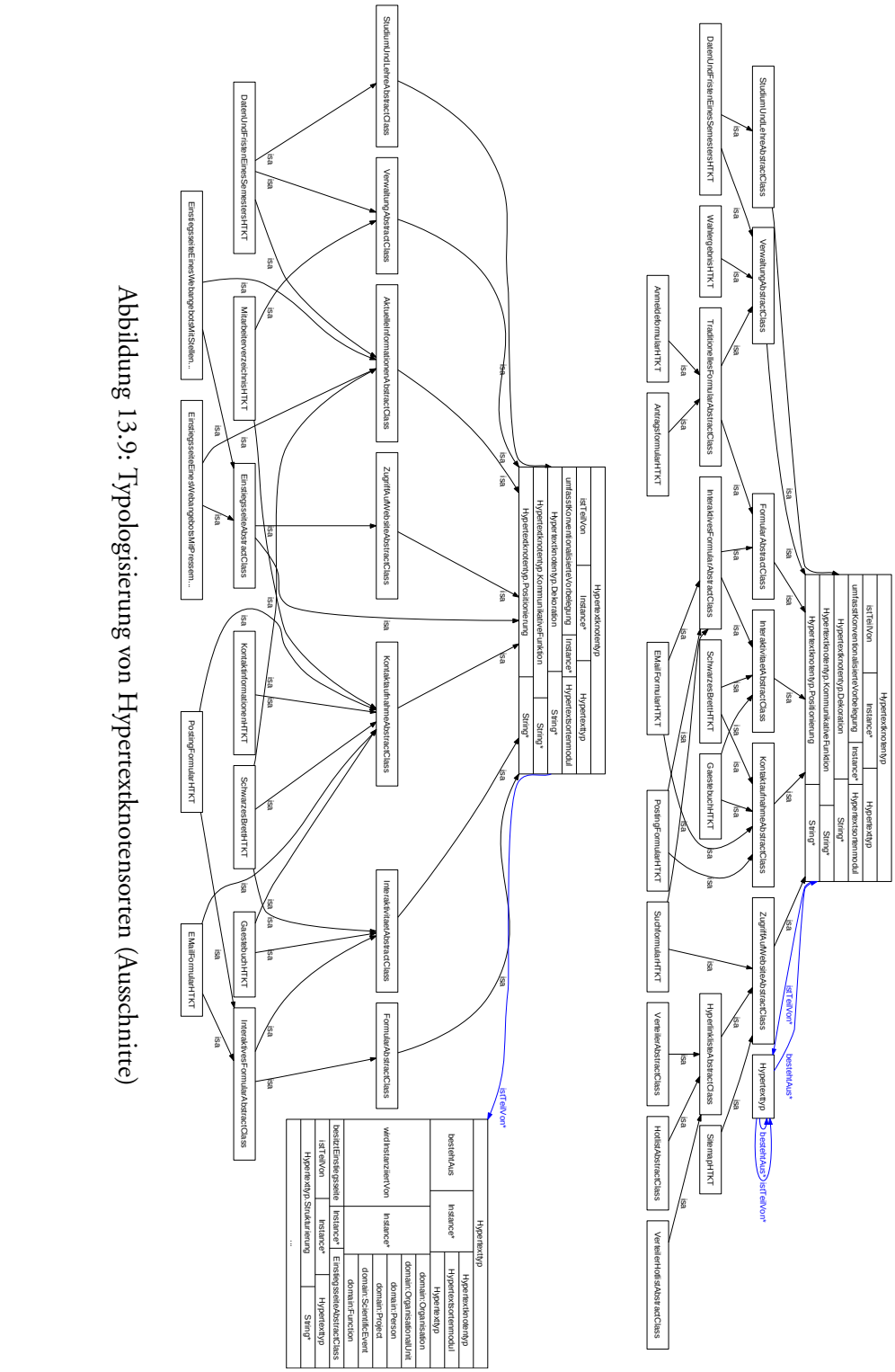


Abbildung 13.9: Typologisierung von Hypertextknotensorten (Ausschnitte)

Die Problematik der Typologisierung von Hypertextsorten und Hypertextknotensorten wurde in Abschnitt 11.8 diskutiert. *Eindeutige* Zuordnungen sind nur in den wenigsten Fällen durchführbar (vgl. die Abbildungen 11.11 und 11.12, S. 526 f., für Beispiele). Die in OWL existierende Relation *is-a* kann zwar diesbezüglich als Einschränkung aufgefasst werden, die jedoch durch den Mechanismus der Mehrfachvererbung aufgehoben wird: Neben den empirisch ermittelten abstrakten Klassen wurden in die Hypertextsortenontologie weiterführende abstrakte Klassen aufgenommen, die unterschiedliche Zuordnungskriterien reflektieren. Falls eine bestimmte Hypertextknotensorte einem derartigen Kriterium zugeordnet werden kann, wird die korrespondierende abstrakte Klasse der jeweiligen Hypertextknotensorte als weiteres Superkonzept hinzugefügt.[27] Abbildung 13.8 umfasst einen kleinen Ausschnitt der oberen Ebenen der Hypertextsortenontologie und zeigt den Einsatz des in OWL integrierten Mechanismus der Mehrfachvererbung zur Typologisierung von Hypertextsorten auf. Abbildung 13.9 stellt zwei weitere Ausschnitte der Ontologie dar, die diese Vorgehensweise innerhalb der Ontologie von Hypertextknotensorten verdeutlichen.[28]

13.5.4 Ebene 2: Hypertextknotentypen und Hypertextknotensorten

Hypertexttypen bzw. -sorten umfassen ein Inventar von Hypertextknotentypen bzw. -sorten. Innerhalb von OWL existieren zwei Möglichkeiten, diese Zusammenhänge zu formalisieren. Zunächst können die Klassendefinitionen mit Restriktionen versehen werden, die die Ausprägung der Relation `bestehtAus` auf die ermittelten Hypertextknotensorten beschränkt. Die zweite Möglichkeit besteht in der Spezifikation einer Relation, die `bestehtAus` untergeordnet wird.[29] Diese Option wurde aus Gründen der Veranschaulichung gewählt, da die zuerst genannte Möglichkeit nicht von *OntoViz* visualisiert werden kann: Abbildung 13.10 zeigt die vier Hypertextsorten *Benutzungshinweise für lokal verfügbare Software*, *Tutorial*, *Lehrwerk, Referenz* sowie *Handbuch, Manual*, die dem Hypertexttyp *Software-Dokumentation* zugehörig sind (vgl. Abschnitt 12.6.4). Zusätzlich zeigt die Grafik die beteiligten Hypertextknotensorten, die über vier untergeordnete Relationen von `bestehtAus` referenziert werden.[30] Sie wurden unmittelbar in den vier zugehörigen Konzepten definiert und ihr Geltungsbereich bezieht sich ausschließlich auf die korrespondierenden Hypertextknotensorten. Darüber werden die Organisationseinheiten dargestellt, die – der fünften Analyse zufolge (vgl. Kapitel 12) – Instanzen des Hypertexttyps *Software-Dokumentation* publizieren.[31]

[27] Abstrakte Klassen bieten sich für den Einsatz in einem IR-Szenario an, so dass der Anwender z. B. eine Suchanfrage auf die von einer spezifischen abstrakten Klasse dominierten Hypertextsorten einschränken könnte.

[28] Abbildung 13.9 zeigt abstrakte Klassen (Suffix: `AbstractClass`) und Hypertextknotentypen bzw. -sorten mit dem Suffix `HTKT`, der der Abgrenzung von abstrakten Klassen dient. Weiterhin wurde die Einführung dieses Suffix notwendig, da identisch benannte Hypertextsortenmodule und Hypertextknotensorten existieren und die Bezeichnungen von Klassen und Relationen innerhalb einer OWL-Ontologie eindeutig sein müssen.

[29] Neben der Klassenhierarchie können in OWL auch die Eigenschaften und Relationen von Klassen in Form einer Hierarchie strukturiert werden (vgl. Abschnitt 13.2.3).

[30] Die Suffixe der vier Relationen bilden sich aus den Anfangsbuchstaben der vier Hypertextsorten.

[31] Die Klasse `WorkGroup` entspricht einer Professur bzw. einer Arbeitsgruppe, `Institute` bezieht sich auf ein Institut oder Seminar und `ComputerCenter` korrespondiert mit dem Rechenzentrum. Die ebenfalls in der Abbildung enthaltenen Relationen `workgroup.subjectarea` und `institute.subjectarea` verweisen auf die Ontologie wissenschaftlicher Themen und Fachgebiete, deren übergeordnete Klasse mit `SubjectArea` etikettiert wurde. Diese Relationen drücken die fachliche Ausrichtung einer Professur oder eines Instituts aus.

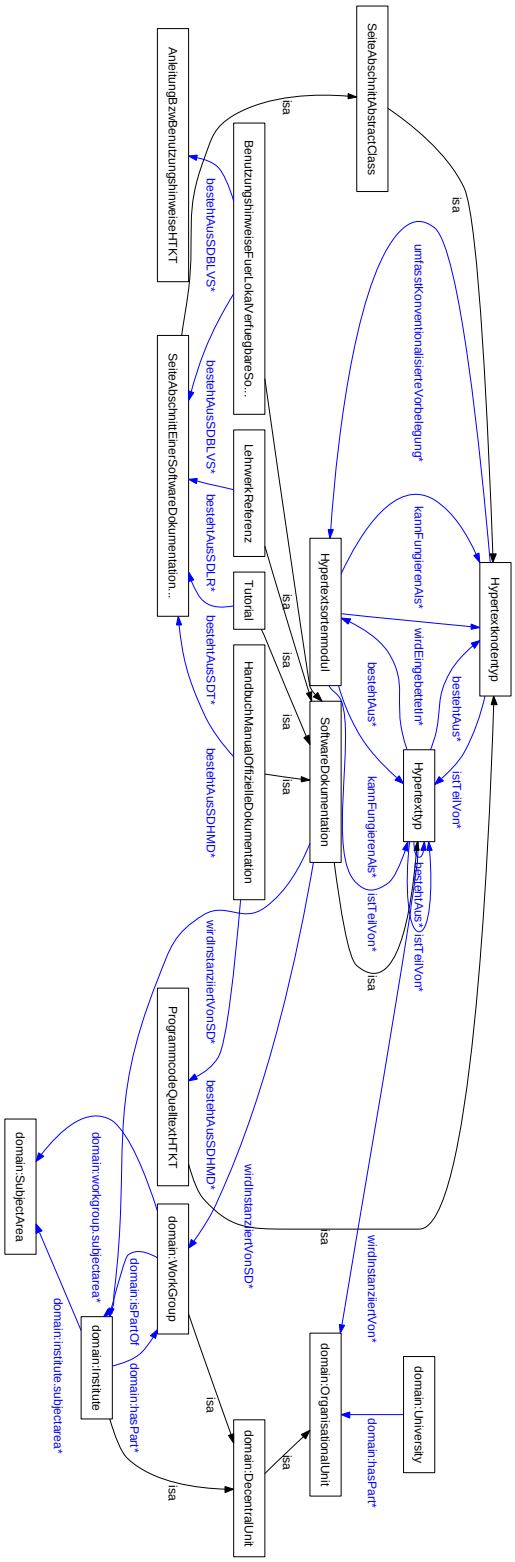


Abbildung 13.10: Die Hypertexttyps *Software-Dokumentation* (Ausschnitt)

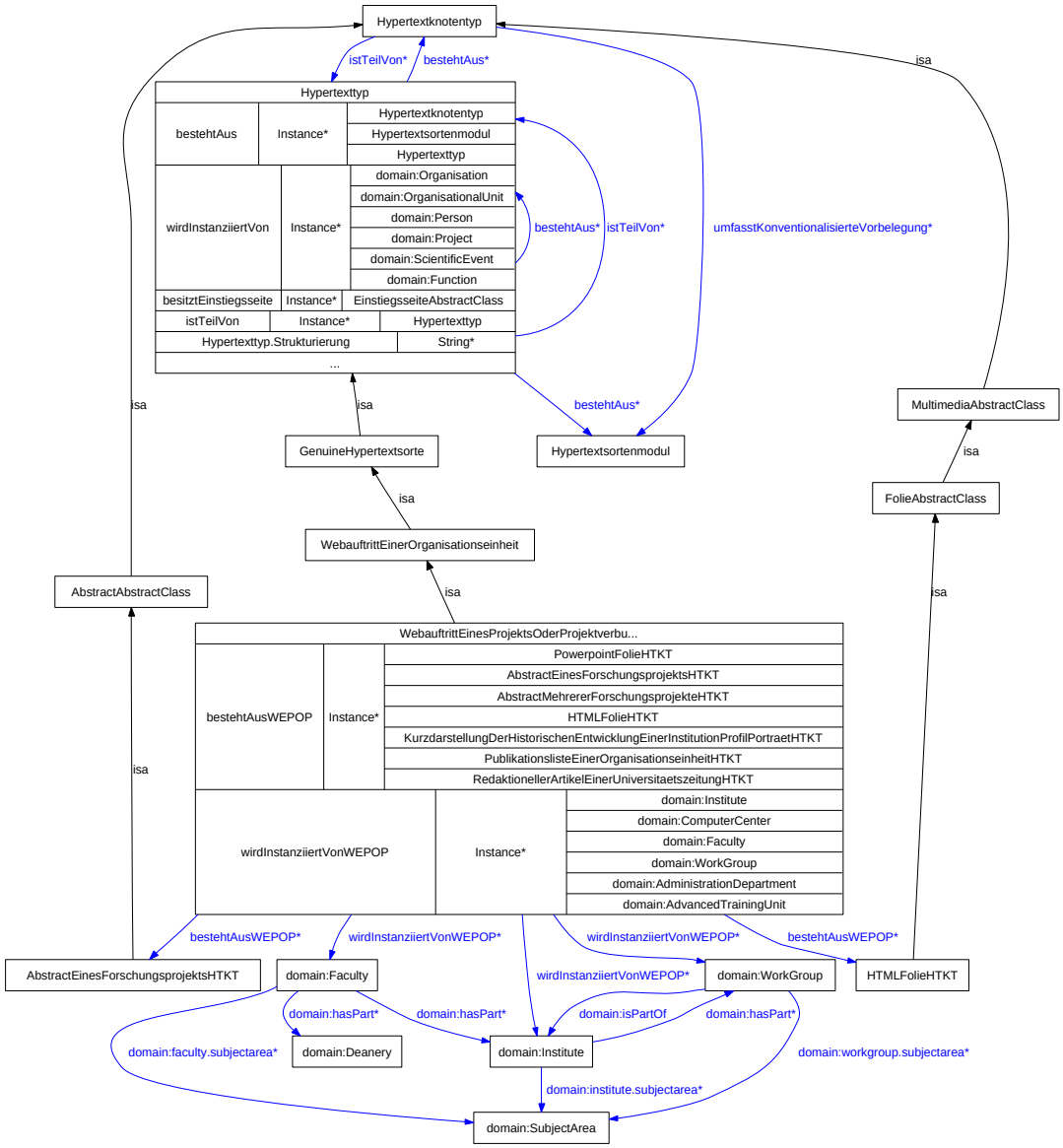


Abbildung 13.11: Der Hypertexttyp *Webauftritt eines Projekts oder Projektverbundes* im Kontext der Domänenontologie und der Hypertextknotentypen (Ausschnitt)

Abbildung 13.11 zeigt mit dem *Webauftritt eines Projekts oder Projektverbundes* (vgl. Abschnitt 12.6.1) ein weiteres Beispiel: Exemplare werden unter anderem von Fakultäten, Instituten bzw. Seminaren und Professuren bzw. Arbeitsgruppen publiziert. Neben den in der Abbildung enthaltenen Klassen der Domänenontologie zeigt die Definition der Hypertextsorte die weiteren Konzepte, die sich im Geltungsbereich der Relation `wirdInstanziiertVonWEPOP` befinden. Zusätzlich zu `HTMLFolieHTKT` und `AbstractEinesForschungsprojektsHTKT` beinhaltet diese Ansicht fünf weitere Hypertextknotensorten, die über den Geltungsbereich der für die Hypertextsorte definierten Relation `bestehtAusWEPOP` spezifiziert werden. Ein Vergleich mit den Geltungsbereichen der abstrakteren Relationen `wirdInstanziiertVon` und `bestehtAus`, die für die übergeordnete Klasse `Hypertexttyp` und alle ihre Subklassen gilt, verdeutlicht den Mechanismus der Angabe spezifischerer Restriktionen zur Repräsentation der distinktiven Merkmale einer Subklasse.

13.5.5 Ebene 3: Hypertextsortenmodule

Die drei Ebenen des Hypertextsortenmodells entsprechen in der Hypertextsortenontologie drei Klassen, die als eigenständige Ontologien mit individuellen Strukturierungen aufgefasst werden können. Die Ergebnisse der in den Kapiteln 9 bis 11 vorgestellten Studien wurden in die Ontologie von Hypertextsortenmodulen integriert und mit den korrespondierenden Konzepten der Hypertextsorten- und Hypertextknotensortenontologien über Relationen verbunden. Weiterhin wurden Verbindungen zur Themen- und Domänenontologie etabliert.

Hypertextsortenmodule prägen primäre und sekundäre Typen aus, die sich auf ein Inventar von sieben Typen beziehen (vgl. Abschnitt 5.6). Diese wurden als einzelne Klassen unterhalb des Superkonzepts `HypertextsortenmodulTyp` in der Ontologie verankert.[32] Ausgehend von der Klasse `Hypertextsortenmodul` referenziert die Relation `besitztTyp` die Klasse `HypertextsortenmodulTyp`. Für diese Relation wurden die beiden untergeordneten Relationen `besitztPrimaerenTyp` und `besitztSekundaerenTyp` definiert, die über Restriktionen innerhalb konkreter Definitionen von Hypertextsortenmodulen mit den zugehörigen primären und sekundären Typen verknüpft werden (vgl. Fußnote 29).

Die Klasse `Hypertextsortenmodul` umfasst die Konzepte `AtomaresHypertextsortenmodul` und `KomplexesHypertextsortenmodul`, denen die korrespondierenden Hypertextsortenmodule zugeordnet wurden. Ein komplexes Hypertextsortenmodul besteht aus mindestens zwei atomaren Hypertextsortenmodulen, die innerhalb der Definition des komplexen Hypertextsortenmoduls über eine Subrelation von `bestehtAusAtomarenHTSM` referenziert werden, in der die atomaren Hypertextsortenmodule den Geltungsbereich der Relation konstituieren.[33] Die Relation `bestehtAusAtomarenHTSM` ist wiederum eine Subrelation von `bestehtAus`.

Die Verbindung zwischen Hypertextsorten sowie Hypertextknotensorten und Hypertextsortenmodulen manifestiert sich in der Relation `umfasstKonventionalisierteVorbelegung` (vgl. Abbildung 13.7), für die mehrere Subrelationen definiert wurden: Über die Relation

[32] Das Superkonzept der Klasse `HypertextsortenmodulTyp` ist `owl:Thing`, d. h. diese Hierarchie wird nicht unmittelbar innerhalb der Ontologie von Hypertextsortenmodulen repräsentiert.

[33] Um das Inventar von Relationen dieser Ontologie überschaubar zu halten, sollte prinzipiell die Option präferiert werden, dieses Wissen als Restriktionen innerhalb der Klassendefinition und nicht als eigenständige Relationen zu repräsentieren. Die Entscheidung für die Subrelation ist ausschließlich in dem Aspekt der gegebenen Visualisierbarkeit durch *OntoViz* begründet (vgl. Abschnitt 13.5.4).

umfasstUniversaleHTSM werden Hypertextsortenmodule referenziert, die in arbiträren Hypertextsorten und Hypertextknotensorten instanziiert werden können, z. B. der Zugriffszähler oder das Datum der letzten Änderung. Darüber hinaus wurden für verschiedene Hypertextknotensorten Relationen wie z. B. umfasstHTSM_EEUWA (für *Einstiegsseite eines universitären Webauftritts*) definiert, die die an der jeweiligen Hypertextknotensorte beteiligten Hypertextsortenmodule spezifiziert. Da sich diese in obligatorische und optionale Hypertextsortenmodule aufteilen, besitzen Relationen wie umfasstHTSM_EEUWA, die ihrerseits Subrelationen von umfasstKonventionalisierteVorbelegung darstellen, zwei Subrelationen, deren Geltungsbereiche die korrespondierenden obligatorischen und optionalen Hypertextsortenmodule enthalten (umfasstObligatorischeHTSM_EEUWA und umfasstOptionaleHTSM_EEUWA).

Abbildung 13.12 verdeutlicht die Vorgehensweise durch zwei Ausschnitte der Hypertextsortenmodule, die an der Hypertextknotensorte *Einstiegsseite eines universitären Webauftritts* beteiligt sind.[34] Darüber hinaus zeigt der obere Ausschnitt eine Binnenstrukturierung, die sich auf die Navigationshilfen bezieht (vgl. Abschnitt 11.4.2): Die Navigationshilfe ist eine Subklasse von AtomaresHypertextsortenmodul und umfasst Konzepte, die unter anderem der primären, sekundären und zielgruppenspezifischen Navigationshilfe entsprechen. Diese besitzen wiederum jeweils eine Subklasse, die die spezifischen Ausprägungen innerhalb der genannten Hypertextknotensorte repräsentieren. Für diese Navigationshilfen können unterschiedliche Typen belegt werden, die sich z. B. auf die komplexe Tabelle oder die einfache Liste mit vertikal angeordneten Listenpunkten beziehen. Diese Typen werden in einer weiteren Ontologie repräsentiert, die sich – ebenso wie HypertextsortenmodulTyp – außerhalb der drei zentralen Ebenen der Hypertextsortenontologie befindet. Über die Relation besitztNavigationshilfeTyp werden die hochfrequenten Subtypen in den spezifischen Ausprägungen der Hypertextsortenmodule über Restriktionen referenziert.

Der in Abbildung 13.12 unten dargestellte Ausschnitt verdeutlicht die implizite Repräsentation genereller und spezifischer Hypertextsortenmodule: Die *Einstiegsseite eines universitären Webauftritts* kann als Hypertextknotensortenvariante der allgemeineren Hypertextknotensorte *Einstiegsseite des Webauftritts einer Institution* aufgefasst werden. Die innerhalb der Analyse der 35 Einstiegsseiten ermittelten generellen Hypertextsortenmodule können sowohl in der übergeordneten Hypertextknotensorte als auch in der Variante verwendet werden, wohingegen die spezifischen Hypertextsortenmodule ausschließlich für die Variante gelten (vgl. Tabelle 11.9, S. 486). Die übergeordnete Hypertextknotensorte wird in der Ontologie als Superkonzept repräsentiert, das die generellen Hypertextsortenmodule über Relationen referenziert. Somit wird dieses Inventar genereller Hypertextsortenmodule über die in OWL vorgesehene Vererbung an die Subklasse EinstiegsseiteEinesUniversitaerenWebauftritts propagiert, während in dieser Hypertextknotensortenvariante lediglich die spezifischen Hypertextsortenmodule zu repräsentieren sind.

[34] Die im Rahmen der Analysen ermittelten Frequenzen von Hypertextsortenmodulen werden in *datatype properties* hinterlegt. Diesbezüglich existiert die Problematik, dass in OWL kein Mechanismus zur Aufnahme konkreter Daten dieser Art in Klassen existiert, weil derartige Informationen eher Individuen bzw. Instanzen zuzurechnen sind. Da in der Hypertextsortenontologie jedoch Klassen als abstrakte Beschreibungen fungieren, die auf empirischen Befunden beruhen, sind die Frequenzangaben jedoch durchaus den Klassen zugehörig, so dass für eine Hypertextknotensorte mehrere Datentypeneigenschaften vom Typ Integer (xsd:int) angelegt wurden, deren Wertebereich auf den Wert eingeschränkt wird, der in der jeweiligen Analyse ermittelt wurde (z. B. EEUWA.AbbildungDerInstitution.frequency mit dem Wert 49).

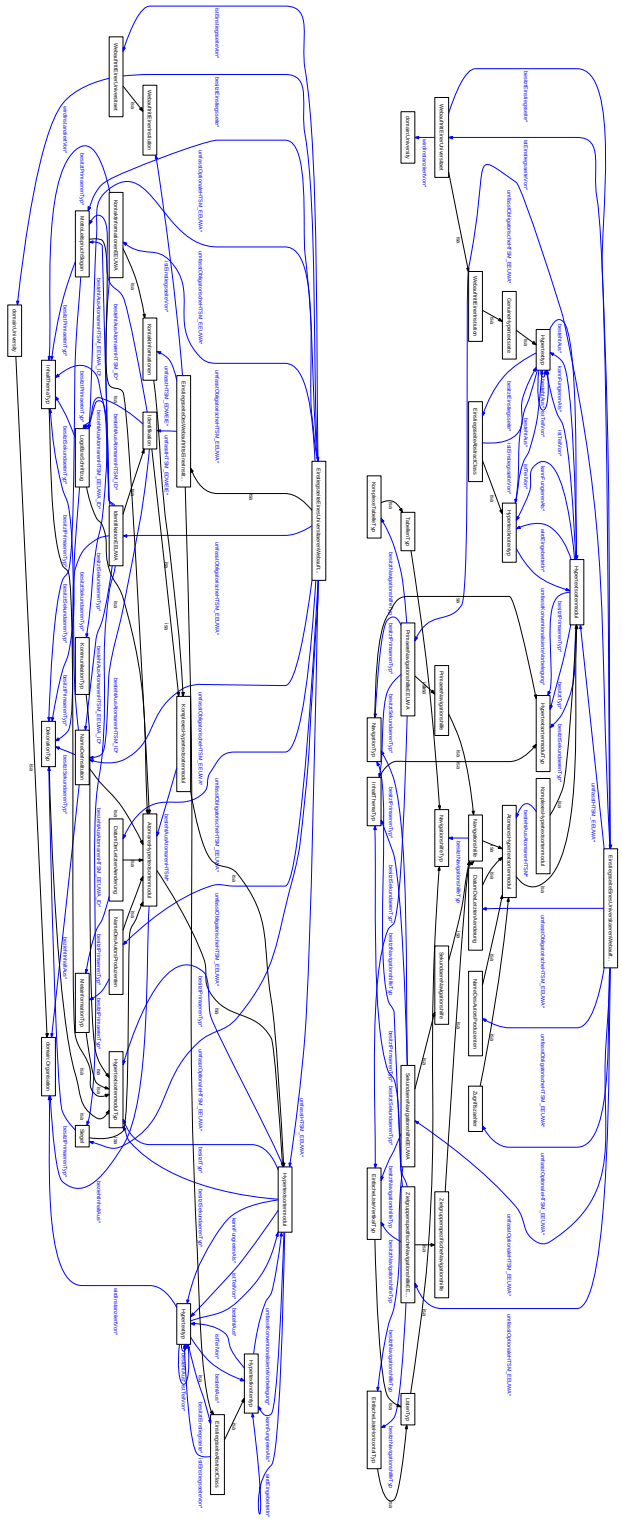


Abbildung 13.12: Die Hypertextknotensorte *Einstiegsseite eines universitären Webauftritts* (Ausschnitte)

13.5.6 Weiterführende Beispiele

Nachfolgend werden mehrere weiterführende Beispiele erläutert, die Spezifika der Hypertextsortenontologie betreffen. Hierzu zählt die Markierung von Einstiegsseiten einer Hypertextsorte, die Verbindung zwischen Hypertextsorten- und Domänenontologie, der Hypertexttyp *Homepage einer Person*, zielgruppenspezifische Navigationshilfen, das Verhältnis zwischen Hypertextsortenmodulen und Hypertextmodulen sowie Sequenzierungen.

Hypertextsorten und Einstiegsseiten

Die in Abbildung 13.12 enthaltene allgemeine Verbindung zwischen Einstiegsseiten und Hypertextsorten stellt sich wie folgt dar: Die Klasse `EinstiegsseiteAbstractClass` wird von der Klasse `Hypertexttyp` über die Relation `besitztEinstiegsseite` referenziert (in umgekehrter Richtung existiert die Inverse `istEinstiegsseiteVon`). Die Hypertextknotensorte `EinstiegsseiteEinesUniversitaerenWebauftritts` ist eine Subklasse des Hypertextknotentyps `EinstiegsseiteDesWebauftrittsEinerInstitution`, die wiederum eine Subklasse von `EinstiegsseiteAbstractClass` ist. Auf diese Weise kann zwischen der zugehörigen Hypertextsorte `WebauftrittEinerUniversitaet` (diese stellt eine Subklasse des Hypertexttyps `WebauftrittEinerInstitution` dar) und der genannten Hypertextknotensorte die spezifischere Ausprägung der bereits auf einer generischen und allgemein gültigen Ebene repräsentierten Relation festgelegt werden.

Verbindungen zwischen Hypertextsorten- und Domänenontologie

Abbildung 13.13 verdeutlicht die Verbindung zur Domänenontologie über die Hypertextknotensorte `VerteilerOrganisationseinheiten`.[35] Ein Exemplar dieser Hypertextknotensorte wird in vielen Fällen von der Instanz der Einstiegsseite eines universitären Webauftritts über einen Hyperlink verknüpft (vgl. Abschnitt 11.4.2). Der nach dem Kriterium der Organisationseinheiten einer Hochschule zusammengestellte Verteiler referenziert wiederum die Webauftritte von Fakultäten, Fachbereichen, Professuren, Arbeitsgruppen sowie zentralen Einheiten. Diese besitzen innerhalb der Domänenontologie korrespondierende Konzepte, die sie als zentrale und dezentrale Organisationseinheiten einer Universität kennzeichnen. Über individuelle Relationen sind diese Klassen mit der Ontologie wissenschaftlicher Themen und Fachgebiete verbunden, so dass der Arbeitsschwerpunkt einer Organisationseinheit repräsentiert werden kann, der sich unmittelbar auf die Inhalte des zugehörigen Hypertextes auswirkt.[36] Die Organisationseinheit bestimmt somit gewissermaßen das Repertoire von Hypertextsorten, das innerhalb ihres Webauftritts verwendet wird.

[35] Siehe hierzu auch Abbildung 5.2 (S. 276), die die typische hierarchische Strukturierung eines universitären Webauftritts auf der Basis der Organisationseinheiten zeigt (vgl. ebenfalls Abbildung 11.4, S. 494).

[36] Ein weiterer Zusammenhang zwischen der Domänenontologie und der Ontologie wissenschaftlicher Themen und Fachgebiete besteht in Restriktionen hinsichtlich des individuellen Arbeitsschwerpunkts der hierarchisch gestuften dezentralen Organisationseinheiten: Fakultäten besitzen oftmals eine inhaltliche Ausrichtung, die sich grob an der ersten Ebene der Themenontologie orientiert (vgl. Abbildung 13.3, S. 587), wohingegen sich Fachbereiche eher auf die zweite, Institute bzw. Seminare auf die zweite oder dritte und Professuren bzw. Arbeitsgruppen auf sehr spezifische weiterführende Ebenen der Themenontologie beziehen.

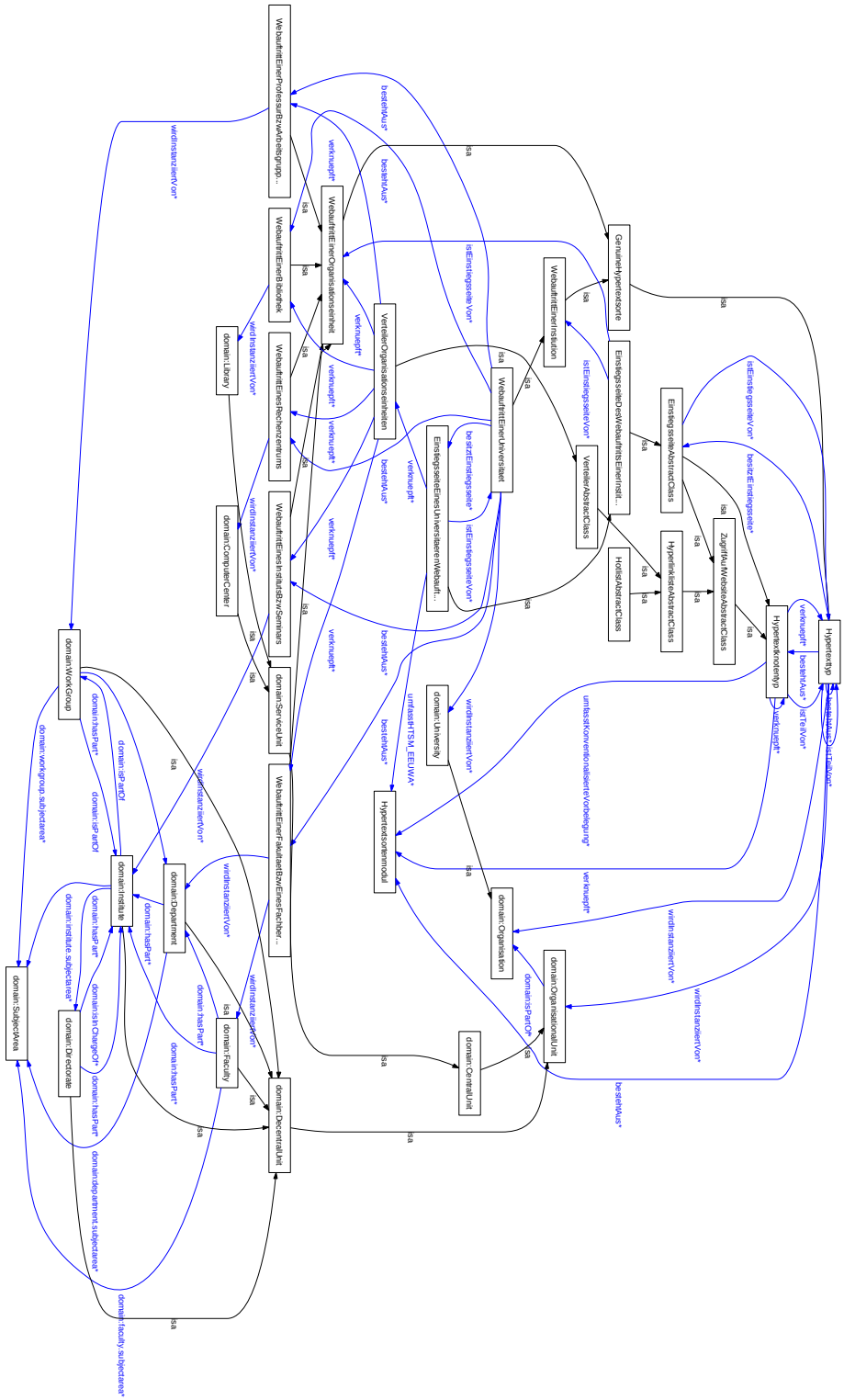


Abbildung 13.13: Relationen zwischen der Hypertextsortenontologie und dem Domänenmodell (Ausschnitt)

Die abstrakte Klasse Homepage einer Person

Abbildung 13.14 stellt die Hypertextsorten des abstrakten Hypertexttyps *Homepage einer Person* sowie die jeweiligen Einstiegsseiten als korrespondierende Hypertextknotensorten dar. Die diesbezüglich in Kapitel 10 präsentierte Typologisierung (vgl. Abbildung 10.7, S. 460) kann innerhalb der OWL-Ontologie unmittelbar übernommen werden und beruht auf einer Differenzierung zwischen privaten und beruflichen Homepages. In Abschnitt 10.6 wird argumentiert, dass der Hypertexttyp *Homepage einer Person* einen prototypischen Kern besitzt, der von den untergeordneten Hypertextsorten übernommen und mit individuellen Spezifika angereichert wird. Gemäß dem Prinzip der Vererbung von Klasseneigenschaften wird dieser Kern in dem Konzept `HomepageEinerPerson` und dem zugehörigen Hypertextknotentyp `EinstiegsseiteDerHomepageEinerPersonAbstractClass` spezifiziert.[37] Der prototypische Kern entspricht wiederum den in den Kapiteln 9 und 10 ermittelten generellen Hypertextsortenmodulen, deren Verwendung nicht auf die jeweils untersuchte Hypertextsorte beschränkt ist. Besitzt ein Hypertextsortenmodul somit einen generellen Status (vgl. Tabelle 9.9, S. 422, sowie Tabelle 10.3, S. 457), wird es als Bestandteil des prototypischen Kerns aufgefasst. Bei einer typischen Ausprägung innerhalb der entsprechenden Einstiegsseite (vgl. hierzu ebenfalls die genannten Tabellen), wird es in dem Hypertextknotentyp `EinstiegsseiteDerHomepageEinerPersonAbstractClass` repräsentiert, bei einer typischen Ausprägung in einem internen Knoten wird das Hypertextsortenmodul innerhalb der Klasse `HomepageEinerPerson` spezifiziert. Die Angabe von Hypertextsortenmodulen erfolgt, wie bereits in Abschnitt 13.5.5 diskutiert, durch die Definition von zwei Relationen, in deren Geltungsbereiche die obligatorischen und optionalen Hypertextsortenmodule aufgenommen werden. Diese Relationen sind Subrelationen einer generischen Relation, der im Falle der *Homepage einer Person* eine besondere Rolle zukommt: Falls ein Hypertextsortenmodul in der einen Hypertextsorte obligatorisch, in der anderen jedoch nur optional ist, kann über die Aufnahme in die Definition der übergeordneten Relation `umfasstHTSM_HPEP` die Gemeinsamkeit obligatorischer und optionaler Konstituenten, d. h. die potenzielle Existenz des Hypertextsortenmoduls in einer beliebigen Instanz einer beliebigen untergeordneten Hypertextsorte markiert werden. Dies gilt auch für diejenigen generellen Hypertextsortenmodule, die von lediglich *einer* Analyse bestätigt wurden, jedoch prinzipiell auch in der anderen Hypertextsorte vorkommen können. In den *beiden* untersuchten Hypertextsorten obligatorisch oder optional existente Hypertextsortenmodule können über die beiden Subrelationen `umfasstObligatorischeHTSM_HPEP` und `umfasstOptionaleHTSM_HPEP` angegeben werden.

Abbildung 13.15 zeigt einen Ausschnitt der an dem prototypischen Kern beteiligten Hypertextsortenmodule und die spezifischeren Ausprägungen der beiden Hypertextsorten *private Homepage eines Studierenden* und *persönliche Homepage eines Wissenschaftlers*. Die abstrakte Klasse `EinstiegsseiteDerHomepageEinerPersonAbstractClass` umfasst z. B. das obligatorische, komplexe und generalisierende Hypertextsortenmodul `Identifikation`, das zwei untergeordnete Ausprägungen besitzt, die sich auf die beiden untersuchten Hypertextsorten beziehen (`IdentifikationPHES` und `IdentifikationPHEW`). Analog umfasst die Klasse `HomepageEinerPerson` das Hypertextsortenmodul `Kontaktinformationen`, das ebenfalls zwei

[37] Bei diesen Klassen bzw. der korrespondierenden Hypertextsorte sowie dem Hypertextknotentyp handelt es sich, wie bereits angemerkt wurde, um abstrakte Klassen, die nicht ausgeprägt werden können.

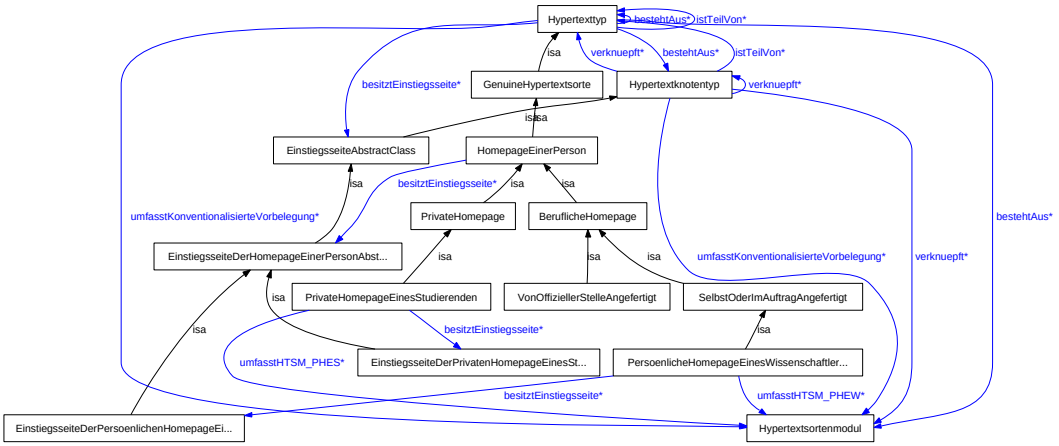


Abbildung 13.14: Die Hypertextsorten des Hypertexttyps *Homepage einer Person*

spezifischere Subklassen besitzt. In diesem Zusammenhang existiert zwischen den Hypertextsorten der Unterschied, dass dieses Hypertextsortenmodul bei der *persönlichen Homepage eines Wissenschaftlers* den primären Typ Inhalt bzw. Thema und den sekundären Typ Kommunikation besitzt, wohingegen im Falle der *privaten Homepage eines Studierenden* aufgrund der individuellen Typen der beteiligten Hypertextsortenmodule der primäre Typ Kommunikation und der sekundäre Typ Interaktion vorliegt. Durch die Modellierung als einzelne Klassen können derartige Spezifika ebenfalls in die Ontologie aufgenommen werden.

Abbildung 13.16 zeigt die diskutierten Bestandteile der Hypertextsorten *private Homepage eines Studierenden* und *persönliche Homepage eines Wissenschaftlers* im Kontext der Domänenontologie. Für die übergeordnete Klasse `Hypertexttyp` gilt, dass sie unter anderem von Personen instanziiert werden kann. Für die genannten Hypertextsorten wurden weiterführende Restriktionen integriert, die sich auf Studierende (`domain:Student`) sowie Hochschulangehörige, die in Forschung und Lehre tätig sind (`domain:ResearchPosition`, `domain:TeachingPosition`), beziehen; der Hypertexttyp *Homepage einer Person* besitzt bezüglich der Instanziierung keine Restriktion, wodurch sein Status als abstrakte Klasse deutlich wird. Die atomaren Hypertextsortenmodule der spezifischen Ausprägungen der beiden abstrakten Klassen `Kontaktinformationen` und `Identifikation` verdeutlichen die Kopplung an die Untersuchungsdomäne.[38] Das innerhalb von `KontaktinformationenPHEW` verwendete atomare Hypertextsortenmodul `Sprechstunde` bezieht sich auf die Klasse `domain:OfficeHours`, die innerhalb der Domänenontologie als `domain:MeetingEvent` (Superklasse: `domain:Event`) repräsentiert wird. Das Ereignis wird von Angehörigen der Klasse `domain:TeachingPosition` durchgeführt, die auch an ihm teilnehmen. Der Ort einer Sprechstunde ist das Büro der jeweiligen Person, das mit einem Raum identisch ist, der sich in einem Gebäude befindet, das eine Straßenadresse besitzt (`domain:StreetAddress`), die wiederum die Instanz des Hypertextsortenmoduls `Strassenadresse` determiniert, das innerhalb des komplexen Hypertextsortenmoduls `KontaktinformationenPHEW` verwendet wird. Die Konzepte `domain:Email`

[38] Die Klasse `AtomaresHypertextsortenmodul` konnte aus Darstellungsgründen nicht in Abbildung 13.16 aufgenommen werden, weshalb z. B. `Strassenadresse` und `Telefonnummer` keine Superklasse besitzen.

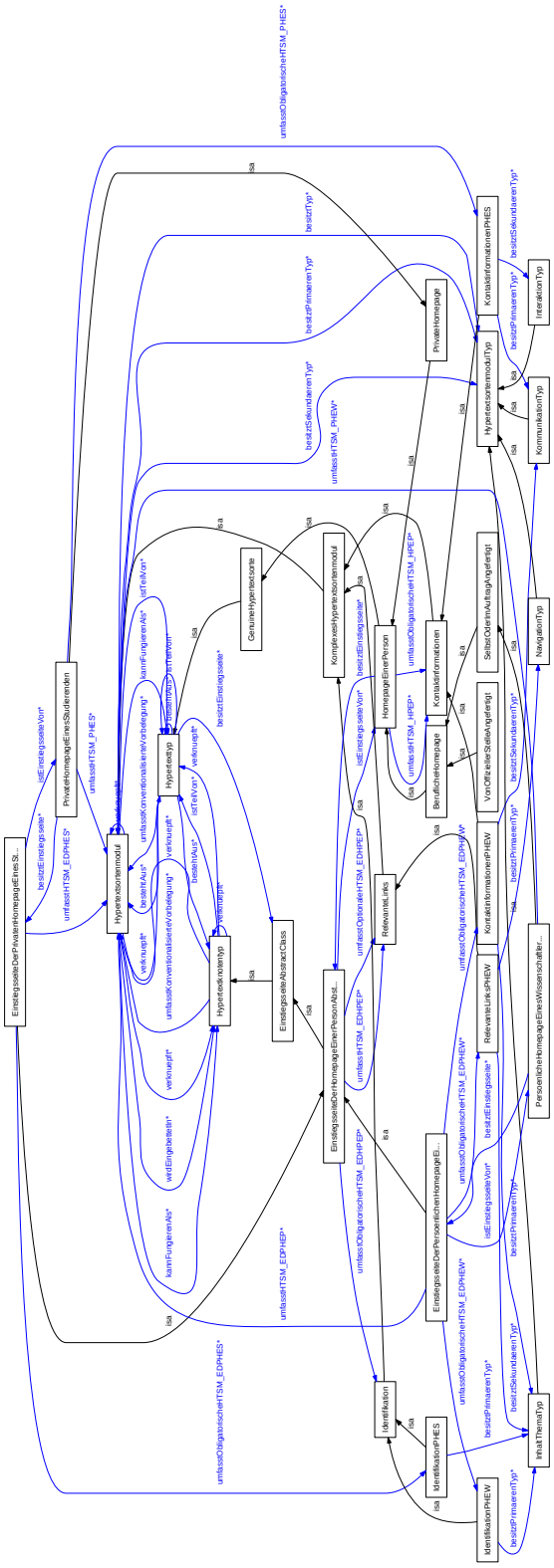


Abbildung 13.15: Die Vererbung generieller und Repräsentation spezifischer Hypertextsortenmodule (Ausschnitt)

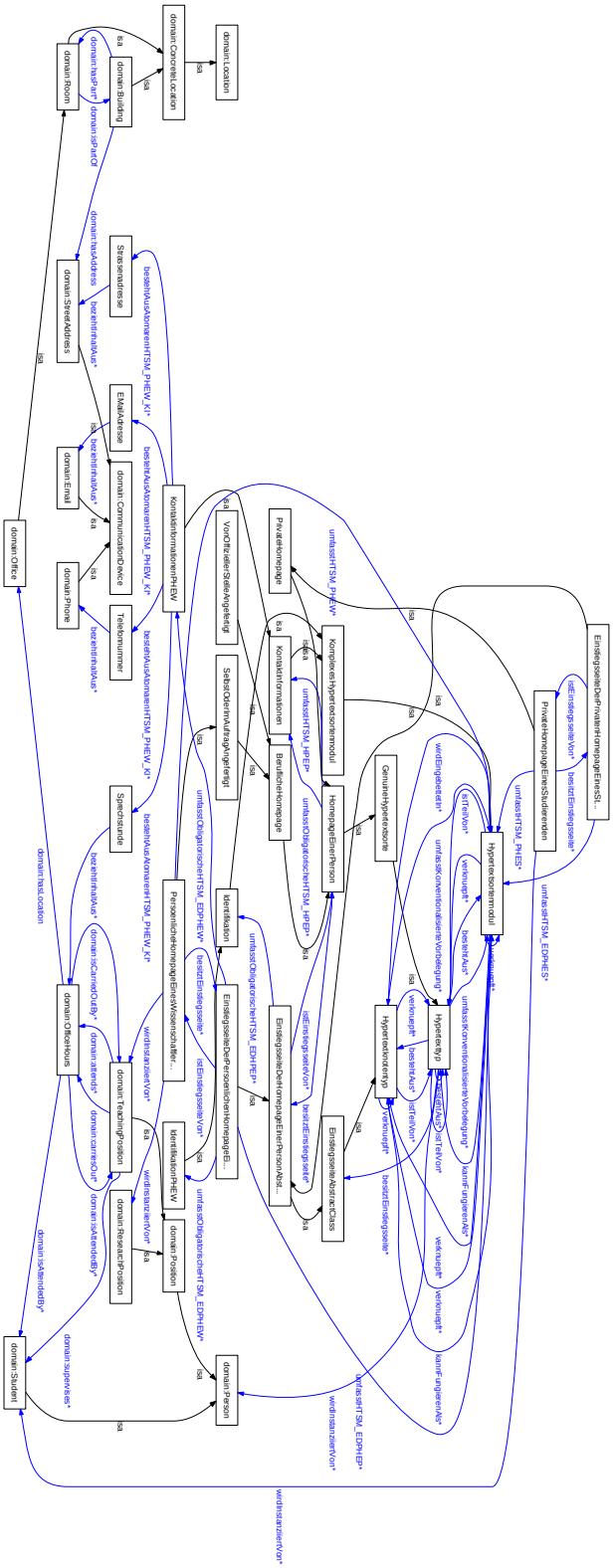


Abbildung 13.16: Die Hypertextsorte *persönliche Homepage eines Wissenschaftlers* im Kontext der Domänenontologie (Ausschnitt)

und domain:Phone sind Subklassen von domain:CommunicationDevice und korrespondieren mit den Hypertextsortenmodulen EMailAdresse und Telefonnummer.[39]

Zielgruppenspezifische Navigationshilfen

Die Websites von Hochschulen richten sich an unterschiedliche Zielgruppen. In einigen Einstiegsseiten universitärer Webauftritte existieren zielgruppenspezifische Navigationshilfen (vgl. Abschnitt 11.4.2), die Hyperlinks enthalten, deren Hyperlinkanzeiger mit der Bezeichnung einer Zielgruppe wie z. B. „Presse“, „Wirtschaft“ oder „Alumni“ etikettiert wurden und den Rezipienten zu einer Instanz des Hypertextknotentyps *Verteiler* führen (vgl. Abschnitt 11.5.3, insbesondere Tabelle 11.12, S. 496). Diese umfassen Listen von Hyperlinks zu weiterführenden Webangeboten, die – nach Ansicht des Produzenten – für Angehörige der Zielgruppen von Relevanz sein könnten. Abbildung 13.17 zeigt die Repräsentation des innerhalb der *Einstiegsseite eines universitären Webauftritts* optionalen, atomaren Hypertextsortenmoduls ZielgruppenspezifischeNavigationshilfeEEUWA. Dieses Hypertextsortenmodul richtet sich – repräsentiert durch die Relation gerichtetAn – unter anderem an ehemalige Studierende (domain:Alumnus), Angehörige der Presse (domain:MediaPosition) und Mitarbeiter von Unternehmen, die z. B. Wissenschaftler mit der Anfertigung eines Gutachtens beauftragen (domain:ResearchAndDevelopmentPosition). Die durch gerichtetAn repräsentierten Beziehungen zwischen diesem Hypertextsortenmodul und Angehörigen von Zielgruppen sind gleichermaßen Bestandteile der kommunikativen Funktion des Hypertextsortenmoduls als auch der Kontextfaktoren (hier: die an der kommunikativen Handlung beteiligten Individuen). Betroffen ist hiervon auch die Strukturierung mehrerer Instanzen eingebetteter Hypertext(knoten)sorten, die entweder in Form eines Verteilers oder eines speziell angefertigten Hypertextes weiterführende Informationen für eine Zielgruppe beinhalten.

Hypertextmodule und Hypertextsortenmodule

Die im vorangegangenen Abschnitt diskutierte Abbildung 13.17 stellt eine sehr abstrakte und vereinfachende Darstellung dar – Abbildung 13.18 hebt verschiedene weiterführende Details hervor: Die bisherigen Ausführungen haben die Ebene der Hypertextmodule bewusst nur am Rande diskutiert (vgl. Abschnitt 5.6.2). Alle spezifischen Subklassen des atomaren Hypertextsortenmoduls Navigationshilfe (z. B. die primäre, sekundäre und zielgruppenspezifische Navigationshilfe) bestehen aus einer Sequenz von Hyperlinks, die auf unterschiedliche Weise angeordnet werden können. Zur Repräsentation dieser Beziehungen wurde die Klasse Hypertextmodul (unterhalb von owl:Thing) in die Ontologie aufgenommen, deren Subklassen den Bausteinen der Textoberfläche entsprechen (vgl. Abschnitt 5.6.2). Auch bezüglich dieser generischen Ebene sind Partonomien zu beobachten, so ist ein einzelner, isoliert positionierter Hyperlink ebenso als Hypertextmodul zu betrachten wie die Liste von Hyperlinks, die aus Hyperlinks besteht (vgl. Abbildung 5.3, S. 286). Diese Relation wird durch bestehtAusHTM und ihre Inverse istTeilVonHTM ausgedrückt, die ihrerseits – ebenso wie die

[39] Die Relationen, die von diesen Konzepten der Domänenontologie zu domain:Office (ein Festnetztelefon befindet sich in einem Büro) und domain:TeachingPosition sowie domain:ResearchPosition führen (Hochschulangehörige verfügen über Kommunikationsmittel), wurden nicht in die Abbildung übernommen.

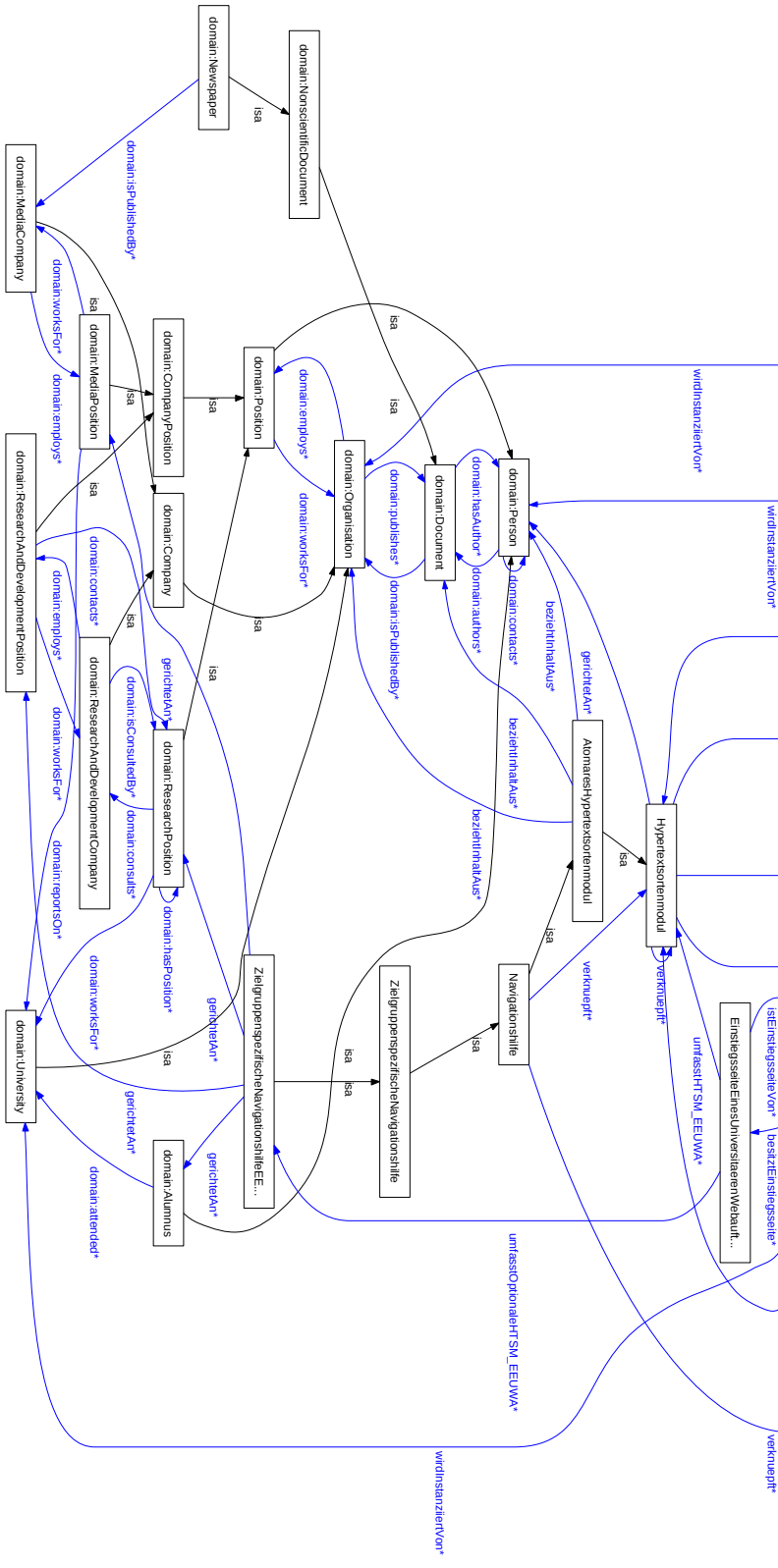


Abbildung 13.17: Die zielgruppenspezifische Navigationshilfe im Kontext des Domänenmodells (Ausschnitt)

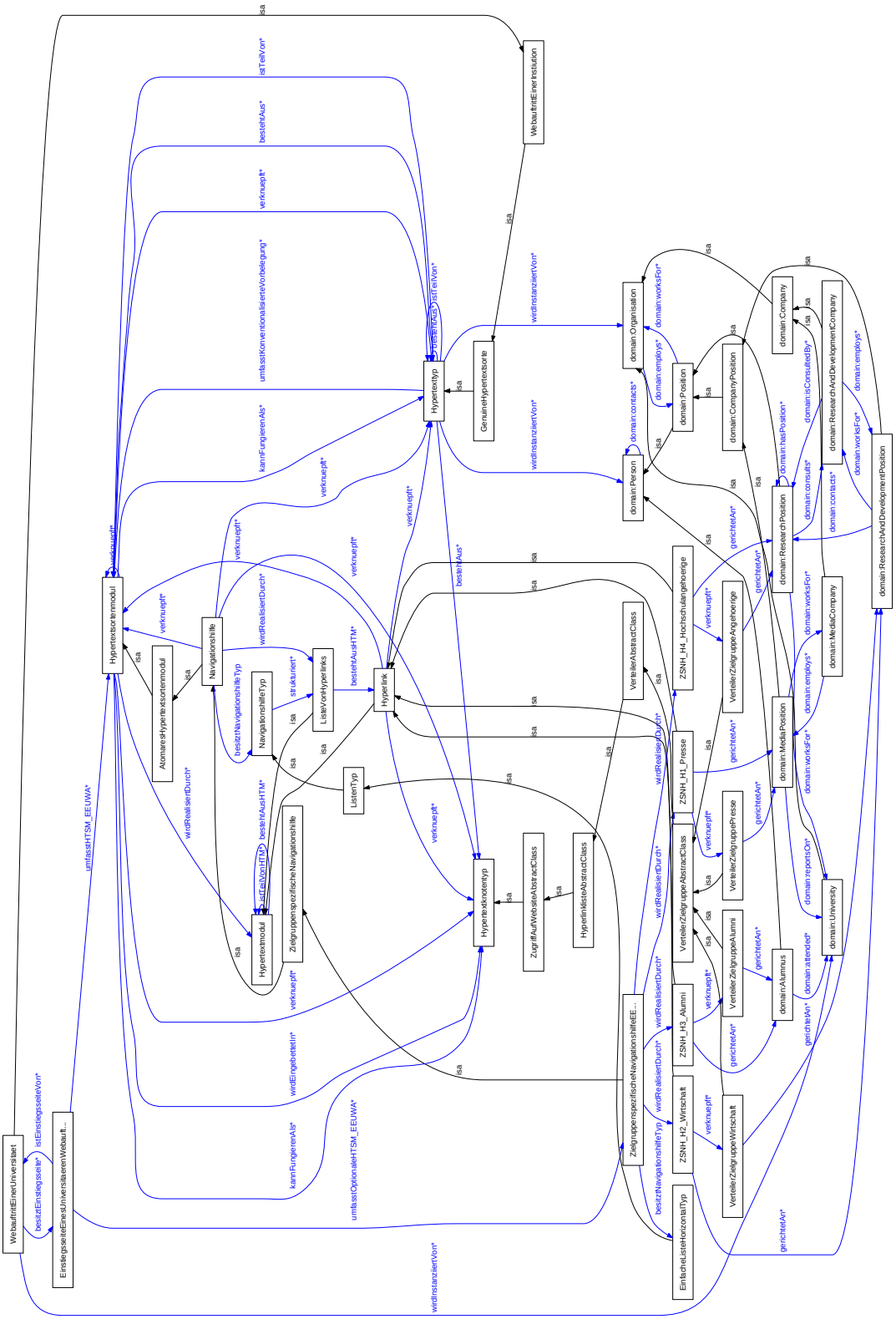


Abbildung 13.18: Die zielgruppenspezifische Navigationshilfe und die beteiligten Hypertextmodule (Ausschnitt)

korrespondierenden Relationen bestehtAus und istTeilVon (vgl. Abschnitt 13.5.2) – Subrelationen von bestehtAusAbstractRelation und istTeilVonAbstractRelation sind, so dass diese beiden Paare partitiver Relationen eine gemeiname Superrelation besitzen. Die Verbindung zwischen den Klassen Hypertextsortenmodul und Hypertextmodul erfolgt durch die Relation wirdRealisiertDurch, so dass z. B. ausgedrückt werden kann, dass das Hypertextsortenmodul Navigationshilfe durch das Hypertextmodul ListeVonHyperlinks realisiert wird. Die Klasse NavigationshilfeTyp (vgl. Abschnitt 13.5.5) bezieht sich auf unterschiedliche Typen von Navigationshilfen, die innerhalb der Instanzen einer bestimmten Hypertextknotensorte ermittelt werden können. Durch diese Differenzierung werden in der Hypertextsortenontologie unterschiedliche Varianten ermöglicht, die jedoch – in Bezug auf Hypertextmodule – auf der gemeinsamen Basis der sequenziellen Anordnung von Hyperlinks fußen. Eine spezifische Subklasse von NavigationshilfeTyp determiniert also die Anordnung von ListeVonHyperlinks, weshalb zwischen diesen Klassen die Relation strukturiert eingeführt wurde. Abbildung 13.18 stellt ausschnittsweise verschiedene Subklassen von Hyperlink dar, die zur Realisierung der zielgruppenspezifischen Navigationshilfe benutzt werden, die in den Einstiegsseiten universitärer Webauftritte Verwendung finden. Diese Hyperlinks (z. B. ZSNH_H1_Presse) richten sich – im Gegensatz zu dem Hypertextsortenmodul ZielgruppenspezifischeNavigationshilfeEEUWA (vgl. Abbildung 13.17) – an genau eine Zielgruppe (beispielsweise Medienvertreter, d. h. domain:MediaPosition) und verknüpfen Instanzen des Hypertexttyps *Verteiler*, die vom Produzenten nach dem Kriterium der Zielgruppenzugehörigkeit zusammengestellt wurden (vgl. Tabelle 11.12).

Hypertextsorten und Sequenzierungen

Die Sequenzierung der Instanzen von Hypertextsorten, Hypertextknotensorten und Hypertextsortenmodulen wird durch die Relation verknuepft markiert. Über diese Relation können prinzipiell beliebige Graphenstrukturen modelliert werden, es ist dabei jedoch die Funktion der Ontologie als maschinenlesbares Format zu beachten: Die drei Ebenen der Hypertextsorten, Hypertextknotensorten und Hypertextsortenmodule werden jeweils durch eigenständige Klassenhierarchien repräsentiert, die generische, d. h. abstrahierende Beschreibungen der in den Studien ermittelten Ergebnisse beinhalten. Somit besitzt die detaillierte Aufnahme von Verknüpfungsstrukturen naturgemäß die Prämisse, eben diese Strukturen in Stichprobenanalysen zu erheben, um somit gegebenenfalls Verallgemeinerungen feststellen und in die Ontologie integrieren zu können. Die in Teil III der Arbeit präsentierten Analysen beziehen sich jedoch primär auf die initiale Identifizierung der Konstituenten, die in den Webangeboten der Untersuchungsdomäne in Bezug auf die drei Ebenen des Hypertextsortenmodells verwendet werden. Darüber hinaus haben insbesondere die Kapitel 9 bis 11 gezeigt, dass umfangreiche Varianzen hinsichtlich der Integration eines Hypertextsortenmoduls in einen heterogenen Knoten oder einer Auslagerung in einen separaten Knoten beobachtet werden können (vgl. etwa Abschnitt 11.5.8), so dass bezüglich der untersuchten Stichproben nur eine begrenzte Anzahl typischer Verknüpfungsmuster (z. B. hierarchische und lineare Sequenzierung) ermittelt und in die Ontologie aufgenommen werden konnte.

Dass OWL jedoch auch für eine detaillierte Repräsentation von Hyperlinks geeignet ist, zeigt die Möglichkeit der Spezifizierung von Subrelationen. Es können z. B. die von Kuhlen

(1991) vorgeschlagene Typologie von Hyperlinks (vgl. Abbildung 3.3, S. 108), die Klassenhierarchie von „link types“ nach Haas und Grams (1998b, vgl. Abbildung 4.6, S. 181) oder das Linkinventar von Mehler et al. (2004) in die Hypertextsortenontologie integriert werden.[40] Für diesen Zweck kann die Relation `verknuepft` als generische Relation aufgefasst werden, die durch Subrelationen angereichert wird, die unmittelbar den unterschiedlichen Typen dieser Typologien entsprechen. Auf diese Weise können maschinelle Analyseverfahren bei der Verarbeitung eines HTML-basierten Hypertextes für jeden Hyperlink den korrespondierenden Linktyp gemäß der Hypertextsortenontologie rekonstruieren und annotieren.

13.5.7 Das Spannungsverhältnis zwischen OWL-Ontologien und XML-Dokumentgrammatiken

In Kapitel 5 wurde die Problematik der Repräsentation einer Typologie von Textsorten mit Hilfe texttechnologischer Formate diskutiert: Mittels einer XML-basierten DTD kann lediglich das Textstrukturmuster einer einzelnen Textsorte modelliert werden (vgl. Abbildung 5.1, S. 262). Die Erfassung texttypologischer Beziehungen, die zwischen *mehreren* auf diese Weise repräsentierten Textsorten existieren, ist innerhalb der korrespondierenden Dokumentgrammatiken *nicht* möglich. In Abschnitt 5.2.1 wird weiter argumentiert, dass die für diesen Zweck notwendige Beschreibungsebene oberhalb der einzelnen DTDs anzusiedeln wäre, um auf diese Weise Relationen zwischen den Textsorten und ihren einzelnen Bestandteilen repräsentieren zu können. Eben diese Funktion übernimmt die Hypertextsortenontologie, in der OWL als Repräsentationsformat für Typologien von Hypertextsorten und ihre Konstituenten gemäß des in Kapitel 5 eingeführten Hypertextsortenmodells verwendet wird.

Da die OWL-basierte Hypertextsortenontologie durch XML-Dokumentgrammatiken zu flankieren ist, erscheint es notwendig, das Spannungsverhältnis zwischen diesen beiden Formalismen genauer zu betrachten. Die primäre Aufgabe von OWL ist die Repräsentation von Ontologien für wissensbasierte Systeme innerhalb der *Semantic Web*-Schichtenarchitektur. Hierfür stehen Elemente und Attribute zur Verfügung, die Klassen und unterschiedliche Typen von Eigenschaften realisieren, die entweder für eine oder zwischen mehreren Klassen bzw. Instanzen gelten. OWL schränkt das Inventar der vom Anwender verwendbaren Auszeichnungskonstrukte ein und bietet drei alternative Versionen der Sprache an, für die jeweils formale Semantiken existieren, so dass die Ontologien maschinellen Verarbeitungsprozessen wie z. B. Inferenzmaschinen zur Verfügung gestellt werden können (vgl. Abschnitt 13.2.3, sowie z. B. Rocha et al., 2004). Bei XML-basierten Formalismen zur Repräsentation von Dokumentgrammatiken handelt es sich hingegen um Metasprachen, d. h. diese Formalismen erlauben es den Benutzern, Auszeichnungssprachen anzufertigen, deren Syntax durch eine Dokumentgrammatik in Form einer XML-DTD oder XML Schema-Beschreibung repräsentiert wird. Eine DTD spezifiziert also unter anderem die Namen von Elementen, Datentypen von Attributen und die Inhaltsmodelle von Elementen, die sich wiederum aus geordneten oder ungeordneten Gruppen weiterer Elemente zusammensetzen, die zusätzlich durch die

[40] Mehler et al. (2004) unterscheiden in ihrem maschinellen Analyseansatz, der in Kapitel 14 diskutiert wird, zwischen „kernel links“ (hierarchische Verknüpfungen), „up“ und „down links“ (hierarchische Verknüpfungen, die über mindestens zwei Ebenen der hierarchischen Strukturierung reichen), „across links“ (Verknüpfungen zwischen zwei Teilbäumen eines hierarchisch sequenzierten Hypertextes) und externen „outside links“.

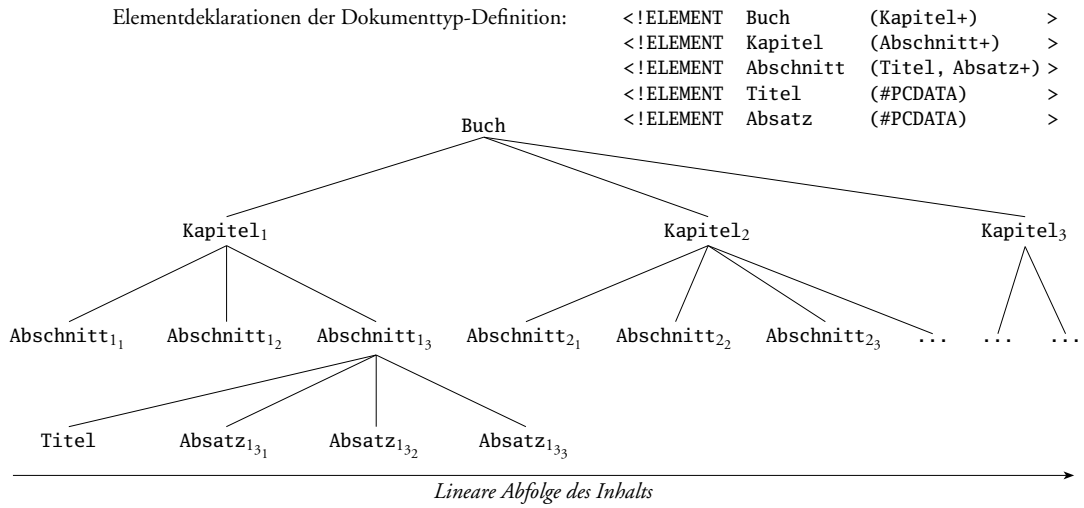


Abbildung 13.19: Spezifizierung der linearen Abfolge in einer Dokumenttyp-Definition

Okkurrenzindikatoren `?`, `+` und `*` als obligatorisch oder optional markiert werden können.[41] Innerhalb einer XML-Instanz, die zu ihrer DTD valide ist, existiert somit eine Baumstruktur, die von den in der Instanz enthaltenen Elementen aufgespannt wird und den in der DTD spezifizierten Inhaltsmodellen entspricht, die wiederum „the document's storage layout and logical structure" (Bray et al., 2004b) reflektieren. Abbildung 13.19 verdeutlicht die Spezifizierung der linearen Abfolge von Elementen in einer DTD, die – im Falle einer validen Instanz – der linearen Abfolge des Inhalts eines zugehörigen Dokuments entsprechen.

OWL basiert zwar ebenfalls auf XML, jedoch zielt sie nicht auf die Funktionalität der Validierung einer OWL-Instanz gegen eine DTD ab (die im Falle von OWL auch nicht existiert), sondern auf die Bereitstellung eines Vokabulars zur Repräsentation von Ontologien. Ein OWL-Dokument besteht im Regelfall aus zwei Teilen: Zunächst werden die Definitionen der Klassen und Relationen aufgeführt, anschließend folgen die Individuen. Listing 13.1 stellt ein gekürztes Beispiel bezüglich der Hypertextsortenontologie dar, das innerhalb der Domänenontologie ein Individuum der Klasse `Professor` enthält, der über eine Instanz der Klasse `Email` verfügt und eine persönliche Homepage pflegt. Die Sequenzierung bzw. Serialisierung der XML-Elemente erfolgt nach dem in RDF existierenden Prinzip der Verknüpfung einer Ressource (z. B. `Email`) mit einem Attribut (`isUsedBy`), das einen bestimmten Wert besitzt (`#PeterMustermann1`).[42] Somit bezieht sich die Reihenfolge der Elemente naturgemäß nicht auf einen korrespondierenden Text, sie ist vielmehr flexibel von den repräsentierten Klassen, Individuen und ihren Eigenschaften abhängig.

In der Hypertextsortenontologie werden unter anderem Hypertextsortenmodule als generische Bausteine einer Hypertextsorte oder einer Hypertextknotensorte durch Klassen reprä-

[41] Der XML-Standard (Bray et al., 2004b) spricht von „choice lists of content particles" (ungeordnete Inhaltsmodelle bzw. Teile von Inhaltsmodellen, deren einzelne Bestandteile durch `|` voneinander abgetrennt werden) und „sequence lists of content particles" (die durch `,` markierte Reihenfolge der Bestandteile ist signifikant).

[42] Der RDF-Standard spricht auch von *subject*, *predicate* und *object*, die wiederum ein RDF *statement* bilden.

```
<PersoenlicheHomepageEinesWissenschaftlers rdf:ID="PersoenlicheHomepageVonPeterMustermann">
  <wirdInstanziiertVon>
    <domain:Professor rdf:ID="PeterMustermann1">
      <domain:person.firstname>Peter</domain:person.firstname>
      <domain:hasOfficeComDevice.email>
        <domain:Email rdf:ID="PeterMustermann1EMail">
          <domain:isUsedBy rdf:resource="#PeterMustermann1"/>
          <domain:email.address>peter.mustermann@uni-stadt.de</domain:email.address>
        </domain:Email>
      </domain:hasOfficeComDevice.email>
      <domain:person.gender>m</domain:person.gender>
      <domain:person.cv>Lebenslauf von Peter Mustermann [...]</domain:person.cv>
      <domain:person.lastname>Mustermann</domain:person.lastname>
    </domain:Professor>
  </wirdInstanziiertVon>
</PersoenlicheHomepageEinesWissenschaftlers>
```

Listing 13.1: Individuen innerhalb einer OWL-Ontologie (gekürzt)

sentiert. Einerseits sind, wie Teil III gezeigt hat, obligatorische von optionalen Hypertextsortenmodulen zu unterscheiden: Wenn ein Hypertextsortenmodul in der Mehrzahl der untersuchten Dokumente beobachtet werden kann, handelt es sich um einen obligatorischen, konventionalisierten Bestandteil einer Hypertext(knoten)sorte. Optionale Hypertextsortenmodule hingegen bilden ihre Peripherie. Bezüglich der Realisierung von Hypertextsortenmodulen existieren auch präferierte Sequenzen, nach denen die Produzenten korrespondierender Hypertextexemplare die Hypertextsortenmodule anordnen: Innerhalb der Hypertextknotensorte *Einstiegsseite eines universitären Webauftritts* wird typischerweise zunächst das Hypertextsortenmodul `Identifikation` ausgeprägt, woraufhin die `PrimaereNavigationshilfe` und `Kontaktinformationen` folgen und abschließend wird die `Fusszeile` realisiert (vgl. Abbildung 4.8, S. 209, und Abbildung 11.3, S. 488). Darüber hinaus existieren auch innerhalb komplexer Hypertextsortenmodule bevorzugte Reihenfolgen der beteiligten atomaren Hypertextsortenmodule. Da OWL nicht der Modellierung abstrakter Textstrukturmuster, sondern der Wissensrepräsentation dient, steht kein Mechanismus zur Verfügung, um derartige Sequenzen von Hypertextsortenmodulen spezifizieren zu können.

Mit Hilfe inkrementell benannter Relationen kann diese Funktionalität nachgebildet werden: Abbildung 13.20 zeigt mehrere Hypertextsortenmodule, die an Instanzen der Hypertextknotensorte *Einstiegsseite der persönlichen Homepage eines Wissenschaftlers* beteiligt sind. Über die Reihenfolge der nach dem Muster `sequenziert_EPHEW_`$_{1...n}$ benannten Relationen, die als Subrelationen von `umfasstHTSM_EPHEW` definiert wurden, kann die typische Realisierungssequenz modelliert werden. Dieses Beispiel zeigt die offensichtlichen Grenzen des Einsatzes von OWL als Formalismus zur Repräsentation von Textstrukturmustern auf, für die die Texttechnologie etablierte Verfahren zur Verfügung stellt (z. B. XML-DTDs und XML Schema-Beschreibungen). Gleichwohl wäre es möglich, eine weiterführende Emulation von Dokumentgrammatiken in OWL vorzunehmen. So könnte die in den Inhaltsmodellen von Elementdeklarationen vorgenommene Hierarchisierung informationeller Komponenten (ein `Buch` umfasst etwa mehrere `Kapitel`-Elemente) durch eine Explizierung der Zusammenhänge dieser Informationseinheiten durch eine Relation wie z. B. `bestehtAus` modelliert werden und die Sequenzierung könnte nach dem oben dargestellten Ansatz erfolgen.

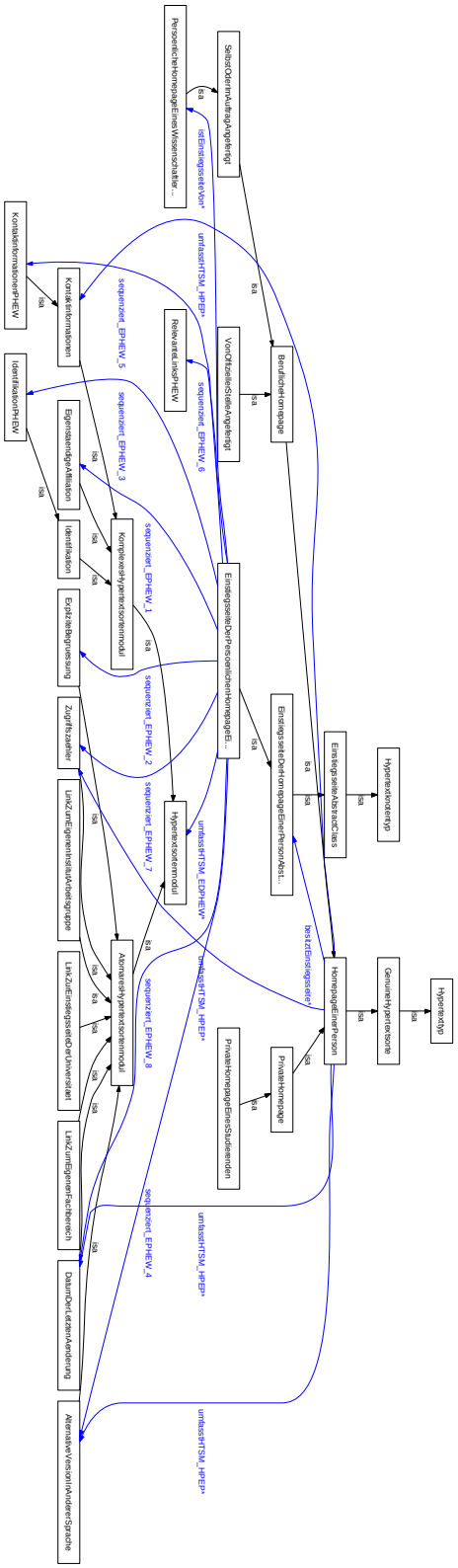


Abbildung 13.20: Lineare Sequenzierung von Hypertextsortenmodulen durch Relationen

Erdmann (2001, S. 136 ff.) verdeutlicht die Problematik, dass die Schachtelung von Elementen innerhalb des Inhaltsmodells einer Elementdeklaration auf *unterschiedlichen* Prinzipien basieren kann, anhand verschiedener Beispiele. Hierzu zählt z. B. die Subklassifizierung (`ComputerProduct` → `ComputerHardware, ComputerSoftware`), die Konzept-Instanz-Relation (`ComputerSoftware` → `Product+`), die *part-of* -Relation (`Buch` → `Kapitel+`), die Aggregation (`DateTime` → `Date, Time`) und die Listenzugehörigkeit (`SupplierList` → `Supplier+`).[43] Da XML keine explizite Semantik besitzt, kann der Verschachtelung von informationellen Einheiten durch Inhaltsmodelle keine formale Bedeutung zugeschrieben werden, so dass sich – bei einer abstrakten Betrachtung der Hypertextsortenontologie – die Frage stellt, nach welchen Prinzipien die Klassen und Relationen ausgewählt werden sollen, die das Gerüst der generischen Dokumentgrammatik einer Hypertextsorte bilden.

13.5.8 Dokumentgrammatische Informationen in der Ontologie

Die Ontologie dient der abstrakten Repräsentation einer Typologie von Hypertextsorten und ihrer Konstituenten. Hypertextsortenmodule als elementare Bausteine können als Klassen definiert und in unterschiedlichen Kontexten (d. h. unterschiedlichen Hypertextsorten) eingesetzt und gegebenenfalls durch die in OWL existenten Vererbungs- und Spezialisierungsmechanismen an ihre individuellen Ausprägungen in einer Hypertextsorte angepasst werden.

Aus Sicht einer texttechnologischen Anwendung sollte die Hypertextsortenontologie als eine Sammlung von Dokumentgrammatiken fungieren, so dass z. B. zunächst die Hypertextsorte eines gegebenen Hypertextes (eine Gruppe zusammengehöriger HTML-Dokumente) identifiziert werden kann, um daraufhin die Instanzen von Hypertextknotensorten und Hypertextsortenmodulen zu bestimmen und auf die korrespondierende Dokumentgrammatik abzubilden, so dass letztlich eine maschinelle Transformation unstrukturierter HTML-Dokumente in strukturierte XML-Dokumente stattfindet. Hierbei fungieren die ermittelten Hypertextsortenmodule, Hypertextknotensorten und die Hypertextsorte selbst als Individuen, die die in der Ontologie definierten Klassen instanziieren.

Der Bedarf zur Anreicherung der Hypertextsortenontologie mit Dokumentgrammatiken bezieht sich, wie Abschnitt 13.5.7 bereits angedeutet hat, auf drei unterschiedliche Schichten, die die Ebenen der Mikro-, Makro- und Superstruktur betreffen: Erstens umfasst OWL keine Möglichkeiten zur Spezifizierung der Sequenzierung einzelner Individuen innerhalb der Makro-, d. h. der Knotenebene, so dass für diesen Zweck die typische Funktionalität von Dokumentgrammatiken herangezogen werden sollte. Zweitens stellen atomare Hypertextsortenmodule zwar die elementaren Bausteine des Hypertextsortenmodells dar (vgl. Kapitel 5), sie besitzen jedoch in vielen Fällen auf der Mikroebene interne Textstrukturmuster, deren Repräsentation durch DTDs oder XML Schema-Beschreibungen durchgeführt werden sollte, sofern die analysierten Stichproben der Webangebote einer bestimmten Untersuchungsdomäne entsprechende Generalisierungen nahe legen.[44] Drittens bezieht sich die Ebene der Superstruktur auf die Sequenzierung der Knoten eines Hypertextes.

[43] Im Rahmen einer Diskussion der Semantik von XML meint Erdmann (2001, S. 138): „Allein für das Element-Nesting konnten in wenigen veröffentlichten DTDs ein halbes Dutzend verschiedener Bedeutungen identifiziert werden. Die automatische Identifikation der Semantik aus der XML-Struktur ist nahezu aussichtslos."

[44] Dieser Aspekt wird besonders bei denjenigen Hypertextsortenmodulen deutlich, die den primären oder sekundären Typ Textstrukturmuster ausprägen (z. B. *Publikationsliste* und *Lebenslauf*, vgl. Tabelle 10.3, S. 457).

Dokumentgrammatiken auf den Ebenen der Super- und Makrostruktur

Bezüglich der Super- und Makrostruktur spezifiziert die Ontologie Hypertextknotensorten und Hypertextsortenmodule, die ihrerseits als Hypertextknotensorten fungieren können. Eine innerhalb des eingangs skizzierten Anwendungsszenarios maschinell erzeugte XML-Dokumentinstanz kann zwar prinzipiell mit den korrespondierenden Klassenetiketten annotiert werden (vgl. Listing 13.1), jedoch besteht hierdurch noch nicht die Möglichkeit, die erzeugte Instanz gegen eine Dokumentgrammatik der Hypertextsorte zu validieren.

Diesbezüglich ist das von Erdmann (2001) präsentierte Verfahren zur (naturgemäß verlustbehafteten) Übersetzung einer Ontologie in eine Dokumentgrammatik von entscheidender Relevanz: Die Basis einer Ontologie sind die Bezeichnungen von Klassen und ihren Eigenschaften, die in ihrer Gesamtheit das Vokabular einer bestimmten Domäne spezifizieren, und eben dieses Vokabular sollte nun in XML-Dokumenten einsetzbar sein, so dass die korrespondierenden Elemente, Attribute und Inhalte als Instanziierungen der Ontologie formal interpretiert werden können. Das Werkzeug DTDMaker ist in der Lage, „aus einer Ontologie eine kanonische XML-Struktur zu generieren und in Form einer DTD zu repräsentieren." (Erdmann, 2001, S. 175).[45] Für diesen Zweck berücksichtigt das Tool zur Erzeugung von Element- und Attributdeklarationen die in der Ontologie definierten Klassen, ihre Eigenschaften und die Beziehungen zwischen Klassen. Für jede Klasse (z. B. `Person`) wird ein Element deklariert, dessen Inhaltsmodell aus Elementen besteht, die den Eigenschaften dieser Klasse entsprechen (z. B. `name` und `knowHow`).[46] Alternativ können Eigenschaften auch in Attributdeklarationen übersetzt werden. Relationen, die in der Ontologie zwischen zwei Klassen gelten, werden innerhalb der DTD mittels unterschiedlicher Linkingmechanismen realisiert (unter anderem durch Attribute vom Typ ID/IDREF). Erdmann (2001, S. 176) sieht zwar die Spezifizierung syntaktischer Restriktionen, z. B. „die Reihenfolge von serialisierten Objekten oder Attributen", als einen Vorteil von DTDs an, der DTDMaker erstellt jedoch ein Inhaltsmodell des `CONTAINER` genannten Wurzelelements der erzeugten DTD, das „aus der *beliebigen Kombination* von Konzept-Elementen (*in beliebiger Reihenfolge und beliebiger Kardinalität*)" besteht (Erdmann, 2001, S. 182; Hervorhebungen hinzugefügt, G. R.). Eine für den Zweck der automatischen Erzeugung einer DTD von Hypertextsorten notwendige Erweiterung dieses Werkzeugs beträfe somit die Spezifizierung präziserer und „härterer" (im Sinne von Yates und Sumner, 1997, vgl. Abschnitt 4.2.2) Inhaltsmodelle, die z. B. durch inkrementell benannte Relationen realisiert werden könnte (vgl. Abschnitt 13.5.7). Hierzu müsste ein Werkzeug wie der DTDMaker mit einem Verfahren ausgestattet werden, die durch die Suffixe der Relationen ausgedrückte Sequenz in korrespondierende Inhaltsmodelle zu überführen. Darü-

[45] Wie bereits in Fußnote 43 angesprochen wurde, beschäftigt sich Erdmann (2001, S. 81) mit der Semantik von XML und fasst Ontologien als „semantisches Schema zur expliziten Repräsentation der Bedeutung von XML-Elementen und -Attributen [auf]. Sie ergänzen damit die syntaktischen bzw. strukturellen Schemata, wie sie durch DTDs oder XML-Schema-Definitionen festgelegt werden." Erdmann verwendet keine XML-basierte Sprache zur Repräsentation von Ontologien, sondern den Formalismus *Frame-Logic*. Der DTDMaker wird von einem Parser flankiert, der XML-Instanzen, die zu einer von DTDMaker erzeugten DTD valide sind, in ein *Frame-Logic*-Programm überführen kann.

[46] Subklasseninformationen werden nur implizit in die DTD übernommen, da sich alle Elemente, die Klassen entsprechen, auf der gleichen Strukturierungsebene befinden. Innerhalb der DTD „erbt" eine Subklasse die Eigenschaften ihrer Superklasse, indem diese – neben den den distinktiven Eigenschaften der Subklasse – in das Inhaltsmodell der Subklasse eingetragen werden.

ber hinaus müssten obligatorische und optionale Hypertextsortenmodule als obligatorische oder optionale Bestandteile der diese Bausteine einbettenden Inhaltsmodelle realisiert werden. Sequenzierungen bezüglich der Hypertextstrukturierung könnten über entsprechende Verlinkungsmechanismen innerhalb von XML ausgedrückt werden (vgl. auch Mehler et al., 2004), deren Repräsentation in der Hypertextsortenontologie über ein Inventar korrespondierender Relationen stattfindet (vgl. auch Abschnitt 13.5.6). Die Hypertextsortenontologie ist mit Informationen in Form proprietärer Relationen anzureichern, die der Kompilierung einer Dokumentgrammatik mit gegebenenfalls signifikanten Abfolgen der einzelnen Elemente durch ein Werkzeug wie DTDMaker dienen können.

Dokumentgrammatiken auf der Ebene der Mikrostruktur

Auf der Ebene der Mikrostruktur existiert, wie bereits dargestellt wurde, der Bedarf, atomare Hypertextsortenmodule, die hinsichtlich des Typs Textstrukturmuster markiert sind, durch eine DTD bzw. ein DTD-Fragment anzureichern, das ihre jeweilige Binnenstruktur spezifiziert. Hierbei kann es sich um manuell erstellte oder maschinell induzierte Dokumentgrammatiken handeln (vgl. Kapitel 14). Die syntaktischen Möglichkeiten von OWL stellen jedoch keinen Mechanismus zum Import einer DTD oder einer XML Schema-Beschreibung als weiterführende, dokumentorientierte Spezifikation einer Klasse bereit, so dass ebenfalls eine proprietäre, den OWL-Standard erweiternde Lösung zu implementieren ist. Für diesen Zweck bietet sich das Element `rdfs:comment` an, in dem für jede Klasse beliebige Kommentare – und somit z. B. auch Referenzierungen von DTDs, DTD-Fragmenten oder XML Schema-Beschreibungen – abgelegt werden können.

Das Anwendungsszenario von Erdmann (2001) bezieht sich auf die Komplementierung einer Ontologie mit einer XML-basierten Dokumentgrammatik, die die Klassen und Eigenschaften der Ontologie durch einen maschinellen Überführungsprozess integriert. Auf diese Weise können die in der DTD spezifizierten Elemente und Attribute auf die Konzepte und Relationen der Ontologie abgebildet werden. Im Kontext der maschinellen Detektion von Hypertextsorten und der Identifikation von Hypertextsortenmodulen kommt der Ebene der Hypertextmodule eine zentrale Rolle zu: Hypertextmodule werden – in Bezug auf einen Einzelknoten – als Bausteine aufgefasst, die die Textoberfläche eines HTML-Dokuments konstituieren (vgl. Kapitel 5). Das Inventar von Hypertextmodulen ist begrenzt, z. B. *Fließtextabschnitt*, (isoliert positionierter) *Hyperlink*, *Textfragment* (d. h. ein isoliert positioniertes Fragment textuellen Inhalts, das nicht aus mehreren Sätzen besteht), *Liste von Textfragmenten* und *Überschrift* (ein Textfragment, das in einer größeren Schriftart dargestellt ist). Zwischen atomaren Hypertextsortenmodulen und Hypertextmodulen existieren verschiedene Korrelationen (vgl. Abschnitt 13.5.6). In Kapitel 14 wird ein Verfahren zur maschinellen Ermittlung von Hypertextmodulen vorgestellt, das die erste Stufe der automatischen Instanziierung von Hypertextsortenmodulen bildet. Für diesen Zweck kann eine konzeptuelle Verbindung zwischen Hypertextmodulen und den Konstituenten der Binnenstruktur atomarer Hypertextsortenmodule hergestellt werden. Ein HTML-Dokument wird zunächst in seine Hypertextmodule partitioniert, d. h. die einzelnen Bestandteile des Dokuments werden ermittelt und entsprechende Etiketten markieren individuelle Hypertextmodule. Abbildung 13.21 verdeutlicht die konzeptuelle Verbindung für den Einsatz der Hy-

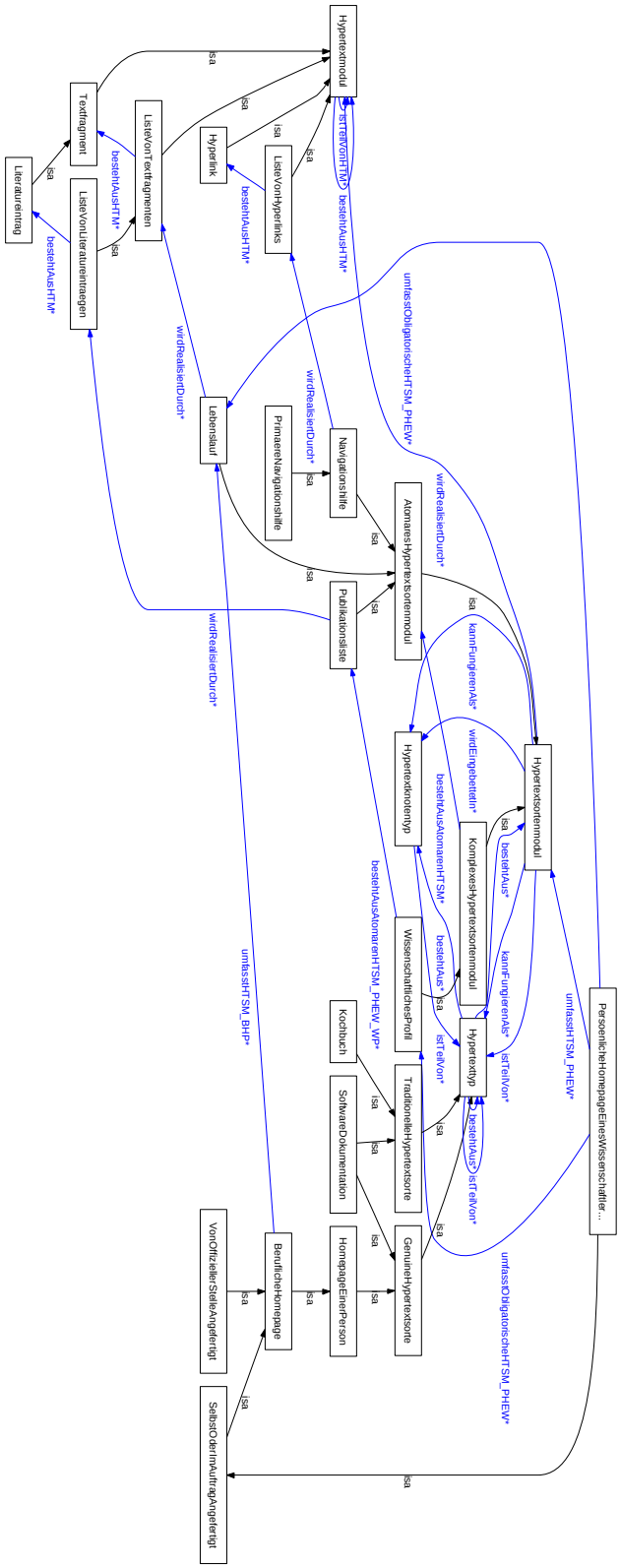


Abbildung 13.21: Das Verhältnis zwischen Hypertextmodulen auf der Textoberfläche und Hypertextsortenmodulen als atomare Bausteine von Hypertextsorten im Kontext texttechnologischer Anwendungen

pertextsortenontologie in der sich anschließenden Verarbeitungsphase: Die einzelnen Konstituenten atomarer Hypertextsortenmodule werden auf Hypertextmodule abgebildet, diese Abbildung findet jedoch nur implizit statt, indem eine spezifizierende Subklasse referenziert wird. Das Hypertextsortenmodul `Publikationsliste` wird realisiert von dem Hypertextmodul `ListeVonPublikationseintraegen` (Subklasse von `ListeVonTextfragmenten`), das wiederum aus mindestens zwei Instanzen des Hypertextmoduls `Literatureintrag` (Subklasse von `Textfragment`) besteht.[47] Auf der Erkennungsseite findet somit zunächst eine Detektion der abstrakten Hypertextmodule statt (z. B. Instanzen von `Textfragment`, die wiederum zu `ListeVonTextfragmenten` aggregiert werden können). Daraufhin können hypertextmodulspezifische Erkennungsregeln eingesetzt werden, um die jeweilige Subklasse (wie z. B. `Literatureintrag`) zu bestimmen. Die abstrakten Hypertextmodule können als das Vokabular einer Analyse-DTD aufgefasst werden, das zur Anreicherung eines zu verarbeitenden HTML-Dokuments verwendet wird. Da die Abfolge der XML-Elemente, die abstrakten Hypertextmodulen entsprechen, nicht spezifiziert werden kann, ist es möglich, diese DTD unmittelbar aus der Hypertextsortenontologie mit dem von Erdmann (2001) dargestellten Verfahren zu generieren. In sich anschließenden Verarbeitungsstufen können erkannte Instanzen der Klassen von Hypertextmodulen – nun jedoch unter Zuhilfenahme der Hypertextsortenontologie und korrespondierender Erkennungsregeln – sukzessive verfeinert werden, da die in der Ontologie herrschenden Relationen den Suchraum einschränken und die Hypertextmodule ebenfalls innerhalb der Hypertextsortenontologie repräsentiert werden, weshalb auch sie Inferenzprozessen zur Verfügung stehen. Diese Einschränkung bezieht sich einerseits auf die Menge der potenziell von einem abstrakten Hypertextmodul instanziierten Hypertextsortenmodule, andererseits können diese Informationen zur Bestimmung der Hypertextsorte eingesetzt werden, da die erkannten Hypertextsortenmodule im Rahmen dieses Anwendungszwecks den Suchraum auf diejenigen Hypertextsorten reduzieren, in denen das erkannte Hypertextsortenmodul verwendet werden kann.

Bezüglich des von Erdmann vorgestellten Verfahrens existiert auf einer zusätzlichen Ebene ein Bedarf zur Erweiterung des Übersetzungsprozesses: Für *eine* Ontologie erzeugt der DTDMaker exakt *eine* DTD.[48] Die Hypertextsortenontologie fungiert jedoch als eine Art *Sammlung modularisierter Dokumentgrammatiken*, da beispielsweise Beziehungen zwischen Hypertextsorten (z. B. betten Instanzen von Hypertextsorte x Instanzen von Hypertextsorte y ein) zu repräsentieren sind, die Vererbungsmechanismen von OWL zur Spezifizierung von Subklassen eingesetzt werden und in mehreren Hypertext(knoten)sorten verwendete Hypertextsortenmodule lediglich einmal definiert und mehrfach referenziert werden können. Aus diesem Grund ist eine Ontologie, die die Definitionen von n Hypertextsorten enthält, nicht auf eine, sondern auf mindestens n Dokumentgrammatiken abzubilden.[49] Abbildung 13.22 veranschaulicht diese Vorgehensweise: Ein Werkzeug, das auf der von Erdmann beschriebe-

[47] Hierdurch wird die strikte Trennung der modellierten Konzepte aufgehoben, so dass es zu einer Vermischung der Ebenen kommt. Diese kann jedoch nicht vermieden werden, da Korrelationen zwischen Hypertextsortenmodulen und den Hypertextmodulen existieren, mit denen sie typischerweise realisiert werden.

[48] Die Struktur und Verknüpfung der DTD kann jedoch durch Kommandozeilenparameter des Werkzeugs beeinflusst werden, um z. B. Eigenschaften von Klassen in Subelemente oder in Attribute zu überführen.

[49] Es müsste zusätzlich die Möglichkeit gegeben sein, diejenigen DTDs, die den Definitionen eingebetteter Hypertextsorten entsprechen, in die DTD einer übergeordneten Hypertextsorte zu integrieren.

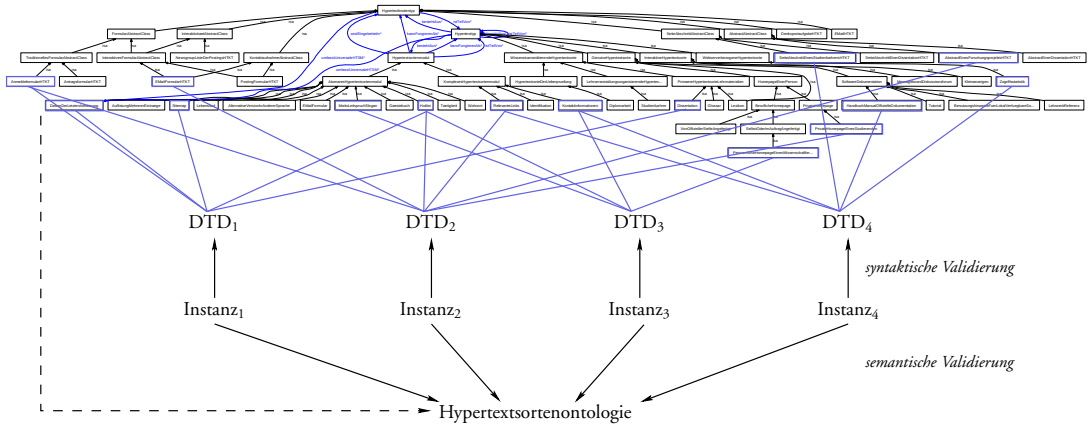


Abbildung 13.22: Generierung von Dokumentgrammatiken aus der Hypertextsortenontologie (schematisch)

nen Komponente basiert, generiert aus der Definition einer Hypertextsorte eine DTD. Ein maschinell mit dem in der DTD spezifizierten Vokabular von Elementbezeichnern annotiertes Dokument kann mit einem XML-Parser syntaktisch gegen diese DTD validiert werden. Bei Erfolg kann daraufhin geprüft werden, ob die XML-Instanz gültige Fakten in Bezug auf die Klassendefinitionen der Hypertextsortenontologie enthält – das maschinell annotierte XML-Dokument entspräche in diesem Fall dem Individuenteil der OWL-Ontologie.[50]

13.6 Zusammenfassung

Dieses Kapitel stellt – als Ergänzung des in Kapitel 5 eingeführten Hypertextsortenmodells – ein maschinenlesbares Repräsentationsformat für Hypertextsorten und Typologien von Hypertextsorten vor, das auf der *Web Ontology Language* (OWL) basiert. Typischerweise wird in Ontologien Domänenwissen in Bezug auf einen bestimmten Weltausschnitt spezifiziert, so dass z. B. Suchmaschinen mit Inferenzmechanismen ausgestattet werden können, um mit Hilfe der Wissensbasis Anfragen einschränken zu können. Mit der Repräsentation der in Teil III ermittelten Hypertextsorten und ihrer Konstituenten in OWL wird somit eine neuartige Wissensebene in das *Semantic Web* aufgenommen.

Die drei Ebenen des Hypertextsortenmodells werden als Subklassen bzw. Subkonzepte der von OWL vorgegebenen Klasse `owl:Thing` definiert, sie konstituieren prinzipiell unabhängige Teilontologien. Die im Hypertextsortenmodell vorgesehenen Beziehungen zwischen einer Hypertextsorte, Hypertextknotensorten, Hypertextsortenmodulen und eingebetteten Hypertextsorten werden über Relationen (Eigenschaften von Klassen) hergestellt, die bezüglich derjenigen Klassen, in denen diese Relationen Verwendung finden und auch hinsichtlich der

[50] Auch der Prozess der semantischen Verarbeitung wird von Erdmann (2001) durch einen Prototypen demonstriert (KBMaker), der auf einem in *Java* implementierten XML-Parser aufsetzt. Der KBMaker übersetzt eine XML-Instanz in eine *Frame-Logic* Wissensbasis (vgl. Fußnote 45).

Klassen, auf die sie sich beziehen, eingeschränkt werden können. Subklassen stellen spezifische Ausprägungen einer Superklasse dar. Durch den Mechanismus der Mehrfachvererbung können Typologisierungen – z. B. von Hypertextsorten – flexibel auf der Basis multipler Kriterien erfolgen. Die zwischen den drei Ebenen existenten Relationen können ebenfalls in Form einer Hierarchie angeordnet werden, so dass generische Relationen, die zwischen Konzepten wie `Hypertextknotensorte` und `Hypertextsortenmodul` gelten, z. B. in Bezug auf Teilhierarchien durch Subrelationen zusätzlich spezifiziert werden können. Innerhalb einer Klassendefinition können über Restriktionen, die sich auf die Anwendung von Relationen beziehen, weitere Spezifizierungen vorgenommen werden, die den distinktiven Merkmalen einer Hypertextsorte, Hypertextknotensorte oder eines Hypertextsortenmoduls entsprechen. Auf der Basis dieses Inventars wurden die in Teil III der Arbeit vorgestellten Ergebnisse nahezu vollständig in die Hypertextsortenontologie integriert.

In die Hypertextsortenontologie werden eine Domänenontologie und eine Ontologie wissenschaftlicher Themen und Fachgebiete importiert. Die erste Ontologie beschreibt den Aufbau einer generischen Hochschule und modelliert die Untersuchungsdomäne, während sich die zweite Ontologie auf eine Taxonomie wissenschaftlicher Themen bezieht und somit die Hypertextsortenontologie – und auch die Prozesse der Klassifikation eines Hypertextes in eine Hypertextsorte sowie die thematische Klassifikation – komplementiert. Teil III zeigt, dass sich die Struktur eines universitären Webauftritts maßgeblich an den Strukturen der instanziierenden Hochschule orientiert. Über die Verknüpfung der Ontologien durch entsprechende Relationen wird diese Verbindung expliziert.

Im Hinblick auf texttechnologische Anwendungen kann die Hypertextsortenontologie als eine formale Beschreibung multipler Dokumentgrammatiken aufgefasst werden. Da OWL jedoch für den Zweck der Modellierung von Domänenwissen und nicht für die Aufgabe der Repräsentation von Dokumentstrukturen ausgelegt ist, müssen Sequenzen – z. B. obligatorischer und optionaler Hypertextsortenmodule innerhalb einer spezifischen Hypertextknotensorte – durch Relationen nachgebildet werden. Diese Relationen könnten von einem Werkzeug eingesetzt werden, um eine OWL-Ontologie in eine Dokumentgrammatik zu überführen. Dabei müssten auch zusätzliche Dokumentgrammatiken berücksichtigt werden, die die Binnenstrukturierungen atomarer Hypertextsortenmodule darstellen. Produzenten von HTML-Dokumenten können ein überschaubares Inventar von Hypertextmodulen einsetzen, um auf der Textoberfläche Komponenten zu realisieren. Hypertextmodule werden ebenfalls in der Hypertextsortenontologie repräsentiert. Spezifizierungen dieser generischen Bausteine in Form von Subklassen können auf die möglichen Realisierungsformen bestimmter Hypertextsortenmodule abgebildet werden. Hierdurch wird eine Verbindung der abstrakten Ebene der Hypertextsortenmodule mit der Ebene der Textoberfläche hergestellt. Insbesondere diese Verbindung stellt den Anknüpfungspunkt der ersten Phase einer maschinellen Erkennung von Hypertextsorten und Hypertextsortenmodulen dar, die im folgenden Kapitel diskutiert wird. Der Ansatz basiert auf der sukzessiven Annotierung eines Dokuments mit Informationen, die sich auf die Ebene der Hypertextmodule beziehen. Da die Hypertextsortenontologie in einem texttechnologischen Standardformat repräsentiert wird, ist es möglich, weiterführende Informationen (z. B. Erkennungsregeln oder von einer Disambigierungskomponente benötigte statistische Daten) unmittelbar in der Ontologie zu hinterlegen. Über diese Annotierungen kann in einem zweiten Schritt auf die Ebene der Hypertextsortenmodule geschlos-

sen werden, die von den erkannten Hypertextmodulen instanziiert werden. Die Klassen und Relationen der Hypertextsortenontologie fungieren also als abstrakte Beschreibungen, die von Dokumenten manifestiert werden. Die in ihnen annotierten Informationen können in Bezug auf die Hypertextsortenontologie als Individuen aufgefasst werden.

13.7 Fazit – Zur Integration der Hypertextsortenontologie in computerlinguistische Anwendungen

Die *Web Ontology Language* stellt, wie die zahlreichen in diesem Kapitel präsentierten Beispiele zeigen, ein sehr flexibles Format für die Repräsentation von Hypertextsorten dar. Der Formalismus kann mit beliebigen weiteren Ressourcen angereichert werden (z. B. DTDs, DTD-Fragmenten oder mittels RDF kodierten Extraktionsregeln, vgl. Potok et al., 2002). Es stellt sich jedoch nun die Frage nach der praktischen Anwendung der Hypertextsortenontologie innerhalb eines computerlinguistischen und texttechnologischen Systems. Ein derartiges System kann sich beispielsweise auf die Prozesse der Erkennung der Hypertextsorte eines gegebenen HTML-basierten Hypertextes, die Identifizierung der instanziierten Hypertextsortenmodule und die Abbildung ihrer Inhalte auf die Hypertextsortenontologie sowie auf die Domänenontologie beziehen. Jede einzelne dieser Verarbeitungsphasen und der benötigten Komponenten und Ressourcen ist mit einer ungemeinen Komplexität verbunden, die letztendlich auf die Variabilität und funktionale Heterogenität der im *World Wide Web* angebotenen Dokumente zurückzuführen ist. Ein weiteres Problem betrifft die derzeit noch mangelnde Verfügbarkeit von Werkzeugen zur Verarbeitung OWL-basierter Ontologien. Mit dem *protégé OWL Plug-in* steht zwar eine ausgereifte Entwicklungsumgebung bereit, jedoch ist bislang keine der Inferenzmaschinen, die OWL-Ontologien verarbeiten können, in der Lage, über Instanzen zu schließen, so dass die Anwendung der Hypertextsortenontologie für die Aufgabe der Reduzierung des Suchraums (z. B. nach der Erkennung von Instanzen spezifischer Hypertextsortenmodule, um die potenziell übergeordneten Hypertextsorten zu ermitteln) derzeit noch verschiedenen Restriktionen unterliegt. Das nachfolgende Kapitel geht bezüglich der Konzeptionierung und Implementierungs eines computerlinguistischen und texttechnologischen Systems zur Erkennung von Hypertextsorten auf zwei wesentliche Aspekte ein: Zum einen wird eine Architektur vorgestellt, die die für diese Aufgabe benötigten Komponenten und Ressourcen identifiziert (vgl. Kapitel 7). Zum anderen wird für die notwendigen Prozesse der aktuelle Forschungsstand dargestellt, der ebenfalls die massive Komplexität eines derartigen Systems verdeutlicht.

14 Computerlinguistische Anwendungen von Hypertextsorten

14.1 Einleitung

Die maschinelle Identifizierung von Hypertextsorten kann in mehreren Anwendungsszenarien eingesetzt werden. Eine zentrale Stellung nimmt die Hypertextsortenerkennung im Rahmen einer Suchmaschine ein (vgl. Kapitel 1). Eine Recherche, die nur auf Stichwörtern und Phrasen beruht, kann unter Umständen tausende von Ergebnissen liefern. Wird es dem Anwender jedoch zusätzlich ermöglicht, Hypertextsorten zu spezifizieren, kann diese Information als Filter eingesetzt werden, um die Treffermenge auf die relevanten Hypertextsorten einzuschränken und in gebündelter Form als Ergebnisse zu präsentieren. In mehreren Arbeiten wird eine solche Anwendung als Desiderat genannt (vgl. z. B. Cronin et al., 1998, Haas und Grams, 2000, und Roussinov et al., 2001), doch liegen bislang nur wenige implementierte Prototypen vor, die zudem zahlreichen Restriktionen unterliegen, so dass mit der Bereitstellung einer derartigen Suchmaschine höchstens langfristig zu rechnen ist.

Dieses Kapitel geht auf diese Anwendung von Hypertextsorten in computerlinguistischen Systemen ein. Zunächst werden in Abschnitt 14.2 die vorliegenden Arbeiten zur maschinellen Identifizierung von Genres und Web-Genres thematisiert und vor dem Hintergrund des Hypertextsortenmodells (Kapitel 5) und der in Teil III präsentierten Ergebnisse einer kritischen Analyse unterzogen. Aus dieser Diskussion wird die Architektur einer Anwendung abgeleitet, die die Hypertextsortenontologie als Ressource einsetzt (Abschnitt 14.3). Die Architektur beinhaltet mehrere Komponenten, die im Anschluss dargestellt werden: Abschnitt 14.4 erläutert die initiale Konvertierung beliebiger HTML-Dokumente nach XHTML, um die entstehenden Dokumente XML-Werkzeugen zur Verfügung stellen zu können. Abschnitt 14.5 präsentiert einen Textparser für arbiträre HTML-Dokumente, der der Ermittlung von Hypertextmodulen dient. Abschnitt 14.6 geht auf diejenigen Prozesse ein, die im Rahmen des noch sehr jungen Forschungszweiges *Web Genre Identification* bislang nicht dis-

kutiert werden, für eine Realisierung der Hypertextsortenidentifizierung auf der Grundlage der zuvor eingeführten Architektur jedoch von essenzieller Bedeutung sind. Abschließend stellt Abschnitt 14.7 verschiedene informations- und sprachtechnologische Anwendungen vor, die durch die maschinelle Identifizierung von Hypertextsorten ermöglicht werden.

14.2 Maschinelle Erkennung von Genres und Web-Genres

Die maschinelle Identifizierung von Hypertextsorten stellt eine vielversprechende Anwendung zur Bewältigung der oft zitierten Informationsflut dar.[1] Anwender einer Suchmaschine besäßen die Möglichkeit, zusätzlich zu den in ein Suchformular eingetragenen Wörtern und Phrasen die Hypertextsorten der zu findenden Dokumente zu spezifizieren.

Abschnitt 14.2.1 geht zunächst auf die thematische Klassifikation bzw. Kategorisierung von Texten und Hypertexten ein. Daraufhin diskutiert Abschnitt 14.2.2 den Prozess der maschinellen Zuordnung von Texten in ihre jeweiligen Genres (vgl. Fußnote 1, S. 155). Abschnitt 14.2.3 stellt die bislang vorliegenden Ansätze zur automatischen Erkennung von Web-Genres vor, die anschließend einer kritischen Analyse unterzogen werden (Abschnitt 14.2.4).

14.2.1 Klassifikation und Kategorisierung von Texten und Hypertexten

Die maschinelle Bestimmung der thematischen Kategorie eines Textes stellt eine Anwendung dar, die mittlerweile zu den sprachtechnologischen Standardverfahren gehört (Jackson und Moulinier, 2002). Zu den Anwendungsszenarien zählt z. B. die Erkennung (und anschließende Tilgung) von *spam*-Mail, die Kategorisierung von Nachrichtenartikeln in Ressorts wie *Politik*, *Sport*, *Wirtschaft*, *Unterhaltung* und *Kultur* sowie die Klassifikation von HTML-Dokumenten in hierarchisch angeordnete Kategoriensysteme, wie sie z. B. von WWW-Katalogen wie *Yahoo!* oder dem *Open Directory Project* gepflegt und permanent erweitert werden.

Grundlegende Verfahren

Es wird zwischen der Textklassifikation und der Textkategorisierung unterschieden (vgl. Mehler, 2004). Das Ziel eines Textklassifikationsverfahrens betrifft die Partitionierung eines Textkorpus, so dass einzelne Klassen (*cluster*) einen möglichst homogenen Charakter besitzen, wobei der Unterschied zwischen Klasssen möglichst groß sein soll. Es liegen *keine* initialen Klassen vor, stattdessen werden diese von dem Algorithmus erstellt, sie können jedoch nur mit zusätzlichem Aufwand auch maschinell etikettiert werden (Dörre et al., 2001). Die Textkategorisierung basiert hingegen auf einem in Form einer Liste oder einer Hierarchie angeordneten, vorgegebenen Kategorienschema (vgl. Brückner, 2001, und Breaux und Reed, 2005). Derartige Prozesse werden oftmals mit maschinellen Lernverfahren durchgeführt (Mitchell, 1997). Für die Textklassifikation kommen Methoden des unüberwachten Lernens zum Einsatz, wohingegen die Textkategorisierung auf überwachten Lernverfahren beruht: Repräsentative Beispieldokumente werden in einem ersten Schritt von einem Experten in das Kategorienschema eingeordnet, woraufhin ein Algorithmus diejenigen Merkmalsausprägungen

[1] Verschiedene in diesem Abschnitt thematisierte Aspekte wurden in Rehm (2004b,c,d) publiziert.

bestimmt, die für die einzelnen Kategorien distinktiv sind. Nach dieser Trainingsphase können unbekannte Texte verarbeitet werden, indem ihre individuellen Merkmalsausprägungen ermittelt und von den maschinell erzeugten Regelsystemen oder auf der Basis der erhobenen statistischen Häufigkeiten kategorisiert werden. Sowohl die Klassifikation als auch die Kategorisierung basieren auf der Analyse von Textmerkmalen, wobei meist ausschließlich die enthaltenen Wörter betrachtet werden – ein Text wird als *bag of words* betrachtet.[2] Diese werden tokenisiert, auf ihre Stammformen reduziert und von Stoppwörtern bereinigt.

Der Vergleich von Texten stellt einen wichtigen Teilprozess dar (Dörre et al., 2001). Hierzu werden meist statistische IR-Verfahren eingesetzt: Zunächst wird der Merkmalsraum einer Kollektion bestimmt und zur Komplexitätsminimierung reduziert, indem hoch- und niedrigfrequente Merkmale entfernt bzw. unterschiedlich gewichtet werden (z. B. mittels einer *tf idf* -Formel, vgl. Baeza-Yates und Riberiro-Neto, 1999). Ein Text wird als Merkmalsvektor repräsentiert, dem für jedes Merkmal (jede Dimension des Merkmalraumes) seine Frequenz in dem Text als Wert zugewiesen wird. Durch die Anwendung eines Distanzmaßes kann die Ähnlichkeit zweier Vektoren und somit die Kategorienzugehörigkeit des unbekannten Textes bestimmt werden, da die Kategorien ebenfalls über solche Vektoren repräsentiert werden.

Es existieren zahlreiche Kategorisierungsalgorithmen mit individuellen Vor- und Nachteilen: kNN (k *Nearest Neighbour*) berechnet in der Trainingsphase die Merkmalsvektoren der Beispiele. Ein unbekanntes Dokument wird zunächst ebenfalls vektorisiert, um es daraufhin mit den Beispielvektoren zu vergleichen. Das Resultat sind die k nächsten Nachbarn und somit die Kategorie(n) des unbekannten Dokuments. Der probabilistische Naive Bayes Ansatz basiert auf den bedingten Wahrscheinlichkeiten von Merkmalen. Diese werden beim Training aufgebaut, Zusammenhänge zwischen einzelnen Merkmalen können jedoch nicht berücksichtigt werden. Das SVM-Verfahren (*Support Vector Machine*) arbeitet mit Vektoren, die – für jeweils eine Kategorie – das Trainingsergebnis sind, und der Ebene, die positive von negativen Beispielen im Vektorraum trennt, am nächsten liegen. Der Merkmalsvektor eines unbekannten Textes wird zur Klassifikation mit diesen Support-Vektoren verglichen. Evaluationen zeigen, dass SVM-basierte Systeme anderen Verfahren bezüglich der Klassifikationsgüte überlegen sind (vgl. Dumais et al., 1998, Yang und Liu, 1999, Goller et al., 2000). Die Präzision kann über eine morphologische Vorverarbeitung der Eingabe verbessert werden (Yang und Chute, 1994). Für jedes Verfahren ist die Auswahl und Reduktion der Merkmale wichtig (Yang und Pedersen, 1997), um die Kategorisierung nicht unnötig komplex zu gestalten und *overfitting* zu vermeiden: Dieses liegt vor, wenn ein Algorithmus derart spezifische Klassenprofile erzeugt hat, dass nur die Trainingsbeispiele korrekt klassifiziert werden.

Information Retrieval im World Wide Web

Im *Information Retrieval*-Kontext können Suchanfragen wie Texte behandelt werden: Die Anfrage wird in einen Vektor überführt und mit allen Merkmalsvektoren einer Kollektion verglichen, um eine Liste der Dokumente mit den ähnlichsten Vektoren zu erstellen und als

[2] Siehe hierzu auch Adamzik (2004, S. 118): „Für den normalen Sprachbenutzer dürfte das »Was?« eines Textes grundsätzlich im Vordergrund [...] stehen; außerdem sind Thema und Inhalt zweifellos das, was man intuitiv am leichtesten erfassen kann. Das liegt daran, dass man sie großenteils unmittelbar am Sprachmaterial ablesen kann; denn die Lexeme geben sozusagen schon die Kategorien vor [...]."

Ergebnis zu präsentieren. Es ist die Aufgabe von Suchmaschinen, einen möglichst umfangreichen Teil des öffentlich zugänglichen WWW zu indexieren, so dass zu Suchanfragen relevante Dokumente geliefert werden können, um spezifische Informationsbedarfe zu decken. Suchmaschinen verwenden ein umfangreiches Inventar heterogener Methoden: Ein Volltextindex umfasst eine alphabetisch sortierte Wortliste, die Verweise auf die HTML-Dokumente enthält, in denen ein Indexterm gefunden wurde (vgl. Dörre et al., 2001). Zusätzlich werden Informationen über die Hyperlinkstruktur einbezogen: Auf diese Weise werden diejenigen Webseiten mit sehr vielen von externen Websites eingehenden Hyperlinks mit einem höheren Gewicht versehen, da es sich offenbar um populäre Dokumente handelt. In der Regel wird auch der Hyperlinkanzeiger einer eingehenden Verknüpfung in den Vektor der Indexterme des Zielknotens aufgenommen (vgl. z. B. Brin und Page, 1998, Attardi et al., 1999, Fürnkranz, 1999, Amitay, 2000b, Glover et al., 2002, und Chen und Chue, 2005).[3] Auf diese Weise können binäre Dateitypen wie z. B. Fotos, Grafiken und Videos indexiert werden, in denen kein analysierbarer Text enthalten ist.

Die Suchmaschine *Google* verwendet nach eigenen Angaben mehr als 100 unterschiedliche Metriken, die in die Bestimmung der Relevanz eines Dokuments für eine Suchanfrage einfließen. Informationen über die beteiligten Algorithmen, Verfahren und Gewichtungen werden von Suchmaschinenbetreibern nur in sehr rudimentärer Form publiziert, da sie die Basis ihrer Geschäftsmodelle darstellen. Eine detaillierte Betrachtung der hochgradig komplexen Methoden, die an der Erstellung, Pflege und Verarbeitung von Indizes beteiligt sind, die mehrere Milliarden HTML-Dokumente beinhalten, ist an dieser Stelle nicht möglich (vgl. z. B. Gudivada et al., 1997, Baeza-Yates und Riberiro-Neto, 1999, Belew, 2000, Hirai et al., 2000, Arasu et al., 2001, Hu et al., 2001, sowie die in Kapitel 7 diskutierten verwandten Aspekte). Es kann festgehalten werden, dass für jedes HTML-Dokument – neben dem enthaltenen Text – eine Vielzahl von Merkmalen erhoben wird, die für die Berechnung seiner Relevanz für eine Suchanfrage eingesetzt werden (*relevance ranking*). Hierzu zählen z. B. spezifische HTML-Auszeichnungen, die zur Gewichtung textueller Inhalte eingesetzt werden (vgl. Kim und Zhang, 2003, und Kwon und Lee, 2003). Sobald ein robustes Verfahren zur Bestimmung der Hypertextsorte eines Hypertextes vorliegt, kann auch diese Information in die Berechnung der Relevanz eines Dokuments für eine Anfrage einbezogen werden.

Klassifikation und Kategorisierung von Dokumenten im World Wide Web

Für das WWW werden Prozesse der Klassifikation und der Kategorisierung in Bezug auf zahlreiche Fragestellungen realisiert, weshalb an dieser Stelle nur ein für das vorliegende Kapitel relevanter Ausschnitt dargestellt werden kann. Derartige Ansätze werden meist unter dem Schlagwort *Web Mining* subsumiert, das sich in die Gebiete *Web Content Mining* (Analyse der Inhalte von HTML-Dokumenten), *Web Usage Mining* (Entdeckung von Mustern im Benutzerverhalten, vgl. Cooley et al., 1997) und *Web Structure Mining* (Analyse von Hyper-

[3] Auf der Basis der Hyperlinkstruktierung können „authorities“ (Knoten mit vielen Informationen zu einem Thema und zahlreichen eingehenden Hyperlinks) von „hubs“ (verweisen zu einer Vielzahl von Dokumenten zu einem Thema) unterschieden werden (Kleinberg, 1998, Chakrabarti et al., 1999). Da solche Differenzierungen in nahezu allen Systemen verwendet werden, bezeichnet Géry (2002a) dieses Verfahren – in Anlehnung an den traditionellen „bag of words“-IR-Ansatz – als „bag of links-approach“.

linkstrukturen, vgl. Chakrabarti et al., 1998) aufteilt (vgl. Kosala und Blockeel, 2000).[4] Von diesen Teilgebieten sind im Kontext der Entwicklung von Verfahren zur Identifizierung von Hypertextsorten das *Web Content Mining* und das *Web Structure Mining* von Bedeutung.[5]

Asirvatham und Ravi (2001) gehen davon aus, dass *jedem* HTML-Dokument einer der Typen „information page", „research page" und „personal home page" zugeordnet werden kann. Die Kategorisierung erfolgt durch die Ermittlung typenspezifischer Merkmale. Hierzu zählen „textual information" und „image information", z. B. der Umfang des Textinhalts sowie die Verhältnisse zwischen der Anzahl von Zeichen in Hyperlinkanzeigern und sonstigen Zeichen. Ebenso wird die Anzahl der Farben eingebetteter Bilder analysiert, die auf ihren Inhalt hindeuten: Bilder auf „information pages" umfassen den Verfassern zufolge mehr Farben als diejenigen auf „personal home pages". Bilder in „research pages" stellen wiederum eher Diagramme und Strichzeichnungen mit nur wenigen Farben dar. Ein Test mit etwa 4 000 Dokumenten zeigt, dass ca. 88% korrekt kategorisiert werden.

Pierre (2001) stellt ein Verfahren vor, das keine Analyse einzelner HTML-Dokumente durchführt, sondern vollständige Websites von Firmen einer inhaltlichen Kategorisierung unterzieht.[6] Die Klassenhierarchie beinhaltet 21 industrielle Sparten (z. B. „Construction", „Finance and Insurance" und „Health Care and Social Assistance"), die Kategorisierung basiert insbesondere auf den Inhalten von `meta`-Elementen. Falls in der Einstiegsseite keine Metadaten enthalten sind, werden weitere Dokumente einbezogen, wobei diejenigen Hyperlinks präferiert werden, deren Anzeiger Schlüsselwörter enthalten („product", „services", „about", „info", „press" etc.).[7] Auf diese Weise werden Dokumente einbezogen, die für eine Klassifikation der industriellen Sparte geeignet sind. Falls sie ebenfalls keine `meta`-Elemente beinhalten, werden *alle* besuchten HTML-Dokumente zu *einem* synthetischen Dokument konkateniert und zum Aufbau des Merkmalsraumes eingesetzt. Eine Evaluation zeigt, dass die Verwendung der Inhalte von `meta`-Elementen eine höhere Präzision erreicht als der Textinhalt. Kwon und Lee (2003) präsentieren ein ähnliches Verfahren: Zur Reduktion der Komplexität werden zunächst alle Dokumente einer Website mit einem Gewicht versehen, so dass lediglich eine Teilmenge der Dokumente berücksichtigt wird. Jedes Dokument wird von einem kNN-Algorithmus thematisch kategorisiert, woraufhin die ermittelten Daten auf die gesamte Website ausgedehnt werden. Eine Evaluation belegt, dass dieser Ansatz präziser ist als die alleinige Kategorisierung der Einstiegsseite einer Website.

Die meisten überwachten Lernverfahren ignorieren hierarchische Kategoriensysteme und verwenden flache Listen. Dumais und Chen (2000) beschäftigen sich mit der hierarchischen

[4] Kosala und Blockeel (2000) umschreiben *Web Mining* als „converging research area from several research communities, such as database, IR, and AI research communities especially from machine learning and NLP."

[5] Es werden hierbei nur Ansätze diskutiert, die überwachte Lernverfahren einsetzen. Modha und Spangler (2000) stellen eine Anwendung von Clustering-Algorithmen auf Sammlungen von HTML-Dokumenten dar.

[6] Amitay et al. (2003) verfolgen ein ähnliches Ziel, verwenden jedoch ausschließlich die Hyperlinkstrukturierung einer Website und eingehende Hyperlinks. Die Daten wurden von einem Suchmaschinenanbieter bezogen und stammen aus einem *Crawl* über ca. 500 Millionen Dokumente. Amitay et al. gehen von acht „functionality categories" aus: „corporate sites", „content & media sites", „search engines", „Web hierarchies & directories", „portals", „E-stores", „virtual hosting services" und „universities". Experimente mit zwei Algorithmen zeigen, dass die Kategorisierung mit einer Präzision von 54,5% bzw. 59% durchgeführt wird.

[7] Der Umstand, dass hierfür Schlüsselwörter eingesetzt werden können, basiert – in Bezug auf das Hypertextsortenmodell (Kapitel 5) – auf konventionalisierten Linkanzeigern in dem Hypertextsortenmodul *primäre Navigationshilfe* der Hypertextknotensorte *Einstiegsseite* des zugehörigen Hypertexttyps.

Kategorisierung, die den Vorteil besitzt, dass eine Differenzierung zunächst auf der Ebene der obersten Kategorien getroffen werden kann, woraufhin jeweils spezifischere Verfahren für die Subkategorien eingesetzt werden können. Dieser Ansatz kann z. B. bei der Erstellung von WWW-Katalogen von Nutzen sein oder zur Präsentation von Suchmaschinenergebnissen verwendet werden, indem diese nicht als einfache Sequenz aufgelistet, sondern zunächst inhaltlich kategorisiert und anschließend als Liste thematischer Kategorien mit zugehörigen Treffern aufgeführt werden. Dumais und Chen benutzen hierfür die von Suchmaschinen generierten Kurzzusammenfassungen, die mit SVM-Verfahren auf der ersten Ebene in 13 und auf der zweiten Ebene in insgesamt 150 Kategorien sortiert werden (vgl. auch Labrou und Finin, 1999, Tiun et al., 2001, und Davidov et al., 2004). Die Tests zeigen jedoch, dass die hierarchische Kategorisierung nur minimale Vorteile aufweist.

Yang et al. (2002) evaluieren die Qualität überwachter Lernverfahren (z. B. Naive Bayes und kNN) bei der Anwendung auf drei Korpora, die aus den Websites von Firmen und *computer science*-Instituten US-amerikanischer Hochschulen bestehen. Dem letztgenannten Korpus wurden die sieben Kategorien „student“, „course“, „faculty“, „project“, „staff“, „department“ und „other“ zugewiesen (ähnlich bei Quek, 1997, und Yu et al., 2002). Die *fallback*-Kategorie „other“ umfasst mehr als 3 000 Dokumente, wohingegen *student* mit ca. 500 Dokumenten die häufigste Inhaltskategorie darstellt. Auch Yang et al. verwenden unterschiedliche Merkmale (z. B. die Inhalte von `meta`- und `title`-Elementen) und kommen zu dem Ergebnis, dass die Auswahl der einzusetzenden Merkmale von der Domäne, der Kategorisierungsaufgabe, dem Algorithmus und den zu analysierenden Daten abhängig ist, wobei nur umfangreiche Tests dazu in der Lage sind, die jeweils geeignetste Kombination zu bestimmen.

Zusammenfassend können verschiedene Aspekte festgehalten werden: Überwachte Lernverfahren benötigen ein flaches oder hierarchisches Kategorienschema und repräsentative Trainingsdaten für jede einzelne Kategorie. Für die thematische Kategorisierung werden oftmals WWW-Kataloge eingesetzt, da sie Inhaltskategorien und zugehörige Dokumente umfassen. In GERHARD wird der Ansatz verfolgt, ein Dokument durch eine computerlinguistische Analyse auf Kategorien der UDK-Hierarchie abzubilden (vgl. Abschnitt 13.3), so dass die Ontologie wissenschaftlicher Themen und Fachgebiete als Basis einer thematischen Kategorisierung benutzt werden kann. Weiterhin liegen in den genannten Arbeiten konzeptuelle Überschneidungen vor: Neben einer ausschließlich thematischen Kategorisierung werden gelegentlich Kategorien angenommen, die eher Hypertextsorten zugeordnet werden können: Die von Asirvatham und Ravi (2001) angenommenen Typen „information page“, „research page“ und „personal home page“ können in Bezug auf die Hypertextsortenontologie als abstrakte Hypertextknotentypen konzeptualisiert werden. Die von Yang et al. (2002) verwendeten Kategorien – z. B. „student“, „course“ und „faculty“ – besitzen ebenfalls Entsprechungen in mehreren Hypertexttypen und Hypertextsorten.

14.2.2 Maschinelle Erkennung von Genres

Im Gegensatz zur thematischen Kategorisierung konnte sich die maschinelle Identifizierung von Textsorten bislang nicht als computerlinguistische Standardanwendung etablieren. Ein wesentlicher Unterschied zur thematischen Kategorisierung betrifft eine umfangreichere Analyse linguistischer Merkmale mit quantitativen Verfahren. Alle Arbeiten basieren auf der An-

nahme, dass das Genre eines Textes durch eine Analyse seiner sprachlichen Eigenschaften (vgl. Abschnitt 2.3.5) ermittelt werden kann.[8] Im zweiten Schritt findet eine Klassifikation statt, die die Merkmale auf eine gegebene Kategorie (d. h. ein Genre) abbildet. Die Genre-Kategorisierung wird als Komplement der thematischen Kategorisierung betrachtet.[9]

Biber (1988) geht von der Beobachtung aus, dass sowohl mündlich als auch schriftlich realisierte Genres charakteristische lexikalische und syntaktische Eigenschaften aufweisen, die mit quantitativen Verfahren ermittelt werden können. Die Untersuchung basiert auf einem Korpus von 481 Texten, die den *Lancaster-Oslo-Bergen* und *London-Lund* Korpora sowie einer Sammlung von Briefen entnommen wurden; das Korpus umfasst 17 schriftlich (z. B. „press reportage", „religion", „biographies" und „general fiction") und sechs mündlich realisierte Genres (z. B. „face-to-face conversation", „broadcast" und „planned speeches"). Genres können in Bezug auf unterschiedliche Dimensionen der sprachlichen Variation verglichen werden (z. B. formell vs. informell und literarisch vs. umgangssprachlich), die eher den Polen eines Kontinuums als binären Merkmalen ähneln (vgl. Abschnitt 2.2.7). Biber (1988, S. 13) bildet die linguistischen Merkmale nicht auf eine a priori bestimmte Menge von Dimensionen ab, sondern geht davon aus, dass eine derartige sprachliche Dimension auf einem Muster basiert, das durch eine Kookkurrenz sprachlicher Merkmale gebildet wird. Bestimmte, kookkurrent in mehreren Texten in Erscheinung tretende Merkmale konstituieren also eine sprachliche Dimension. Quantitative Verfahren werden zur Identifizierung derartiger Merkmalsgruppen eingesetzt und diese Gruppen werden anschließend interpretiert, um die Kookkurrenz zu erklären. Biber (1988, S. 73 ff.) verwendet 67 lexikalische und syntaktische Merkmale (z. B. „past tense", „first person pronouns", „pronoun it", „by-passives", „that verb complements", „infinitives" und „mean word length"), deren Frequenzen durch die Identifizierung derjenigen Wörter maschinell ermittelt werden, die ein spezifisches Merkmal eindeutig signalisieren. So werden z. B. für das Merkmal „causative adverbial subordinators" nur die Vorkommen von „because" erhoben, da es sich um den einzigen Subordinator handele, der ausschließlich kausal verwendet wird (ebd., S. 236). Mit Hilfe eines statistischen Verfahrens, das auf den Merkmalsfrequenzen operiert, ermittelt Biber sechs Dimensionen, die mit Etiketten versehen werden. Das eigentliche Ziel der Arbeit besteht in der Beschreibung der Unterschiede zwischen Schriftsprache und gesprochener Sprache im Englischen. Biber (1988, S. 199) kommt zu dem Schluss, dass kein einzelnes differenzierendes Merkmal vorliegt, vielmehr existieren Dimensionen der Variation und bestimmte Typen von „speech" und „writing" ähneln sich im Hinblick auf die sechs ermittelten Dimensionen (z. B. „Informational versus Involved Production" und „Narrative versus Non-Narrative Concerns"), auf denen die 23 Genres verortet werden können.

[8] Einige der im Folgenden diskutierten Arbeiten verwenden Methoden aus dem Bereich der *Authorship Attribution*, d. h. der maschinellen Ermittlung des Autors eines Textes, um z. B. den oder die Verfasser anonymer Texte oder auch Plagiate zu ermitteln. Verwandte Ansätze werden im Bereich *Document Understanding* entwickelt, z. B. die Erkennung von Seitentypen wie „cover", „title" und „table of contents" durch die Erhebung geometrischer und typografischer Merkmale eines eingescannten Textes (vgl. z. B. Shin et al., 2001).

[9] Zum Beispiel grenzen Karlgren und Cutting (1994, S. 1071) ihren Ansatz explizit von einer Erkennung des Textthemas ab: „[Genre identification] should not be confused with the issue of identifying topics [...]. Although not orthogonal to genre-dependent variation, the variation that relates directly to content and topic is along other dimensions. [...] Texts about certain topics may only occur in certain genres, and texts in certain genres may only treat certain topics".

Karlgren und Cutting (1994) präsentieren einen statistischen Ansatz zur „text genre recognition“ und verwenden Texte aus dem *Brown*-Korpus, das Genres wie z. B. „Press: reportage“, „Press: editorial“, „Religion“ und „General Fiction“ enthält. Die Identifizierung basiert auf einer Methode aus der deskriptiven Statistik: Im Rahmen der Diskriminanzanalyse werden kategorisierte Objekte und Informationen über ihre Parameterausprägungen zur Ermittlung von Diskriminanzfunktionen eingesetzt. Diese können zur Bestimmung der Kategorien unbekannter Texte verwendet werden.[10] Als Parameter fungieren 20 Merkmale, z. B. die Vorkommen von „therefore“, „me“, „it“, „that“, „which“ und „I“ sowie die Frequenzen von Adverbien, Präpositionen, Zeichen und Sätzen, die mit Hilfe eines *Part-of-Speech*-Taggers ermittelt werden können. Karlgren und Cutting führen drei Experimente durch: Zunächst werden 500 Texte in die Kategorien „informative“ und „imaginative“ eingeteilt, wobei 22 fehlerhafte Zuordnungen beobachtet werden (4%). Anschließend werden die Texte in die Klassen „press“, „non-fiction“, „fiction“ und „misc.“ einsortiert, wofür 134 Fehler berichtet werden (27%). Die Klasse „miscellaneous“ stellt „a loose grouping of different informative texts“ dar (ebd., S. 1072), weshalb gerade diese Kategorie von vielen fehlerhaften Zuweisungen betroffen ist (47%). Abschließend wird eine Klassifizierung in 15 Kategorien vorgenommen, wobei in 48% aller Zuweisungen Fehler vorliegen. Wenn für die Subkategorien von „fiction“ eine einzelne Klasse angenommen wird, reduzieren sich die Fehler auf 35%. Es wird deutlich, dass die Fehlerrate mit zunehmender Kategorienanzahl steigt.

Kessler et al. (1997) stellen ein ähnliches Verfahren vor, das ebenfalls mit dem *Brown*-Korpus evaluiert wird. Diesem wurden 499 Texte entnommen und in ein Trainings- (402 Texte) sowie ein Evaluationskorpus (97 Texte) aufgeteilt.[11] Sie gehen davon aus, dass ein Genre über ein Bündel von Eigenschaften beschrieben werden kann, die als „generic facets“ bezeichnet werden (vgl. Crowston und Kwasnik, 2004, sowie Abschnitt 5.2.2). Diese werden mit 55 maschinell extrahierbaren Merkmalen („generic cues“) assoziiert, die in strukturelle (z. B. Nominalisierungen, Passivkonstruktionen und Wortartfrequenzen), lexikalische (Anrede, spezifische Affixe, Datumsangaben etc.), zeichenbezogene (z. B. Interpunktions- und Trennzeichen und Abkürzungen) und abgeleitete „cues“ (Kombinationen lexikalischer und zeichenbezogener Merkmale) aufgeteilt werden. Die Experimente beziehen sich auf drei Textfacetten: (i) „brow“ charakterisiert den intellektuellen Hintergrund und umfasst die Ebenen „popular“, „middle“, „upper-middle“ und „high“; (ii) das binäre Merkmal „narrative“ kennzeichnet, ob es sich um einen narrativen Text handelt; (iii) „genre“ bezieht sich – ähnlich wie bei Karlgren und Cutting – auf die Kategorien „reportage“, „editorial“, „scitech“, „legal“, „nonfiction“ und „fiction“. Zur Klassifikation werden statistische Verfahren und ein neuronales Netz eingesetzt, die auf lexikalischen, zeichenbezogenen und abgeleiteten „cues“ operieren. Die Ergebnisse zeigen, dass Oberflächenmerkmale erfolgreich zur Kategorisierung von Texten eingesetzt werden können: Mit diesen wird eine Präzision von 79% erreicht, die

[10] Jedoch verwenden Karlgren und Cutting (1994, S. 1073) sowohl zur Generierung der Diskriminanzfunktionen als auch zur Evaluation der Funktionen die identische Kollektion von 500 Texten.

[11] Nach Ansicht von Kessler et al. (1997, S. 32) werden die problematischen Aspekte einer maschinellen Genre-Klassifikation erst bei der Verarbeitung sehr großer und heterogener Textbestände wie dem WWW deutlich (vgl. Abschnitt 1.1). Das *Brown*-Korpus enthält Texte, die ursprünglich in einer nicht-digitalen Form publiziert wurden. Ein Demonstrator namens „Northrop“, der den von Kessler et al. (1997) entwickelten Ansatz verwendet, steht unter `http://www.parc.com/istl/groups/qca/` zur Verfügung.

sich beim Einsatz der Merkmale von Karlgren und Cutting auf 66% reduziert. Für die binäre Kategorisierung einzelner Genres wird eine Präzision zwischen 90% und 100% angegeben; für „editorial", „nonfiction" und „legal" sind dabei jedoch zahlreiche Fehler zu verzeichnen.

Stamatatos et al. (2001) gehen ebenfalls davon aus, dass ein Textstil von linguistischen Merkmalen gekennzeichnet ist. Im Gegensatz zu den von Karlgren und Cutting sowie Kessler et al. gewählten Ansätzen setzt die Vorgehensweise von Stamatatos et al. (2001) eine computerlinguistische Analyse voraus, die die Ermittlung von Satzgrenzen und Wortarten sowie ein partielles syntaktisches Parsing betrifft. Auf dieser Basis werden 22 „style markers" erhoben, z. B. Frequenzangaben bezüglich Sätzen, Interpunktionszeichen sowie Nominal-, Verbal- und Präpositionalphrasen. Es wird ein Korpus griechischer Texte verwendet, das für zehn stilistisch homogene Genres (z. B. „press editorial", „reportage", „academic prose", „literature" und „recipes") jeweils 25 Beispiele enthält.[12] Zwei statistische Verfahren werden mit jeweils zehn Texten trainiert, woraufhin die verbleibenden Texte zur Evaluation eingesetzt werden. Als Fehlerrate werden für beide Verfahren 18% angegeben; zusätzlich werden zwei alternative Merkmalsgruppen (Umfang des Vokabulars und Anzahl hochfrequenter Wörter) getestet, die jeweils schlechtere Ergebnisse liefern. Fehlerhafte Zuordnungen betreffen z. B. Texte des Genres „press editorial", die häufig als „press reportage" kategorisiert werden, sowie „curricula vitae" und „official documents". Stamatatos et al. (2001, S. 493) kommen zu dem Ergebnis, dass die Textlänge eine zentrale Rolle einnimmt; sie sollte mindestens 1 000 Wörter betragen, um eine „improved performance" gewährleisten zu können. Zukünftige Arbeiten sollten sich auf die Ermittlung von Textstrukturen beziehen, um diejenigen Textteile zu bestimmen, „where the useful information is more likely to be found." (ebd.).

In einer früheren Studie verwenden Stamatatos et al. (2000) die Häufigkeiten hochfrequenter Wörter als „stylistic markers". Für die vier Genres „Editorials", „Letters to the editor", „Reportage" und „Spot news" werden dem *Wall Street Journal Corpus* jeweils 40 Texte entnommen und in ein Trainings- und ein Evaluationskorpus aufgeteilt. Zur Bestimmung der hochfrequentesten Wörter des Englischen wird die Frequenzliste des *British National Corpus* benutzt. Etwa 75% der hochfrequenten Wörter des WSJ-Korpus sind auch hochfrequente Wörter im BNC. Erneut wird eine Diskriminanzanalyse eingesetzt, die auf den Frequenzen der häufigsten Wörter des BNC in den Trainingstexten der vier Genres beruht. Die in Bezug auf ein Genre individuellen Frequenzen werden zur Ermittlung des Genres eines unbekannten Textes eingesetzt. Die geringste Fehlerrate (2,5%) wird bei der Verwendung der 30 hochfrequentesten Wörter aus dem BNC erreicht: Weniger Wörter (z. B. zehn) erreichen keine ausreichende Differenzierung, und deutlich mehr Wörter erzeugen *overfitting*. In einem zweiten Schritt werden acht Interpunktionszeichen in den Kategorisierungsprozess integriert (Punkt, Komma, Semikolon, Klammern, Fragezeichen etc.): Mit dieser Kombination erreicht das Verfahren Stamatatos et al. zufolge eine Präzision von mehr als 97%.

[12] Das von Stamatatos et al. (2001) aufgebaute Korpus besteht zwar aus Texten, die aus dem WWW heruntergeladen wurden, es fließen jedoch *keine* WWW-spezifischen Merkmale in die Kategorisierung ein. Zudem werden nur traditionelle Genres (bzw. Textsorten) betrachtet, so dass dieses Verfahren keinesfalls, wie Santini (2004c, S. 12) annimmt, dem Bereich „Web genre identification" zugeordnet werden kann. Stamatatos et al. (2001, S. 481) haben bei der Korpuszusammenstellung einen möglichst umfassenden Abdeckungsgrad angestrebt, „trying to cover as many genres as possible." Weiterhin wurden alle Bestandteile der HTML-Dokumente, die nicht dem eigentlichen Text zugehörig sind, manuell entfernt. Diese Vorverarbeitung hebt die Notwendigkeit einer Analyse der Makrostruktur von HTML-Dokumenten hervor (vgl. Abschnitt 14.5).

Dewdney et al. (2001) untersuchen zwei Merkmalsgruppen, die mit überwachten Lernverfahren getestet werden, um sieben Genres zu detektieren („Advertisement", „Bulletin Board", „FAQ", „Message Board", „Radio News", „Reuters Newswire" und „Television News"). Für jedes Genre stehen zwischen 998 und 2 000 Trainingstexte zur Verfügung. Die erste Gruppe von Texteigenschaften („word features") basiert auf IR-Verfahren, um einen aus Wörtern bestehenden Merkmalsraum aufzubauen. Die zweite Gruppe („presentation features") besteht aus 89 Merkmalen. Hierzu gehören linguistische Eigenschaften (z. B. POS-Frequenzen und Tempora), geschlossenen Wortklassen (z. B. Wochentage, Monate und Sternzeichen), die durchschnittliche Länge von Sätzen und Wörtern sowie Interpunktions- und Tabulatorzeichen. Die Experimente zeigen zwei Aspekte auf: Erstens können bereits die „presentation features" erfolgreich zur Kategorisierung von Texten in Genres eingesetzt werden, zweitens liefert eine Kombination der beiden Gruppen eine Verbesserung der Resultate. Das SVM-Verfahren erreicht hiermit eine durchschnittliche Präzision von 92%.

Santini (2004b) verwendet Trigramme von Wortarten zur Klassifikation von Genres. Derartige syntaktische Informationen wurden – mit Ausnahme von Stamatatos et al. (2001) – in bisherigen Arbeiten ignoriert, da sie eine POS-Annotierung voraussetzen. Santini verwendet zehn im BNC enthaltene Genres und jeweils 15 Texte. Es werden vier Gruppen von Merkmalen eingesetzt: Diese umfassen (i) 835 POS-Trigramme (ohne Interpunktionszeichen) und (ii) 1 033 POS-Trigramme (mit Interpunktion). Sie wurden gesammelt, indem zunächst für jede der zehn Kategorien diejenigen Trigramme ermittelt wurden, die innerhalb eines Genres eine Frequenz zwischen 30 und 100 besitzen. Anschließend wurden alle Trigramme betrachtet und diejenigen entfernt, die in mehr als drei Genres gefunden werden, um somit eher gebräuchliche und sehr untypische POS-Sequenzen zu entfernen. Weiterhin wurde (iii) eine Liste von 65 Trigrammen aus der ersten Liste, sowie (iv) eine Liste von 74 Trigrammen aus der zweiten Liste erzeugt. Zur Kategorisierung wird das überwachte Lernverfahren Naive Bayes mit den vier Merkmalsgruppen eingesetzt. Neben einer Unterscheidung aller zehn Genres führt Santini Kategorisierungen durch, die sich nur auf sechs schriftlich bzw. vier mündlich realisierte Genres beziehen. Weiterhin wird eine binäre Differenzierung zwischen diesen beiden Gruppen vorgenommen. Die Verfasserin ermittelt für diese Szenarien eine Kategorisierungsgüte zwischen 78,6% und 99,3%. Die Trigrammmerkmale ohne Informationen über Interpunktionszeichen liefern in nahezu allen Fällen bessere Werte. Die Kategorisierung aller zehn Genres erfolgt mit einer Präzision von 87% (gefilterte Merkmale ohne Interpunktionsdaten), so dass Santini zu dem Schluss kommt, dass die bislang vorgeschlagenen Merkmalsinventare durch derartige Sammlungen von POS-Trigrammen ergänzt werden sollten, da mittlerweile für zahlreiche Sprachen leistungsfähige POS-Tagger erhältlich sind.

Tabelle 14.1 stellt die Ansätze zur Identifizierung von Genres[13] im Überblick dar.[14] Es wird deutlich, dass maschinelle Verfahren durchaus in der Lage sind, für eine überschaubare Anzahl von Genres eine hohe bis sehr hohe Präzision zu erzielen. Hierfür werden in

[13] Viele der in den Arbeiten verwendeten Genres können nicht unmittelbar als Textsorten aufgefasst werden. Aufgrund des sehr umfangreichen Geltungsbereichs von Kategorien wie z. B. „informative", „imaginative", „literature", „press" und „fiction" handelt es sich vielmehr um Texttypen.

[14] Die Arbeit von Biber (1988) kann dieser Anwendung zwar nicht zugeordnet werden, doch bieten sich die 67 Merkmale zur Erkennung von Genres an, wie die anderen Ansätze, die sich alle auf Biber (1988) beziehen, zeigen. Santini (2004c) geht in einer kritischen Überblicksdarstellung auf weitere Studien ein.

Autoren	Korpus	Methoden	Genres	Präzision
Biber (1988)	481 englische Texte aus dem *Lancaster-Oslo-Bergen Corpus*, dem *London-Lund Corpus of Spoken English* und einer Sammlung von Briefen	Statistische Verfahren (Faktorenanalyse) zur Ermittlung textueller Dimensionen, die als Bündel spezifischer Merkmalsausprägungen definiert werden; es werden 67 linguistische Merkmale verwendet	(1) Press reportage; (2) Editorials; (3) Press reviews; (4) Religion; (5) Skills and hobbies; (6) Popular lore; (7) Biographies; (8) Official documents; (9) Academic prose; (10) General fiction; (11) Mystery fiction; (12) Science fiction; (13) Adventure fiction; (14) Romantic fiction; (15) Humor; (16) Personal letters; (17) Professional letters; (18) Face-to-face conversation; (19) Telephone conversation; (20) Public conversations, debates, and interviews; (21) Broadcast; (22) Spontaneous speeches; (23) Planned speeches	k. A.
Karlgren und Cutting (1994)	500 englische Texte aus dem *Brown* Korpus	Statistische Verfahren (Diskriminanzanalyse) mit 20 Merkmalen („noun“, „it“, „adverb“ etc.); es werden drei Experimente durchgeführt	1. (1) Informative; (2) Imaginative 2. (1) Press; (2) Fiction; (3) Miscellaneous; (4) Non-Fiction 3. (1) Press: reportage; (2) Press: editorial; (3) Press: reviews; (4) Religion; (5) Skills and Hobbies; (6) Popular Lore; (7) Belles Lettres, Biographies etc.; (8) Government documents & misc.; (9) Learned; (10) General Fiction; (11) Mystery; (12) Science Fiction; (13) Adventure and Western; (14) Romance; (15) Humor	1. ca. 96% 2. ca. 73% 3. ca. 52% (mit sechs Subkategorien von „Fiction“), ca. 65% (ohne die „Fiction“-Subkategorien)
Kessler et al. (1997)	499 englische Texte aus dem *Brown* Korpus	Statistische Verfahren und neuronale Netze, 55 Merkmale (Anrede, Interpunktion, spezifische Affixe, Datumsausdrücke, Frequenzangaben von Fragezeichen, Ausrufezeichen, Worttrennungen etc.)	(1) Reportage; (2) Editorial; (3) Scitech; (4) Legal; (5) Nonfiction; (6) Fiction	Das erfolgreichste Experiment operiert mit 79%; je nach Verfahren und Genre-Kategorie werden Werte zwischen 58% und 100% erreicht.
Stamatatos et al. (2000)	160 englische Texte aus dem *Wall Street Journal* Korpus (nicht annotiert)	Statistische Verfahren (Diskriminanzanalyse) auf der Basis der Häufigkeiten hochfrequenter Wörter des Englischen sowie den Frequenzen von acht Interpunktionszeichen	(1) Editorial; (2) Letter to the editor; (3) Reportage; (4) Spot news	ca. 97%
Stamatatos et al. (2001)	250 griechische Texte, die im World Wide Web veröffentlicht wurden (jeweils 25 Texte für zehn Genres)	Statistische Verfahren (multiple Regressions- sowie Diskriminanzanalyse) auf der Grundlage von Worthäufigkeiten, Interpunktionszeichen und linguistischen Merkmalen (insgesamt 22 Merkmale)	(1) Press editorial; (2) Press reportage; (3) Academic prose; (4) Official documents; (5) Literature; (6) Recipes; (7) Curricula vitae; (8) Interviews; (9) Planned speeches; (10) Broadcast news, scripted	Fehlerrate: 18% (beide Verfahren)
Dewdney et al. (2001)	9 705 englische Texte	Überwachte Lernverfahren, die auf zwei unterschiedlichen Merkmalsgruppen operieren (die 323 „word features“ des Korpus und 89 verschiedene „presentation features“)	(1) Advertisement; (2) Bulletin board; (3) FAQ; (4) Message board; (5) Radio news; (6) Reuters newswire; (7) Television news	Je nach Verfahren und verwendeten Merkmalen bis zu 92,1%
Santini (2004)	150 englische Texte aus dem *Britisch National Corpus*	Überwachtes Lernverfahren auf der Basis unterschiedlicher Konfigurationen von POS-Trigrammen und Interpunktionsdaten	(1) Conversation; (2) Interview; (3) Public debate; (4) Planned speech; (5) Academic prose; (6) Advert; (7) Biography; (8) Instructional; (9) Popular lore; (10) Reportage	Je nach Verfahren und verwendeten Merkmalen werden Werte zwischen 78,6% und 99,3% erreicht

Tabelle 14.1: Die Ansätze zur maschinellen Erkennung von Genres im Überblick

allen Arbeiten linguistische Merkmale eingesetzt, deren Ermittlung unterschiedlich komplexe sprachtechnologischen Analysen voraussetzt. Diesbezüglich ist zu berücksichtigen, dass bereits in der Phase der Erhebung der einzelnen Merkmalsausprägungen fehlerhafte Werte ermittelt werden können, da Prozesse wie z. B. die Erkennung von Phrasengrenzen für freie Texte bislang noch nicht mit einer Präzision von 100% durchgeführt werden können (Stamatatos et al., 2001); somit sollte die nachfolgende Phase der Berechnung eines Genres robust konzipiert werden, um fehlerhafte Merkmalsausprägungen kompensieren zu können. Zudem werden ausschließlich ASCII-Texte untersucht. Sämtliche Ansätze gehen davon aus, dass ein Text aus einer Sequenz von Wörtern bzw. Sätzen besteht.

Es stellt sich nun die Frage nach der Übertragbarkeit der vorgeschlagenen Verfahren auf die maschinelle Identifizierung von Hypertextsorten. Diesbezüglich kann *nicht* vereinfachend davon ausgegangen werden, dass ein Hypertext aus einer Sequenz von Wörtern oder Sätzen besteht, denn die Instanz einer Hypertextsorte kann eingebettete Hypertextsorten umfassen, d. h. es sind die Grenzen von Hypertexten einzubeziehen. Auf der Ebene des Knotens existieren autarke Hypertextsortenmodule, deren Inhalte *nicht* in abstrahierender Weise konkateniert und mit derartigen Verfahren analysiert werden können, weil Hypertextsortenmodule flexibel in einem Knoten kombiniert oder in mehrere Knoten separiert werden.[15]

14.2.3 Maschinelle Erkennung von Web-Genres

In einer IR-Architektur kann die maschinelle Erkennung von Hypertextsorten als zusätzlicher Filter eingesetzt werden. Eine derartige Funktion wird bislang von keiner der großen Suchmaschinen im WWW angeboten – die Gründe hierfür werden an späterer Stelle diskutiert (vgl. Abschnitt 14.9). Mittlerweile liegen jedoch mehrere Forschungsprototypen vor, die eine „automatic genre classification of web documents“ (Lim et al., 2005a,b) oder ein „World Wide Web retrieval by document type classification“ (Matsuda und Fukushima, 1999) anstreben. Im Folgenden werden diese in Tabelle 14.2 im Überblick dargestellten Verfahren thematisiert und anschließend einer kritischen Analyse unterzogen (vgl. Abschnitt 14.2.4).

Das Interesse der sprachwissenschaftlichen Forschung am *World Wide Web* wurde insbesondere durch persönliche Homepages hervorgerufen. Die *Homepage einer Person* war auch der erste Hypertexttyp, für den spezialisierte Recherchewerkzeuge entwickelt wurden. Shakes et al. (1997) stellen die Suchmaschine *Ahoy!* vor: Der Anwender gibt über ein Formular Vor- und Nachnamen der gesuchten Person ein und spezifiziert zusätzlich die Institution, an der sie tätig ist. *Ahoy!* richtet eine Anfrage an die Metasuchmaschine *Metacrawler*, die eine Trefferliste liefert. Falls keine Institution angegeben wurde, kann sie durch eine parallel stattfindende Anfrage an E-Mail-Verzeichnisse ermittelt werden, in denen nach dem Namen der Person gesucht wird. Die Resultate der Suchmaschine werden einer Filterung unterzogen: Über eine möglicherweise ermittelte E-Mail-Adresse der Person kann auf ihren

[15] Dieser Umstand kann durch einen Vergleich des auf Wortfrequenzen basierenden Ansatzes von Stamatatos et al. (2000) mit der in Kapitel 8 präsentierten Analyse verdeutlicht werden: Stamatatos et al. gehen davon aus, dass ein Token eindeutig dem zu analysierenden Text zugehörig ist. Diese Prämisse besitzt in Bezug auf ein arbiträres HTML-Dokument keine Gültigkeit, weil Token unterschiedlichen Hypertextsortenmodulen zugehörig sein können, die ihrerseits individuelle Bausteine darstellen: Ein Token kann z. B. einer Navigationshilfe, einem Zugriffszähler, einer Begrüßung oder einem Abschnitt von Fließtext entstammen. Die in Hyperlinkanzeigern vorkommenden Trigramme (vgl. Tabelle 8.3, S. 375) veranschaulichen diese Problematik.

	Web-Genre	Methoden	Merkmale	Präzision
Shakes et al. (1997)	Personal Homepage	Einsatz einer Metasuchmaschine; verschiedene Filter- und Rankingkomponenten; Pattern Matching in Bezug auf die Merkmale; Generierung („guessing“) von URLs	URLs von Dokumenten und Inhalte von `title`-Elementen	74%
Rauber und Müller-Kögler (2001)	k. A.	Unüberwachtes Lernverfahren (Cluster-Analyse mittels einer selbstorganisierenden Karte)	Merkmale der Gruppen (a) Textkomplexität, (b) Interpunktions- und Sonderzeichen, (c) Schlüsselwörter und (d) Markup	k. A.
Karlgren et al. (1998)	Non-textual Genres: (1) Informal, private; (2) Public, commercial; (3) Interactive pages; (4) Link collections; (5) Listings, tables; (6) Error messages; Textual Genres: (7) Journalistic materials; (8) Reports; (9) Other running text; (10) FAQs; (11) Discussions	Überwachtes Lernverfahren (Konstruktion von Entscheidungsbäumen mittels Trainingsdaten)	Insgesamt 40 Merkmale, z. B. lexikalische Eigenschaften, Frequenzen von Wörtern und HTML-Elementen	k. A.
Matsuda und Fukushima (1999)	(1) product catalogue; (2) online shop; (3) advertisement for help; (4) Call for papers; (5) links; (6) FAQ; (7) glossary; (8) home page; (9) bulletin board	Gewichtetes Pattern Matching in HTML-Elementen	Schlüsselwörter, Hyperlinks, URLs, Struktur des HTML-Elementbaums, Bilder, OCR, Plug-ins	Durchschnittliche Präzision der Suche mit „document types“: 88,9% (ohne: 31,2%)
Finn et al. (2002)	(1) Kommentare bzw. Editorials („opinion“); (2) Nachrichtenartikel („fact“)	Überwachtes Lernverfahren (Konstruktion von Entscheidungsbäumen mittels Trainingsdaten für drei verschiedene Gruppen von Merkmalen)	(a) Wörter („bag of words“); (b) 36 Wortarten; (c) 76 statistische Merkmale (z. B. Frequenzen von Stoppwörtern, Interpunktionszeichen und langen Wörtern, Satzlänge, Vorkommen spezifischer Schlüsselwörter)	Durchschnittlich zwischen ca. 68% und 72% in Bezug auf die Texte dreier Themenkategorien
Lee und Myaeng (2002, 2004)	(1) Reportage; (2) Editorial; (3) Research articles; (4) Reviews; (5) Homepage; (6) Q&A; (7) Spec	Konstruktion genrespezifischer Vektorbeschreibungen in der Trainingsphase, mit denen die Vektoren unbekannter Texte verglichen werden	Gewinnung von Merkmalen durch Gewichtung themenspezifischer und genrespezifischer Terme	87% (für englische Texte), 90% (für koreanische Texte)
Shepherd et al. (2004)	(1) Personal home page; (2) Corporate home page; (3) Organization home page	Überwachtes Lernverfahren (neuronales Netz)	Es werden 14 von 20 Merkmalen der Gruppen (a) „content“ (z. B. Vorkommen von Telefonnummern, Liste von Schlüsselwörtern), (b) „form“ (z. B. Anzahl von Bildern, Dateigröße) und (c) „functionality“ (z. B. Anzahl von Hyperlinks und HTML-Formularen) verwendet	Durchschnittliches F-Maß für das Experiment mit den präzisesten Ergebnissen (d. h. separate Klassifikatoren ohne „noise pages“): 70,6%
Meyer zu Eissen und Stein (2004)	(1) Help; (2) Article; (3) Discussion; (4) Shop; (5) Portrayal (non-private); (6) Portrayal (private); (7) Link Collection; (8) Download	Überwachte Lernverfahren (SVM, neuronales Netz), die auf zwei Merkmalsgruppen arbeiten	Insgesamt 35 Merkmale (HTML-Elemente, Textstatistiken, Worthäufigkeiten, Interpunktionszeichen, *Part of speech*-Frequenzen etc.)	ca. 70%
Lim et al. (2005a,b)	Non-textual Genres: (1) Personal homepages; (2) Public homepages; (3) Commercial homepages; (4) Bulletin collections; (5) Link collections; (6) Image collections; (7) Simple tables/lists; (8) Input pages; Textual Genres: (9) Journalistic materials; (10) Research reports; (11) Official materials; (12) Informative materials; (13) FAQs; (14) Discussions; (15) Product specifications; (16) Others (informal texts)	Überwachtes Lernverfahren (k Nearest Neighbour-ähnliches Algorithmus)	Insgesamt 329 Merkmale (vgl. Tabelle 14.3)	75,9%

Tabelle 14.2: Die Ansätze zur maschinellen Erkennung von Web-Genres im Überblick

Login geschlossen werden, der wiederum in der URL der persönlichen Homepage verwendet wird (z. B. `g91063@...` → `http://.../~g91063/`). Ein Filter bestimmt, ob es sich bei einem Dokument um eine persönliche Homepage handelt. Diese Heuristik basiert ausschließlich auf den Inhalten des `title`-Elements und der URL. Falls die Suchmaschine keine Treffer liefert, ist *Ahoy!* in der Lage, URLs dynamisch zu generieren. Hierzu werden entsprechende Muster eingesetzt, die typische Permutationen des Namens erzeugen (für „Hans Meier" z. B. `.../~hans/`, `.../~hmeier/`, `.../~meier/`, `.../people/hans/` etc.) und versuchen, unter der generierten URL eine HTTP-Query abzusetzen (für diesen Schritt muss die Institution bekannt sein). Shakes et al. geben an, dass 74% der mit *Ahoy!* durchgeführten Suchanfragen die gewünschte Homepage als erstes Ergebnis liefern, wohingegen *AltaVista* lediglich 58% ermittelt. Die Architektur eignet sich Shakes et al. zufolge insbesondere für diejenigen „classes of pages", deren Mitglieder mit einfachen Mitteln als Vertreter einer Klasse bestimmt werden können und bereits teilweise in Suchmaschinen verzeichnet sind. Neben persönlichen Homepages werden „popular articles or academic papers on a single topic", „product reviews", „price lists" und „transportation schedules" aufgeführt. Diese Beschreibung zeigt bereits die Grenzen von *Ahoy!* auf: Instanzen des Hypertextknotentyps *Einstiegsseite der Homepage einer Person* werden durch typische URL-Muster ermittelt. Wenn eine URL (z. B. `http://www.hansmeier.de`) nicht den implementierten Annahmen entspricht, kann die Homepage entweder nicht gefunden werden oder sie wird als irrelevant eingestuft.

Rauber und Müller-Kögler (2001) präsentieren ein System, das sowohl eine thematische als auch eine Genre-Klassifikation durchführt, um einen intuitiven Zugang zu heterogenen Textkollektionen zu ermöglichen (vgl. Rauber und Merkl, 2003). Beide Prozesse basieren auf selbstorganisierenden Karten (SOMs), d. h. es werden unüberwachte Lernverfahren eingesetzt, die durch Merkmalsvektoren repräsentierte Dokumente eigenständig nach ihrer Ähnlichkeit strukturieren. Wortfrequenzen bilden die Basis des thematischen Merkmalsraumes. Die SOM erzeugt *cluster* ähnlicher Vektoren, die in einem zweidimensionalen Raum verortet werden. Das grafische Interface basiert nicht auf der Auflistung von Dokumenten, sondern auf der Präsentation eines metaphorischen Bücherregals. Das Aussehen eines Dokuments wird durch Metadaten bestimmt, so dass unterschiedliche Dokumenttypen als Bücher, Ordner, Blattsammlungen und Hefter dargestellt werden. Über weitere Metaphern wird der Umfang eines Dokuments (Dicke des Buches) oder sein Alter (Staubschicht, Spinnweben) visualisiert. Die Genre-Klassifikation erfolgt durch eine Analyse, die sich auf unterschiedliche Merkmalsgruppen bezieht: Die Textkomplexität wird über einfache Statistiken wie z. B. die durchschnittliche Anzahl von Wörtern pro Satz und Sätzen pro Absatz oder die Wortlänge ermittelt. Weiterhin werden die Vorkommen von Interpunktions- und Sonderzeichen sowie von Schlüsselwörtern erhoben. Abschließend erfolgt eine Analyse des Markups, um z. B. die Anzahl von Gleichungen und Literaturverweisen zu bestimmen. Diese Merkmale werden in Vektoren zusammengefasst, die ebenfalls mit einer SOM verarbeitet werden, die Dokumente mit ähnlichen „stylistic characteristics" (Rauber und Müller-Kögler, 2001, S. 6) in Clustern bündelt. Hierdurch entsteht z. B. in der Karte ein Bereich, der sehr lange Dokumente mit einer hohen Textkomplexität umfasst, ein weiterer Bereich beinhaltet Dokumente mit kürzeren Sätzen und einer großen Anzahl von Doppelpunkten und Anführungszeichen. Diese Cluster besitzen keine Etiketten wie etwa „Interview" für das zuletzt genannte Areal. Die Integration der Genre-Analyse in die grafische Oberfläche erfolgt durch die Abbildung des

zweidimensionalen Raumes auf spezifische Farben, mit denen die Objekte im Bücherregal eingefärbt werden, so dass einander ähnliche Farben ähnliche Genres visualisieren. Das Verfahren wurde mit 1 000 HTML-Dokumenten getestet, die 14 Websites von österreichischen Zeitungen und Magazinen entnommen wurden.[16] Die Genre-Analyse deckt verschiedene Cluster auf, z. B. enthält ein Bereich der Karte unterschiedliche Typen von Interviews, ein anderer umfasst Ankündigungen von Radio- und Fernsehsendungen. Problematisch ist, dass das tatsächliche Genre eines Dokuments durch den Einsatz der Farbmetapher nicht intuitiv erkennbar ist. Von Vorteil ist jedoch die Repräsentation gradueller Wechsel zwischen Genres, die durch unterschiedliche Farbtöne visualisiert werden.

Zur Unterstützung aufgabenorientierter Suchanfragen kombinieren Matsuda und Fukushima (1999) erstmals ein IR-Verfahren mit einer „document type classification", wozu Strukturcharakteristika von HTML-Dokumenten verwendet werden. Es werden Regeln eingesetzt neun Dokumenttypen wie z. B. „product catalogue", „links", „FAQ" und „home page" zu identifizieren.[17] Die Charakteristika werden über Beschreibungen erfasst, die sich auf die Merkmale „keyword", „link", „URL", „structure", „image", „OCR" und „plugin" beziehen (vgl. Abschnitt 14.6.2): Mit ihrer Hilfe können für den „product catalogue" z. B. die Schlüsselwörter „product", „service" oder „system" innerhalb von `h2`-Elementen oder „trademark" und „company" innerhalb von `body` detektiert werden. Weitere Merkmale können zur Beschreibung der minimalen oder maximalen Anzahl interner Hyperlinks oder zum Mustervergleich in URLs eingesetzt werden. Jedes Merkmal besitzt ein Gewicht, so dass für jedes der neun Regelsysteme ein *score* berechnet werden kann, der den „document type" des Dokuments darstellt. Matsuda und Fukushima geben an, dass die Suche nach Dokumenttypen in konkreten Problemlösungsszenarien eine Präzision von 88,9% aufweist, wohingegen die Suche nach Schlüsselwörtern lediglich 31,2% ergibt. Diese Daten beruhen auf einem Experiment, das sich auf die Dokumenttypen „product catalogue", „links", „investigation report", „advertisement for help", „prize" und „update software" bezieht. Von diesem Ansatz können somit nur diejenigen Instanzen von Dokumenttypen erfasst werden, für die die von den Verfassern antizipierten und intuitiv mit Gewichten versehenen Regeln zutreffen.[18]

Finn et al. (2001, 2002) testen Gruppen von Merkmalen zur Kategorisierung der Texteigenschaften von HTML-Dokumenten. Verwendet werden ein *bag of words*-Ansatz, POS-Statistiken und manuell zusammengestellte, statistische Merkmale (ähnlich Kessler et al., 1997), die für die Konstruktion von Entscheidungsbäumen eingesetzt werden. Exemplifiziert werden die Methoden an der Differenzierung zwischen Kolumnen bzw. Kommentaren („opinion") und Nachrichtenmeldungen („fact"; vgl. auch Dimitrova und Kushmerick, 2003). Da HTML-Dokumente peritextuelle Elemente besitzen, wird der eigentliche Artikeltext über eine Heuristik ermittelt. Diese Vorverarbeitung ist notwendig, da die Nutzung des

[16] Rauber und Müller-Kögler (2001, S. 7) berichten, dass hierzu eine „cleansing procedure" implementiert wurde, die für jede Website die „characteristic formatting structures" entfernt (z. B. Werbebanner, Fußzeilen und Navigationshilfen), „as these would unduely interfere with the stylistic analysis" (vgl. Fußnote 12, S. 635).

[17] Matsuda und Fukushima (1999, S. 110) erläutern das Konzept „document type" wie folgt: „Most Internet users can recognize a certain document type to which a particular Web page belongs just by casually looking at it. [... U]sers [...] evaluate a page not from its contents but from its various format and design information."

[18] Die Anzahl der in den präsentierten Tests verwendeten Dokumenttypen schwankt, so dass nicht deutlich wird, von wie vielen Typen Matsuda und Fukushima tatsächlich ausgehen. Die erstaunlich hoch erscheinende Klassifikationspräzision könnte durchaus von diesem Umstand profitiert haben.

vollständigen Dokumentinhalts zu fehlerhaften Kategorisierungen führte (vgl. die Fußnoten 12 und 16). Finn et al. verwenden Kommentare und Nachrichtenmeldungen aus den Themenbereichen Fußball, Politik und Finanzen, für die ein Korpus von 796 Artikeln zusammengestellt wurde. Es ist das Ziel, einen Klassifizierer zu erstellen, der sowohl innerhalb einer derartigen Kategorie als auch themenübergreifend präzise Resultate produziert. Die Evaluierung des Verfahrens mit den Merkmalsgruppen zeigt, dass innerhalb eines Themas eine Kategorisierungsgüte zwischen 85% und 88% erreicht werden kann. Die themenübergreifende Evaluierung erfolgt durch die Anwendung eines themenspezifischen Klassifikators auf die beiden anderen Kategorien; die Präzision schwankt hierbei zwischen 58% und 91%. Es zeigt sich, dass POS-Merkmale konsistent gute Ergebnisse liefern. Der *bag of words*-Ansatz liefert bei einer Anwendung auf andere Themenkategorien schlechte Resultate, da er domänenabhängig und somit nicht generalisierbar ist. Finn et al. kommen zu dem Schluss, dass die Wahl einer Merkmalsgruppe von jedem zu verarbeitenden Genre abhängig ist, wodurch die Skalierbarkeit einer Genre-Kategorisierung beeinträchtigt wird. Es wird zwar eine „genre classification" angestrebt, Finn et al. beschränken sich jedoch auf nur *eine* textuelle Dimension (im Sinne von Biber, 1988). Die Differenzierung zwischen „opinion" und „fact" kann auch auf andere Textsorten sowie Textteile der Vertreter spezifischer Textsorten angewendet werden: Wissenschaftliche Artikel stellen im Wesentlichen Fakten dar, doch wird in Abschnitten wie „Verwandte Arbeiten" oder „Diskussion" die Meinung der Verfasser dargestellt. Daher muss ein System zur Identifizierung von Hypertextsorten zwangsläufig weitere Textdimensionen sowie eine Analyse der Dokumentstruktur berücksichtigen.

Karlgren et al. (1998) und Bretan et al. (1998) kombinieren ebenfalls traditionelle IR-Verfahren mit der Detektion von Genres. Der Prototyp bezieht Dokumente von einer Suchmaschine, an die die Anfragen der Nutzer weitergereicht werden und verarbeitet die Ergebnisliste (ähnlich wie Shakes et al., 1997). Die Identifizierung des Genres bezieht sich auf die in Abbildung 4.7 (S. 185) dargestellte Hierarchie (vgl. Abschnitt 4.4.6) und verwendet 40 Merkmale (z. B. einige der von Biber, 1988, vorgeschlagenen Charakteristika sowie die Frequenzen von Bildern und Hyperlinks). Mit Hilfe von Trainingsdaten wird ein Entscheidungsbaum erzeugt (Quinlan, 1993), der Zuordnungen auf der Basis von „a few dozen rules" vornimmt (Karlgren et al., 1998, S. 89). Es werden keine Angaben zur Präzision des Ansatzes gemacht, jedoch verläuft die Kategorisierung auf der ersten Ebene der Hierarchie („textual" vs. „non-textual") „quite well, something like a ninety per cent success rate, while the subsplits make the wrong choice somewhere between once in three or four times [sic]." (ebd.).

Lee und Myaeng (2002, 2004) untersuchen den Einfluss des Themas auf die Identifizierung von Genres. Das Korpus beinhaltet 7 828 koreanische und 7 615 englische HTML-Dokumente, die den Genres „Reportage", „Editorial", „Research articles", „Reviews", „Homepage", „Q&A" sowie „Spec" und auch thematischen Kategorien zugeordnet wurden.[19] Die Methodik entspricht überwachten Lernverfahren: Es werden die Merkmalsausprägungen in den Trainingsdaten bestimmt, woraufhin ein Algorithmus unbekannten Dokumenten Genres zuweist. Das Spezifikum besteht in der Ermittlung der zu verwendenden Merkmale und

[19] Lee und Myaeng (2004) gehen (anders als z. B. Stamatatos et al., 2001) nicht auf die Frage der Vorverarbeitung der Dokumente ein. Implizit wird jedoch deutlich, dass offenbar der gesamte Textinhalt der Dokumente extrahiert und für die Kategorisierung eingesetzt wird. Die Arbeit von Lee und Myaeng (2004) wurde nicht in Abschnitt 14.2.2 dargestellt, weil das Inventar betrachteter Genres die Kategorie „Homepage" beinhaltet.

ihrer Gewichtung: Es wird beobachtet, dass die Frequenz eines Terms in den Dokumenten einer Genre-Kategorie sehr hoch sein kann, weil der Term themen- und nicht genrespezifisch ist. Die Verfasser nehmen also an, dass ein Zusammenhang zwischen Thema und Genre besteht und dieser Umstand fließt in die maschinelle Extraktion der Merkmale ein, die aus den Termen der Dokumente bestehen: Zur Gewichtung eines Merkmals wird bestimmt, wie viele Trainingsdokumente eines Genres den Term enthalten und wie gleichmäßig er in den thematischen und in den Genre-Kategorien verteilt ist, ob es sich also bei einem Term um ein distinktives Merkmal eines Genres oder eines Themas handelt; letztere werden ignoriert, erstere werden positiv gewichtet. Lee und Myaeng nehmen an, dass ein Merkmal, das für ein spezifisches Genre charakteristisch ist, in mehreren Themenkategorien aber nur in einer Genre-Kategorie enthalten ist. Aus den gewichteten Merkmalen jeder Genre-Kategorie wird ein Vektor erzeugt (vgl. Abschnitt 14.2.1). Zur Kategorisierung eines Dokuments werden dessen Merkmalsausprägungen bestimmt, in einen Vektor überführt und mit den Genre-Vektoren verglichen. Die Verfasser ermitteln für dieses Verfahren eine Präzision von 87% für die englischen und 90% für die koreanischen Texte. Es wurde auch ein überwachtes Lernverfahren (Naive Bayes) getestet, für das die Werte 83% und 75% angegeben werden. Im Kontext einer vollautomatisch durchgeführten Identifizierung von Web-Genres ist dieser Ansatz kritisch einzuschätzen, denn er setzt eine thematische Kategorisierung der Trainingsdaten voraus. Mit steigender Anzahl zu berücksichtigender Web-Genres nimmt die Komplexität der thematischen Kategorisierung zu, so dass sich, und dies gilt für sämtliche hier besprochenen Arbeiten, die Frage stellt, ob das Verfahren auch bei der Verarbeitung von mehr als 100 Web-Genres (bzw. Hypertextknotensorten) zufriedenstellende Ergebnisse liefert.

Shepherd et al. (2004) knüpfen an Shepherd und Watters (1999) an und führen Experimente zur Erkennung der Einstiegsseiten von Organisationen, Unternehmen und persönlichen Homepages durch (ähnlich bei Meyer zu Eissen und Stein, 2004, die sich jedoch auf andere Web-Genres beziehen). Es wird argumentiert, dass bisherige Ansätze lediglich Merkmale der Ebenen „content“ und „form“ einsetzen, weshalb Shepherd et al. zusätzlich Eigenschaften der „functionality“-Komponente berücksichtigen (vgl. Abschnitt 4.4.3). Diesen drei Ebenen werden 20 Merkmale zugeordnet, die z. B. die Anzahl verwendeter `meta`-Elemente, Schlüsselwörter, die Existenz von Telefonnummern („content“), die Größe einer Datei, die Anzahl der Wörter („form“) und die Anzahl von Hyperlinks („functionality“) umfassen. Das Korpus besteht aus 321 HTML-Dokumenten und besteht in etwa zu gleichen Teilen aus den drei eingangs genannten Kategorien und 77 „noise pages“. Die Experimente werden mit neuronalen Netzen durchgeführt und beziehen sich auf das Trainieren *eines* Netzes, das alle drei Kategorien verarbeiten kann und *unterschiedlicher* Netze, die jeweils ein Web-Genre kategorisieren; für die Ergebnisse der Experimente werden F-Maße[20] angegeben. Shepherd et al. gelangen zu mehreren Schlussfolgerungen: Die Kategorisierung der persönlichen Homepages erfolgt mit der höchsten Präzision; die Ergebnisse für die kommerziellen Einstiegsseiten fallen minimal schlechter aus, wohingegen die Kategorisierung der „organization home pages“ deutlich abfällt. Als Grund wird angeführt, das derartige HTML-Dokumente keinen „specific style that is unique to them“ besitzen, vielmehr ähneln sie – je nach Produzent – ent-

[20] Für IR-Verfahren werden die *precision* (p, Genauigkeit der extrahierten Einzelinformationen) und der *recall* (r, Vollständigkeit der extrahierten Einzelinformationen bezüglich der Gesamtmenge) gemessen; das Maß F wird eingesetzt, um *recall* und *precision* in einem Wert ungewichtet abzubilden, z. B. mittels $F = 2rp/(r + p)$.

Feature	Description
U1	Depth of URL *The set of features related with URL*
U2	Document type (document extension): the value is one of { HTML, SCRIPT, DOC, OUTPUT, and MIX }
U3	Is '/~' used in URL?
U4	Is filename in { index, default, main, home, main_default }? Or is filename omitted?
U5	Domain area: *com*, *org*, *edu*, *net*, *gov*, *ac.kr*, *co.kr*, *go.kr*, *re.kr*, *ne.kr*, *or.kr*, *pe.kr*, etc.
U6	Number of URLs (= number of frames in a document)
UL1–UL35	Is it used in URL? For 35 lexical terms: *faq*, *news*, *board*, *detail*, *list*, *qna*, *index*, *shop*, *data*, *go*, *view*, *front*, *main*, *company*, *item*, *paper*, *bbslist*, *product*, *read*, *papers*, *start*, *file*, *gallery*, *introduction*, *info*, *login*, *search*, *research*, *bbs*, *link*, *intro*, *people*, *profile*, *photoi*, *photo*
H1	Frequency of links to the same domain/total frequency of tags used in a document *The set of features related with HTML tags*
H2	Frequency of links to the different domain/total frequency of tags used in a document
H3	Frequency of links/total number of characters in a document
H4–H75	Frequency of tag/total frequency of tags used in a document for 72 html tags: *col*, *textarea*, *input*, *frame*, *iframe*, *select*, *img*, *area*, etc.
F1	Number of characters *The set of features for frequencies/ratio of textual and POS tokens*
F2	Number of words
F3	Number of candidate sentences
F4	Number of detected sentences/number of candidate sentences
F5	Average number of words per sentence
F6	Average number of characters per word
F7	Number of candidate sentences/number of characters
F8–F13	Number of TYPE-words/total number of words for TYPE: *hangul*, *hanja*, *alphabet*, *digit*, *punctuation*, *symbol*
T1–T9	Number of POS words/total number of words for 9 POSs: *noun*, *pronoun*, *adjective*, *verb*, *adverb*, *interjection*, *modifier*, *postposition*, *verbal-ending*
T10	Average number of morphological results per word (morphological ambiguities)
T11–T15	Number of DICTINFO-words/total number of words for DICTINFO: *sino*, *foreign*, *proper*, *onomatopoeic/mimetic*, *title*
MC1–MC50	Frequency of CONTENT words/total freq. of content words for 50 most freq. used content words *The set of features related with lexical entries*
MF1–MF50	Frequency of FUNCTION words/total frequency of function words; for 50 most frequently used function words
MP1–MP32	Frequency of PUNCTUATION/total frequency of punctuation for 32 most frequently used punctuation marks
S1	Number of usual words/total number of words (frequency of usual word > 1000 in the training corpus)
S2	Number of unusual words/total number of words (frequency of unusual words = 1 in the training corpus)
V1	Unique number of words/total number of words (Vocabulary richness)
P1	Number of declarative sentences/number of candidate sentences *The set of features related with structural information*
P2	Number of imperative sentences/number of candidate sentences
P3	Number of question sentences/number of candidate sentences
P4	Number of sentence with parsing failure/number of candidate sentences in a document
P5	Average number of syntactic trees per sentence (syntactic ambiguities)
P6–P22	Number of phrase/total number of phrases in a document for 17 phrases: *NP*, *VP*, *AJP*, *AUXP*, *AVP*, *CONJP*, *SENT*, *IMPR*, etc.
P23–P39	Average number of words per phrase for 17 phrases: *NP*, *VP*, *AJP*, *AUXP*, *AVP*, *CONJP*, *SENT*, *IMPR*, etc.
C1–C11	Number of chunks for 11 expressions: *date*, *time*, *postal address*, *telephone number*, *money*, *unit*, *Copyright*, *e-mail*, *personal names*, *abbreviation*, *numeric*

Tabelle 14.3: Die von Lim et al. (2005b) eingesetzten Merkmale

weder einer persönlichen Homepage oder einer kommerziellen Einstiegsseite (ebd., S. 249; vgl. auch Kennedy und Shepherd, 2005). Darüber hinaus wird mit der Einbeziehung der „functionality"-Merkmale nur in wenigen Fällen eine Verbesserung der Präzision erreicht.

Lim et al. (2005a,b) präsentieren die sowohl im Hinblick auf die Anzahl der Kategorien als auch bezüglich der Menge der Merkmale ausführlichste Studie zur Identifizierung von Web-Genres. Basierend auf den Arbeiten von unter anderem Karlgren und Cutting (1994), Kessler et al. (1997) und Stamatatos et al. (2001) wird ein umfangreicher Katalog von insgesamt 329 Merkmalen erstellt (vgl. Tabelle 14.3). Zusätzlich zu den Eigenschaften traditioneller Genres beziehen Lim et al. Merkmale ein, die HTML-Dokumente kennzeichnen (vgl. Abschnitt 14.6.2 und Fußnote 59, S. 679). Bezüglich des Inventars von 16 Web-Genres setzen Lim et al. eine modifizierte Version der Hierarchie von Dewe et al. (1998) ein (vgl. Abbildung 4.7, S. 185).[21] Insgesamt 1 328 koreanische HTML-Dateien fungieren

[21] Lim et al. (2005b, S. 1267) präsentieren eine Tabelle, die die 16 „Web-Genres" und ihnen zugeordnete „Samples" enthält, deren Status nicht deutlich wird, so wird z. B. „Resume" für das Web-Genre „Personal homepages" aufgeführt, und „Informative materials" umfasst „Recipes, lecture notes, encyclopedic information".

als Korpus, das unter anderem auf der Grundlage sehr populärer Suchanfragen zusammengestellt wurde, um eine ausgeglichene Verteilung zu gewährleisten. Experten haben diesen Dokumenten die korrespondierenden Web-Genres des angesprochenen Inventars zugewiesen, wobei eine Überprüfung durch zwei weitere Personen stattfand.[22] Die Tests werden mit einem kNN-ähnlichen Verfahren durchgeführt. Die Extraktion der Merkmale erfolgt in mehreren Schritten: Zunächst werden die HTML-Elemente und Satzgrenzen ermittelt, woraufhin eine morphologische und syntaktische Analyse stattfindet.[23] Die Experimente dienen der Bestimmung der Präzision der insgesamt 12 Merkmalsgruppen (U, UL, H, F, T, MC, MF, MP, S, V, P, C; vgl. Tabelle 14.3) und operieren auf unterschiedlichen Teilen der HTML-Dokumente: Diese werden in die Bereiche TM (Inhalte von `title` und `meta`), ANCH (alle Hyperlinkanzeiger) und BODY (der verbleibende Text) sowie ihre vier Kombinationen eingeteilt (TM+ANCH, ANCH+BODY, TM+BODY und TM+ANCH+BODY). Die individuelle Präzision der 12 isoliert angewendeten Merkmalsgruppen liegt jeweils unterhalb von 50%, jedoch liefern die HTML-Elemente einen Wert von 55,1% (bezüglich TM+ANCH). Lim et al. (2005b, S. 1272 ff.) zeigen, dass einige Web-Genres durch spezifische Merkmalsausprägungen generalisierbar sind, so kann z. B. „FAQ“ mit einer hohen Präzision kategorisiert werden, weil 36 der 54 im Korpus enthaltenen FAQ-Dokumente die Zeichenkette „faq“ innerhalb ihrer URL besitzen. Wenn alle Merkmale kombiniert werden, ergibt sich eine Präzision von etwa 74%, wobei die Performanz der WWW-spezifischen Merkmale „slightly better“ ist als die der traditionellen Texteigenschaften (ebd., S. 1274); durch eine maschinelle Auswahl der Merkmale kann die Präzision auf 75,7% gesteigert werden. Die Verarbeitung der „textual“ Web-Genres (vgl. Tabelle 14.2) erfolgt mit einer höheren Kategorisierungsgüte, so besitzen z. B. „research reports“, „journalistic materials“ und „discussions“ für nahezu alle Merkmalsgruppen distinktive Ausprägungen, so dass sie präzise detektiert werden können. Die Web-Genres „input pages“ und „simple tables/lists“ (beide „non-textual“) sowie „others“ sind am anderen Ende des Spektrums angesiedelt, ihr Abdeckungsgrad wird von Lim et al. (2005b, S. 1274) als „awful“ bezeichnet: „It means there are no distinguishable properties in the documents in these genres. Indeed many web documents contain tables, input windows within their pages by default [sic]. Consequently, we must look into more carefully whether or not these classes are indispensable for the web genres [sic].“ Diese Problematik basiert auf einer eindeutig zu identifizierenden Ursache: Das von Lim et al. verwendete Inventar von Web-Genres besitzt eine unzureichende theoretische Fundierung, was bereits durch die Präsenz der Kategorie „others“ deutlich wird. Weder „input pages“ noch „simple tables/lists“ können als Web-Genres bezeichnet werden, vielmehr handelt es sich – in Bezug

[22] Nach eigenen Angaben mussten durch diesen „double-check“ die Web-Genres von lediglich acht Dokumenten modifiziert werden (Lim et al., 2005b, S. 1267). Dieser Umstand steht einerseits in einem Widerspruch zu den in Abschnitt 4.4 präsentierten Analysen zur Sammlung von Web-Genres, in denen oftmals berichtet wird, dass sich die Zuweisung eines Dokuments zu einem Web-Genre problematisch gestaltet hat (was insbesondere für Zuweisungen durch mehrere Personen gilt). Andererseits ähneln sich einige der von Lim et al. (ebd.) für *unterschiedliche* Web-Genres angeführten „Samples“ sehr, z. B. „complaint board“, „notice board“ (Web-Genre: „Bulletin collections“) und „pages in news group“, „page for question“, „page for answer“ („Discussions“).

[23] Im Hinblick auf die in Abschnitt 14.5 diskutierte Notwendigkeit, einen Textparser zur Ermittlung der makrostrukturellen Bausteine von HTML-Dokumenten einzusetzen, thematisieren Lim et al. (2005b) *nicht*, wie die Satzgrenzenerkennung und die morphologischen und syntaktischen Analysen durchgeführt werden oder mit welcher Präzision die eingesetzten Verfahren arbeiten.

auf das Hypertextsortenmodell (Kapitel 5) – um Hypertextmodule, die vermutlich aufgrund ihrer häufigen Verwendung und auffälligen Darstellungscharakteristika von Lim et al. als eigenständige „Web-Genres“ interpretiert werden. Zugleich zeigt diese Problematik (und die Tatsache, dass die Web-Genres der Gruppe „non-textual“ nur mit einer schlechten Präzision erkannt werden können) die Notwendigkeit einer Analyse der Makrostrukturbausteine von HTML-Dokumenten auf: „input pages“ und „simple tables/lists“ sind für die Erkennung von Hypertextknotensorten selbstverständlich „unverzichtbar“, schließlich beziehen sich diese Bezeichnungen auf eine andere Strukturebene (Hypertextmodule), die ihrerseits innerhalb von Dokumenten funktional oder inhaltlich-thematisch markiert sind und somit als Hypertextsortenmodule fungieren (z. B. als primäre Navigationshilfe oder als „search box“).

14.2.4 Fazit – Kritische Einschätzung der Arbeiten

Die in den Abschnitten 14.2.2 und 14.2.3 dargestellten Arbeiten zeigen, dass es prinzipiell möglich ist, Systeme zur maschinellen Identifizierung von Genres und Web-Genres zu konstruieren. Die Verfahren, die zur Bestimmung traditioneller Genres vorgeschlagen wurden, basieren auf strukturellen und linguistischen Merkmalen, die durch eine effiziente Verarbeitung der Textoberfläche ermittelt werden können. Die meisten Ansätze zur Ermittlung von Web-Genres knüpfen an diese Arbeiten an und ergänzen WWW-spezifische Eigenschaften. Vor dem Hintergrund des Hypertextsortenmodells (Kapitel 5) und der in Teil III präsentierten Analysen können fünf Problemkreise identifiziert werden:

1. *Auswahl und Granularität der verwendeten Web-Genres* – Die Ansätze zur maschinellen Erkennung von Web-Genres beschränken sich auf Inventare, die zwischen zwei und 16 Kategorien umfassen. Alle Listen wurden ad hoc erstellt, basieren nicht auf zuvor durchgeführten empirischen Analysen und sind nicht textlinguistisch motiviert; auch Santini (2004c, S. 13) empfindet diese Problematik als „common limitation of Web genre detection projects“. Die meisten eingesetzten Inventare von Web-Genres beinhalten mit unterschiedlichen Typen von Homepages, Hyperlinklisten und dem FAQ die offensichtlichen Kategorien. Die Existenz anderer Klassen wie z. B. „link collections“ (Karlgren et al., 1998), „advertisement for help“ (Matsuda und Fukushima, 1999), „input pages“ und „simple tables/lists“ (Lim et al., 2005a,b) wird weder begründet noch theoretisch reflektiert.

2. *Restriktion des Merkmalinventars* – In fast allen verwandten Arbeiten (Ausnahmen stellen Stamatatos et al., 2001, und Lim et al., 2005a, 2005b, dar) werden hinsichtlich der eingesetzten Erkennungsmerkmale Restriktionen vorgenommen, die sich auf die Implementierbarkeit korrespondierender Methoden beziehen, für die gleichzeitig die Notwendigkeit besteht, dass sie sowohl effizient als auch hochgradig präzise operieren müssen: Wenn z. B. die durchschnittliche Anzahl Wörter pro Satz als Merkmal zur Verfügung stehen soll, ist es notwendig, ein Verfahren zur maschinellen Satzgrenzenerkennung zu verwenden, das so präzise wie möglich arbeitet. Die Effizienz bezieht sich auf die Problematik, dass für einzelne Dokumente bis zu 329 Merkmale ermittelt werden (bei Lim et al., 2005a,b), so dass für deren Erhebung komplexe und rechenintensive sprachtechnologische Prozesse nicht in Frage kommen; Kessler et al. (1997) verzichten z. B. aufgrund der hiermit verbundenen „computational cost“ auf POS-Informationen. Meyer zu Eissen und Stein (2004) unterscheiden zwischen Merkmalen mit einem geringen (Textstatistiken), mittleren (wortbezogene Merkmale, die z. B. einen

Lexikon-*lookup* voraussetzen) und hohen „computational effort“ (z. B. POS-Tagging, syntaktisches Parsing). Lim et al. (2005b, S. 1264) merken an, dass die Auswahl von Erkennungsmerkmalen, die in der Lage sind, möglichst eindeutige Differenzierungen zwischen einzelnen Web-Genres vorzunehmen, „the core of automatic genre classification“ ist; Shepherd et al. (2004) bezeichnen die Auswahl einer geeigneten Menge von Merkmalen als „most important open question“. Im Gegensatz dazu vertritt Santini (2004c, S. 19) die Auffassung, dass die Verwendung einer „restricted number of linguistic and [...] layout features“ der zentrale Schwachpunkt der aktuellen Ansätze sei; der ausschließliche Einsatz linguistischer Merkmale (im Sinne von Biber, 1988) wird von Santini (2004c, S. 20) in Bezug auf das WWW als „completely inadequate“ bezeichnet. Die vorliegenden Arbeiten zeigen, dass Kombinationen unterschiedlicher Typen von Merkmalen erfolgversprechend sind und präzisere Kategorisierungen vornehmen als isolierte Merkmale. Von Bedeutung sind zwei Aspekte: Zunächst sollte die „computational cost“ eines Verfahrens zur Ermittlung eines Merkmals in dieser initialen Phase der Erprobung von Möglichkeiten zur Realisierung neuer Technologien kein beschränkender Faktor sein, zum anderen scheint es – gerade angesichts der großen Anzahl existierender Hypertextsorten (vgl. Teil III) – notwendig zu sein, kaskadierte Verfahren zu entwickeln, die auf *multiplen* Merkmalsräumen operieren. Sobald einander ähnliche Hypertextknotensorten in ein Inventar aufgenommen werden, ist es notwendig, Merkmale für ihre maschinelle Differenzierung zu ermitteln. Ein solches kaskadiertes Verfahren könnte z. B. auf der ersten Ebene eine grobe Kategorisierung nach Hypertextknotentypen vornehmen, woraufhin speziell angepasste Verfahren und auch Merkmalsgruppen zur Differenzierung der jeweiligen Hypertextknotensorten eingesetzt werden (vgl. Dumais und Chen, 2000).

3. *Mangelnde theoretische Fundierung* – Das wohl zentralste Problem sämtlicher Ansätze betrifft ihre mangelnde theoretische Fundierung. Einerseits linguistische und andererseits erkennungsbezogene Aspekte von Genres bzw. Textsorten werden unmittelbar in das WWW übertragen, ohne die Frage zu diskutieren, welche Spezifika Web-Genres besitzen, was sie letzten Endes von Genres unterscheidet und welche dieser Eigenheiten für die maschinelle Erkennung relevant sind bzw. diese negativ beeinflussen könnten. Das Konzept „Genre“ wird in der Regel lediglich als Komplement des Textthemas aufgefasst (vgl. auch Santini, 2004c, S. 6). Shepherd et al. (2004) und Kennedy und Shepherd (2005) beziehen ihr Merkmalsinventar zwar auf die drei Dimensionen „content“, „form“ und „functionality“ (vgl. Abschnitt 4.4.3), diese groben Ebenen können jedoch keinesfalls als theoretisches Fundament eines adäquaten Kategorisierungsverfahrens betrachtet werden. Die mangelnde theoretische Fundierung hängt unmittelbar mit der nachfolgend diskutierten Problematik zusammen.

4. *Das einzelne HTML-Dokument als Analyseeinheit* – Alle bislang vorgeschlagenen Verfahren verwenden einzelne HTML-Dokumente als atomare Analyseeinheit, d. h. weder die übergreifende Ebene des zugehörigen Hypertextes noch die untergeordnete Ebene eingebetteter Bausteine wird beachtet. Diese Kritik bezieht sich nicht auf den notwendigen Einsatz von Hyperlinks als Erkennungsmerkmale (z. B. die Frequenzen interner oder externer Verknüpfungen), sondern auf die Analyse zugehöriger Dokumente. Anders formuliert: Alle Verfahren operieren *ausschließlich* auf der Ebene der Hypertextknotentypen und -sorten. Die korrespondierenden Hypertextsorten werden ebenso ignoriert wie Hypertextsortenmodule. Sowohl die Darstellung des Hypertextsortenmodells (Kapitel 5) als auch die in Teil III dargestellten Analysen zeigen allerdings, dass auch hinsichtlich der maschinellen Erkennung eine Differen-

zierung dieser drei Ebenen zwingend notwendig ist (vgl. Abschnitt 4.5). Erste Ansätze zur Durchbrechnung dieses monolithischen Paradigmas finden sich bei Lim et al. (2005a,b), die zwar HTML-Dokumente in unterschiedliche Bestandteile partitionieren (TM, ANCH, BODY), jedoch keine Analyse der Makrostruktur vornehmen, um „subatomare" Einheiten zu ermitteln. Diese (textuellen und peritextuellen) Bausteine der Makrostruktur werden vor der Verarbeitung durch entsprechende Filter entfernt (z. B. von Rauber und Müller-Kögler, 2001, vgl. auch die Fußnoten 12 und 16) oder führen bei einer Berücksichtigung *sämtlicher* Teile eines HTML-Dokuments oftmals zu fehlerhaften Kategorisierungen: Wenn das maschinelle Lernverfahren für ein spezifisches Web-Genre – insbesondere gilt diese Problematik für die unterschiedlichen Typen von Einstiegsseiten – Trainingsdokumente einbezieht, die typische Konfigurationen bezüglich der ausgeprägten Hypertextsortenmodule besitzen, kann ein Dokument, das diesem Web-Genre zugehörig ist, möglicherweise nicht korrekt kategorisiert werden, weil eine Variation bezüglich der unmittelbaren Ausprägung von Hypertextsortenmodulen vorliegt. Mit anderen Worten: Wenn ein Hypertextsortenmodul nicht in einer zu verarbeitenden Einstiegsseite enthalten ist, sondern durch einen Hyperlink verknüpft wird, kann dieser Umstand zu einer fehlerhaften Kategorisierung führen – durch eine derartige Auslagerung in eigenständige Knoten ändert sich jedoch das Web-Genre (d. h. die Hypertextknotensorte) des ursprünglichen HTML-Dokuments nicht.

5. *Die verwendeten Methoden und ihre Skalierbarkeit* – Maschinelle Lernverfahren werden in nahezu allen Arbeiten zur Generalisierung der als repräsentative Vertreter fungierenden Trainingsdokumente eingesetzt, um anschließend unbekannte Dokumente mit den Kategorienprofilen zu vergleichen und ihnen ein Web-Genre zuzuweisen. Für ein begrenztes Inventar von Web-Genres liefern diese Verfahren befriedigende bis gute Resultate. Die Präzision sämtlicher Lernverfahren nimmt jedoch bei steigender Kategorienanzahl ab, so dass sich z. B. bezüglich der in Kapitel 12 dargestellten Ergebnisse die Frage stellt, mit welchen Methoden eine zufriedenstellende Kategorisierung von ca. 120 Hypertexttypen bzw. Hypertextsorten und ebenso vielen Hypertextknotensorten erzielt werden kann. Neben der Notwendigkeit, neuartige Erkennungsmerkmale zu konzipieren, ist zusätzlich abzusehen, dass ein einzelner Klassifikator auf der Basis *eines* Lernverfahrens nicht in der Lage sein wird, mit derartig umfangreichen Kategorieninventaren umgehen zu können. Zudem zeigen die von Lim et al. eingesetzten Merkmale C1–C11 (vgl. Tabelle 14.3), dass die oftmals durchgeführte Verwendung von POS-Frequenzen von zusätzlichen Verfahren komplementiert werden muss. Lim et al. erheben die Anzahl der Vorkommen von „expressions" – hierbei handelt es sich vermutlich um einfache reguläre Ausdrücke zur Erkennung von Datums- und Zeitausdrücken, Adressen, Telefonnummern etc. Diese Merkmale zeigen ebenfalls die Tendenz zur Ermittlung „subatomarer" Bausteine auf, zugleich verdeutlichen sie, dass die Identifizierung von Hypertextsorten auf Methoden aus dem Bereich der Informationsextraktion (IE) angewiesen ist, die sich zur Identifizierung von Hypertextsortenmodulen anbieten.

Abschließend kann festgehalten werden, dass die maschinelle Ermittlung von Web-Genres „a very difficult task" ist (Shepherd et al., 2004, S. 239), dessen Komplexität durch die zusätzliche Annahme eines empirisch erhobenen Inventars von Kategorien und die Einbeziehung der zentralen Variationsebenen um ein *Vielfaches* steigt. Der nachfolgende Abschnitt stellt eine Systemarchitektur vor, die auf den Ergebnissen der vorangegangenen Kapitel beruht. Daraufhin werden die Verarbeitungsprozesse erläutert.

14.3 Eine Architektur zur computerlinguistischen Erkennung und Verarbeitung von Hypertextsorten

Dieser Abschnitt präsentiert eine Architektur (vgl. Abbildung 14.1) für ein System zur computerlinguistischen Erkennung und weiterführenden texttechnologischen Verarbeitung von Hypertextsorten bzw. Instanzen von Hypertextsorten. Die Architektur wurde einerseits aus der Kritik der verwandten Arbeiten abgeleitet (vgl. Abschnitt 14.2.4), andererseits basiert sie auf den bislang in dieser Arbeit dargestellten Ergebnissen. Darüber hinaus knüpft sie an die Funktionalität der Korpusdatenbank an (vgl. Kapitel 7).

Da sich die Architektur auf das Paradigma der Texttechnologie bezieht (vgl. Lobin und Lemnitzer, 2004), ist es im ersten Schritt notwendig, die HTML-Dokumente zu normalisieren, da Webseiten oftmals fehlerhafte Auszeichnungen auf der Ebene der HTML-Struktur enthalten. Die Normalisierung erfolgt durch eine maschinelle Konvertierung der Dokumente nach XHTML (Pemberton, 2002). Da XHTML eine Anwendung von XML darstellt, liegt im Falle einer erfolgreichen Konvertierung eine wohlgeformte XML-Instanz vor, die mit beliebigen XML-Werkzeugen verarbeitet werden kann (vgl. Myllymaki, 2001).

An die Konvertierung schließt sich ein Textparser an, der in der Lage sein sollte, auf der Grundlage des HTML-Elementbaums und der textuellen Inhalten die auf der Textoberfläche verwendeten Hypertextmodule zu ermitteln (vgl. Abschnitt 5.6.2). In dieser Verarbeitungsstufe können weitere computerlinguistische Komponenten hinzugezogen werden: Da unterschiedliche Typen von Hypertextmodulen existieren, kann auf der Basis dieser Differenzierung ein POS-Tagger z. B. nur auf diejenigen Hypertextmodule angewendet werden, die auch tatsächlich Text (im Sinne einer Sequenz von Wörtern) enthalten, wohingegen Hypertextmodule, die eher Daten enthalten, mit spezifischen Erkennungsausdrücken untersucht werden (Lim et al., 2005a,b). Der Tokenisierer sollte eine maschinelle Satzgrenzenerkennung beinhalten (vgl. etwa Grefenstette und Tapanainen, 1994, und Palmer, 1994, 2000), so dass nachfolgenden Komponenten Textstatistiken wie z. B. die Anzahl der Sätze eines Absatzes oder die Anzahl Wörter pro Satz als Merkmale zur Verfügung gestellt werden können (vgl. Abschnitt 14.2). Das Annotierungsformat der ermittelten Hypertextmodule entspricht einer Analyse-DTD, die das in der Hypertextsortenontologie enthaltene Vokabular zur Repräsentation von Hypertextmodulen reflektiert (vgl. Abschnitt 13.5.8). Da nun nicht mehr HTML-, sondern XHTML-Dokumente verarbeitet werden, können die Analyseergebnisse durch die Einführung eines zusätzlichen Namensraumes direkt in das Dokument integriert werden, indem XML-Elemente und -Attribute die Resultate der Analyse verkapseln.

Sobald für eine Gruppe von HTML-Dokumenten, die einen Hypertext konstituieren, alle Hypertextmodule ermittelt sind, können diese Bausteine auf Hypertextsortenmodule abgebildet werden. Hierfür können die in Abschnitt 13.5.8 dargestellten Verfahren eingesetzt werden, d. h. die Hypertextsortenontologie dient als semantische Ressource für Inferenzverfahren, um für einen gegebenen Hypertext sukzessive die Instanzen von Hypertextsortenmodulen zu bestimmen, aus denen wiederum auf die verwendete Hypertextsorte geschlossen werden kann. Die Komponenten zur Verarbeitung der drei Ebenen sollten ineinander greifen: Wenn z. B. vor der Detektion der Hypertextsortenmodule die Hypertextsorte eines Hypertextes bestimmt werden kann, sollte diese Information auch den anderen Komponenten zur Einschränkung des Suchraumes zur Verfügung stehen. Dies gilt auch für die Erkennung von

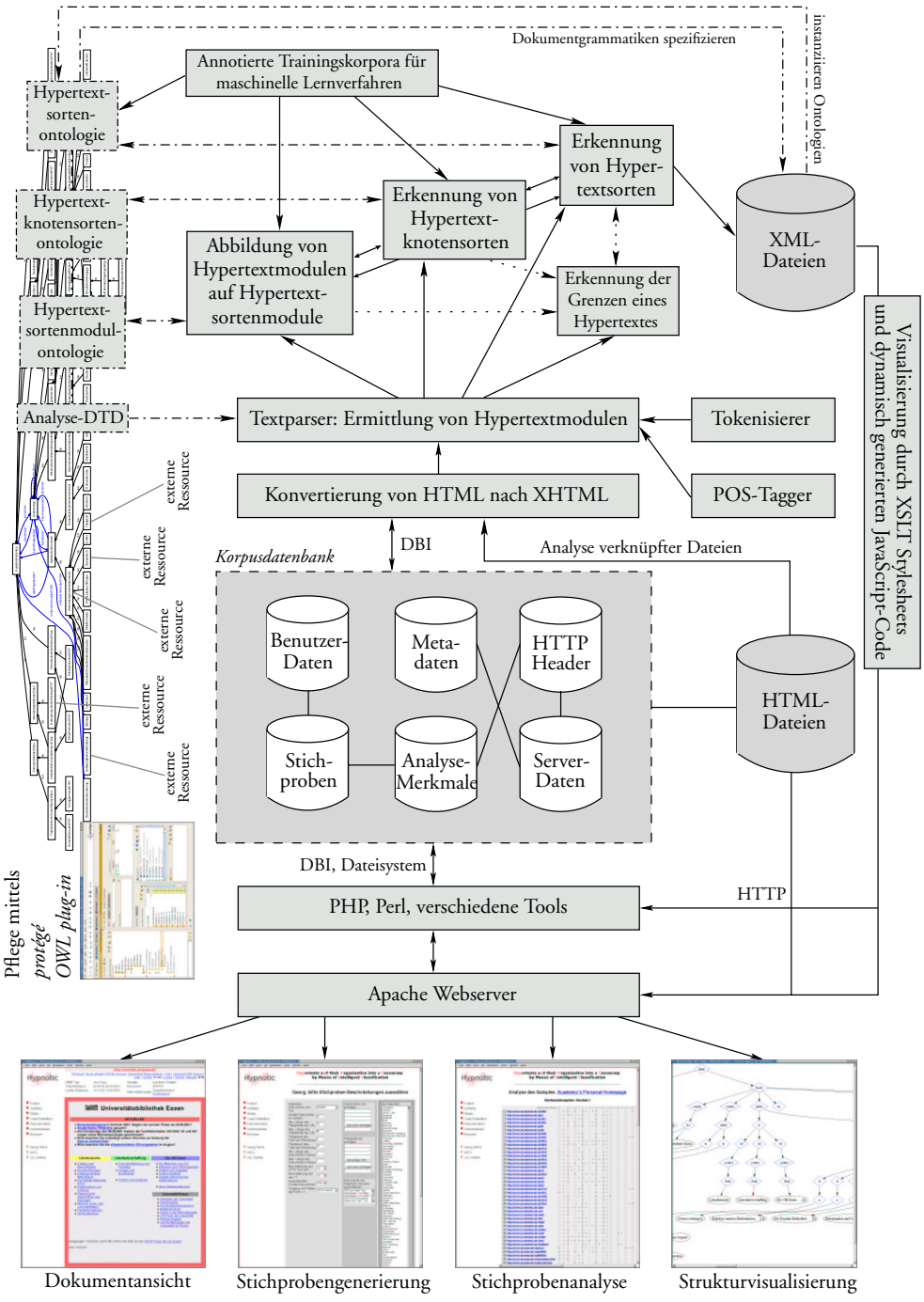


Abbildung 14.1: Architektur der Erkennung und Verarbeitung von Hypertextsorten

Hypertextknotensorten, denn wenn für ein Dokument bereits die Information vorliegt, dass es sich um eine Einstiegsseite handelt, kann – je nach Hypertextsorte – eine spezifische Konfiguration von Hypertextsortenmodulen antizipiert und andere Hypertextsortenmodule ausgeschlossen werden. Umgekehrt können auch zunächst sämtliche Hypertextsortenmodule bestimmt werden, um daraufhin die jeweiligen Hypertextknotensorten und schließlich die Hypertextsorte zu ermitteln. Da zu erwarten ist, dass für unterschiedliche Hypertexttypen und -sorten unterschiedliche Verfahren einzusetzen sind, sollte eine Implementierung maximale Flexibilität gewährleisten. Die Hypertextsortenontologie kann in diesem Zusammenhang nicht nur zur Repräsentation von Informationen über die Konstituenten von Hypertextsorten verwendet werden. Über Annotationen können Klassen der Ontologie mit externen Ressourcen verknüpft werden. Diese ebenfalls in texttechnologischen Formaten (z. B. RDF) repräsentierten Ressourcen können – ähnlich der von Potok et al. (2002) beschriebenen Vorgehensweise – beliebige Informationen enthalten (z. B. Wortfrequenzen, spezifische Kollokationen, charakteristische Schlüsselwörter, reguläre Ausdrücke etc.), die zur Erkennung von Konstituenten eingesetzt werden können. Gerade bei der Abbildung von Hypertextmodulen auf Hypertextsortenmodule sollten derartige Ressourcen benutzt werden, um z. B. von der Erkennung eines abstrakten Eigennamenmusters wie „Prof. Dr. *Vorname Nachname*“ in dem Hypertextmodul *Überschrift* auf das atomare Hypertextsortenmodul *Name des Homepage-Besitzers* (Teil des komplexen Hypertextsortenmoduls *Identifikation*, vgl. Tabelle 10.3, S. 457) zu schließen. Die Hypertextsortenontologie und die enthaltenen weiterführenden Informationen (hierzu gehören z. B. auch manuell annotierte Trainingskorpora für Lernverfahren) werden als *domänenabhängige Ressourcen* verstanden, wohingegen das Analysesystem *domänenunabhängige Verfahren* implementiert. Die Adaption des Systems an eine neue Domäne erforderte somit die Untersuchung zugehöriger Webauftritte, die Aggregierung der Ergebnisse in Form einer Hypertextsortenontologie bzw. ihre Integration in eine bestehende Hypertextsortenontologie und die Spezifizierung externer Ressourcen.

Die Erkennung der Grenzen von Hypertextsorten besitzt einen besonderen Stellenwert. Eine derartige Komponente bestimmt – gegebenenfalls unterstützt durch partielle Ergebnisse der drei Prozesse zur Hypertextsortenerkennung – diejenigen Knoten, die ein Hypertext umfasst. Für eine umfangreiche Kollektion von HTML-Dokumenten, z. B. sämtliche Dokumente eines universitären Webauftritts – müsste dieses Werkzeug also Hypertexte und *eingebettete Hypertexte* auf beliebigen Ebenen präzise identifizieren können, um die jeweiligen Knoten den weiterführenden Prozessen zur Verfügung zu stellen. Ein derartiges Werkzeug kann von unterschiedlichen Merkmalen Gebrauch machen (z. B. die URLs der Dokumente und rekurrent verwendete Strukturmuster, etwa identische Kopf- oder Fußzeilen).

Parallel zu den hier nur sehr abstrakt skizzierten Komponenten könnte eine themenbasierte Klassifikation stattfinden, die HTML-Dokumenten inhaltliche Kategorien zuweist, die z. B. der Ontologie wissenschaftlicher Themen und Fachgebiete entnommen werden. Zur Realisierung von Recherchen sollte ebenfalls ein IR-Verfahren integriert werden, das die Filterung nach Hypertextsorten erlaubt, so dass zunächst sämtliche Hypertextsorten, Hypertextknotensorten und Hypertextsortenmodule einer Kollektion ermittelt und anschließend dem IR-Verfahren als zusätzliche Metadaten zur Verfügung gestellt werden.[24]

[24] Diese beiden Prozesse werden in Abbildung 14.1 nicht dargestellt.

Der Textparser reichert ein XHTML-Dokument mit Informationen an, die als XML-Elemente und -Attribute in den Elementbaum integriert werden. Die Abbildung von Hypertextmodulen auf Hypertextsortenmodule erfolgt durch eben diese Informationen, so dass eine zusätzliche Anreicherung des XHTML-Dokuments stattfindet. Diese kann sich ebenfalls auf die Ebenen der Hypertextknotensorte und der Hypertextsorte beziehen, d. h. letztendlich strebt die in Abbildung 14.1 dargestellte Architektur eine Überführung von HTML-Dateien in XML-Dateien an, die als Informationsextraktionsproblem konzeptualisiert werden kann. Die zugehörigen Dokumentgrammatiken können maschinell aus der Hypertextsortenontologie gewonnen werden (vgl. Abschnitt 13.5.8). Die vollständige Erkennung und Annotierung der Instanz einer Hypertextsorte entspräche der Aggregierung aller beteiligten HTML-Dokumente in Form einer synthetischen XML-Dokumentinstanz, die alle zugehörigen Knoten bündelt. Bezogen auf die Web-Oberfläche der Korpusdatenbank können nun Visualisierungsverfahren angewendet werden, um dem Benutzer die annotierten Strukturen zu präsentieren oder sie weiteren Verarbeitungsprozessen zur Verfügung zu stellen.

Die erste Phase der in diesem Abschnitt vorgestellten Architektur wurde als *proof-of-concept* implementiert. Es existiert jedoch noch eine Vielzahl ungelöster und unbearbeiteter Probleme, die eine vollständige, für ein Produktionssystem geeignete Implementierung derzeit noch verhindern. Zu diesen Problemen zählt die Tatsache, dass bislang kein Verfahren zur Ermittlung der Grenzen von Hypertexten und insbesondere eingebetteten Hypertexten vorliegt. Darüber hinaus beinhaltet die Architektur zahlreiche Prozesse und Komponenten, die für eine extrem hohe Komplexität des Gesamtsystems verantwortlich sind, so dass die Realisierung einer *robusten* Implementierung mit einem *hohen* Abdeckungsgrad derzeit fraglich erscheint (vgl. auch die Abschnitte 13.5.8 und 13.7). Die nachfolgenden Abschnitte gehen auf die einzelnen Komponenten ein, die zur Hypertextsortenerkennung benötigt werden.

14.4 Konvertierung von HTML-Dokumenten nach XHTML

Die im vorangegangenen Abschnitt thematisierte Systemarchitektur basiert auf einer texttechnologischen Methodologie zur Identifizierung und weiterführenden Verarbeitung von Hypertextsorten.[25] Die erste Komponente führt eine Konvertierung beliebiger HTML-Dokumente nach XHTML durch. Es entstehen wohlgeformte XML-Dokumente, die mit XML-Werkzeugen verarbeitet werden können. Abschnitt 14.4.1 geht auf die Implementierung des Konverters ein, woraufhin Abschnitt 14.4.2 die Visualisierung der Baumstruktur erläutert.

14.4.1 Implementierung des Konverters

Das *Perl*-Modul `Hypnotic::HTML2XHTML` ist in der Lage, arbiträre HTML-Dokumente nach XHTML zu konvertieren (vgl. Myllymaki, 2001) und wird für den Textparser (vgl. Abschnitt 14.5), in der Web-Oberfläche der Korpusdatenbank (vgl. Abschnitt 7.3) und beim indirekten Korpuszugriff eingesetzt (vgl. Abschnitt 7.4). Das Modul verkapselt eine Heuristik, die eine modifizierte und speziell konfigurierte Version des Werkzeugs *Tidy* und das

[25] Verschiedene Aspekte der in diesem Abschnitt sowie dem nachfolgenden Abschnitt 14.5 vorgestellten Ansätze wurden in Rehm (2004b) publiziert.

```
1  package Hypnotic::HTML2XHTML;
2
3  use strict;
4  use Exporter;
5  use vars qw(@ISA @EXPORT @EXPORT_OK $VERSION);
6
7  @ISA        = ('Exporter');
8  @EXPORT     = qw(&html2xhtml);
9  @EXPORT_OK  = qw(&html2xhtml);
10 $VERSION    = 0.02;
11
12 my $tidycfg = "/home/georg/Dissertation/perllib/Hypnotic/tidy.cfg";
13 my $tidy    = "/usr/local/bin/tidy";  # path to our tidy binary
14 my $errout  = "";                     # may contain errors from 2 passes
15
16 sub recover {
17     my ($html_in) = @_;
18     use HTML::TreeBuilder;
19     my $t = HTML::TreeBuilder->new(); # construct new tree object
20     $t->store_comments(1);            # we want to keep comments
21     $t->no_space_compacting(1);       # do not compact whitespace
22     $t->parse($html_in);              # parse the data
23     $t->eof();
24     my $html_out = $t->as_HTML;       # convert
25     $t->delete;
26     return $html_out;
27 }
28
29 sub postprocess {
30     my ($input) = @_;
31     $input =~ s/\014/ /g;             # ^L -> " "
32     $input =~ s/\x0c/ /g;             # ^L -> " "
33     $input =~ s/\cL/ /g;              # ^L -> " "
34     $input =~ s/\r/ /g;
35     $input =~ s/\x00/ /g;             # ^@ -> " "
36     $input =~ s/\c[/ /g;              # ^[ -> " "
37     $input =~ s/\x0b/ /g;             # ^V -> " "
38     $input =~ s/\013/ /g;             # ^V -> " "
39     return $input;
40 }
```

Listing 14.1: Das Perl-Modul `Hypnotic::HTML2XHTML` (Fortsetzung in Listing 14.2)

im *Comprehensive Perl Archive Network* (CPAN, `http://www.cpan.org`) erhältliche Modul `HTML::TreeBuilder` einsetzt (vgl. die Listings 14.1 und 14.2).[26]

Das Werkzeug *Tidy* ist in der Lage, HTML-Dokumente nach XHTML konvertieren zu können.[27] *Tidy* und `HTML::TreeBuilder` beinhalten Regeln für einen robusten Umgang mit fehlerhaften HTML-Strukturen, die im WWW sehr häufig vorliegen, da praktisch alle Browser fehlertolerante Parsing-Strategien für – aus Sicht der mit Hilfe von SGML (ISO 8879) spezifizierten HTML-DTDs – defekten HTML-Code beinhalten. Da *Tidy* diesen Konvertie-

[26] Im Zuge der Auswahl der Werkzeuge wurden auch Alternativen getestet, etwa die *Perl*-Module `XML::PYX` und `XML::Driver::HTML` sowie das UNIX-Tool *recode*, die jedoch keine zufriedenstellenden Resultate liefern.

[27] Siehe hierzu auch die Studie von Dollar Consulting (2002), in der verschiedene technische Möglichkeiten zur dauerhaften Archivierung der zahlreichen „virtual exhibits“ der Smithsonian Institution diskutiert werden, die als HTML-Dokumente vorliegen. Als Archivformat wird XHTML kombiniert mit *tar* (*tape archiver*) vorgeschlagen, zur Migration empfehlen die Autoren *Tidy* in der DOS-Ausführung.

```
41 sub html2xhtml {
42     use IO::Select;
43     use IO::Handle;
44     use IPC::Open3;
45     my ($html, $c) = @_;              # the HTML-document
46     my $xhtml  = "";                  # construct bidirectional pipe
47     my $pid = open3(\*W, \*R, \*E, "$tidy -config $tidycfg -") or
48         die "Couldn't open3() $tidy:$!";
49     my $stdout = IO::Select->new();
50     my $stderr = IO::Select->new();
51     $stdout->add(\*R);
52     $stderr->add(\*E);
53     print W "$html";
54     close(W);
55     while (1) {                       # stdout is ready for reading
56         if (my ($who) = $stdout->can_read(0)) {
57             $who->read(my $out, 10000);
58             last if (!($out));        # readable but empty == close
59             $xhtml .= $out;           # collect stdout
60         }
61         if (my ($who) = $stderr->can_read(.01)) {
62             $who->read(my $err, 10000);
63             $errout .= $err;          # collect stderr
64         }
65         last if (my ($ex) = $stdout->has_exception(0));
66         last if (my ($ex) = $stderr->has_exception(0));
67     }
68     waitpid $pid, 0;                  # wait until child is finished
69     $stdout->remove(\*R);             # unregister these handles
70     $stderr->remove(\*E);
71     if ($errout) {
72         if (defined($c)) {            # finished 2nd pass through tidy
73             if ($c == 1) {
74                 my $error = $errout;
75                 $errout = "";
76                 $xhtml = postprocess($xhtml);
77                 return ($xhtml, $error);
78             }
79         }
80         else {                        # tidy complained about errors
81             $html = recover($html);   # activate error-recovery
82             html2xhtml($html,1);      # run data through tidy again
83         }
84     }
85     else {
86         $xhtml = postprocess($xhtml);
87         return ($xhtml, $errout);     # errout _must_ be empty now
88     }
89 }
```

Listing 14.2: Fortsetzung von Listing 14.1 (gekürzt)

rungsprozess mit einer höheren Präzision als `HTML::TreeBuilder` ausführt, wird zunächst versucht, das HTML-Dokument mit Hilfe dieses Werkzeugs einzulesen und nach XHTML zu konvertieren. Gelingt dies nicht, wird `HTML::TreeBuilder` als *fallback*-Komponente eingesetzt, um ein fehlerfreies HTML-Dokument zu erzeugen, das daraufhin erneut mittels *Tidy* nach XHTML konvertiert wird. Eine Evaluierung mit 10 000 zufällig ausgewählten HTML-Dokumenten zeigt, dass auf diese Weise 98,7% aller Dateien korrekt konvertiert werden, wobei in 270 Fällen `HTML::TreeBuilder` eingesetzt wurde. Die Wohlgeformtheit der XHTML- bzw. XML-Dokumente wurde daraufhin mit dem nicht validierenden XML-Parser *expat* überprüft. Lediglich fünf der 9 872 erfolgreich konvertierten Dokumente erzeugen dabei eine Fehlermeldung, die in fast allen Fällen auf Binärzeichen zurückzuführen sind, die einen Konflikt mit dem Unicode-Zeichensatz UTF-8 hervorrufen. Abschließend wurde versucht, die Dokumente mit dem XML/SGML-Parser `onsgmls`[28] gegen die XHTML 1.0 Transitional-DTD zu validieren. Fehler, die sich auf die Schachtelung von Elementen beziehen, sind kaum aufgetreten; das Gros der Fehlermeldungen resultierte aus unerlaubten Attributwerten (sehr häufig z. B. bei `valign`) oder nicht existenten und folglich im Standard nicht deklarierten Elementen (z. B. `blink` und `spacer`; vgl. Abschnitt A.4.2).

Die Listings 14.1 und 14.2 stellen die Funktionen des Moduls `Hypnotic::HTML2XHTML` in gekürzter Form dar. Die Routine `html2xhtml()` nimmt als Argument eine HTML-Datei entgegen, vollzieht die Konvertierung nach XHTML und liefert das XML-Dokument zurück (Zeile 87). Es wird ein *Tidy*-Prozess erzeugt, der mittels Interprozesskommunikation (bereitgestellt durch `IPC::Open3`) in Form bidirektionaler Pipelines angesteuert wird.[29] Zur Umlenkung der Datei-Handles werden `IO::Select` und `IO::Handle` benutzt: Von der Standardeingabe nimmt der *Tidy*-Prozess das HTML-Dokument entgegen (Zeile 53), von der Standardausgabe (Zeilen 56–60) wird sukzessive die Ausgabe – d. h. das XHTML-Dokument – eingelesen und auf der Standardfehlerausgabe (Zeilen 61–64) werden etwaige Fehlermeldungen abgefangen. Falls Fehler vorliegen (Zeilen 80–83), ist *Tidy* nicht ohne Weiteres in der Lage, die Eingabe nach XHTML zu konvertieren, weshalb das HTML-Dokument an `recover()` übergeben wird. Diese Funktion erzeugt ein Objekt vom Typ `HTML::TreeBuilder`, das mit einer alternativen Strategie versucht, die nicht wohlgeformte HTML-Quelle zu korrigieren. Anschließend wird erneut *Tidy* aufgerufen, um das auf diese Weise korrigierte Dokument nach XHTML zu konvertieren. Die Funktion `postprocess()` ersetzt verschiedene Steuerzeichen durch Leerzeichen, um in nachgeschalteten Werkzeugen (z. B. *expat*) einen Konflikt bezüglich der Zeichensatzkodierung UTF-8 zu vermeiden.

14.4.2 Visualisierung des Elementbaumes

HTML-Elemente spannen eine Baumstruktur auf. Da der Textparser eine Manipulation dieser Struktur vornimmt, war es notwendig die manipulierte mit der ursprünglichen Baumstruktur vergleichen zu können. Für diesen Zweck wird das *PHP*-Skript `showtree.php` eingesetzt, das das *Perl*-Skript `h2x2g.pl` aufruft. Dieses Skript produziert aus der HTML-Quelle

[28] Hierbei handelt es sich um eine erweiterte Version des Parsers `nsgmls`, der Bestandteil des von James Clark implementierten *SP*-Pakets ist (vgl. `http://openjade.sourceforge.net` und `http://www.jclark.com/sp/`).

[29] Der Einsatz eines ausführbaren Programms war notwendig, da sich die Library-Version von *Tidy* (*TidyLib*), die von einem kapselnden *Perl*-Modul flankiert wird, zum Zeitpunkt der Implementierung des hier vorgestellten Verfahrens noch in der Entwicklung befand (vgl. `http://tidy.sourceforge.net`).

eine GIF-Datei, die die Baumstruktur visualisiert. Da `showtree.php` einen HTTP-Response-Header generiert, der den nachfolgenden Datenstrom als `Content-Type: image/gif` kennzeichnet, kann die Ausgabe von `h2x2g.pl` unmittelbar an die Standardausgabe erfolgen.

Das *Perl*-Skript `h2x2g.pl` arbeitet in zwei Schritten: Zunächst findet eine Konvertierung nach XHTML-Transitional (Pemberton, 2002) statt. Das *Perl*-Modul `GraphViz` ist in der Lage, Graphen zu erzeugen, wozu das Werkzeug *dot* aus dem *GraphViz*-Paket[30] benutzt wird. Das *Perl*-Skript verwendet eine modifizierte Fassung des Moduls `GraphViz::XML`, um die Struktur einer XML-Datei in Form eines als GIF-Datei realisierten Baumes zu erzeugen. Bei dieser Darstellung handelt es sich um eine vereinfachte grafische Repräsentation der DOM-Struktur (*Document Object Model*, Hors et al., 2000) eines XML-Dokuments.[31]

Die zentralen Bestandteile des Skriptes `h2x2g.pl` sind in den Listings 14.3 und 14.4 dargestellt. Das HTML-Dokument wird auf der Standardeingabe übergeben, woraufhin die Konvertierung nach XHTML stattfindet. Der XML-Parser *expat* wird daraufhin in Form des Moduls `XML::Parser` eingesetzt, um das XHTML-Dokument für das Modul `GraphViz::XML` vorzubereiten. Mittels `setHandlers()` werden Funktionen mit Ereignissen verknüpft. Ein `Start`-Event wird z. B. ausgelöst, wenn *expat* auf ein öffnendes Element trifft. Die Vorverarbeitung umfasst die Tilgung verschiedener Sonderzeichen, die Erzwingung einer Latin 1-Kodierung[32] sowie die Reduktion des enthaltenen Markups und auch Textes, d. h. Attribute und die Inhalte von `script`- und `style`-Elementen werden entfernt und in Elementen verfügbare textuelle Informationen werden mit Hilfe der Funktion `flush_bag()` auf die ersten drei Wörter reduziert, wobei die Anzahl der getilgten Wörter in eckigen Klammern notiert wird (vgl. Abbildung 14.2).[33] Zusätzlich wird `HTML::Entities` eingesetzt, um HTML-Entitäten dekodieren zu können, so dass in der zu generierenden GIF-Datei die referenzierten Sonderzeichen enthalten sind. Die vier Funktionen, die an die Event-Handler des XML-Parsers gebunden sind, erzeugen eine reduzierte Kopie der Eingabe-Datei, die in der Skalarvariable `$xmlpp` abgelegt wird. Dieses XML-Dokument kann dann mit Hilfe des Moduls `GraphViz::XML` und der Methode `as_gif()` in eine GIF-Datei überführt werden.

14.5 Generisches Textparsing arbiträrer HTML-Dokumente

Um den Bestandteilen eines HTML-Dokuments Hypertextsortenmodule zuweisen zu können, ist es notwendig, zunächst die Elemente der Textoberfläche zu ermitteln. Die *Hypertext Markup Language* (Raggett et al., 1999) stellt zwar eine Anwendung von SGML dar

[30] Das Paket *GraphViz* (Gansner und North, 2000) ermöglicht die Generierung von Graphen (`http://www.graphviz.org`, `http://www.research.att.com/sw/tools/graphviz/`). In der Graphenbeschreibungssprache DOT werden Knoten und Kanten spezifiziert, woraufhin die Werkzeuge *dot* (für gerichtete Graphen) und *neato* (für ungerichtete Graphen) mit Hilfe von Layout-Algorithmen visuelle Repräsentationen erzeugen.

[31] Es handelt sich um eine vereinfachte DOM-Darstellung, weil lediglich die Baumstruktur sowie die Namen von Elementen und Ausschnitte des enthaltenen Textes dargestellt werden. Weitere Informationen, die in DOM eigene Knoten darstellen, werden aus Platzgründen nicht visualisiert, da derartige Grafiken – auch ohne die Aufnahme von z. B. Entitäts- oder Attributinformationen – selbst bei kleinen HTML-Dateien sehr große Abmessungen annehmen (z. B. 14 656 × 591 Pixel für das Dokument aus Abbildung 7.8, S. 347).

[32] Mit Hilfe des Moduls `Unicode::String` sind Konvertierungen zwischen den Zeichensatzkodierungsarten Latin 1 und Unicode (benutzt wird UTF-8) möglich (vgl. auch Sasaki und Witt, 2004).

[33] Abbildung 14.2 basiert nicht auf einer GIF-Datei. Stattdessen wurde `h2x2g.pl` so modifiziert, dass das `GraphViz`-Modul eine Postscript-Ausgabe generiert, da hierdurch eine bessere Darstellungsqualität erzielt wird.

```
#!/usr/local/bin/perl -w

use strict;
use XML::Parser;                    # to preprocess the document
use HTML::Entities;
use GraphViz::myXML;                # uses "latin1" as input_filter
use Unicode::String qw (utf8 latin1);
use Hypnotic::HTML2XHTML;           # HTML -> X(HT)ML conversion

my ($threshold, $trim) = (3, 1);    # put three words in one node
my ($input, $xmlpp)    = ("", "");  # original and processed documents
my $proc = 1;                       # to escape "script" and "meta"
my @bag  = [];                      # token buffer

while (<>) { $input .= $_; }        # read HTML document from stdin

my ($xml, $error) = &html2xhtml($input);

my $p = new XML::Parser(ProtocolEncoding => 'ISO-8859-1');

$p->setHandlers(Start   => \&start,   End  => \&end,
                Default => \&default, Char => \&char);

$p->parse($xml);                    # construct abridged XML document

my $g = GraphViz::XML->new($xmlpp); # generate graph

print $g->as_gif();                 # print GIF to stdout

sub default {
    my ($p, $string) = @_;
    if ($string =~ /^\&/) {         # process entities only
        $string = decode_entities($string);
        if (@bag > 0) {             # put entity in @bag if it exists
            push(@bag,"$string");
        }
        else {
            $xmlpp .= "$string";
        }
    }
}

sub start {
    my ($p, $tag, @attr) = @_;
    &flush_bag();
    $tag = utf2lat($tag);           # enforce latin1
    $proc = 0 if (($tag eq "script") || ($tag eq "style"));
    $xmlpp .= "<$tag>\n";           # delete attributes
}

sub end {
    my ($p, $tag) = @_;
    $tag = utf2lat($tag);           # enforce latin1
    $proc = 1 if (($tag eq "script") || ($tag eq "style"));
    &flush_bag();
    $xmlpp .= "</$tag>\n";
}
```

Listing 14.3: Das Perl-Skript `h2x2g.pl` (Fortsetzung in Listing 14.4)

```
58 sub char {
59     my ($p, $text) = @_;
60     return if ($text =~ /^\s+$/);
61     if ($proc == 1) {
62         $text = utf2lat($text);          # enforce latin1
63         while ($text =~ /([^\s]{1,})/g) {
64             push(@bag,$1);               # put words into bag
65         }
66     }
67 }
68 
69 sub utf2lat {
70     my ($in) = @_;
71     my $out = utf8($in);
72     return $out->latin1;                 # enforce latin1
73 }
74 
75 sub flush_bag {
76     if (@bag > 0) {                      # bag of words is filled ...
77         my ($b, $count) = ("", 0);
78         my $current_threshold = $threshold;
79         $current_threshold = @bag - 1 if (@bag < $threshold);
80         if (@bag > $threshold) {
81             $count = @bag - $threshold;  # number of deleted words
82         }
83         else {
84             $count = 0;
85         }
86         for (my $c = 1; $c <= $current_threshold; $c++) {
87             $b .= $bag[$c] . " ";
88         }
89         if ($count > 0) {
90             $b .= " ... [$count]";
91         }
92         $xmlpp .= $b;                    # put text into XML document
93         @bag = [];                       # initialize these two again
94     }
95 }
```

Listing 14.4: Fortsetzung von Listing 14.3 (gekürzt)

(ISO 8879), doch handelt es sich nicht um eine Sprache, die ausschließlich der strukturellen Auszeichnung eines Dokuments dient. HTML basiert *nicht* auf dem für dokumentorientierte SGML- und XML-basierte Auszeichnungssprachen typischen Paradigma der Trennung von Struktur und Präsentation, da mehrere Ebenen der Textauszeichnung vermischt werden und eine Manipulation der Dokumentgestaltung möglich ist (vgl. auch Etzioni, 1996, und Walker, 1999). Aus diesem Grund existiert das Phänomen des *tag abuse* (Barnard et al., 1996), d. h. HTML-Elemente und -Attribute werden oftmals aufgrund ihres typografischen Effekts bezüglich der Darstellung im Browser und nicht zur Explizierung desjenigen Strukturmerkmals verwendet, das von einem Element markiert wird (vgl. Abschnitt 5.3.2).[34]

[34] Siegel (1999b, S. 8) erläutert: „Designer stellen die Gestaltung über die Struktur. Sie haben das strukturorientierte HTML angenommen und dann schrecklich »umgebogen«, damit es ihren visuellen Ansprüchen genügte." Siegel (1999b, S. 75) zufolge sind die von HTML vorgegebenen Möglichkeiten zur Realisierung von Listen ungeeignet: „Deshalb verwende ich auch keine HTML-Tags zur Anzeige von [...] Listen. Statt dessen baue ich Listen visuell auf und habe so die Sicherheit, daß sie so dargestellt werden, wie ich es will."

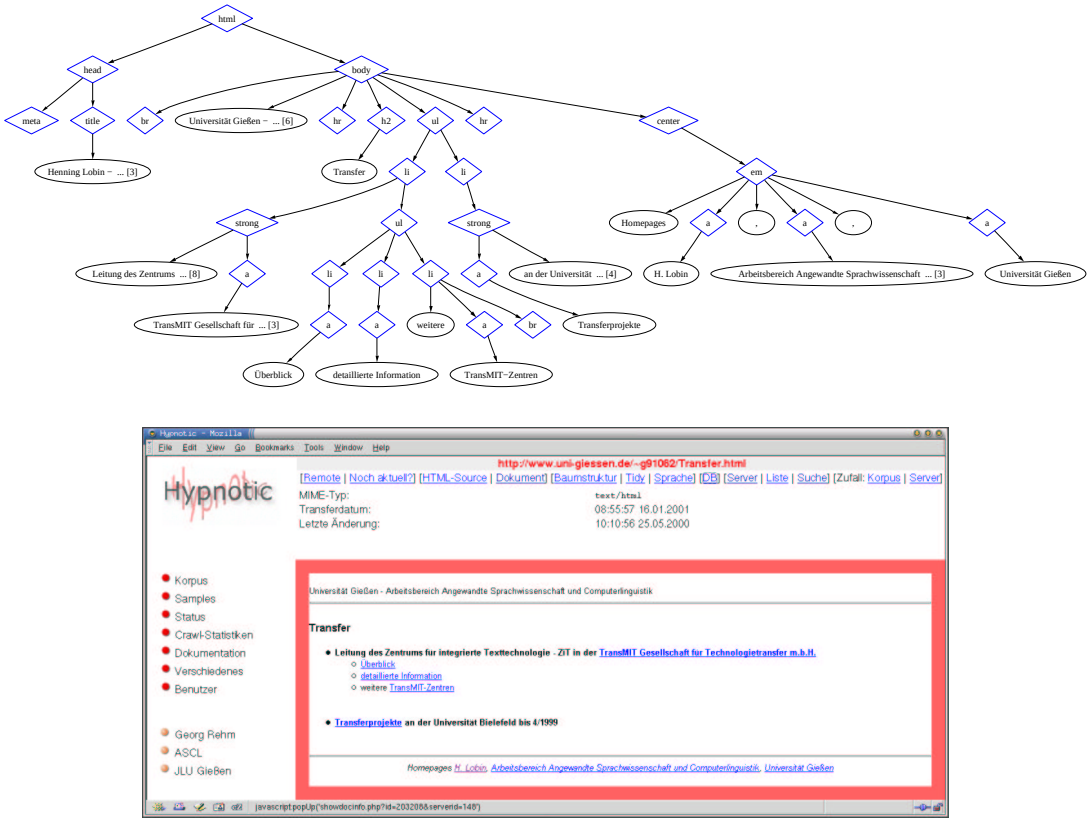


Abbildung 14.2: Die Visualisierung der Baumstruktur eines HTML-Dokuments

Im Folgenden wird eine prototypische Implementierung der Identifizierung von Hypertextmodulen vorgestellt. Abschnitt 14.5.1 vermittelt zunächst einen Überblick über den Textparser, woraufhin verwandte Ansätze zur DOM-basierten Verarbeitung von HTML-Dokumenten diskutiert werden (Abschnitt 14.5.2). Anschließend werden die Funktionsweise des Werkzeugs (Abschnitt 14.5.3) und der Algorithmus thematisiert (Abschnitt 14.5.4). Die durch den Textparser analysierten Informationen können von der Web-Oberfläche der Korpusdatenbank auf unterschiedliche Weise visualisiert werden (Abschnitt 14.5.5).

14.5.1 Der Textparser im Überblick

Die *tag abuse*-Problematik macht es erforderlich, HTML-Dokumente einer Analyse zu unterziehen, die textstrukturell und funktional identische, aber mit *unterschiedlichen* HTML-Elementen realisierte Bausteine ermittelt. Es ist notwendig, die vom Produzenten intendierten Makrostruktureinheiten zu rekonstruieren, da sie den unter Umständen vom *tag abuse* betroffenen HTML-Elementen nicht unmittelbar entnommen werden können (vgl. Eiron und McCurley, 2003, S. 93). Dieser Rekonstruktionsprozess bezieht sich auf eine Analyse der „visuellen Semantik“ (vgl. Fußnote 34) der korrespondierenden HTML-Elemente und -Attribute. Zur Bestimmung der intendierten Funktion eines Bausteins sind insbesondere

typografische Merkmale zu untersuchen. Das in HTML zur Verfügung stehende Inventar ist sehr begrenzt, es sollten jedoch unter anderem Informationen über Navigationshilfen, Überschriften, Fließtext, Textfragmente und Logos zur Verfügung stehen und diese Informationen sollten – bezogen auf ein heterogenes Korpus von Dokumenten – durch ein konsistentes und überschaubares Vokabular von XML-Elementen zugreifbar sein.

Da die initiale Konvertierung in wohlgeformten XML-Dokumenten resultiert (vgl. Abschnitt 14.4), kann die Repräsentation der ermittelten Bausteine unmittelbar in den analysierten Dokumenten erfolgen. Hierfür wird eine Multiebenen-Annotation eingesetzt (vgl. Witt, 2004): In das Dokument wird ein zusätzlicher Namensraum integriert, der auf die Analyse-DTD der Hypertextsortenontologie verweist und eine weitere Auszeichnungsschicht identifiziert (vgl. Abbildung 14.1). Auf diese Weise stehen die Primärdaten zu jedem Zeitpunkt der Verarbeitung zur Verfügung und werden sukzessive um weitere Informationen angereichert, die von dem meist hochgradig komplexen Elementbaum abstrahieren.

Die Analyse basiert auf einer rekursiven Untersuchung der DOM-Struktur (Hors et al., 2000) des im ersten Schritt entstandenen XHTML-Dokuments. Neben der von den HTML-Elementen aufgespannten Baumstruktur werden die *Frequenzen* von Token in ihren Teilbäumen berücksichtigt, es handelt sich also um ein sprachunabhängiges Verfahren. Die DOM-Verarbeitung erfolgt unter anderem durch XPath-Ausdrücke, die für einzelne Teilbäume mehrere Merkmale berechnen, die zur Bestimmung ihrer Textstrukturfunktionen eingesetzt werden können. Sobald eine solche Funktion ermittelt wurde, wird der Teilbaum durch ein XML-Element des angesprochenen Namensraumes verkapselt. Nach Ablauf der Analyse liegt das ursprüngliche XHTML-Dokument vor, das eine zusätzliche Auszeichnungsschicht umfasst, die die vom Produzenten implizit markierten Textoberflächenkomponenten expliziert.

14.5.2 Verwandte Arbeiten

Es existiert eine Vielzahl von Ansätzen zum Parsing und zur Verarbeitung von HTML-Dokumenten. Diese Arbeiten lassen sich in unterschiedliche Kategorien einteilen, die wiederum partiell mit den Komponenten der in Abbildung 14.1 dargestellten Architektur korrespondieren. Für die Ermittlung von Hypertextmodulen sind insbesondere diejenigen Verfahren relevant, die eine domänenunabhängige Verarbeitung von HTML-Dokumenten anstreben, z. B. zur Verbesserung von Such- und Navigationsverfahren und als Vorverarbeitungsstufe für Methoden des Textzusammenfassens. Die korrespondierenden Werkzeuge operieren in der Regel mit texttechnologischen Standards auf den Baumstrukturen, die von den in einem Dokument enthaltenen HTML-Elementen aufgespannt werden. Daneben existieren domänenspezifische Ansätze zur Informationsextraktion – im Hinblick auf HTML-Dokumente ist diesbezüglich von *Wrapping* die Rede (vgl. Rehm, 2004d). Diese Verfahren werden im Kontext der Erkennung von Hypertextsortenmodulen diskutiert (Abschnitt 14.6.3).

Carchiolo et al. (2002, 2003) verarbeiten den HTML-Elementbaum auf der Grundlage von „primary tags“, die „primary nodes“ konstituieren (`table`, `p`, `map`, `hr`, `a`, `frame`, `iframe`, `head`, `form`, `select`, `input`, `td`, `tr`, `option` und `area`). Nach der Berechnung gewichteter Merkmale für jeden primären Knoten (z. B. die relative und absolute Tiefe im Baum, die Anzahl der Blätter und Attribute) werden „collections“ (rekurrente Sequenzen von HTML-Elementen) konstruiert, die abschließend auf „logical sections“ abgebildet werden („document“, „hea-

der“, „footer“, „index“, „logical data“ und „interactive“ sections). Diese werden definiert als „parts of a page each of which collects related information“ (Carchiolo et al., 2003, S. 341). Die Verfasser streben insbesondere die Verbesserung von Such- und Navigationsverfahren an.

Chan und Yu (1999) schildern einen Ansatz zur Extraktion von Eigenschaften des Webdesigns. Zunächst wird eine „canonical form“ des HTML-Markups erzeugt, die auf „primitive“ (`br`, `b`, `i`, `font`, `table` etc.) und „compound tags“ (`h1`–`h6`, `p`, `ol` etc.) beruht. Daraufhin werden „objects“ identifiziert, wozu „object tags“ und Tokenfrequenzen eingesetzt werden. Auf der Grundlage dieser Daten werden drei „design knowledge bases“ konstruiert, die „site layout and navigation“, „web page objects“ und „web page layouts“ umfassen.[35]

In Anlehnung an das *Document Object Model* stellen Chen et al. (2001) das „Function-based Object Model“ (FOM) vor, das zur Konvertierung von HTML-Dokumenten nach WML entwickelt wurde. Da mobile Endgeräte aufgrund ihrer kleinformatigen Displays oftmals nicht in der Lage sind, Webseiten vollständig darstellen zu können, ist es notwendig, redundante Objekte zu entfernen und den Inhalt zu kondensieren (vgl. auch Chen et al., 2003a). Auf der Grundlage der Merkmale eines HTML-Dokuments und der enthaltenen Elemente und Attribute, werden statistische Methoden, Entscheidungsbäume und ein visueller Mustervergleich angewendet (ähnlich bei Song et al., 2004), der auf einer intern erzeugten Darstellung eines Dokuments beruht, um auf der Basis des „basic FOM“ einzelner Objekte ihr „specific FOM“ zu bestimmen.[36] Diese gehören unterschiedlichen Kategorien („specific FOM“) an, z. B. „information object“, „navigation object“, „interaction object“, „special function object“ und „decoration object“. Abschließend werden die Objekte, die als funktional wichtig und nicht redundant eingestuft werden, zu einer WML-Version aggregiert.

Gupta et al. (2003) ermitteln den Inhalt arbiträrer Webseiten, wozu unterschiedliche Filter eingesetzt werden, die die DOM-Struktur rekursiv verarbeiten, um Knoten zu entfernen oder zu modifizieren, so dass ausschließlich die Inhalte beibehalten werden. Der Filter namens „advertisement remover“ entfernt eingebettete Grafiken, deren URLs in einer Liste der Webserver von Werbungsanbietern enthalten sind. Der „link list remover“ entfernt Listen von Hyperlinks, die sich in Tabellenzellen befinden, so dass sich der Leser auf den Inhalt des Dokuments konzentrieren kann. Gupta et al. nennen das automatische Textzusammenfassen, die Anpassung von Dokumenten an mobile Endgeräte und die Aufbereitung einer Webseite für die Sprachsynthese als mögliche Anwendungsgebiete.

Chen et al. (2003b) passen HTML-Dokumente mit einem dreispaltigen Layout an mobile Endgeräte an. Das Ziel betrifft die Rekonstruktion der „content structure based on the clues the author embeds in a web page“, so dass ein einzelnes Dokument in mehrere Webseiten unterteilt werden kann, die auf kleinformatigen Displays effizient navigiert werden können. In der DOM-Analyse wird die „semantic structure“ identifiziert, indem jeder Elementknoten in eine der „content block categories“ namens „header“, „footer“, „left side bar“, „right side bar“ und „body“ klassifiziert wird. Weiterhin werden explizite (z. B. `hr`, `table`, `td` und `div`)

[35] Im Kontext der Ermittlung von Informationen über das Webdesign geben Chan und Yu (1999, S. 550) an: „[W]e currently extract useful object design templates from the object trees by defining “*object genres*.” The genre of an object is extracted by object-specific analysis. [...] Object genres are used to help classify page layouts.“ (Hervorhebung hinzugefügt, G. R.). Dieser Aspekt wird jedoch weder exemplifiziert noch konkretisiert.

[36] Das „basic FOM“ von „basic objects“ („a non-breakable element within two tags“, Chen et al., 2001, S. 590) umfasst funktionale Merkmale wie z. B. „Presentation“, „Decoration“ und „Interaction“. Mehrere „basic objects“ bilden ein „composite object“, denen ebenfalls ein „basic FOM“ zugewiesen wird.

und implizite Grenzen („blank areas created intentionally“) detektiert. Daraufhin wird ein Dokument in der „page adaptation stage“ in mehrere Webseiten aufgeteilt.

Buyukkokten et al. (2001a,b) präsentieren ein Verfahren zur Navigation umfangreicher HTML-Dokumente auf kleinformatigen Displays und zum Textzusammenfassen, das auf der „Akkordeon“-Metapher beruht.[37] Mit dem „Page Parser“, dem „Partition Manager“ und dem „Organization Manager“ werden die „semantic textual units“ ermittelt und zu einer Baumstruktur rearrangiert, die aus den Eigenschaften der Einheiten konstruiert wird. Die reduzierte Dokumentversion enthält lediglich die „semantic textual units“ der obersten Ebene.

Das HTML-Element `table` nimmt eine zentrale Position für das Webdesign ein – es handelt sich um die wohl extremste Form des *tag abuse*, da oftmals nicht tabellarische Daten präsentiert werden. Stattdessen werden komplexe, unsichtbare Tabellenstrukturen konstruiert, die das Layout-Gerüst der Inhalte konstituieren. Es existieren jedoch auch Zwischenstufen, in denen zwar tabellarische Daten präsentiert werden, diese werden jedoch nicht über die korrekte, sondern eine Layout-orientierte Anwendung von `table` realisiert. Mehrere Ansätze beschäftigen sich ausschließlich mit der Verwendung des Elements `table`: Hurst (2002) und Cohen et al. (2002) streben die Differenzierung genuiner Tabellen auf der Basis überwachter Lernverfahren an. Die verwendeten Merkmale basieren auf einer DOM-Repräsentation und umfassen unter anderem „single HTML row“, „single HTML column“ und „bag of tags“. Die Autoren geben einen Recall von 92–96% und eine Präzision von 95–100% an (diese Evaluierung wurde mit 89 positiven und 250 negativen Trainingsbeispielen durchgeführt). Chen et al. (2000), Penn et al. (2001), Wang et al. (2000) und Wang und Hu (2002) beschreiben ähnliche Verfahren.[38] Alam et al. (2003) identifizieren Tabellen, die zur Realisierung von Listen eingesetzt werden. Lim und Ng (1999, S. 466) versuchen, die „intended hierarchy of the data content of the table“ mit DOM- und XPath-ähnlichen Verfahren zu rekonstruieren.

14.5.3 Funktionsweise des Textparsers

Der in *Perl* implementierte Textparser analysiert Merkmale der nach XHTML konvertierten HTML-Dokumente.[39] Hierfür wird das Modul `XML::LibXML` eingesetzt, das eine vollständige Implementierung von DOM Level 2 (Hors et al., 2000) enthält. Für die Visualisierung werden unter anderem `GraphViz::XML` und `XML::LibXSLT` verwendet.[40] Die DOM-Implementierung wird von einem XPath-Prozessor (Clark und DeRose, 1999) flankiert, so dass die Adressierung und Lokalisierung einzelner Knoten und Teilbäume möglich ist.

Der Textparser wurde auf der Grundlage verschiedener Prinzipien entwickelt: Zunächst sollen sämtliche Algorithmen und Methoden so einfach, robust und generell wie möglich

[37] Buyukkokten et al. (2001a, S. 214) erläutern diese Metapher folgendermaßen: „We call our summarization strategy *accordion summarization* because a page can be shrunk or expanded much like an accordion.“

[38] Beispielsweise kategorisieren Penn et al. (2001) eine Tabelle als genuin, wenn das `table`-Element keine weiteren `table`-Elemente enthält, die Anzahl Wörter in den Tabellenzellen kleiner als ein Schwellwert ist und sich in den einzelnen Zellen maximal jeweils ein HTML-Element befindet, wobei jedoch z. B. Listen, Frames und Formulare nicht erlaubt sind. Dieser Ansatz operiert mit einer Präzision von etwa 90%.

[39] Der Textparser kann mit Visualisierungsfunktionen über die Web-Oberfläche der Korpusdatenbank aufgerufen werden. Er wurde primär mit den Dokumenten der dritten Analyse (Kapitel 10) entwickelt und evaluiert.

[40] Das *Perl*-Modul `XML::LibXML` verkapselt die C Bibliothek *libxml* (Version 2.4.26), die, ebenso wie *libxslt*, in der Desktop-Oberfläche *Gnome* entwickelt und eingesetzt wird (vgl. `http://xmlsoft.org`).

realisiert werden, so dass im besten Falle *beliebige* HTML-Dokumente erfolgreich verarbeitet werden können. Zusätzlich ist die Domänen- und Sprachunabhängigkeit des Werkzeugs ein wichtiger Aspekt, die Funktionalität soll nicht auf spezifische Dokumente beschränkt sein. Die Annotierung der Analyseergebnisse innerhalb eines zusätzlichen Namensraumes erfolgt nicht destruktiv, d. h. die Primärdaten (der Dokumentinhalt und das HTML-Markup) sollen zu jedem Verarbeitungszeitpunkt zur Verfügung stehen. Weiterhin sollte die Ermittlung möglichst vieler implizit vorhandener Textmerkmale implementiert werden.

Die Strukturanalyse beginnt im Wurzelelement, woraufhin rekursive Funktionen Eigenschaften für jeden Knoten berechnen.[41] Hierzu gehören vor allem die Merkmale, die sich auf das Markup beziehen, z. B. die Anzahl der Kind-Elemente, der prozentuale Anteil der Elemente eines Teilbaums im Verhältnis zum Gesamtbaum sowie vergleichbare Daten für die Anzahl der Wörter, die in den Textknoten eines Teilbaums enthalten sind. Hyperlinks werden bezüglich ihrer Ziele analysiert, die als `external` (Webserver in anderer Domain), `samedomain` (anderer Server in gleicher *second level*-Domain, falls etwa ein Link von `www.uni-giessen.de` nach `opac.uni-giessen.de` vorliegt) sowie `internal` (gleicher Webserver) kategorisiert werden.[42] Mittels `img` referenzierte Bilddateien werden von den entfernten Webservern in das Analysesystem übertragen. Es werden die physikalischen Abmessungen ermittelt, wobei explizite Angaben durch die `img`-Attribute `height` und `width` eine höhere Präzedenz besitzen.[43]

Da die Verarbeitung auf einer rekursiven Traversierung von DOM-Repräsentationen basiert (vgl. auch Gupta et al., 2003), werden Analyseergebnisse unmittelbar in dieser Datenstruktur abgelegt. Zu Beginn der Analyse wird in dem Knotenobjekt, das das Wurzelelement `html` repräsentiert, ein Namensraum (Bray et al., 2004a) mit dem Präfix `hypnotic:` deklariert (zusätzlich zum Default-Namespace von XHTML) – die Elemente dieses Namensraumes korrespondieren mit der in Abbildung 14.1 dargestellten „Analyse-DTD“. Weiterhin findet ein *Part-of-Speech*-Tagging statt. Die Analyseergebnisse können nun als Attribute dieses Namensraumes unmittelbar in bestehende HTML- bzw. XHTML-Elemente eingetragen werden.[44] Makrostrukturkomponenten werden durch XML-Elemente markiert. Mit Hilfe dieser parallelen Auszeichnung wird der Einsatz von XML-Techniken ermöglicht. Durch diese Anreicherung vergrößert sich ein XHTML-Dokument beim aktuellen Stand der Implementierung etwa um den Faktor 25. Neben dem Zugriff auf bereits ermittelte Analyseergebnisse noch

[41] Während in diesem Ansatz die von *Tidy* erzeugten XHTML-Versionen beliebiger HTML-Dokumente als normalisierte Repräsentation eingesetzt werden, überführen Chan und Yu (1999, S. 548) zu verarbeitende Dokumente in eine „canonical form“, die aus einer Art Teilmenge von HTML besteht: Das Element `h2` wird z. B. auf die Elementsequenz `<br><br><font size="5">` abgebildet.

[42] Bezüglich dieser Hyperlinktypen liegt eine Vereinfachung vor, die auf den Umstand zurückzuführen ist, dass kein Verfahren zur maschinellen Erkennung der Grenzen von Hypertexten vorliegt (vgl. Abschnitt 14.6.1).

[43] Hierzu wird das Perl-Modul `Image::Size` benutzt, das verschiedene UNIX-Werkzeuge kapselt. Es könnten weitere Merkmale in die Analyse einfließen, z. B. die Anzahl der Farben oder gegebenenfalls vorhandene Transparenz, um etwa eine Aussage darüber treffen zu können, ob es sich bei einer Datei eher um eine Strichzeichnung oder ein Foto handelt (vgl. Asirvatham und Ravi, 2001). Für diese Funktion sind Erkenntnisse aus dem Bereich des Dokumentverstehens und der Dokumentanalyse relevant (vgl. Shin et al., 2001).

[44] Nagao und Hasida (1998) schlagen eine „global document annotation“ vor. Autoren von HTML-Dokumenten sollen Werkzeuge zur Verfügung gestellt werden, um die Dokumente mit linguistischen Informationen anzureichern, die durch XML-Elemente ausgedrückt werden. Da jedoch nicht zu erwarten ist, dass Autoren weitere Editoren in ihre Produktionsprozesse einbetten, muss das avisierte Ziel, die „globale und interkulturelle Kommunikation zu revolutionieren“ (Nagao und Hasida, 1998, S. 921), als unrealistisch eingestuft werden.

während der Verarbeitung besteht ein weiterer Vorteil darin, zu jedem Zeitpunkt wohlgeformte XML-Dateien ausgeben zu können, die gefiltert oder visualisiert werden können.

Es ist das Ziel der Analyse, den meist sehr komplexen Elementbaum in abstraktere Strukturen zu partitionieren, damit diese zur Abbildung auf Hypertextsortenmodule eingesetzt werden können. Die Analyse legt eine Art Metasicht auf den Elementbaum, die durch die Elemente und Attribute der Analyse-DTD repräsentiert wird. Die Daten, die nach der Verarbeitung vorliegen, umfassen vor allem Einheiten wie z. B. Überschriften, Listen, Textabschnitte, Trennelemente und interaktive Bereiche. Eine Liste sollte zwar mit Hilfe der HTML-Elemente `ul` (*unnumbered list*), `ol` (*ordered list*) oder `dl` (*definition list*) ausgezeichnet werden, jedoch finden sich auch häufig lediglich einzelne Textabschnitte (`p` oder `br`), die nur wenige Worte umfassen und auf der linken Seite von einer kleinformatigen Grafik, die einen *bullet point* darstellt, flankiert sind. Durch den hochfrequenten Einsatz derartiger Techniken des *tag abuse* muss der Textparser zwangsläufig über eine Vielzahl robuster Methoden verfügen, um die grundlegenden Textstrukturmerkmale eines HTML-Dokuments zu erkennen und entsprechend zu annotieren. Das Element `hr` (*horizonal rule*) erzeugt z. B. eine horizontale Linie und wird praktisch immer als räumlicher Trenner einzelner Informationsteile benutzt. Hierzu kann jedoch auch eine Grafik benutzt werden, deren Existenz sich mittels einer Analyse ihrer Abmessungen ermitteln lässt, woraufhin ein solches Vorkommen – ebenso wie das Element `hr` – mit dem Attribut-Wert-Paar `hypnotic:TagGroup="separator"` markiert wird. Auf diese Weise können auch Werbebanner und Icons (häufig 16×16 oder 32×32) ausgezeichnet werden. Zwei weitere wichtige Werte für das Attribut `TagGroup` sind `inline` und `block`. Diese werden bereits in den HTML-DTDs unterschieden und umfassen einerseits Elemente, die sich nur auf einzelne Wörter eines größeren Blocks beziehen (`strong`, `em`, `u` etc.) und andererseits Elemente, die ihrerseits größere Blöcke konstituieren (`p`, `code`, `table` etc.).

14.5.4 Der Textparsing-Algorithmus

Der Textparsing-Algorithmus beruht auf mehreren kaskadiert ablaufenden Stufen mit steigender Komplexität. Das Verfahren basiert auf drei zentralen Merkmalen: Hierzu zählen diejenigen HTML-Elemente, die einen Absatzwechsel verursachen, wenn das Dokument von einem Browser dargestellt wird, eine Analyse der in beliebigen Knoten der DOM-Struktur verwendeten Schriftgrößen und eine rudimentäre Grafikanalyse.

Stufe 1: Grundlegende Annotierung von Knoten

Die initiale Verarbeitung des DOM-Baumes betrifft einen rekursiven Abstieg, im Zuge dessen Analyseergebnisse mit neuen Elementen und Attributen des eingeführten Namensraumes annotiert werden. Jedem Element des Baumes wird das Attribut `hypnotic:Path` hinzugefügt, das eine absolute XPath-ähnliche Pfadspezifikation als Wert enthält (beispielsweise `/xhtml:html[1]/xhtml:body[1]/xhtml:h1[2]`). Weiterhin werden Hyperlinks und Wörter (definiert als beliebige durch Zwischenraum getrennte Zeichenketten) sowohl auf der lokalen Ebene als auch in Bezug auf den Teilbaum gezählt, den ein Knoten gegebenenfalls dominiert. Für jedes Element wird die aktuelle Schriftgröße berechnet, die zur generellen Schriftgröße eines Dokuments in Bezug gesetzt wird. Diese kann in HTML über das Element `basefont`

angegeben werden.[45] Die Größe der Grundschrift wird auf den Wert `100` abgebildet. Relative Wechsel der Schriftgröße werden relativ zu dieser Basis berechnet, wohingegen explizite Größenangaben in absoluten Werten resultieren: Die Elemente für Überschriften (`h1` bis `h6`; vgl. auch Tatsumi und Asahi, 2005) korrespondieren mit den Werten `160` bis `110`. Die Elemente `big` und `small` inkrementieren oder dekrementieren die aktuelle Schriftgröße um den Wert `10`. Die Elemente `strong` und `b` sowie `em` und `i` erhöhen die Schriftgröße um den Wert `5` (innerhalb einer Überschrift jedoch nur um den Wert `2`). Relative und absolute Schriftgrößenwechsel durch das Element `font` werden in ähnlicher Weise verarbeitet.[46] Die Analyse von Grafiken bezieht sich auf eine Untersuchung der `img`-Elemente (vgl. Hu und Bagga, 2003, für einen ähnlichen Ansatz). Zunächst werden die Werte der Attribute `src`, `height` und `width` extrahiert. Anschließend wird die referenzierte Grafik mittels HTTP von dem entfernten Webserver auf das Analysesystem übertragen, woraufhin ihre Abmessungen ermittelt werden.[47] Daraufhin weist eine Heuristik der Grafik eine Kategorie zu: Wenn `x` und `y` jeweils weniger als 6 betragen, wird die Grafik als *spacer* benutzt, d. h. das Bild dient der Realisierung eines pixelgenauen Layouts. Wenn `x` und `y` zwischen 6 und 45 betragen, wird die Grafik als Icon kategorisiert. Falls der Quotient $\frac{x}{y} > 10$, handelt es sich um einen räumlichen Trenner. Falls die Abmessungen einer Grafik in einer Liste der standardisierten Größen von Werbebannern gefunden werden (468×60, 156×60, 137×60 etc.), wird das Bild entsprechend eingeordnet. Das jeweilige `img`-Element wird von einem der folgenden Elemente verkapselt: `hypnotic:Separator`, `hypnotic:Icon`, `hypnotic:Banner` und `hypnotic:Spacer`.[48]

Stufe 2: Kapselung von Textknoten

Diese Stufe führt eine rekursive Verarbeitung des DOM-Baumes durch, um jeden Textknoten mit dem Element `hypnotic:Text` zu verkapseln. Jedem `hypnotic:Text`-Knoten werden mehrere Attribute zugewiesen (z. B. die Schriftgröße).

Stufe 3: Erkennung von Textblöcken

Diese Stufe ermittelt und markiert Textblöcke anhand der `hypnotic:Text`-Elemente. Ein Textblock ist definiert als ein absatzähnliches Objekt mit konsistenter Schriftgröße, wobei Abweichungen von ± 5 zugelassen sind. Die Funktion `findTextBlocks` basiert auf einem *top-down* und *depth-first* operierenden *tree walker*, der die Funktion `markTextBlocks` aufruft, sobald ein Textblock gefunden wurde. Listing 14.5 stellt die Funktion in verkürztem

[45] Das Element `basefont` wird in 0,67% der Korpusdokumente verwendet (vgl. Tabelle A.9, S. 740).

[46] Die Analyse von CSS-Informationen zur Manipulation der Schriftgröße wird nicht unterstützt, weil diese im Korpus in lediglich 21,7% aller Dokumente existieren (vgl. Abschnitt A.4.9). Im Gegensatz dazu existieren jedoch in einem Dokument durchschnittlich 310 Wörter sowie 120,57 HTML-Elemente und 236,04 HTML-Attribute; diese Angaben verdeutlichen den Bedarf für eine umfassende Untersuchung der Auszeichnungsstruktur. Eine CSS-Analyse kann jedoch mit einfachen Mitteln realisiert werden, da diverse *Perl*-Module zur Verarbeitung von CSS-Dateien existieren (vgl. auch Buyukkokten et al., 2001a, S. 216). Ein weiterführendes Desiderat betrifft die Untersuchung von *JavaScript*-Code, um dynamische und interaktive Elemente von HTML-Dokumenten einbeziehen zu können (vgl. Carchiolo et al., 2002).

[47] Falls über die Attribute `width` und `height` explizite Angaben zu den Abmessungen einer Grafik gemacht werden, werden die physikalischen Abmessungen der Datei – wie von jedem Browser – ignoriert.

[48] Eine weiterführende Grafikanalyse könnte z. B. auf der Anzahl verwendeter Farben operieren, um Strichzeichnungen und Diagramme von Fotos zu unterscheiden.

```
@s: Umfasst die Sequenz von hypnotic:Text-Elementen
@b: Absatzelemente: <p>, <ul>, <ol>, <dl>, <td>, <blockquote>, <div>
$n: Das aktuelle Element
$p: Das vorherige Element

while (next element) {
  if ($n == "hypnotic:Icon") {
    push(@s, $n);
  }
  if ($n == "hypnotic:Separator") {
    markTextBlocks(@s);
  }
  if (($n == "<br>") && ($p == "<br>")) {
    markTextBlocks(@s);
  }
  if ($n is element of @b) {
    markTextBlocks(@s);
  }
  if ($n ist kein Bestandteil eines "offenen" Teilbaums, der von einem in @b enthaltenen
      Knoten dominiert wird) {
    markTextBlocks(@s);
  }
  if ($n == "hypnotic:Text") {
    if (!($n->FontSize == $p->FontSize +/- 5)) {
      markTextBlocks(@s);
    }
    push(@s, $n);
  }
}
```

Listing 14.5: Die Funktion `findTextBlocks($r)` in Pseudo-Perl Code

Pseudo-*Perl*-Code dar. Mehrere Zustände entsprechen Grenzen zwischen Textblöcken und können einen Aufruf von `markTextBlocks` auslösen: (i) Zwei oder mehr aufeinander folgende `br`-Elemente; (ii) ein `hypnotic:Separator`-Element; (iii) falls das aktuelle Element zu einer Gruppe von Elementen gehört, die im Browser einen Absatzwechsel verursachen (`@b` in Listing 14.5); (iv) falls der *tree walker* einen unmarkierten Teilbaum verlässt, der von einem der in `@b` enthaltenen Elemente dominiert wird; (v) ein signifikanter Wechsel der Schriftgröße.

Die Funktion `markTextBlocks` bezieht sich auf das Array, das aus `hypnotic:Text`-Elementen besteht und unter Umständen eingebettete `hypnotic:Icon`-Elemente umfasst. Die in dieser Liste enthaltenen Knoten werden durch das Element `hypnotic:TextBlock` markiert, indem dieses Element als übergeordneter Knoten in den Baum eingefügt wird. Diese Einfügung findet auf einer möglichst hoch angeordneten Ebene innerhalb der DOM-Struktur statt, so dass der Teilbaum, der von dem hinzugefügten `hypnotic:TextBlock`-Element dominiert wird, möglichst viele Knoten umfasst. Hierfür werden die Elternknoten der `hypnotic:Text`-Elemente untersucht, wozu die `hypnotic:Path`-Attribute eingesetzt werden. Falls ein Elternknoten keine zusätzlichen `hypnotic:Text`-Kindknoten umfasst, kann der Elternknoten des Elternknotens untersucht werden etc. Auf diese Weise kann der *least common node* (LCN) ermittelt werden. Dabei handelt es sich um denjenigen Knoten, der alle `hypnotic:Text`-Elemente der aktuellen Sequenz und keine weiteren `hypnotic:Text`-Knoten dominiert. Dem LCN kann nun der `hypnotic:TextBlock`-Knoten als weiteres Kindelement hinzugefügt werden, woraufhin der oder die verbleibenden Teilbäume als Kindknoten von `hypnotic:`

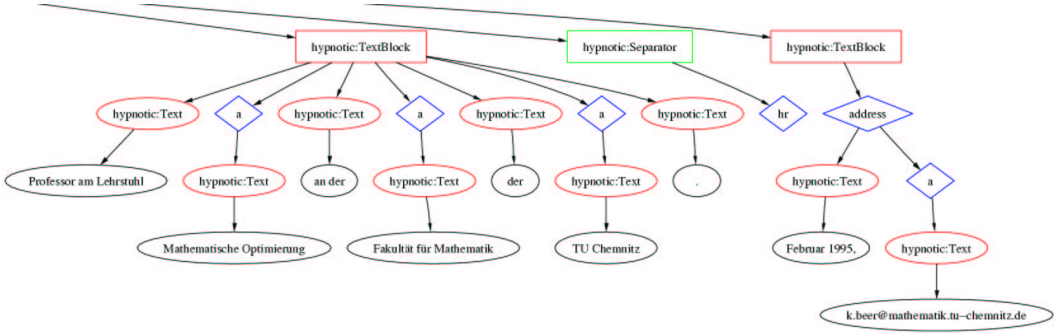


Abbildung 14.3: Ein DOM-Baum nach der Aufnahme zweier `hypnotic:TextBlock`-Knoten

`TextBlock` verschoben werden. Unter Umständen existiert jedoch kein LCN, so dass für den neuen Knoten zunächst die korrekte Position auf der Kindebene eines gemeinsamen Vorgängers der Knoten in `@s` zu ermitteln ist. Daraufhin können die einzelnen Teilbäume, die die in `@s` enthaltenen Knoten umfassen, unter den neuen Knoten verschoben werden. Abbildung 14.3 zeigt einen Ausschnitt eines DOM-Baumes, der auf diese Weise modifiziert wurde.

Stufe 4: Verarbeitung von Textblöcken

Die dritte Stufe trägt eine neue Strukturebene in das Dokument ein, die im ursprünglichen Dokument nur implizit enthalten ist. Die vierte Stufe operiert auf dieser neuen Ebene, indem die von `hypnotic:TextBlock` dominierten Teilbäume untersucht werden, um zusätzliche Strukturen zu ermitteln. Hierfür werden $1 \ldots n$ `hypnotic:TextBlock`-Knoten aggregiert. Der implementierte Prototyp ist in der Lage, Überschriften auf mehreren Ebenen, Fußzeilen (Textblöcke, die eine kleinere Schriftgröße als die Grundschrift umfassen) sowie unterschiedliche Typen expliziter und impliziter Listenstrukturen zu detektieren und zu annotieren.

Die Detektion von Listen, die explizit von HTML-Elementen realisiert werden, erfolgt durch die Bestimmung von `hypnotic:TextBlock`-Elementen, die `ul`-, `ol`- oder `dl`-Knoten dominieren, die wiederum mindestens ein `li`- bzw. `dt`/`dd`-Element beinhalten. Für die Erkennung derartiger Strukturen werden XPath-Ausdrücke eingesetzt. Die Annotierung geschieht durch das Element `hypnotic:List`. Einen Sonderfall stellen multiple Textblöcke innerhalb von `li`-Elementen dar. Die Identifizierung impliziter Listen ist ein sehr komplexer Prozess, der von einem weiteren *top-down* arbeitenden *tree walker* implementiert wird, der spezifische `hypnotic:TextBlock`-Sequenzen, die gegebenenfalls auch `hypnotic:Icon`-Elemente umfassen, als `hypnotic:List` markiert. Beim aktuellen Stand der Implementierung werden die folgenden Listentypen erkannt: (i) *itemized lists*, in denen der *bullet point* als kleinformatige Grafik realisiert ist (Sequenzen der Struktur `hypnotic:Icon`, $1 \ldots n$ `hypnotic:TextBlock`, `hypnotic:Icon`, $1 \ldots n$ `hypnotic:TextBlock`, ...); (ii) implizite *itemized lists*, in denen der *bullet point* typografisch realisiert ist (z. B. konsistent als - oder * zu Beginn eines `hypnotic:TextBlock`-Elements); (iii) implizite Aufzählungslisten, in denen die *labels* aus aufsteigenden Zahlen in arabischer oder (iv) römischer Notation bestehen (vgl. Abbildung 14.4 für

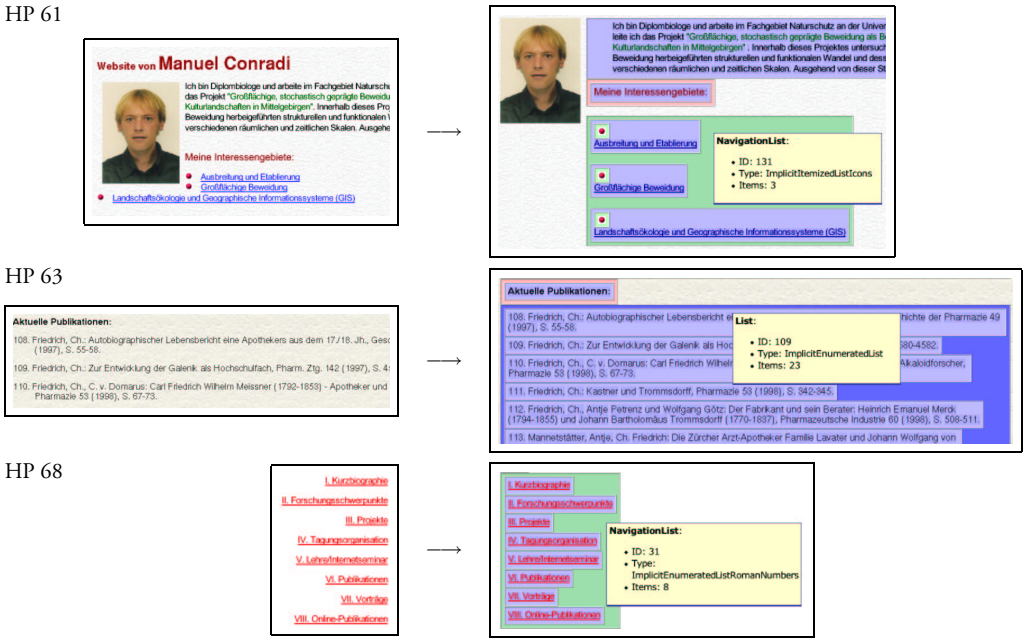


Abbildung 14.4: Beispiele für die Erkennung impliziter Listen durch den Textparser

Beispiele).[49] Das Einfügen des `hypnotic:List`-Knotens erfolgt nach dem Algorithmus zur Ermittlung des LCN einer gegebenen Sequenz von Knoten.

Die Erkennung von Überschriften und Fußzeilen operiert auf den Schriftgrößen benachbarter Textblöcke. Falls sich diese in Bezug auf einen Schwellwert unterscheiden, wird die jeweilige Sequenz als `hypnotic:Headline` oder `hypnotic:Footnote` markiert. Beide Elemente erhalten das Attribut `hypnotic:Level`, das – bezüglich der Schriftgröße – die relative Ebene der Überschrift bzw. der Fußzeile beinhaltet; die Überschrift mit der größten Schriftart wird somit als Überschrift erster Ebene markiert. Weiterhin werden auch Überschriften in Textblöcken ermittelt: Da die Schriftgröße innerhalb eines Textblocks um ± 5 variieren kann, wird ein initialer `hypnotic:Text`-Knoten, der eine um den Wert 5 größere Schriftart als die verbleibenden `hypnotic:Text`-Elemente verwendet, als eingebettete Überschrift markiert.

Stufe 5: Detektion unterschiedlicher Typen von Hyperlinklisten

Die fünfte Stufe umfasst den Einsatz verschiedener Heuristiken zur Bestimmung unterschiedlicher Typen von Hyperlinklisten. Hierfür werden alle Vorkommen von `hypnotic:List` ermittelt, bei denen jeder Listenpunkt mindestens einen Hyperlink enthält. Da die be-

[49] Abbildung 14.4 zeigt auf der linken Seite Ausschnitte der im Korpus enthaltenen Dokumente, wie sie von dem Browser *Mozilla* dargestellt werden. Die rechte Seite enthält die Ausgabe einer Visualisierungskomponente. Der Textparser kann im Modus der Dokumentansicht innerhalb der Web-Oberfläche der Korpusdatenbank aktiviert werden. Die Visualisierung hebt die Ergebnisse der Analyse farblich hervor (vgl. Abschnitt 14.5.5). Die Dokumente stammen aus der in Kapitel 10 analysierten Stichprobe (vgl. Tabelle 10.2, S. 428).

nötigten Informationen über die Anzahl der in einer Liste enthaltenen Hyperlinks und die jeweiligen Typen von Hyperlinks bereits vorliegen, kann die Bestimmung auf dem prozentualen Anteil eines Hyperlinktyps beruhen: Wenn mehr als zwei Drittel aller Hyperlinks vom Typ `internal` sind, wird angenommen, dass es sich um Verknüpfungen innerhalb des aktuellen Hypertextes handelt, so dass das Element `hypnotic:List` durch `hypnotic:NavigationList` (Navigationshilfe) ersetzt wird. Analog wird für die verbleibenden Typen `external` (erzeugt `hypnotic:Hotlist`), `thispage` (`hypnotic:TableOfContents`) und auch `samedomain` (`hypnotic:Dispenser`, d. h. Verteiler, vgl. Abschnitt 11.5.3) vorgegangen. Abbildung 14.4 zeigt Beispiele für Navigationshilfen (vgl. auch Rehm, 2004b).

Stufe 6: Integration von Part-of-Speech-Informationen

Die letzte Stufe beinhaltet die Integration eines robusten syntaktischen Parsers. Mit Hilfe eines XPath-Ausdrucks werden alle `hypnotic:TextBlock`-Elemente ermittelt. In jedem dieser Elemente werden sämtliche `hypnotic:Text`-Elemente gesammelt und die enthaltenen Zeichenketten werden zur Konstruktion eines Arrays eingesetzt, dessen Felder Token umfassen, die durch *whitespace* getrennt sind. Die gesamte Zeichenkette wird an eine *Perl*-Funktion übergeben, die die Software *Machinese Syntax German* des Herstellers *Connexor* aufruft. Diese Software liefert *Part-of-Speech*-Informationen und führt eine syntaktische Analyse durch. Es erwartet eine Sequenz von Wörtern als Eingabe, tokenisiert und analysiert diese und gibt die Ergebnisse der Analyse als XML-Instanz aus. Da ein nicht dokumentierter Algorithmus zur Tokenisierung verwendet wird, ist es notwendig, eine Abbildung der *Connexor*-Tokenisierung (z. B. „[Paul] [lacht] [.]“) auf den *whitespace-separated-token*-Ansatz vorzunehmen („[Paul] [lacht.]“), da für *jedes* Wort innerhalb der DOM-Struktur ein neues Element namens `hypnotic:Token` konstruiert wird, das die ermittelten Analyseinformationen als Attributwerte aufnimmt.[50] Es wurde ein Parser implementiert, der die Tokenisierung des *Connexor*-Tools sukzessive verarbeitet und einen Hash erzeugt, der das gewünschte Tokenisierungsparadigma beinhaltet, indem z. B. Interpunktionszeichen an vorhergehende Wörter angefügt (siehe oben) oder aus einer öffnenden Klammer und dem nachfolgenden Wort ein neues Token hergestellt werden. Dieser Prozess basiert auf einem Vergleich der *Connexor*-Token mit den aus der Eingabe erzeugten *whitespace-separated-tokens*.

14.5.5 Visualisierung und Beispiele

Die Web-Oberfläche der Korpusdatenbank ermöglicht in der Dokumentansicht einen Zugriff auf den Textparser, der das HTML-Dokument verarbeitet und durch eine modifizierte Version ersetzt. Nach der Anreicherung und Serialisierung des DOM-Baums als XHTML-Dokument wird es von einem XSLT-Stylesheet transformiert. Dieses überführt alle Elemente,

[50] Das *Connexor*-Werkzeug (vgl. `http://www.connexor.com`) wird auf alle in einem Dokument enthaltenen Zeichenketten angewendet. Die Identifizierung unterschiedlicher Hypertextsortenmodule kann zwar zur Aktivierung spezialisierter syntaktischer Parser eingesetzt werden, jedoch entstünde hierdurch eine Problematik, die sich auf Lernverfahren bezieht: Wenn z. B. eine Kategorisierung der Inhalte von Hypertextsortenmodulen auf der Basis von POS-Informationen durchgeführt werden soll (vgl. Abschnitt 14.2), ist es notwendig, *alle* Inhalte konsistent mit *einem* Werkzeug zu annotieren, um vergleichbare Ergebnisse zu erhalten. Die derzeitige Implementierung übergibt alle Inhalte eines `hypnotic:Text`-Knotens an das *Connexor*-Werkzeug. Hierdurch entstehen zahlreiche fehlerhafte Analysen, da nicht garantiert ist, dass vollständige Sätze verarbeitet werden.

HP 29

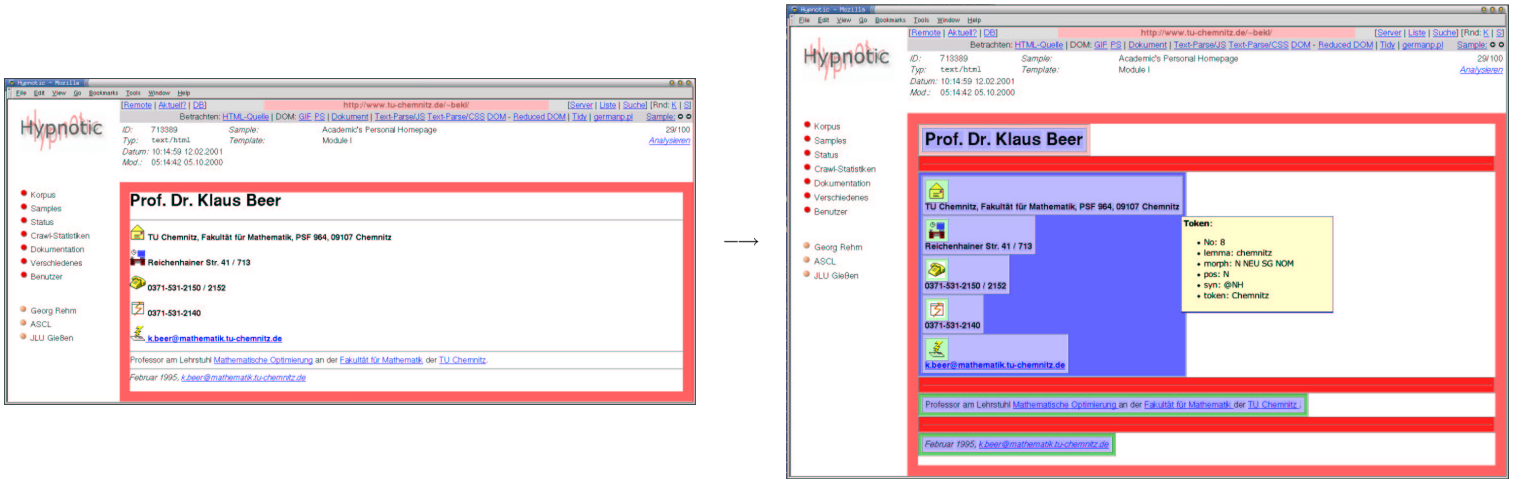


Abbildung 14.5: Links: Quelldokument in der Web-Oberfläche; Rechts: von einem XSLT-Stylesheet visualisierte Analyse des Textparsers (Hellblau: Textblock, Dunkelblau: Liste, Rot: Trenner, Hellrot: Überschrift, Grün: Fußzeile, Hellgrün: Icon)

die durch den hypnotic:-Namensraum gekennzeichnet sind, in Tabellen mit charakteristischen Hintergrundfarben, so dass die Ergebnisse der Analyse betrachtet werden können (Abbildung 14.4 zeigt drei partielle Beispiele; Abbildung 14.5 stellt, ebenso wie Abbildung 14.8, S. 698, ein vollständiges Beispiel dar). Der Textparser generiert dynamisch *JavaScript*-Code-Fragmente, die in die DOM-Struktur integriert werden, so dass – nach der Serialisierung und Anzeige des XHTML-Dokuments – das Bewegen des Mauszeigers über eine analysierte Einheit die Darstellung eines Pop-up-Fensters mit weiterführenden Informationen bewirkt (vgl. die Abbildungen 14.4 und 14.5).[51] Diese Funktion arbeitet kaskadiert: Wenn sich der Mauszeiger über einem Token-Element befindet, werden die korrespondierenden POS-Analyseergebnisse angezeigt. Wenn er in das umgebende TextBlock-Element bewegt wird, werden die übergeordneten Ergebnisse dargestellt.[52] Auf der nächsten Ebene existieren z. B. unterschiedliche Typen von Listen oder Überschriften, zu denen weitere Informationen präsentiert werden. Diese intuitive Visualisierung war insbesondere in der Entwicklungsphase des Textparsers ein immenser Vorteil, da keine Notwendigkeit bestand, umfangreiche und extrem komplexe XHTML-Dokumente in ihrer serialisierten Form zu betrachten. Der Textparser annotiert bei gewissen HTML-Strukturen rekursiv einen Textblock innerhalb eines Textblocks. Derartige Annotationsfehler können durch entsprechende Templates innerhalb des Stylesheets ermittelt und visualisiert werden: Wenn bei der Transformation ein solches Fehlererkennungstemplate ausgelöst wird (in diesem Beispiel durch den Ausdruck hypnotic:TextBlock//hypnotic:TextBlock), fügt das Stylesheet zu Beginn des darzustellenden XHTML-Dokuments eine Fehlermeldung ein, die auf dieses Problem hinweist.

Neben der Visualisierung der Oberflächeneinheiten kann auch die angereicherte DOM-Baumstruktur des XHTML-Dokuments und eine Version *ohne* XHTML-Elemente dargestellt werden, die ausschließlich die Elemente des hypnotic:-Namensraumes umfasst. Diese Funktionen ermöglichen die Überprüfung, dass die neuen Knoten an den korrekten Positionen eingefügt wurden. Die um XHTML-Elemente reduzierte Version wird durch ein XSLT-Stylesheet erzeugt, das alle Elemente des xhtml:-Namensraumes und alle Vorkommen von hypnotic:Text entfernt (vgl. Abbildung 14.6). Hierbei findet unter Umständen eine weitere Modifikation statt: Falls ein XHTML-Element wie z. B. xhtml:p zwei hypnotic:TextBlock-Elemente dominiert, wird xhtml:p zur Strukturerhaltung durch das Element hypnotic:Node ersetzt. Eine Restrukturierung des Elementbaumes, die die logische Struktur der ermittelten Komponenten reflektiert, findet beim aktuellen Stand der Implementierung nicht statt.[53]

14.5.6 Fazit

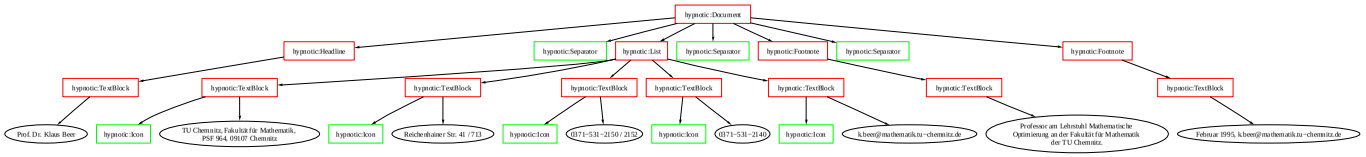
Die in dem Textparser implementierten Algorithmen arbeiten stabil und fehlerfrei. Von den 100 zur Entwicklung und Evaluation eingesetzten Dokumenten (vgl. Abschnitt 14.5.3) werden jedoch vier fehlerhaft annotiert. Der im vorangegangenen Abschnitt angesprochene und für dieses Problem verantwortliche Fehler ist ein Resultat des Umstands, dass noch eine Viel-

[51] In den hier präsentierten, unter *Linux* angefertigten Bildschirmabzügen ist der Mauszeiger selbst aus technischen Gründen nicht enthalten. Die Realisierung der Pop-up-Fenster erfolgt durch die *JavaScript*-Bibliothek *overlib* (http://www.bosrup.com/web/overlib/).

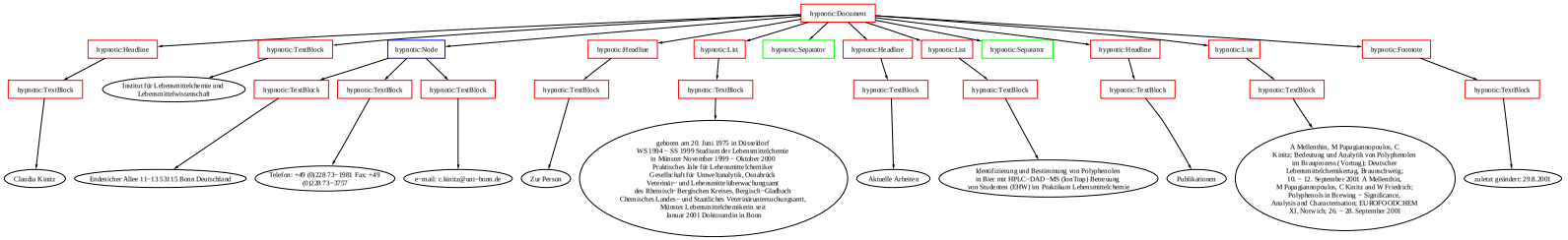
[52] Anhang F auf der CD ROM demonstriert diese Funktionalität anhand zweier Bildschirmvideos.

[53] Chung et al. (2002) zeigen verschiedene Regeln, anhand derer ein solcher Prozess implementiert werden kann.

HP 29 (vgl. Abbildung 14.5, S. 670):



HP 42 (vgl. Abbildung 10.5, S. 456, und Abbildung 14.8, S. 698):



HP 68 (vgl. Abbildung 14.4, S. 668):

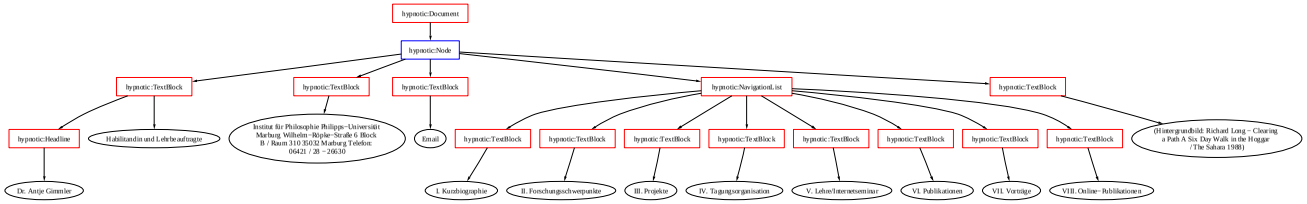


Abbildung 14.6: Reduktion der DOM-Struktur auf die Analyseinformationen durch XSLT-basierte Tilgung der HTML-Elemente

zahl zusätzlicher Erkennungsregeln zu implementieren ist: Die Aufgabe des Textparsers betrifft die Detektion der von Autoren oftmals eingesetzten Strategien zur Realisierung spezifischer Textstrukturkomponenten.[54] Da eine Vielzahl derartiger Strategien existiert, ist es notwendig, diese innerhalb des Textparsers durch korrespondierende Erkennungsregeln zu antizipieren, was in der vorliegenden Implementierung nur in Ausschnitten demonstriert werden konnte. Eine mögliche Erweiterung dieses Ansatzes betrifft die manuelle Annotierung von Beispieldokumenten mit den zugehörigen Makrostrukturobjekten. Eine auf diese Weise annotierte Kollektion könnte als Trainingskorpus für ein maschinelles Lernverfahren eingesetzt werden, das die regelbasierte Erkennung unterstützt.

In vielen Dokumenten werden `table`-Elemente zur Realisierung komplexer Layouts eingesetzt, die von dem Textparser nur partiell verarbeitet werden können. Diese Problematik wurde aus zwei Gründen ignoriert: Einerseits handelt es sich bei dem Textparser lediglich um ein *proof-of-concept*, andererseits wird auch die Analyse von `table`-Strukturen naturgemäß mit der Aufgabe konfrontiert, dass die von Autoren verwendeten Strategien zumindest in Teilen als Regeln rekonstruiert werden müssen. Zudem ist der Textparser mit einem Algorithmus zur Erkennung genuiner Tabellen auszurüsten (vgl. Abschnitt 14.5.2). Ebenfalls implementiert werden muss eine Erkennung unterschiedlicher Typen von Textblöcken. Hierzu zählen insbesondere Abschnitte, die Fließtext oder Textfragmente (alle Textblöcke in Abbildung 14.5) enthalten, da diese Information in der Lage ist, das zu berücksichtigende Inventar von Hypertextsortenmodulen einzuschränken.

Der Textparser wurde als initiale Komponente des in Abbildung 14.1 dargestellten Systems zur Erkennung von Hypertextsorten implementiert und zeigt bereits auf dieser grundlegenden Ebene der Identifizierung der Hypertextmodule die Komplexität eines derartigen Systems auf. Es existieren weiterführende Anwendungsszenarien: Die hierarchischen Informationen, die implizit von den eingefügten Elementen ausgedrückt werden (Überschrift, Fußzeile, Textblock, Liste etc.), können zur Restrukturierung des DOM-Baumes eingesetzt werden, so dass die logische Dokumentstruktur reflektiert wird (Chung et al., 2002). Zur Ermittlung der Hypertextsortenmodule ist eine derartige Restrukturierung notwendig, weil ein Hypertextsortenmodul nicht notwendigerweise von genau einem Hypertextmodul realisiert wird: Instanzen der Hypertextsortenmodule *Publikationsliste* oder *Lebenslauf bzw. biografische Angaben* umfassen z. B. häufig eine Überschrift, mehrere Listen und Zwischenüberschriften. Über einen statistischen Vergleich der einzelnen Einträge kann ihre Ähnlichkeit ermittelt werden (z. B. Durchschnittsumfang eines Listeintrags oder Verwendung von Interpunktionszeichen), so dass zusammengehörige Hypertextmodule zu einem übergeordneten Baustein aggregiert werden können. Eine derartige Funktion könnte auch das Betrachten eines Dokuments auf einem mobilen Endgerät erleichtern, um z. B. initial nur die Überschrift und den ersten Absatz darzustellen, woraufhin der Rezipient entscheiden kann, ob und welche weiterführenden Bestandteile des Dokuments betrachtet werden sollen, indem der Baum sukzessive geöffnet wird (vgl. Buyukkokten et al., 2001a,b, Rahman et al., 2001). Für das automatische Textzusammenfassen kann der Textparser als initiale Komponente eingesetzt werden, so dass spezifische Verfahren auf einzelne Makrostrukturbausteine angewendet werden können.

[54] Siehe hierzu auch Chen et al. (2003b), die die Funktionsweise ihres Verfahrens (vgl. Abschnitt 14.5.2) skizzieren: „[T]oday's natural language technology is still far from delivering a satisfactory result. Our approach is to perform the analysis by inferring the content structure of a web page embedded by the web author."

Auch im Kontext traditioneller Suchmaschinen kann der Textparser Verwendung finden, indem z. B. Wörter, die in Überschriften erster Stufe enthalten sind, stärker gewichtet werden als die Bestandteile einer Fußzeile (vgl. auch Cutler et al., 1997). Explizit annotierte Satzgrenzen können zur Extraktion von Testsätzen eingesetzt werden, um den Abdeckungsgrad traditioneller Parser zu evaluieren (Volk, 1998). Diese Anwendung bezieht sich darauf, das *World Wide Web* als Korpus für linguistische Untersuchungen zu betrachten („Web as corpus approach", vgl. Kilgarriff, 2001, und Rehm, 2004d).

Das zuletzt genannte Anwendungsszenario zeigt einen methodologischen Aspekt auf, der die in Teil III dargestellten Studien betrifft: Diese beziehen sich unter anderem auf Hypertextsortenmodule. Korpuslinguistische Untersuchungen beziehen sich jedoch in der Regel auf eine mikrostrukturelle Ebene, d. h. es werden z. B. syntaktische Strukturen oder die Frequenzen von Wortarten ermittelt. Die Funktionalität des Textparsers verdeutlicht, weshalb derartige Untersuchungen in Bezug auf HTML-Dokumente nicht ohne Weiteres durchgeführt werden können: Traditionelle Korpora besitzen typischerweise eine Annotierung der Textstrukturkomponenten (Überschriften, Absätze etc.) und werden mit computerlinguistischen Werkzeugen (z. B. Tokenisierern, POS-Taggern und Parsern) verarbeitet, um linguistisch annotierte Korpora zu erzeugen (Sasaki und Witt, 2004), die anschließend mit korpuslinguistischen Tools untersucht werden können. Aufgrund der mehrfach angesprochenen Heterogenität der *Hypertext Markup Language* können arbiträre HTML-Dokumente jedoch *keinesfalls* als einfache Sequenzen von Sätzen aufgefasst und von computer- oder korpuslinguistischen Standardwerkzeugen verarbeitet und ausgewertet werden. Zunächst ist es zwingend notwendig, eine Bestimmung der enthaltenen Hypertextmodule vorzunehmen, so dass ein syntaktischer Parser z. B. nur auf `TextBlock`-Elemente vom Typ „Fließtext" angewendet wird. Erst wenn eine derartige Differenzierung vorliegt, können medienspezifische Varietäten ermittelt werden, um z. B. die syntaktischen Strukturen von Überschriften in den Instanzen spezifischer Hypertextsorten mit den Überschriften in verwandten Exemplaren traditioneller Textsorten zu vergleichen. Der Textparser erlaubt nun zumindest einen rudimentären Zugriff auf diese Ebene, die in HTML-Dokumenten nicht explizit zur Verfügung steht.

14.6 Zur maschinellen Identifizierung von Hypertextsorten

Die Diskussion der vorliegenden Arbeiten zur Erkennung von Web-Genres hat mehrere Aspekte verdeutlicht: Zunächst fehlt sämtlichen Ansätzen eine theoretische Basis, die die Charakterisierung des Konzepts „Web-Genre" erlaubt (vgl. Abschnitt 14.2.4). Die Betrachtung eines *einzelnen* HTML-Dokuments als *vollständiges und abgeschlossenes* Exemplar eines *einzelnen* Web-Genres stellt in diesen Studien eine konzeptionell und technologisch bedingte Restriktion zur Durchführung dieser initialen Experimente dar. Die Limitierung ist jedoch nicht in der Lage, die Komplexität der realen Gegebenheiten in Bezug auf die unterschiedlichen Ebenen der Konstituenz von Hypertextsorten erfassen zu können. Mit dem in Kapitel 5 eingeführten Hypertextsortenmodell und dem zugehörigen Repräsentationsformat (Kapitel 13) liegt nun zwar eine derartige Basis vor, die hochgradige Komplexität der benötigten Methoden und Ressourcen verhindert jedoch – zumindest hinsichtlich des Rahmens der vorliegenden Arbeit – eine vollständige prototypische Realisierung eines derartigen Systems.

Nachfolgend werden die zentralen Prozesse und Komponenten thematisiert, die in einer vollständigen Implementierung einzusetzen sind. Hierbei handelt es sich um ein Verfahren zur Erkennung der Grenzen von Hypertexten, da die Zuordnung eines Knotens zu einem Hypertext und die Identifizierung eingebetteter Hypertexte eine Prämisse der Detektion von Hypertextsorten darstellt (Abschnitt 14.6.1). Dieser Prozess kann, wie die eingangs diskutierten Arbeiten zeigen, als Anwendung von Lernverfahren konzeptualisiert werden. Daher stellt Abschnitt 14.6.2 einen Katalog von Merkmalen vor, die bei der Erkennung von Hypertextsorten und Hypertextsortenmodulen zu berücksichtigen sind (Abschnitt 14.6.3).

14.6.1 Erkennung der Grenzen von Hypertexten

Die Instanz einer Hypertextsorte umfasst Hypertextsortenmodule, die in einzelnen Knoten ausgeprägt werden. Für die automatische Erkennung einer Hypertextsorte ist es von zentraler Bedeutung, diejenigen Knoten (d. h. HTML-Dokumente) zu bestimmen, die einen Hypertext konstituieren, so dass sie in das Erkennungsverfahren einbezogen werden können. Diese Aufgabe sieht sich insbesondere mit der Problematik konfrontiert, dass weder auf der Ebene von HTML noch in Bezug auf HTTP explizite Informationen über die Knoten eines Hypertextes vorliegen (vgl. Kwon und Lee, 2003). Zudem können Hypertextsorten weitere Hypertextsorten einbetten, so dass zusätzlich die Grenzen der untergeordneten Hypertexte zu bestimmen sind. Weiterhin bieten zahlreiche Websites mehrere Hypertexte parallel an, so dass einfache Heuristiken zur Detektion dieser Grenzen nicht eingesetzt werden können (z. B. *alle* HTML-Dokumente *einer* Website als Konstituenten *eines* Hypertextes aufzufassen oder die hierarchische Verzeichnisstruktur eines Webservers mit eingebetteten Hypertexten gleichzusetzen, vgl. Géry und Chevallet, 2001, Thelwall, 2002).[55]

Ester et al. (2002) weisen darauf hin, dass sich nahezu alle bislang vorgelegten Ansätze zur Klassifizierung von HTML-Dokumenten auf die Ebene der einzelnen Datei beschränken. Daher wird die Klassifizierung vollständiger Websites vorgeschlagen, die auf einer Verarbeitung der HTML-Dokumente basiert, die eine Website umfasst.[56] Eine solche Anwendung setzt ein Verfahren zur Bestimmung der „borders of a site“ (Ester et al., 2002, S. 250) voraus, wobei sich die Verfasser jedoch darauf beschränken, *alle* HTML-Dokumente, die auf einem Webserver per HTTP erreichbar sind, als Bestandteile *einer* Website zu betrachten.[57] Die Entwick-

[55] Die Betrachtung derjenigen HTML-Dokumente, die sich auf dem Webserver ein Verzeichnis teilen, als (ausschließliche) Konstituenten *eines* Hypertextes liegt nahe, doch kann eine derartige Heuristik lediglich zur initialen Hypothesenbildung eingesetzt werden: Sobald z. B. ein *Content Management System* eingesetzt wird, liegen üblicherweise keine Angaben über Verzeichnisstrukturen vor, da HTML-Dokumente dynamisch aus Datenbanken generiert werden. Eiron und McCurley (2003, S. 88) betrachten derartige URLs ebenfalls als äußerst problematisch und ignorieren sie im Rahmen der Evaluierung ihrer Heuristiken (siehe im Folgenden).

[56] Für diese Klassifizierung werden zwei Verfahren eingesetzt: Innerhalb des „superpage“-Ansatzes werden sämtliche HTML-Dokumente als einzelnes Dokument – als „superpage“ – betrachtet und gemeinsam klassifiziert. Der zweite Ansatz basiert auf der Klassifizierung von Bäumen mit maschinellen Lernverfahren. Die Bäume werden aus der Verknüpfungsstruktur ermittelt und enthalten etikettierte Knoten. Diese Etiketten beziehen sich wiederum auf abstrakte Konzepte, die grob mit Hypertextknotensorten vergleichbar sind: Es wird eine Evaluierung mit den Websites mittelständischer Unternehmen durchgeführt, weshalb „page classes“ wie z. B. „company philosophy“, „online contact“, „places and opening hours“, „products and services“, „references and partners“, „employees“, „vacancies“ und „other“ angenommen werden.

[57] Siehe hierzu Ester et al. (2002, S. 251): „The *web site* of a domain D is a directed graph $G_D(N, E)$. A Node $n \in N$ represents an HTML-page whose URL starts with D. [...] Thus, every HTML-document under the same

lung eines Verfahrens zur Ermittlung der tatsächlichen Grenzen einer Website wird von Svatek und Vacura (2003) angestrebt, wofür insbesondere die Sequenzierung einer Gruppe von Dokumenten eingesetzt werden soll. Zusätzlich kann die „logical website boundary identification" Svatek und Vacura zufolge auf rekurrenten HTML-Fragmenten beruhen, die ein Autor zur Realisierung eines konsistenten Webdesigns verwendet (vgl. auch Abschnitt 14.6.2). Eiron und McCurley (2003) stellen Heuristiken zur Bestimmung von „compound documents" im WWW und ihren Einstiegsseiten vor, die als *Teil* eines umfassenderen Hypertextes betrachtet werden. Die Verfahren sollen der Verbesserung von IR-Verfahren dienen.[58] Vornehmlich entsprechen derartige Dokumente Exemplaren traditioneller Textsorten (vgl. Fußnote 72, S. 696), die von Autoren oder Konvertern in mehrere Knoten aufgeteilt werden. Die Heuristiken basieren auf manuell erstellten Regeln, die auf unterschiedlichen Typen von Verknüpfungen – z. B. „outside links", „across links", „down links" und „up links" – im Dokumentbaum eines Webservers operieren. Obwohl diese Heuristiken den Verfassern zufolge „very high success rates" (ebd., S. 88) liefern, liegt ihnen letztlich ebenfalls die Annahme zugrunde, dass die Verzeichnisstrukturen die Hypertextstrukturierungen reflektieren. Untypische Strukturen (wenn sich z. B. mehrere „compound documents" ein Verzeichnis teilen oder dynamisch generierte URLs eingesetzt werden) können somit nicht erfasst werden, was ebenfalls für die Problematik der Einbettung gilt. Mit „compound documents" (bzw. „cDocs") beschäftigen sich auch Dmitriev et al. (2005), die manuell kategorisierte Trainingsbeispiele und überwachte Lernverfahren zur Erkennung einsetzen. Die Resultate werden von den Verfassern zwar als „good" und „ok" eingeschätzt (ebd., S. 1125), jedoch existieren drei Probleme, die Fehler verursachen: (i) Zunächst handelt es sich um „*singletons*", die als „cDocs consisting of only one page" (ebd.) aufgefasst und von dem Algorithmus nicht erfasst werden. (ii) Das zweite Problem sind „so-called *contents pages*. These are the pages resembling a contents page of a journal" (ebd.), sie verknüpfen zahlreiche „cDocs". (iii) Das letzte Problem ist auf „so-called *vocabulary pages*" zurückzuführen: „These are pages resembling a vocabulary." (ebd.). Dmitriev et al. gehen davon aus, dass ein „singleton" ein „*compound* document" darstellt, das aus nur *einem* HTML-Dokument besteht. Diese Vermischung der konzeptuellen Ebenen kann mit dem Hypertextsortenmodell adäquater beschrieben werden: Einige Dokumente sind als monosequenzierte Instanzen eingebetteter Hypertextsorten („compound documents"), andere als Exemplare von Hypertextknotensorten („singletons") ausgeprägt. Der Problemfall „contents page" zeigt, dass eine Identifizierung von Hypertextknotensorten (hier: *Inhaltsverzeichnis*) in die Erkennung der Grenzen von Hypertexten einzubeziehen ist. Dies gilt ebenso für „vocabulary pages", bei denen jedoch nicht deutlich wird, welche Hypertextknotensorte mit diesem Begriff gemeint ist.

Li et al. (2000, S. 123) stellen ein Regelsystem zur Bestimmung der „logical domains" (im Gegensatz zu „physical domains" wie z. B. `.uni-giessen.de`) einer Website vor, worunter

domain name is a node in the site graph of the domain". Die Verfasser bezeichnen diesen Ansatz als „rather simple definition" (ebd.), die für die in der Evaluation eingesetzten Websites (vgl. Fußnote 56) jedoch in der Regel keine Einschränkungen mit sich bringt.

[58] Es wird eine Kombination mehrerer Heuristiken eingesetzt, denn „there is *no simple formulation of a single technique* that will identify such documents. The problem of reconstructing compound documents can be based on discerning clues about the document authoring process, or by structural relationships between URLs and their content." (Eiron und McCurley, 2003, S. 87; Hervorhebung hinzugefügt, G. R.). In Bezug auf die Erkennung von Hypertextsorten verschärft sich diese Problematik um ein Vielfaches.

„a group of pages that has a specific semantic relation and a syntactic structure that relates them“ verstanden wird (z. B. „a user home page“, „a project group“ und „a tutorial on XML“; ebd.). Zudem besitzt eine „logical domain“ eine „particular function or is self-contained as an atomic information unit“ (ebd., S. 124). Die teilweise domänenabhängigen Erkennungsregeln operieren auf Metadaten, dem Inhalt von `title`-Elementen, URLs, Hyperlinks und der Hyperlinkstruktur und identifizieren zunächst Einstiegsseiten, woraufhin die Grenzen einer „logical domain“ bestimmt werden. Zusätzlich wird ein rudimentäres Inventar von Typen unterschiedlicher Websites angenommen („Entry page for navigation“, „Personal site“, „Topic site“ und „Popular site“). Jede der 11 Regeln besitzt ein positives oder negatives Gewicht und bezieht sich auf einen der genannten Aspekte (z. B. „incoming link anchor text“ oder „title“), woraufhin der vorliegende Wert entweder numerisch (etwa „outgoing link count > 20“) oder mit einem regulären Ausdruck verglichen wird, um z. B. eine Zeichenfolge wie `/~.*/` am Ende einer URL oder `welcome` im Inhalt von `title` zu erkennen, wodurch das der URL zugehörige HTML-Dokument als mögliche Einstiegsseite einer persönlichen Homepage klassifiziert wird. Da derartige manuell erstellte Regelsysteme zwangsläufig Einschränkungen unterliegen, schlagen Li et al. den Einsatz von Lernverfahren zur Gewichtung der Regeln vor, wodurch jedoch nicht die Frage des Abdeckungsgrades der Regeln geklärt wird. Nachdem mit Hilfe des Regelsystems alle URLs eines Webservers bearbeitet wurden, werden diejenigen Adressen, denen ein hoher *score* zugewiesen wurde, als Einstiegsseiten aufgefasst, woraufhin die verbleibenden URLs den „logical domains“ zugeordnet werden. Für diesen Zweck werden Informationen über die Hyperlinkstruktur und die hierarchische Strukturierung des Dateisystems eines Webservers eingesetzt. Insgesamt zeigt das Verfahren Li et al. (2000, S. 131) zufolge „good results“, jedoch ist auch dieser Ansatz, um nur einen problematischen Aspekt zu nennen, nicht in der Lage, eingebettete Hypertexte erkennen zu können.

Die Ansätze zur Erkennung der Grenzen von Hypertexten verwenden einfache Heuristiken, die sowohl auf den Verzeichnishierarchien als auch auf der Verknüpfungsstruktur operieren und die „borders of a site“, die „logical website boundary“, „compound documents“ oder „logical domains“ ermitteln. Keiner der Ansätze ist in der Lage, eingebettete Hypertexte erkennen zu können, doch ist gerade diese Information für die Erkennung von Hypertextsorten von zentraler Bedeutung: Eine Instanz der *persönlichen Homepage eines Wissenschaftlers* umfasst das atomare Hypertextsortenmodul *Publikationsliste*. Kapitel 10 hat gezeigt, dass dieses Hypertextsortenmodul in unterschiedlichen Ausprägungen realisiert werden kann. Zunächst kann es, gemeinsam mit weiteren Hypertextsortenmodulen (etwa dem *Lebenslauf*), unmittelbar in die Einstiegsseite (oder einen anderen Knoten) der Instanz eingebettet sein oder es kann als Hypertextknotensorte fungieren, d. h. ein HTML-Dokument der Instanz enthält ausschließlich die *Publikationsliste*. Sie kann jedoch auch, z. B. im Falle einer umfangreichen Liste von Veröffentlichungen, als Hypertext realisiert sein, der nach hypertextsortenmodulspezifischen Prinzipien sequenziert wird (vgl. Abschnitt 10.5.5). Derartige Hypertexte sind zwar als Konstituenten des übergeordneten Hypertextes aufzufassen, stellen aber dennoch eigenständige Hypertexte dar, weil sie typischerweise eine Einstiegsseite besitzen und ein spezifisches Thema des übergeordneten Hypertextes beinhalten. Derartige Publikationslisten enthalten häufig Hyperlinks zu digitalen Volltextversionen einiger oder aller Veröffentlichungen. Dieses Beispiel zeigt auf, dass eingebettete Hypertexte ihrerseits eingebettete Hypertexte enthalten können: Wenn z. B. ein ursprünglich mit dem Textsatzsystem

TEX/LATEX erstelltes Buch mit Hilfe eines Konverters in einen (hierarchisch sequenzierten) Hypertext überführt (vgl. Abschnitt 3.3.6) und der Titel dieses Buches im entsprechenden Knoten der eingebetteten Instanz der *Publikationsliste* mit der Einstiegsseite des maschinell erzeugten Hypertextes verknüpft wird, liegt zwangsläufig eine weitere Einbettungsebene vor. Der Produzent der übergeordneten Instanz unterliegt jedoch keinerlei Restriktionen bezüglich der Anordnung der Dateien in der zur Verfügung stehenden Verzeichnisstruktur, so dass die *Publikationsliste* zwar als übergeordneter Hypertext fungiert, beide Hypertexte jedoch z. B. in *parallelen* Verzeichnissen gepflegt werden (beispielsweise `.../~paul/publications/` und `.../~paul/textbook/`). Daher kann die Verzeichnisstrukturierung nur für eine sehr rudimentäre Ermittlung der Grenzen von Hypertexten eingesetzt werden.

In Bezug auf die Untersuchungsdomäne der universitären Webangebote betrifft die Einbettungsproblematik insbesondere die Hypertexte, die von einzelnen Organisationseinheiten einer Hochschule gepflegt werden: Der *Webauftritt einer Universität* umfasst z. B. mehrere Instanzen der Hypertextsorte *Webauftritt eines Fachbereichs*, die ihrerseits Exemplare der Hypertextsorte *Webauftritt eines Instituts* beinhalten (vgl. Abbildung 5.2, S. 276). Diese Strukturierung entspricht dem typischen Aufbau einer Hochschule, der in der Domänenontologie durch *has-part-* und *part-of* -Relationen modelliert wird. Hinsichtlich der maschinellen Identifizierung der Grenzen dieser Hypertexte wäre es möglich, den Klassen der Domänenontologie, die Organisationseinheiten modellieren, charakteristische Merkmale (z. B. *cue phrases* wie „Fachbereich“ in der Überschrift eines Dokuments) zuzuweisen. Mit diesem Verfahren könnte zwar eine grobe Differenzierung der einzelnen Hypertexte universitärer Webauftritte erzielt werden, jedoch ist ein domänenunabhängiges Verfahren zu präferieren, weil die Domänenontologie z. B. nicht in der Lage ist, untypische Strukturen detektieren zu können.

14.6.2 Merkmale zur Identifizierung von Hypertextsorten

Kategorisierungsverfahren basieren auf einem Kategoriensystem, auf das die zu klassifizierenden Objekte abgebildet werden. Hierfür ist es notwendig, die Werte spezifischer Merkmale der zu verarbeitenden Objekte zu ermitteln und mit den Merkmalsausprägungen zu vergleichen, die die Einheiten des Kategorisierungsschemas beinhalten. Die Identifizierung der Hypertextsorte kann somit als Kategorisierungsaufgabe konzeptualisiert werden: Die in der Ontologie enthaltenen Hypertextsorten bilden das Kategoriensystem, auf das zu verarbeitende Hypertexte abzubilden sind, indem Merkmalswerte maschinell erhoben und mit den korrespondierenden Merkmalen in der Ontologie verglichen werden. Hierfür können, wie Abschnitt 14.2.3 gezeigt hat, verschiedene Methoden eingesetzt werden, z. B. Lernverfahren oder manuell erstellte Abbildungsregeln mit individuellen Gewichtungen.

Die Kategorisierung von Texten in *thematische* Kategorien beruht – abstrakt formuliert – auf Lexemen, die für spezifische Themen charakteristisch sind (vgl. Abschnitt 14.2.1). Die Detektion des Genres eines Textes basiert primär auf linguistischen Merkmalen (vgl. Abschnitt 14.2.2), die wiederum für Beispieldokumente spezifischer Genres charakteristisch sind. Die Kategorisierung von Hypertexten in ihre Hypertextsorten wird diesbezüglich mit einer Problematik konfrontiert, die die Komplexität des Merkmalsraumes betrifft, denn es kann bzw. muss innerhalb eines derartigen Kategorisierungsprozesses eine *Vielzahl* heterogener Eigenschaften berücksichtigt werden, um einen hohen Abdeckungsgrad zu erzielen

(vgl. Abschnitt 14.2.3). Nachfolgend werden die wesentlichen automatisch[59] extrahierbaren Merkmale, Merkmalgruppen und Dokumenteigenschaften, die von einem System zur Identifizierung von Hypertextsorten eingesetzt werden können, vorgestellt.[60]

Metainformationen und Metadaten

Diese erste Gruppe von Merkmalen bezieht sich auf Informationen über ein HTML-Dokument bzw. eine Gruppe von Dokumenten (d. h. einen Hypertext).

URL eines Dokuments – Die in Teil III präsentierten Studien haben mehrfach den Umstand thematisiert, dass URLs oftmals „sprechende" Datei- und Verzeichnisnamen besitzen (vgl. Chi et al., 1999, und Shih und Karger, 2004), von denen wiederum auf die korrespondierende Hypertextknotensorte oder Hypertextsorte geschlossen werden kann. Von besonderer Bedeutung ist ein initial verwendetes ~-Zeichen, das in vielen Fällen zur Adressierung von Instanzen des Hypertexttyps *Homepage einer Person* verwendet wird. Häufig werden auch Dateinamen wie z. B. `publikationen.html` oder `kontakt.html` benutzt, die ebenfalls zur Erkennung eingesetzt werden können. Diese Beispiele zeigen, dass ein Bedarf besteht, die URL in die Analyse einfließen zu lassen. Zusätzlich könnte ein Eigennamenerkenner eingesetzt werden, um z. B. die einer Tilde folgende Zeichenkette als Eigennamen zu klassifizieren, wodurch ein stärker gewichteter Indikator für den Hypertexttyp *Homepage einer Person* vorläge (Shakes et al., 1997). Eine Kategorisierung, die *ausschließlich* auf URL-Mustern beruht, liefert jedoch keine präzisen Ergebnisse (vgl. Heißing, 2000, und Kan, 2004).

HTTP-Response-Header – Das Feld `Last-Modified:` (vgl. Abschnitt 7.2.3) enthält häufig einen Zeitstempel, der das Datum der zuletzt durchgeführten Modifikation markiert. Falls andere Komponenten ein HTML-Dokument z. B. als Instanz der Hypertextknotensorte *aktuelle Meldungen und Informationen* markieren, kann dieser Zeitstempel untersucht werden, um zu ermitteln, ob tatsächlich eine kürzlich durchgeführte Modifikation vorliegt, wodurch die bereits erfolgte Kategorisierung zusätzlich gewichtet werden kann. Durch die Analyse anderer Felder des HTTP-Response-Headers kann bestimmt werden, ob ein Dokument von einem CGI-Skript (falls z. B. `Expires:` vorliegt) oder dynamisch (`Last-Modified:` liegt *nicht* vor)

[59] Der Aspekt der maschinellen Extrahierbarkeit der im Folgenden diskutierten Merkmale bezieht sich auf die prinzipielle Möglichkeit der Implementierung korrespondierender Methoden. Einige der Merkmale können mit einfachen Mitteln bestimmt werden; im Rahmen der vorliegenden Arbeit wurden diesbezüglich mehrere prototypische Werkzeuge implementiert. Verschiedene hier aufgeführte Merkmale werden in der Literatur bislang nicht berücksichtigt und dürften nur mit sehr aufwändigen Verfahren maschinell zu ermitteln sein. Ivory und Hearst (2002) verwenden in diesem Zusammenhang 157 Merkmale, um die Qualität einer Website maschinell zu bestimmen; sie teilen die Merkmale in die Gruppen „Text Elements" (31 Merkmale), „Link Elements" (6), „Graphic Elements" (6), „Text Formatting" (24), „Link Formatting" (3), „Graphic Formatting" (7), „Page Formatting" (27), „Page Performance" (37) und „Site Architecture" ein. Das zur Ermittlung der Merkmale implementierte Werkzeug arbeitet mit einer Präzision von 84%. Eine der im Folgenden präsentierten Liste vergleichbare Aufstellung von Merkmalen zur „automatic genre classification of web documents" wird von Lim et al. (2005b, S. 1264) präsentiert, die darauf abzielen, „as many features as possible to classify genres for web documents" aufzuführen (vgl. Abschnitt 14.2.3). Die sehr hohe Zahl von 329 Merkmalen erklärt sich durch die Granularität der Liste, so werden z. B. 17 unterschiedliche Phrasentypen (NP, VP etc.) und 72 verschiedene HTML-Elemente jeweils als einzelne Merkmale aufgeführt (vgl. Tabelle 14.3, S. 644).

[60] Diese Liste von Detektionsmerkmalen ist nicht als abgeschlossen aufzufassen. Eine initiale Version der Liste wurde ausschnittsweise in Rehm (2002b) publiziert.

generiert wurde. Hierdurch kann die Gruppe zu berücksichtigender Hypertext(knoten)sorten ebenfalls eingeschränkt werden. Mit Hilfe des Feldes `Set-cookie:` kann ein Cookie im Browser des Rezipienten gesetzt werden (vgl. Fußnote 100, S. 104). Diese Erzeugung eines Zustandes wird häufig in E-Commerce-Anwendungen eingesetzt und kann ebenfalls der Restriktion auf eine spezifische Gruppe von Hypertext(knoten)sorten dienen.

Dokumenttitel – Der innerhalb des HTML-Elements `title` enthaltene Dokumenttitel umfasst in vielen Fällen einen Hinweis auf die Hypertextsorte oder Hypertextknotensorte. Die Titel der *persönlichen Homepages von Wissenschaftlern* lauten z. B. in vielen Fällen „homepage", „Dr. *Vorname Nachname*" und „Homepage von *Vorname Nachname*". Auch der `title` sollte auf die Existenz von Eigennamen und Schlüsselwörtern untersucht werden.

Metadaten-Elemente – Explizite Metadaten in Form von `meta`-Elementen stehen zwar in den Dokumenten, die im Korpus enthalten sind, nur in den wenigsten Fällen zur Verfügung (vgl. die Abschnitte A.4.4 und 14.2.1), jedoch kann das von der *Dublin Core Metadata Initiative* vorgeschlagene Inventar von Metadaten (vgl. Abschnitt 3.6.6) ebenfalls zur Bestimmung der Hypertextsorte eingesetzt werden.

Dokumentumfang – Der Umfang eines Dokuments kann über einfache Maße wie z. B. die Anzahl Wörter oder die Anzahl Sätze ausgedrückt werden (vgl. Abschnitt 14.2.2). Hierdurch können, um extreme Beispiele zu nennen, Instanzen der Hypertextknotensorten *Kontaktinformationen* und *wissenschaftlicher Artikel* unterschieden werden.

Die HTML-Struktur eines Dokuments

Der zentrale Stellenwert einer Analyse der HTML-Auszeichnungen eines Dokuments wurde in Abschnitt 14.5 ausführlich thematisiert: Diese ist zwingend notwendig, um die Bausteine der Textoberfläche eines Knotens ermitteln und anschließend auf Hypertextsortenmodule abbilden zu können. Somit müssen Informationen über die Makrostruktur ebenfalls in den Kategorisierungsprozess einfließen. Wenn ein Dokument z. B. lediglich eine Liste von Hyperlinks zu externen Webangeboten enthält, kann diesem Knoten unmittelbar die Hypertextknotensorte *Hotlist* zugeordnet werden. Zu diesen Merkmalen zählen auch typografische Eigenschaften (Formatierung von Text und Hyperlinks, Einsatz von Tabellen und Farben etc.). Die nachfolgenden Abschnitte gehen auf spezifischere Merkmale ein, von denen die meisten ebenfalls der HTML-Struktur zugehörig sind.

In einem Knoten oder einem Hypertext enthaltene Hyperlinks

Die Analyse von Hyperlinks – siehe auch das Beispiel zur *Hotlist* im vorangegangenen Abschnitt – spielt bei der Erkennung von Hypertextsorten eine zentrale Rolle (vgl. Crowston und Williams, 1999, 2000, Haas und Grams, 2000, Roussinov et al., 2001). Dieser Abschnitt geht auf Merkmale ein, die sich sowohl auf einzelne Knoten als auch auf einen Hypertext beziehen. Externe Verknüpfungen sind ebenfalls einzubeziehen (Amitay et al., 2003).

Anzahl Hyperlinks – Eine sehr große oder sehr kleine Anzahl von Hyperlinks in einem Dokument kann – in Relation zur Anzahl der enthaltenen Wörter – für spezifische Hypertextknotensorten bzw. Hypertextsortenmodule charakteristisch sein. Ein Dokument, das *keine* Hyperlinks enthält, kann als Blattknoten des übergeordneten Hypertextes aufgefasst werden,

so dass diejenigen Hypertextsortenmodule, die als Hypertextknotensorten fungieren können und Hyperlinks enthalten müssen, von der Kategorisierung ausgeschlossen werden können.

Seiteninterne, lokale und externe Hyperlinks – Seiteninterne Hyperlinks führen zu einem Ziel, das sich innerhalb des aktuellen Knotens befindet. Lokale Hyperlinks zeigen zu anderen Knoten des aktuellen Hypertextes, und externe Hyperlinks führen zu externen Angeboten. Zusätzlich kann in bestimmten Untersuchungsdomänen eine weitere Ebene angenommen werden, denn Hyperlinks können ebenfalls zu einem anderen Hypertext der übergeordneten Institution verweisen. Diesbezüglich existieren Korrelationen, die sowohl den Ziel- als auch den Ausgangsknoten eines Hyperlinks betreffen, wovon unmittelbar die Analyse der URL eines Knotens betroffen ist: Die Hypertextknotensorte *Publikationsliste* wird z. B. nicht in Dokumenten mit einer URL wie `http://www.uni-stadt.de` instanziiert. Wenn mehrere Knoten eines Hypertextes rekurrent auf *einen* übergeordneten Knoten verweisen, kann geschlossen werden, dass es sich um die Einstiegsseite dieses Hypertextes handelt, falls dieser Knoten seinerseits auf die untergeordneten Knoten verweist.

Sequenzierung und Strukturierung eines Hypertextes – Dieses Merkmal betrifft die Verwendung allgemeiner Sequenzierungsmuster (vgl. Abschnitt 3.5.1). Durch eine Traversierung aller Knoten, die von einer Einstiegsseite dominiert werden, kann die Graphstruktur eines Hypertextes ermittelt werden (vgl. Ester et al., 2002, sowie Mehler et al., 2004). Auf diese Weise kann auch der Status eines Knotens in Bezug zum zugehörigen Hypertext bestimmt werden (z. B. *Sitemap* vs. *Seite/Abschnitt*). Zudem zeigt die globale Sequenzierungsebene, dass weitere Typen von Hyperlinks existieren, die z. B. in einem maschinell nach HTML konvertieren Buch zum *nächsten* („next“, „forward“) oder *vorherigen* („previous“, „back“) Abschnitt oder Kapitel verweisen. Eine derartige Analyse kann etwa der Ermittlung einer linearen (*guided tour*, *forced march*) hierarchischen oder ungeordneten Sequenzierung dienen, jedoch werden im WWW sehr häufig Mischformen eingesetzt (vgl. Gillenson et al., 2000), so dass das Kriterium der Robustheit für ein solches Verfahren eine zentrale Rolle spielt: Für seine Implementierung ist es notwendig, ein Inventar grundlegender Sequenzierungstypen festzulegen (zur Repräsentation kann die Hypertextsortenontologie eingesetzt werden), den Repräsentationen von Hypertextsorten einen oder mehrere dieser Typen zuzuweisen, um im Rahmen der Erkennung eine beobachtete Strukturierung auf einen der Typen abzubilden (vgl. Amitay et al., 2003). Eine präzise Erkennung des Sequenzierungstyps kann zur Einschränkung des Suchraumes bezüglich der in Frage kommenden Hypertextsorten eingesetzt werden.

Methode bzw. Protokoll – Die Methode bzw. das verwendete Protokoll stellt den ersten Teil einer URL dar, im Regelfall wird HTTP eingesetzt. Obwohl nur selten andere Methoden benutzt werden (vgl. Abschnitt A.4.3), können sie als Indikatoren für Hypertextsorten und Hypertextsortenmodule dienen: `ftp:`-Hyperlinks, die in Listenform präsentiert werden, deuten auf das Hypertextsortenmodul *Download-Liste* hin, mit deren Hilfe z. B. Software-Archive zur Verfügung gestellt werden. Die Protokolle `https:` und `ftps:` stellen verschlüsselte Varianten dar und werden insbesondere für sensitive Daten eingesetzt (z. B. E-Commerce-Angebote und Online-Shops im Falle von `https:` sowie Download-Listen für Applikationen, die online bezahlt werden, bezüglich der Verwendung von `ftps:`).

Dateityp des Hyperlinkziels – Der Dateityp wird im Allgemeinen mit einem Suffix wie z. B. `.pdf`, `.zip`, `.tar.gz` oder `.ps.gz` markiert. Über eine HTTP-Anfrage kann der Medien-

typ einer Datei bestimmt werden (dieser wird auch gelegentlich über das Attribut `type` des Elements `a` markiert; vgl. Abschnitt A.3.4). Falls ein Hyperlink also nicht auf ein HTML-Dokument (Medientyp `text/html`) verweist, kann diese Information zur Bestimmung des Hypertextsortenmoduls *Download-Liste* (siehe *Methode bzw. Protokoll*) verwendet werden, falls die Hyperlinks auf komprimierte Archivdateien verweisen. Bei einer Liste von Hyperlinks zu PDF-Dokumenten kann es sich um die Instanz einer *Publikationsliste* handeln, die zusätzlich zu den Literatureinträgen auf digitale Versionen der Veröffentlichungen verweist.

Hypertextknotensorte des Hyperlinkziels – Die Hypertextknotensorte des Dokuments, auf das ein Hyperlink verweist, kann zur Bestimmung der Hypertextknotensorte des Ausgangsknotens eingesetzt werden. Falls der Zielknoten z. B. bereits als *Publikationsliste* identifiziert werden konnte, kann die Gruppe der potenziellen Hypertextknotensorten des aktuellen Knotens auf diejenigen Kategorien eingeschränkt werden, die typischerweise eine Publikationsliste referenzieren. Diese Vorgehensweise kann auch als *fallback*-Strategie eingesetzt werden, falls für den aktuellen Knoten kein eindeutiges Ergebnis ermittelt werden kann.

Hyperlinkanzeiger – Der Hyperlinkanzeiger kann mit computerlinguistischen Verfahren untersucht werden. Wenn z. B. „Andreas Neumann“ verwendet wird, kann das Ergebnis eines Eigennamenerkenners benutzt werden, um die *Einstiegsseite der Homepage einer Person* als potenziellen Hypertextknotentyp des Ziels zu bestimmen. Die Analyse der zugehörigen URL (z. B. `http://www.uni-trier.de/~neumann/`) kann diese Bestimmung bestätigen. Zudem entspricht der erkannte Nachname dem in Verzeichnisnamen der URL.

Funktion und Position des Hyperlinks – Von Bedeutung sind auch Funktion und Position eines Hyperlinks (vgl. die Abschnitte 3.5.5 und 4.4.4). Durch Heuristiken kann einem Link ein Typ wie z. B. „navigation“, „expansion“ oder „resource“ oder eine Positionseigenschaft wie „isolated“, „embedded“ oder „labeled“ zugewiesen werden (vgl. Haas und Grams, 1998a). Da Korrelationen zwischen dem Typ, der Position und dem Hyperlinkanzeiger existieren, ist es prinzipiell möglich, die Funktion maschinell zu bestimmen.

In ein HTML-Dokument eingebettete Bilddateien

Da zahlreiche HTML-Dokumente eingebettete Bilddateien enthalten, ist es notwendig, diese in den Prozess der Identifizierung einer Hypertext(knoten)sorte zu integrieren.

Abmessungen – Die Relevanz der Abmessungen von Bilddateien wurde bereits in Abschnitt 14.5 diskutiert: Kleinformatige GIF-Grafiken stellen oftmals Icons oder *bullet points* dar (z. B. eine Datei namens `red_dot-1.gif` mit den Dimensionen 14×14). Bei einer animierten Grafik der Größe 468×60 handelt es sich aller Wahrscheinlichkeit nach um ein Werbebanner (vgl. Shih und Karger, 2004). Die Abmessungen können über die Attribute `height` und `width` des Elements `img` oder durch eine Analyse der Datei ermittelt werden.

Datei- und Verzeichnisname – Da der eigentliche Inhalt einer Bilddatei nicht zur Verfügung steht (siehe unten), kann durch die Analyse der Datei- und Verzeichnisnamen zumindest eine approximative Bestimmung erfolgen. Logos besitzen oftmals Dateinamen wie z. B. `logoFB4.gif` oder `logo-neu.jpg`, mit Digitalkameras angefertigte Fotos heißen häufig `IMG_1234.JPG` und Icons werden oft aus zentralen Sammlungen bezogen, die sich in entsprechend benannten Verzeichnissen befinden (`.../icons/`). Wenn sich derartige Verzeichnisse außerhalb des Dokumentraumes befinden, dem der aktuelle Hypertext zugehörig ist,

kann weiterhin geschlossen werden, dass es sich möglicherweise um eine offizielle Liste von Icons oder Logos handelt. Über die Korpusdatenbank können zudem populäre Dateinamen von Grafiken ermittelt werden, indem z. B. die Dateinamen der Einträge bestimmt werden, die vom Medientyp `image/gif` sind und eine Dateigröße < 1024 Bytes aufweisen. Bei den 1 346 058 in der Korpusdatenbank verzeichneten Bilddateien, die diesen Angaben entsprechen, handelt es sich nahezu ausschließlich um Icons, die Dateinamen wie z. B. `next.gif` (8 857 Vorkommen), `prev.gif` (7 720), `mail.gif` (2 343), `redball.gif` (1 929), `pfeil.gif` (1 872), `german.gif` (1 146) und `english.gif` (1 067) besitzen. Hochfrequente Dateinamen können über Regeln auf ein Inventar von Icontypen abgebildet werden, die zur Ermittlung der Funktion eines als Hyperlink realisierten Icons eingesetzt werden können.

Inhalt eines Bildes – Bilddateien können unter anderem Fotos, Logos, Zeichnungen, Gemälde, Diagramme, Schriftzüge, Menüeinträge, Hintergrundtexturen oder Comics enthalten. Einige Grafiken bestehen lediglich aus *einem* transparenten Pixel und werden für Layout-Zwecke eingesetzt. Eine grobe Einschätzung des Inhalts einer Bilddatei kann durch die Abmessungen bzw. den Dateinamen (siehe oben) oder das Dateiformat (siehe unten) erfolgen. Im Bereich des *Image Retrieval* wurden Methoden zur Inhaltsbestimmung einer Grafik entwickelt (vgl. Chen et al., 1996, Rui et al., 1999). Falls ein solches Verfahren mit einem sehr breiten Abdeckungsgrad zur Verfügung steht, könnte ein Bild als Porträtfoto kategorisiert werden, wodurch wiederum auf diejenigen Hypertext(knoten)sorten geschlossen werden kann, die typischerweise ein Porträtfoto beinhalten (vgl. Asirvatham und Ravi, 2001).

Alternativer Text – Das Element `img` besitzt das Attribut `alt`, in das eine textuelle Beschreibung eines Bildes eingetragen werden kann (z. B. „Foto von *Vorname Nachname*“). Häufig umfassen die Inhalte von `alt`-Attributen Schlüsselwörter, die zur approximativen Bestimmung des Inhalts oder der Funktion eines Bildes eingesetzt werden können.

Grafikformat – Es existieren zahlreiche Formate von Bilddateien: GIF- und PNG-Dateien werden aufgrund der Spezifikationen dieser Formate vornehmlich für Icons und Diagramme eingesetzt, wohingegen JPEG eher für Fotos benutzt wird (vgl. Abschnitt A.3.4).

Anzahl und Positionierung von Bildern – Die Anzahl und Positionierung von Bildern kann zur Bestimmung von Hypertextknotensorten wie z. B. *Fotogalerie* eingesetzt werden, die *thumbnail*-Bilder enthält, welche wiederum durch Hyperlinks mit hochauflösenden Versionen der Fotos verknüpft sind. Die *thumbnails* werden typischerweise in Form einer Tabelle angeordnet. Navigationshilfen können ebenfalls auf dem Typ der einfachen Tabelle basieren (vgl. Abschnitt 11.4.2) und Hyperlinkanzeiger enthalten, die als Grafiken eingebettet sind, so dass zur Differenzierung zwischen einer Fotogalerie und einer auf diese Weise realisierten Navigationshilfe eine Analyse der eingesetzten Hyperlinks notwendig ist.

Interaktive Komponenten

Die Erkennung interaktiver Komponenten bezieht sich insbesondere auf die Analyse weiterführender Merkmale der *Hypertext Markup Language*.

Formulare – In Bezug auf HTML-Formulare (vgl. Bauer et al., 2000) sind mehrere Aspekte einzubeziehen: Hierzu gehören die Abmessungen des Formulars (erstreckt es sich z. B. über das gesamte Dokument oder beschränkt es sich auf die untere rechte Ecke), Anzahl, Typen

und Anordnung der Formularbestandteile (über das Element `input` realisierte Eingabemasken, Menüs etc.) und die Beschriftung des „submit"-Knopfes, die z. B. „suchen", „kaufen", „bestellen", „Informationsmaterial anfordern" oder „Nachricht abschicken" lauten kann.

JavaScript-Code – Mittels *JavaScript* können statische HTML-Dokumente mit zusätzlichen Interaktionsebenen angereichert werden. Dazu zählt z. B. das farbliche Hervorheben eines Menüeintrags, sobald er vom Mauszeiger erreicht wird oder das Realisieren von *drag and drop*-Übungen im Kontext des WWW-basierten Fremdsprachenlernens. Zur Identifizierung von Hypertextsorten ist es notwendig, den Grad der Interaktivität (vgl. auch Goertz, 1995, S. 102) eines Knotens bzw. Hypertextes zu bestimmen, da Hypertextsorten existieren, die typischerweise keine oder auch sehr umfassende interaktive Komponenten beinhalten.

Java Applets – Über das Element `applet` werden *Java applets* referenziert, deren Größe oftmals von den Attributen `height` und `width` gesteuert wird. Fast alle *applets* stellen interaktive Komponenten dar, so dass ihre Abmessungen ebenfalls in die Berechnung des Interaktivitätsgrades eines Knotens einbezogen werden sollten. Die maschinelle Bestimmung des Inhalts einer *Java*-Anwendung kann nicht erfolgen (um z. B. ein als *applet* realisiertes Werbebanner von einer *Java*-basierten Tabellenkalkulation zu unterscheiden), jedoch kann eine Approximation durch die URL des *applets* und seine Dateigröße stattfinden.

Browser Plug-ins Mit Hilfe von *Plug-ins* kann die Funktionalität eines Webbrowsers erweitert werden, um z. B. unmittelbar im Browser Videodateien anzuzeigen. Dateien, die nur durch ein *Plug-in* dargestellt werden können, werden über die Elemente `embed` und `object` referenziert. Ein HTML-Dokument, das lediglich ein Logo, eine Navigationshilfe, eine E-Mail-Adresse und ein `embed`-Element umfasst, das sich auf das Seitenzentrum auswirkt, kann dem Hypertextknotentyp *Multimedia* zugeordnet werden.

Dokumentübergreifende Merkmale eines Hypertextes

Dokumentübergreifende Merkmale können zur Bestimmung der Grenzen eines Hypertextes und somit auch zur Identifizierung seiner Hypertextsorte dienen. Die Knoten eines Hypertextes können rekurrente Merkmale besitzen, z. B. identische Navigationshilfen, Kopf- und Fußzeilen oder konsistent in der oberen linken Ecke positionierte Logos. Übergeordnete Muster können sich auch auf die in URLs verwendeten Dateinamen beziehen (z. B. `meier.html`, `mueller.html` und `schmidt.html` in einem Verzeichnis namens `.../projekt/`). Die Grenzenbestimmung kann auch auf die Analyse von CSS-Stylesheets ausgedehnt werden: Falls ein Hypertext ein externes Stylesheet einsetzt, ist es als wahrscheinlich anzusehen, dass *sämtliche* Knoten des Hypertextes eben dieses CSS-Stylesheet referenzieren (vgl. auch Gibson et al., 2005). Dies gilt auch für Hintergrundgrafiken und Farbräume.

Linguistische Merkmale

Wörter, Phrasen, Sätze, Texte und Textfragmente stellen – neben grafischen und weiteren multimedialen Inhalten – die zentralen Bestandteile von HTML-Dokumenten dar, weshalb sie naturgemäß in die maschinelle Hypertextsortenidentifizierung einzubeziehen sind. Die nachfolgend dargestellten Merkmale können mit computerlinguistischen Methoden ermittelt werden und sollten – zusätzlich zu den bereits in Abschnitt 14.2.2 diskutierten Merkmalen

(z. B. *Part-of-Speech*-Frequenzen, Wortfrequenzen, Anzahl Sätze, Interpunktion etc.) – in der Implementierung eines derartigen Systems Verwendung finden.[61]

Charakteristische Schlüsselwörter und -phrasen – Zahlreiche Hypertextsortenmodule und Hypertextknotensorten besitzen charakteristische Phrasen (*cue phrases*), z. B. „Frequently Asked Questions", „Lebenslauf", „Persönliche Homepage von *Vorname Nachname*" – oder lediglich „*Vorname Nachname*" – (vgl. Teil III). Neben Überschriften können derartige Spezifika unter anderem auch in Hyperlinkanzeigern und URLs beobachtet werden (vgl. Liu et al., 2003). Somit können individuelle Schlüsselwörter und charakteristische Muster in die Erkennung von Hypertextsorten einfließen. Neben Eigennamen (vgl. Meyer zu Eissen und Stein, 2004), die den alleinigen Inhalt der Überschrift einer Einstiegsseite bilden, können sich *cues* auch auf die Textstruktur beziehen: Datumsangaben sind zwar nicht standardisiert, werden aber mit einem begrenzten Inventar formatiert. Eine Liste von Hyperlinks, die einen initialen Datumsausdruck besitzen, kann auf das Hypertextsortenmodul *neue oder modifizierte Dokumente einer Website* abgebildet werden (sofern es sich um lokale Hyperlinks handelt). In gleicher Weise werden häufig E-Mail-Adressen, die *nicht* als `mailto:`-Link realisiert sind, in Instanzen der Hypertextknotensorte *Mitarbeiterverzeichnis* verwendet (vgl. de Saint-Georges, 1998, Toms und Campbell, 1999, Haas und Grams, 2000, und Roussinov et al., 2001).

Grad der konzeptionellen Mündlichkeit bzw. Schriftlichkeit – Zahlreiche Studien im Bereich der *Computer-Mediated Communication* betonen, dass zugehörige Kommunikationsformen von einer Tendenz zur konzeptionellen Mündlichkeit geprägt sind. Dieses Phänomen kann, wie die Kapitel 8 und 9 gezeigt haben, auch in HTML-Dokumenten beobachtet werden. Auf der Grundlage des in Kapitel 8 eingesetzten Werkzeugs zur maschinellen Ermittlung dieser Merkmale in Kombination mit traditionellen korpuslinguistischen Verfahren ist es möglich, Dokumente auf einem Kontinuum zu verorten, das sich zwischen den Polen der konzeptionellen Mündlichkeit und Schriftlichkeit aufspannt – ein derartiges Kontinuum kann als textuelle Dimension im Sinne von Biber (1988) verstanden werden. Mit einem solchen Verfahren kann Hypertextsorten wie z. B. *private Homepage eines Studierenden*, *Gästebuch* und *Messageboard/Diskussionsforum* eine ausgeprägtere konzeptionelle Mündlichkeit zugeschrieben werden als z. B. einem *Protokoll* oder einem *wissenschaftlichen Artikel*.

Rechtschreibung – Neben Merkmalen für konzeptionelle Mündlichkeit weisen unter anderem private Homepages von Studierenden oftmals Rechtschreibfehler auf, die sich aus der Kommunikationssituation, den Produktionsbedingungen und dem Umstand erklären, dass keine Kontrollinstanz vorliegt. Auf der Basis einer Rechtschreibhilfe wie etwa *ispell* ist es möglich, Fehler maschinell zu detektieren und den Prozentsatz fehlerhaft geschriebener Wörter zu ermitteln, die z. B. für die Hypertextsorten *private Homepage eines Studierenden*, *Gästebuch*, *Messageboard/Diskussionforum* und *E-Mail* charakteristisch sind. Zudem weisen viele HTML-

[61] Stamatatos et al. (2001, S. 473 ff.) klassifizieren die bislang vorgeschlagenen Maße: (i) „token-level measures" können mit einfachen Mitteln berechnet werden und betrachten einen Text als Menge von Sätzen, die Wörter umfassen, es können Maße wie z. B. die Anzahl Wörter, Anzahl Sätze, Anzahl Zeichen pro Wort und Anzahl Interpunktionszeichen erhoben werden. (ii) Die „syntactic annotation" setzt einen POS-Tagger oder einen Parser voraus und betrifft Maße wie die Anzahl von Passivkonstruktionen, Nominalisierungen und Wortarten (vgl. Santini, 2005b). (iii) Die „vocabulary richness" kann z. B. mit einer „type-token ratio" ermittelt werden, die die in einem Text verwendeten Wörter (*token*) in ein Verhältnis zum verwendten Vokabular (*types*) setzt. (iv) Die „common word frequencies" beziehen sich auf Vorkommen spezifischer Wörter (vgl. Abschnitt 14.2.2).

Dokumente auch fehlerhafte Namen von HTML-Elementen und -Attributen auf (vgl. Abschnitt A.4.2). Eine Ermittlung derartiger Fehler, die nur bei der manuellen Erstellung mit einem ASCII-Editor entstehen können (z. B. `adress` statt `address`, vgl. Abschnitt 3.3.6), kann, wie Walker (1999) anmerkt, eingesetzt werden, um den Grad der Spontaneität der Erstellung eines Dokuments zu bestimmen (vgl. Rosso, 2005, S. 78).

Grad der Linearität – Dieses Merkmal betrifft unmittelbar den Einsatz von Hyperlinks: Liegen im Extremfall keine Hyperlinks (in einem Fließtextabschnitt) vor, handelt es sich um eine lineare Ausprägung, verweisen jedoch zahlreiche als Hyperlinkanker realisierte Wörter auf weitere Knoten eines (unsequenzierten) Hypertextes, kann ein entsprechend geringer Grad der Linearität markiert werden (vgl. hierzu aus medienübergreifender Perspektive Goertz, 1995, S. 111 f.). Ein korrespondierendes Analyseverfahren müsste in der Lage sein, ebenfalls die bereits bei den Hyperlinks angesprochenen Merkmale verarbeiten zu können.

Fazit – Zur Kombination der Merkmale

Selbst wenn die traditionellen linguistischen Eigenschaften *nicht* berücksichtigt werden, steht einer Komponente zur maschinellen Identifizierung von Hypertextsorten eine sehr große Zahl von Merkmalen zur Verfügung. In Abschnitt 14.5.2 wurden verschiedene Ansätze zur Ermittlung makrostruktureller Komponenten von HTML-Dokumenten vorgestellt. Diese Verfahren operieren, ebenso wie der Textparser, auf Kombinationen mehrerer Merkmale: Zunächst werden ihre Ausprägungen bzw. die spezifischen Eigenschaften eines Dokuments bestimmt, woraufhin sie z. B. in Regelsystemen oder maschinellen Lernverfahren eingesetzt werden. Bereits die Auszeichnung eines Dokuments mit HTML-Elementen und -Attributen kann also als komplexe Menge von Merkmalen aufgefasst werden (vgl. Kruschwitz, 2001).

Die Realisierung eines Systems zur automatischen Identifizierung von Hypertextsorten setzt Verfahren voraus, die in der Lage sind, zahlreiche der in diesem Abschnitt vorgestellten Merkmale erheben zu können. Die Merkmalsbestimmung muss sowohl mit einer hohen Präzision als auch mit einer umfassenden Robustheit erfolgen, die einzelnen Komponenten müssen also auch bei unerwarteten Eingaben zufriedenstellende Ergebnisse in Form verwertbarer Merkmalsausprägungen liefern. In einem nachfolgenden Schritt sind die Merkmale zur Kategorisierung einzusetzen, die sich entweder (i) im Falle eines Knotens auf die Ermittlung der Hypertextknotensorte bzw. (ii) der Hypertextsortenmodule oder (iii) auf die Hypertextsorte eines Hypertextes bezieht. Zusätzlich zu der Problematik, dass diese drei unterschiedlichen Anwendungen im besten Fall interagierend[62] realisiert werden sollten (vgl. Abbildung 14.1, S. 650), existiert diesbezüglich die Schwierigkeit der Bestimmung *charakteristischer Merkmale*. Für *jede* maschinell zu kategorisierende Entität werden umfangreiche Sammlungen von Trainingsbeispielen benötigt. Die Literatur zum Bereich des maschinellen Lernens weist zudem darauf hin, dass die *Auswahl der Merkmale*, die in Lernverfahren eingesetzt werden, von entscheidender Bedeutung zur Erreichung einer hohen Präzision ist. Ein weiteres Problem betrifft die *Anzahl der Kategorien*. Im Rahmen der empirischen Analysen wurde eine

[62] Falls für einen Hypertext bereits eine Hypertextsorte ermittelt werden konnte, könnten einige Dokumente verbleiben, deren Hypertextknotensorten noch nicht identifiziert werden konnten. Somit kann die Information, dass eine spezifische Hypertextsorte vorliegt, eingesetzt werden, um den Suchraum bezüglich der Hypertextknotensorten der verbleibenden Dokumente entsprechend zu restringieren. Entsprechend können vorliegende Hypertextknotensorten zur Suchraumeingrenzung bei der Bestimmung der Hypertextsorte eingesetzt werden.

Vielzahl von Hypertextsorten, Hypertextknotensorten und Hypertextsortenmodulen ermittelt (vgl. Teil III). Mit zunehmender Kategorienanzahl verringert sich jedoch die Präzision maschineller Lernverfahren. Hypertextsortenmodule als grundlegende Bausteine von Hypertextknotensorten stellen für die maschinelle Erkennung ein weiteres Problem dar.

14.6.3 Identifizierung von Hypertextknotensorten und Hypertextsortenmodulen

Sowohl die Beschreibung als auch die maschinelle Erkennung von Hypertextsorten kann sich nicht auf Einzeldokumente beschränken. Vielmehr ist es notwendig, einerseits die Ebene des gesamten Hypertextes, andererseits die Binnenstruktur der beteiligten Knoten zu untersuchen, um die Ausprägungen von Hypertextsortenmodulen zu ermitteln. Um ein bereits genanntes Beispiel erneut aufzugreifen: Das Hypertextsortenmodul *Publikationsliste* kann in der Einstiegsseite der persönlichen Homepage eines Wissenschaftlers instanziiert werden, der Produzent kann es auch in einen eigenständigen Knoten auslagern, so dass es als Hypertextknotensorte fungiert. Sollte es sich um eine sehr umfangreiche Auflistung handeln, kann sie als eingebetteter Hypertext realisiert werden. Hierbei werden Strukturierungsprinzipien angewendet, die sich auf das korrespondierende Thema oder Textstrukturmuster beziehen.

Die maschinelle Identifizierung von Hypertextsortenmodulen stellt die zentrale Komponente eines Systems zur Bestimmung von Hypertextsorten dar. Der Prozess sollte auf den Ergebnissen des Textparsers basieren, um Hypertextmodule auf Hypertextsortenmodule abzubilden. Es werden zwei Aspekte deutlich: Zum einen ist es notwendig, eine Wissensbasis zu integrieren, die die Hypertextsortenmodule verschiedener Hypertextsorten umfasst. Dieses Wissen wird empirischen Analysen entnommen und ist Teil der Hypertextsortenontologie. Zum anderen setzt diese Abbildung Methoden voraus, die über Merkmalsvektoren hinausgehen: Die bisherigen Ansätze berechnen für jedes HTML-Dokument die individuellen Ausprägungen eines Inventars von Merkmalen, fügen sie zu einem Vektor zusammen und vergleichen ihn mit den in der Trainingsphase erstellten Kategorienbeschreibungen (vgl. Abschnitt 14.2.3). Diese Vorgehensweise kann zwar auf die Ermittlung von Hypertextsortenmodulen angewendet werden, hierzu müssten jedoch umfangreiche Sammlungen von Trainingsdaten angefertigt und manuell annotiert werden. Die vorliegenden Arbeiten setzen entweder maschinelle Lernverfahren ein (z. B. Shepherd et al., 2004, und Lim et al., 2005a, 2005b) oder verwenden Regeln (Matsuda und Fukushima, 1999). Da zu erwarten ist, dass die Identifizierung von Hypertextsortenmodulen zumindest teilweise durch Regelapparate realisiert werden kann, ist eine hybride Systemarchitektur vorzusehen, die beide Methoden kombiniert, d. h. die in Abschnitt 14.6 aufgeführten Merkmale sind nicht auf Lernverfahren zu beschränken, sondern zusätzlich um symbolische bzw. heuristische Verfahren zu ergänzen.[63] Die Identifizierung der Hypertextsortenmodule ermittelt explizite Informationen über die

[63] Kwasnik et al. (2001) streben die Implementierung eines Systems zur „automatic genre identification" an und präferieren ebenfalls einen „hybrid approach", der den „heuristic approach" und den „machine learning approach" kombiniert: Ersterer bezieht sich z. B. auf Regeln, die die Überschrift „Frequently Asked Questions" mit dem Genre „FAQ document" oder eine Tilde in der URL mit dem Genre „home page" assoziieren. Der „machine learning approach" betrifft in diesem Vorhaben die manuelle Etikettierung der Genres von 1 000 Dokumenten, um diese als Trainingsdaten einsetzen zu können.

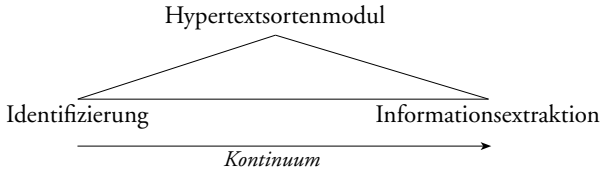


Abbildung 14.7: Erkennung von Hypertextsortenmodulen

Makrostrukturbausteine eines HTML-Dokuments, die von XML-Elementen, die mit Konzepten der Hypertextsortenontologie korrespondieren, in den analysierten Dokumenten reflektiert werden. Es liegt nahe, diesen Prozess als Problem der Informationsextraktion (IE) aufzufassen, wobei unterschiedliche Grade der Granularität vorliegen können: Im Gegensatz zu genuinen IE-Anwendungen ist es bei einem gegebenen Hypertextmodul nicht notwendig, z. B. Vorname, Nachname und Titel einer Person präzise zu extrahieren; zur Zuweisung eines Hypertextsortenmoduls wie *Name des Homepage-Besitzers* sind die Informationen ausreichend, dass es sich bei dem Inhalt eines Hypertextmoduls um den Namen einer Person handelt und dass es zu Beginn der Einstiegsseite eines Hypertextes aufgeführt wird. Die Identifizierung eines Hypertextsortenmoduls kann also als minimale Form der Informationsextraktion aufgefasst werden (vgl. Abbildung 14.7), denn der Prozess der Erkennung extrahiert bereits Informationen über einen Baustein eines HTML-Dokuments. Diese können wiederum eingesetzt werden, um auf zusätzliche, spezifischere Informationen zuzugreifen.

Unterschiedliche Hypertextsortenmodule setzen unterschiedliche Informationen, unterschiedliche Regeln und unterschiedliche Ressourcen voraus, um sie identifizieren zu können. Eine korrespondierende Implementierung, die einen sehr hohen Abdeckungsgrad anstrebt, wird mit einem Komplexitätsproblem konfrontiert, denn es existiert eine Vielzahl von Hypertextsortenmodulen (vgl. Teil III), die aufgrund der individuellen Realisierungen durch einzelne Produzenten auf zahlreiche unterschiedliche Weisen instanziiert werden. Diese individuellen Ausprägungen müssen jedoch entweder durch Regelsysteme, eine entsprechende Anzahl repräsentativer Trainingsbeispiele für ein maschinelles Lernverfahren oder ein hybrides Verfahren abgedeckt werden, um identifiziert werden zu können. Die immense Komplexität eines solchen Vorhabens wird durch eine Betrachtung verwandter IE-Verfahren deutlich, die sich auf HTML-Dokumente beziehen: Derartige Systeme werden als *Wrapper* oder auch *Screen-Scraper* bezeichnet und beziehen sich in der Regel auf die Extraktion der Informationen eines *einzelnen* Makrostrukturbausteins (z. B. Produktinformationen in Online-Katalogen). Die maschinelle Identifizierung von Hypertextsortenmodulen könnte durch eine Vielzahl von *Wrappern* realisiert werden, die im Falle der erfolgreichen Erkennung eines Bausteins die erfolgreiche Identifizierung eines Hypertextsortenmoduls markieren.

Informationsextraktion im World Wide Web

Die traditionelle Informationsextraktion beschäftigt sich mit der robusten Verarbeitung freier Texte zum Zwecke der domänenspezifischen Filterung von Informationen (vgl. Cowie und Wilks, 2000, und Grishman, 2003).[64] Meist werden „flache“ computerlinguistische Metho-

[64] Dieser Abschnitt basiert in Teilen auf Rehm (2004d).

den eingesetzt, da diese weniger fehleranfällig sind als die komplexen Komponenten textverstehender Systeme. Die zu extrahierenden Informationen beziehen sich meist auf generische Ereignisse, zu denen die Fragen „wer?“, „was?“, „wem?“, „wann?“, „wo?“ und eventuell „warum?“ beantwortet werden sollen (Neumann, 2001). Modelliert und repräsentiert werden diese generischen Ereignisse als Templates, die in Form von Attribut-Wert-Paaren unterschiedliche Informationstypen spezifizieren. Ein IE-System ermittelt – z. B. über eine syntaktische Analyse, die sich unter anderem auf Subkategorisierungsinformationen und Eigennamen bezieht – Wörter und Phrasen, die in das korrespondierende Template eingetragen werden. Oftmals werden auch Kategorisierungsverfahren verwendet, um etwa den Typ eines Ereignisses zu bestimmen. Ein IE-System umfasst mehrere Komponenten, z. B. einen Sprachenidentifizierer (vgl. Abschnitt 7.2.2), einen Duplikatserkenner (vgl. Abschnitt 7.2.5), einen Tokenisierer und Satzsegmentierer sowie einen syntaktischen Parser. Neben diesen domänenunabhängigen Werkzeugen erledigen domänenabhängige Interpretationsverfahren die eigentliche Extraktion. Aufgrund der Abhängigkeit von domänenspezifischen Verfahren weisen IE-Systeme eine schlechte Adaptivität und Skalierbarkeit auf (Wilks und Catizone, 1999).

Während IE-Systeme mit ASCII-Texten arbeiten, fokussieren WWW-basierte IE-Anwendungen eine Analyse der HTML-Struktur (vgl. Abschnitt 14.5) und den Einsatz von Lernverfahren (Eikvil, 1999). Ein Experte sammelt typische Beispiele und annotiert diejenigen (Typen von) Informationen, die das System mit automatisch erstellten Regeln aus unbekannten Texten extrahieren soll. Die in den annotierten Beispielen enthaltenen Beobachtungen werden also in Form von Klassifikationsregeln generalisiert. Der Vorteil derartiger Methoden liegt in ihrer schnelleren Verfügbarkeit, da keine sprach- oder textverstehenden Verfahren benötigt werden, jedoch setzen sie eine sehr große Menge annotierter und repräsentativer Trainingsdaten voraus. Da HTML-Dokumente in vielen Fällen nicht ausschließlich aus Fließtextabschnitten bestehen, sondern Mischungen aus Textfragmenten, Listen, einzelnen Wörtern, Phrasen, Sätzen und Absätzen umfassen, können IE-Anwendungen nicht ohne Weiteres beliebige HTML-Dokumente verarbeiten, weil sprachlich motivierte Extraktionsregeln aufgrund dieser „textuellen Irregularitäten“ nicht anwendbar sind: Ein *Wrapper* für eine Informationsquelle – ein Dokumenttyp oder alle Dokumenttypen einer Website – wird mit Regeln ausgestattet, die die Positionen der Informationen innerhalb des HTML-Elementbaumes spezifizieren. Daraufhin ist der *Wrapper* in der Lage, diese zur Extraktion der Informationen zu verwenden.[65] *Wrapper* sollten sowohl effizient als auch robust arbeiten, um Strukturänderungen (Pitkow, 1998) kompensieren können. Sie operieren vornehmlich auf dem Elementbaum und werden in vielen Fällen auf Websites angewendet, die ihre Dokumente dynamisch aus den Beständen einer Datenbank generieren. Da hierbei Daten in vorgefer-

[65] Bei diesen Informationen handelt es sich oftmals nicht um Aspekte generischer Ereignisse, sondern vielmehr um Daten (z. B. Produkt, Hersteller und Preis oder Wohnung, Anzahl Räume, Miete und Lage), weshalb viele IE-Systeme, die sich auf das WWW-beziehen, auch als *Data Extraction*-Anwendungen bezeichnet werden. Unter dem Aspekt der Datenbanksicht auf das WWW wurden zahlreiche Ansätze vorgelegt, die sich mit SQL-ähnlichen Anfragesprachen an Webseiten bzw. semistrukturierte Daten beschäftigen (vgl. Florescu et al., 1998, und Abiteboul et al., 2000). Query-Sprachen werden auch häufig für IE-Zwecke eingesetzt. Ein weiterer Aspekt betrifft die Integration von Daten, die aus *unterschiedlichen* Quellen stammen. Hierfür werden Mediatoren eingesetzt. Ein Mediator nimmt eine Benutzeranfrage entgegen, verteilt diese auf mehrere *Wrapper*, nimmt die extrahierten Informationen entgegen und bündelt diese in Form eines Antwortdokuments (vgl. Ashish und Knoblock, 1997, May und Lausen, 2000).

tigte Schablonen eingefügt werden, besitzen die Informationen immer eindeutige Positionen innerhalb des Elementbaums und können präzise extrahiert werden.[66]

Wrapper werden mit manuellen, halb- oder vollautomatische Methoden erstellt (vgl. Eikvil, 1999). Die einfachste Möglichkeit besteht in der Analyse der HTML-Strukturierung und der manuellen Implementierung. Dieser Ansatz ist zwar sehr effizient (vgl. Soderland, 1997), problematisch ist jedoch seine Fehleranfälligkeit, da es keineswegs trivial ist, aus einigen typischen HTML-Dokumenten eine robuste Implementierung zu generalisieren. Eine Erleichterung stellen spezielle Programmiersprachen (z. B. JEDI oder token-templates, Huck et al., 1998, Thomas, 2000) dar. Die „Web Language" bietet etwa Mechanismen zum Umgang mit HTTP, HTML und XML sowie mit XPath vergleichbare Zugriffsmöglichkeiten (Kistler und Marais, 1998). Myllymaki (2001) konvertiert HTML-Dokumente nach XHTML, um Informationen durch XSLT-Stylesheets extrahieren und nach XML überführen zu können: Falls ein XSLT-Template mit einem XPath-Ausdruck wie z. B. `td[contains(.,'Last Trade')]` zutrifft, kann der Inhalt von `b`, das sich innerhalb von `td` befindet, selektiert und durch das neue Element `price` verkapselt werden. Paradis (2000) schlägt den Einsatz einer Dokumentgrammatik vor, die reguläre Ausdrücke enthält. Mit Hilfe solcher Grammatiken ist es möglich, im Kontext eines IR-Systems lediglich die relevanten Abschnitte derjenigen Dokumente zu extrahieren und in einem „virtuellen Dokument" zu aggregieren, die als Treffer zurückgeliefert werden. RDF-Beschreibungen werden zur Konfiguration von *Wrappern* in dem Projekt VIPAR (Potok et al., 2002) eingesetzt, dessen Ziel es ist, XML-annotierte Artikel aus den Websites verschiedener Tageszeitungen zu erzeugen.[67] Die Autoren geben an, dass Benutzer dieses Systems, die über keine Programmierkenntnisse verfügen, in der Lage sind, in wenigen Stunden mehrere neue RDF-Beschreibungen anzulegen. Auf ähnliche Weise gehen Seo und Choi (2001) vor: In dem System XTROS werden *Wrapper* von XML-Dateien repräsentiert, die Konfigurationsinformationen enthalten; weitere XML-Dateien enthalten domänenspezifische Schlüsselwörter (z. B. „bathroom", „bath" und „baths" für die „real estate domain"), die zur Ermittlung von Positionsdaten eingesetzt werden. Halbautomatische Verfahren erleichtern die Spezifizierung von Positionen durch grafische Werkzeuge, die sich z. B. als *Plug-in* in den Webbrowser integrieren lassen. Da der Code automatisch generiert wird, werden Programmierfehler ausgeschlossen, jedoch bleibt das Problem der mangelnden Adaptivität bestehen. Ein Beispiel ist die „World Wide Web Wrapper Factory" (vgl. Sahuguet und Azavant, 2001): Grafische Assistenten unterstützen den Benutzer auf den Ebenen des Retrievals,

[66] Hsu und Dung (1998) unterscheiden drei Ebenen der Strukturiertheit: Sie bezeichnen ein HTML-Dokument als *strukturiert*, wenn Informationen listenförmig präsentiert werden und eine Informationseinheit eindeutig – getrieben durch syntaktische Hinweise – extrahiert werden kann. Eine Webseite ist *unstrukturiert*, wenn sprachliches Wissen benötigt wird, um eine Informationseinheit korrekt extrahieren zu können. Die dritte Ebene betrifft *semistrukturierte* Dokumente und umfasst diejenigen HTML-Dateien, die nicht unstrukturiert sind und an bestimmten Positionen zuviele oder zuwenige Informationseinheiten oder Permutationen aufweisen, Rechtschreibfehler enthalten oder nicht exakt für ein strukturiertes Format geeignet sind. Etzioni (1996, S. 67) bezeichnet das „labeling problem" in diesem Zusammenhang als „major obstacle", denn „data is abundant on the Web, *but it is unlabeled*." (Hervorhebung hinzugefügt, G. R.).

[67] Embley et al. (1998, 2002) verwenden einen ähnlichen Ansatz zur Extraktion von Informationen aus Stellenanzeigen und Kleinanzeigen für Gebrauchtwagen: Eine Ontologie beschreibt die jeweilige Domäne und verknüpft Konzepte mit regulären Ausdrücken, die als Extraktionsregeln dienen. Die Ontologie wird ebenfalls eingesetzt, um eine Datenbanktabelle zu generieren, die sukzessive mit den Informationen aufgefüllt wird, die das IE-System aus den zu verarbeitenden Dokumenten extrahiert.

der Extraktion und der Abbildung auf ein XML-Format. Alle weiteren Schritte werden automatisch von einer maschinell erzeugten *Java*-Klasse durchgeführt, die den *Wrapper* verkapselt und in eigenen Anwendungen eingesetzt werden kann. Ähnliche Ansätze werden in den Projekten Lixto (Baumgartner et al., 2001) und XWRAP verfolgt (Liu et al., 2000, Buttler et al., 2001) und z. B. an der Überführung von Wettervorhersagen in ein XML-Format exemplifiziert. Vollautomatische Ansätze schließlich benutzen Lernverfahren, um anhand annotierter Trainingsbeispiele generalisierte Extraktionsregeln zu lernen (Kushmerick et al., 1997). Die Beispiele sollten das gesamte Spektrum der zu erwartenden Varietät bezüglich Inhalt und Struktur der Dokumente umfassen und können mit Hilfe grafischer Werkzeuge annotiert werden (Sigletos et al., 2003). Gao und Sterling (1999) benutzen Lernverfahren zur Induktion von *Wrappern* sowie domänenunabhängige Heuristiken zum Vergleich von HTML-Dokumenten, die als Beispiele fungieren. Die Autoren geben eine Präzision von 90% an, die bei einer Evaluierung anhand von 20 Websites aus der Gebrauchtwagen-Branche ermittelt wurde: Für 18 Websites konnte das System erfolgreich *Wrapper* erzeugen, die in der Lage sind, die Hersteller, Modelle und Preise aus einer Liste von Fahrzeugen zu extrahieren.

Wrapper beziehen sich in der Regel auf die Extraktion von Informationen, die *einem* spezifischen Hypertextsortenmodul zugeordnet werden können. Ein *Wrapper* wird meist auf die Instanzen lediglich *eines* Dokumenttyps *einer* Website angewendet. Robustere Verfahren richten sich an konventionalisierte Hypertextsortenmodule, die in sehr ähnlicher Weise von *unterschiedlichen* Instanzen einer Hypertextsorte instanziiert werden. Prinzipiell können *Wrapper* somit zur Extraktion von Informationen bzw. zur Identifizierung von Instanzen derjenigen Hypertextsortenmodule eingesetzt werden, die einen sehr umfassenden Grad der Konventionalisierung besitzen. Für diese können Trainingsbeispiele gesammelt, manuell annotiert und zur Konstruktion einer Gruppe von Kategorisierern verwendet werden. Es ist zu beachten, dass der Textparser (vgl. Abschnitt 14.5) eine makrostrukturelle Analyse des zu verarbeitenden Dokuments liefert, so dass sich *Wrapper* nicht nur auf die HTML-Quelle, sondern zusätzlich (oder ausschließlich) auf die parallel zur Verfügung stehende Makrostrukturebene beziehen können, wodurch die Robustheit der Extraktion gesteigert wird.

Weiterführende Anwendungen von Methoden zur Informationsextraktion

Eine Erweiterung des Wrapping-Ansatzes betrifft die Extraktion von Informationen mit dem Ziel der anschließenden Integration in Wissensbasen (*knowledge extraction*). Das System WebKB (Craven et al., 1998, 2000) basiert auf einer sehr einfachen Domänenontologie, die Institute beschreibt und die Klassen „Activity“ (mit den Subklassen „Research.Project“ und „Course“), „Person“ (Subklassen: „Faculty“, „Staff“, „Student“), „Department“ und „Other“ enthält. Craven et al. definieren drei Ziele: Das Erkennen von Klasseninstanzen durch die Kategorisierung von HTML-Dokumenten, das Erkennen von Relationsinstanzen durch die Kategorisierung von Hyperlinkpfaden und das Erkennen von Instanzen durch die Extraktion von Textfragmenten. Für die Experimente wird vereinfachend angenommen, dass jede Instanz einer Klasse durch genau *ein* Textsegment oder genau *ein* HTML-Dokument repräsentiert wird. Das verwendete Korpus besteht aus zwei Gruppen von jeweils ca. 4 100 Webseiten, die mehreren Webauftritten von Informatik-Instituten entnommen und manuell bezüglich ihrer Klasse und den enthaltenen Relationen annotiert wurden. Von besonderer Bedeu-

tung ist das erste Ziel: Im Hinblick auf die Erkennung von Instanzen kategorisieren Craven et al. (1998) z. B. studentische Homepages in die Klasse „Student", wofür ein Naive Bayes-Verfahren eingesetzt wird, das auf den enthaltenen Wörtern operiert. Diese Kategorisierung kann mit der Erkennung der Hypertextknotensorte *Einstiegsseite der privaten Homepage eines Studierenden* verglichen werden: Während Craven et al. eine unmittelbare Abbildung von HTML-Dokumenten auf Klassen ihrer Domänenontologie vornehmen, nimmt Kapitel 13 diesbezüglich eine Trennung vor, da zunächst die Konstituenten von Hypertextsorten zu ermitteln sind, die wiederum auf Konzepte der Domänenontologie abgebildet werden können. Die Kategorisierung in die Klasse „Student" erfolgt mit einer Präzision von lediglich 43%, wenn jedoch ein geringerer Abdeckungsgrad angenommen wird, erhöht sie sich auf 67%.[68] Craven et al. berichten, dass sich fehlerhafte Kategorisierungen oftmals auf die falschen Subklassen von „Person" beziehen, so werden z. B. 80% der Instanzen der Klasse „Staff" in die allgemeinere Klasse „Person" sortiert. Alternativ wird ein Verfahren zum Lernen von Klassifikationsregeln eingesetzt, das zusätzlich Informationen über Hyperlinks einbezieht; dieser Ansatz zeigt eine höhere Präzision aber einen geringeren Abdeckungsgrad. Hinsichtlich der Verfahren, die für eine vollständige Identifizierung von Hypertextsorten benötigt werden, zeigen die Ergebnisse von Craven et al. auf, dass eine Kategorisierung, die HTML-Dokumente als atomare Einheit der Analyse auffasst, nicht erfolgversprechend ist.

Wrapper extrahieren Informationen, die in den dynamisch generierten Dokumenten einer einzelnen Website oder in eng verwandten Websites enthalten sind. Da *Wrapper* Restriktionen[69] hinsichtlich ihrer Robustheit und der Skalierbarkeit auf andere Domänen bzw. andere Websites aufweisen, wurden Ansätze mit einem größeren Skopus entwickelt: Es wird nicht nur die Extraktion spezifischer Informationen angestrebt, sondern zusätzlich die maschinelle Ermittlung makrostruktureller Komponenten (vgl. Abschnitt 14.5.2). Interessanterweise lehnen sich diese Arbeiten – ohne diesen Umstand jedoch zu thematisieren – an Konzepte an, die sich *unmittelbar* auf Charakteristika von Hypertextsorten beziehen: Davulcu et al. (2003) präsentieren ein System, das in der Lage ist, die Produktinformationen in Katalogen von Online-Shops zu ermitteln und zu etikettieren. Hierzu müssen auf Seiten der Website verschiedene Bedingungen[70] erfüllt sein, wozu z. B. die Organisation in Form einer Taxonomie gehört: Diese wird typischerweise in Form einer Navigationshilfe am linken Rand der Einstiegsseite dargestellt und enthält Produktrubriken. Die Hyperlinks verweisen auf Produktübersichten, die auf einzelne Produkte verweisen (Bucher, 2004, spricht von Fortsetzungserwartungen; vgl. Fußnote 61, S. 87). Das System erhält als Eingabe die URL der Einstiegsseite, woraufhin ein „page segmentation algorithm" das HTML-Dokument in mehrere Areale partitioniert, die durch Segmentgrenzen (spezifische HTML-Elemente) separiert werden. Durch eine Identifizierung der Navigationshilfe können Produktübersichten aufgefunden werden, indem die Ähnlichkeit ihrer URLs ermittelt wird. In den Dokumenten, die

[68] Andere Klassen liefern deutlich schlechtere Ergebnisse, so wird z. B. „Faculty" mit einer Präzision von 17,9% und „Research.Project" lediglich mit 13,0% kategorisiert (Craven et al., 2000, S. 81).

[69] Probleme bereiten dem *Wrapping*-Ansatz HTML-Dokumente, die mit ASCII-Editoren gepflegt werden (vgl. Abschnitt 3.3.6), da sie nicht notwendigerweise reguläre Formate besitzen. Ein zentrales Problem ist ihre mangelnde Adaptivität, da Layout-Modifikationen zwangsläufig Änderungen des HTML-Markups bewirken, wodurch *Wrapper* nicht mehr in der Lage sind, die gewünschten Informationen zu lokalisieren.

[70] Diese Bedingungen entsprechen im der Ausprägung spezifischer Konventionen, die sich insbesondere auf das Merkmal der Strukturierung der Hypertextbasis und spezifische Hypertextknotensorten beziehen.

die Produktübersichten umfassen, werden reguläre Ausdrücke verwendet, um z. B. Preise zu finden. Falls einem Preis ein Hyperlink zugeordnet werden kann, wird dieser von dem integrierten *Crawler* verfolgt und das Dokument mit den Kerninformationen eines Produkts analysiert. Dieses kann nun mit einem zweiten Dokument verglichen werden, um identische und abweichende Segmente zu identifizieren. Letztere werden als Produktdaten interpretiert und extrahiert. Mit Trainingsdaten, die ähnliche Merkmale einsetzen wie die in Abschnitt 14.2 dargestellten Ansätze zur Genre-Kategorisierung können diese Datensätze mit Etiketten versehen und z. B. in eine Datenbank importiert werden. Eine Evaluation zeigt, dass nur bei zwei von neun Websites die Produktübersichten nicht ermittelt werden konnten; die Hierarchien und einzelnen Produkte konnten in allen Webangeboten identifiziert werden. Diese von Davulcu et al. (2003) präsentierte Vorgehensweise identifiziert – ausgehend von der Einstiegsseite – die Instanzen von drei Hypertextsortenmodulen (*primäre Navigationshilfe*, *Produktübersicht*, *Kerndaten eines Produkts*) der zugehörigen Hypertextsorte, nutzt die konventionalisierte Strukturierung korrespondierender Hypertextbasen aus und verfolgt das Ziel der Ermittlung von Produkten und Preisen. Bei einem begrenzten Inventar von Hypertextsortenmodulen kann also eine erfolgreiche Identifizierung geleistet werden, sofern die Instanz der Hypertextsorte eine prototypische Ausprägung besitzt – eben dieses Merkmal gilt für die zwei von dem System fehlerhaft verarbeiteten Websites *nicht* (vgl. Davulcu et al., 2003, S. 13).

Ein weiterer Trend betrifft den Einsatz zusätzlicher Ressourcen. Ciravegna et al. (2003) schlagen ein Verfahren zur Integration von Informationen vor, die aus mehreren verlässlichen Quellen stammen: Webangebote, die auf Datenbanken basieren (z. B. digitale Bibliotheken), können mit sehr einfachen Mitteln verarbeitet werden. Die gefundenen Daten können zur Verarbeitung von Websites eingesetzt werden, die komplexere Verfahren benötigen etc. Hierzu nutzen Ciravegna et al. die Redundanz der im WWW verfügbaren Informationen aus: Wenn z. B. ein Hypertext H, der Publikationslisten enthält, von einem *Wrapper* verarbeitet werden soll, um die Titel und Autoren von Veröffentlichungen zu ermitteln, kann ein einfacherer *Wrapper* eingesetzt werden, um zunächst Informationen aus Suchresultaten der Datenbank *CiteSeer* zu extrahieren, die unter anderem Veröffentlichungen umfassen, die auch in H enthalten sind. Diese nun explizit vorhandenen Informationen können zur *automatischen* Annotation von Trainingsbeispielen aus H eingesetzt werden. Ciravegna et al. beschäftigen sich mit der Aufgabe, die Namen, Veröffentlichungen und Kontaktinformationen der Mitarbeiter von Informatik-Instituten in deren Webauftritten zu identifizieren. Neben *CiteSeer* wird eine weitere digitale Bibliothek, eine Suchmaschine für persönliche Homepages, ein Werkzeug zur Personennamenerkennung und *Google* eingesetzt. Über eine Liste von Vornamen können reguläre Ausdrücke benutzt werden, um Namenskandidaten zu finden (z. B. Eintrag der Liste gefolgt von einem Wort mit initialem Großbuchstaben). Diese werden in den zu untersuchenden Daten annotiert, woraufhin die externen Ressourcen benutzt werden, um zusätzliche Informationen über die Kandidaten zu erheben. Sobald z. B. *CiteSeer* für eine spezifische Kombination aus Vor- und Nachname sechs Publikationen zurückliefert, können deren Titel in den zu verarbeitenden Daten annotiert werden, um im nächsten Schritt als Trainingsdaten für spezifischere *Wrapper* zu fungieren. Ciravegna et al. zufolge arbeitet dieses Verfahren, das auf dem Zyklus der maschinellen Annotierung von Trainingsdaten und des Lernens spezifischerer *Wrapper* beruht, für einfache Aufgaben mit einer sehr hohen Zuverlässigkeit, setzt jedoch adäquate und repräsentative initiale Trainingsbeispiele voraus. Zudem werden Da-

tenbank-gestützte Websites benötigt, die die spezifische Aufgabe des Systems mit weiteren Informationen unterstützen, so dass die Menge der potenziell annotierbaren Informationen von der Verfügbarkeit derartiger Datenbanken abhängt. Dennoch könnte ein solches Verfahren benutzt werden, um eine rudimentäre Anreicherung von Stichproben oder vollständigen Korpora mit zusätzlichen Informationen zu realisieren, die wiederum von Komponenten zur Identifizierung von Hypertextsortenmodulen wie z. B. *Publikationsliste*, *Name des Homepage-Besitzers* oder *Kontaktinformationen* eingesetzt werden.

Zur Erkennung von Hypertextknotensorten und Hypertextsortenmodulen

Chung et al. (2001, 2002) konvertieren HTML-Dokumente, die als „themenspezifisch" bezeichnet werden, mit Hilfe einer DOM-Verarbeitung in XML-basierte Formate. Als Beispiele für Themen werden „product descriptions", „bibliographies" und „resumés" angeführt. Zunächst werden die Inhalte der Textknoten tokenisiert und auf vordefinierte Konzepte abgebildet, wozu maschinelle Lernverfahren und Mustervergleiche eingesetzt werden; Chung et al. setzen eine rigide Strukturierung voraus, da die Tokenisierung ausschließlich auf Interpunktionszeichen basiert. Anschließend wird der Elementbaum restrukturiert, wobei die Knoten der Zwischenrepräsentation so angeordnet werden, dass die Zielstruktur das in dem Dokument enthaltene Layout reflektiert. Die DOM-Knoten, die HTML-Elementen entsprechen, werden sukzessive durch XML-Elementknoten ersetzt (z. B. `institution`, `degree` und `thesis` in Bezug auf das Thema „resumé"). Regeln ordnen die annotierten Elemente unterhalb von „group tags" wie z. B. `h1`, `p` und `table` an und entfernen Listen (`ul`, `ol` und `dl`). Chung et al. berichten, dass für einen Test mit 50 HTML-Dokumenten, die Lebensläufe enthalten, durchschnittlich 3,9 Fehler auftreten; diese Angabe bezieht sich jedoch nur auf fehlerhafte hierarchische Strukturierungen in den erzeugten XML-Instanzen. Hinsichtlich der Identifizierung von Hypertextsortenmodulen sind verschiedene Aspekte hervorzuheben: Die allgemeine Vorgehensweise entspricht dem in Abschnitt 14.5 vorgestellten Textparser, d. h. ein HTML-Dokument wird sukzessive um XML-Elemente angereichert. Der Textparser bezieht sich auf die Annotierung makrostruktureller Komponenten. Chung et al. verzichten auf diese Ebene und annotieren unmittelbar „information carrying objects", so dass dieser Ansatz im Falle von *tag abuse* nicht eingesetzt werden kann, da er sich auf ein vorgegebenes Inventar von HTML-Elementen bezieht. Diese Abstraktion beeinträchtigt die Skalierbarkeit des Verfahrens, zudem wird nicht thematisiert, ob die Restrukturierungsregeln generalisierbar sind. Die zu verarbeitenden Dokumente werden zwar als „topic specific" bezeichnet, jedoch handelt es sich streng genommen um Knoten, in denen das Hypertextsortenmodul *Publikationsliste* als Hypertextknotensorte fungiert. Da es hinsichtlich seines Textstrukturmusters markiert ist, bietet sich dieses Hypertextsortenmodul für Informationsextraktionsprozesse an.

Yang et al. (2003) präsentieren einen DOM-basierten *bottom-up*-Ansatz zur Ermittlung von „semantically meaningful clusters" innerhalb des Elementbaums. Hiermit sind die einzelnen Bestandteile eines implizit vorhandenen Schemas gemeint, das in einer Website verwendet wird, in der Datenbankinhalte mit Templates zu HTML-Dokumenten kombiniert werden. Der Ansatz verfolgt eine ähnliche Funktion wie der Textparser (vgl. Abschnitt 14.5), der jedoch auf die Verarbeitung arbiträrer HTML-Dokumente ausgerichtet ist. Die angesprochenen Bestandteile stellen Partitionen innerhalb des Elementbaums dar, die sich auf seman-

tische Konzepte beziehen. Die Teilbäume werden durch eine Analyse der DOM-Struktur lokalisiert, wobei die räumliche Nähe, die die Teile eines Clusters besitzen, durch einen Vergleich der Pfade von der Wurzel zu einem Knoten berechnet werden. Externe Ressourcen werden zur Ermittlung weiterführender Informationen verwendet, so wird das lexikalisch-semantische Netzwerk WordNet eingesetzt, um verwandte Konzepte zu ermitteln, die als zusätzliche Restriktionen der Strukturanalyse dienen. Die gefundenen Partitionen werden durch heuristische und domänenspezifische Verfahren etikettiert. Hierzu kann z. B. der Text verwendet werden, der sich innerhalb des Knotens befindet, der einen Teilbaum dominiert. Alternativ wird ein thematischer Kategorisierer eingesetzt, der die Inhalte einer Partition auf ein Konzept einer Ontologie abbildet. Yang et al. verwenden hierfür die im *Open Directory Project* enthaltene Thementaxonomie (`http://dmoz.org`). Da jede Kategorie dieses Katalogs eine Liste von HTML-Dokumenten beinhaltet, können sie als Trainingsdaten für ein Lernverfahren fungieren. Wie bereits angesprochen, ähnelt diese Vorgehensweise dem in dieser Arbeit präsentierten Textparser, sie beschränkt sich jedoch nicht auf die Ermittlung der Partitionen, sondern zusätzlich auf ihre inhaltliche Benennung, d. h. eine Navigationshilfe wird nicht als solche, sondern vielmehr als „News“, „Opinion“ oder „Features“ annotiert. Eine weitere Restriktion bezieht sich auf die Verarbeitung dynamisch generierter Dokumente. Auf die Skalierung dieses Ansatzes oder seine Präzision gehen Yang et al. nicht ein.

Im Kontext der Identifizierung von Hypertextsortenmodulen ist die Studie von Mehler et al. (2004) von besonderer Relevanz. Es werden vier abstrakte Ebenen der logischen Hypertextstruktur angenommen, die durch Webauftritte instanziiert werden: (i) Auf der untersten Ebene befinden sich „types of hypertext building blocks“, die als „elementary self-contained units“ von HTML-Dokumenten aufgefasst werden; als Ausprägungen werden `p`-, `hr`-, `ul`- und `table`-Elemente genannt. (ii) Auf der zweiten Ebene befinden sich „hypertext document types“, die von mindestens einem HTML-Dokument instanziiert werden. (iii) Die dritte Ebene betrifft „compound hypertext document types“ (im Sinne von Eiron und McCurley, 2003, vgl. Abschnitt 14.6.1), die als Netzwerke homogener Webseiten der zweiten Ebene betrachtet werden, deren Kohärenz durch die einheitliche Intention der Autoren gegeben ist. (iv) Die letzte Ebene bezieht sich auf „compound document network types“, die durch Websites instanziiert werden. Diese besitzen Mehler et al. zufolge – und eben dieser Aspekt betrifft einen Kernpunkt des Hypertextsortenmodells – zwei zentrale Eigenschaften: Die Realisierung eines „hypertext document types“ durch Websites, HTML-Dokumente oder „building blocks“ ist nicht deterministisch, d. h. es können funktionale Äquivalente existieren; Mehler et al. sprechen von „realizational ambiguity“: Auf der Website einer Konferenz (vgl. Abschnitt 4.6.2) kann der *Call for Papers* z. B. vollständig in einem Einzeldokument realisiert werden; die Anfertigung separater Dokumente, die etwa die Themengebiete einer Tagung erläutern, stellt ein funktionales Äquivalent dar. Das Konzept „polymorphism“ ist als Komplement der Realisierungsambiguität zu betrachten: Es besagt, dass ein einzelnes Dokument mehrere „hypertext document types“ realisieren kann.[71] Mehler et al. gehen von der Hypothese aus, dass diese beiden Phänomene zentrale Charakteristika von „web-based units“ darstellen, so dass es für die Anwendung der Kategorisierung von Websites bzw. HTML-Dokumenten zwin-

[71] Als Beispiel nennen Mehler et al. (2004) die Website einer Konferenz, die als ein einzelnes HTML-Dokument ausgeprägt ist (vgl. auch hierzu Kapitel 5). Das in Abbildung 12.4 (S. 564) dargestellte Dokument D 749 kann in diesem Sinne als eine Extremausprägung von Polymorphismus aufgefasst werden.

gend notwendig ist, zuvor eine Strukturanalyse durchzuführen. Zur Überprüfung dieser Hypothese werden 13 481 HTML-Dokumente von 1 000 Konferenz-Websites mit Hilfe eines SVM-Kategorisierers in 13 Kategorien (z. B. „submission and author instructions“, „call for papers“, „important dates“, „program“ und „venue“) sortiert. Die Ergebnisse zeigen, dass präzise und eindeutige Zuordnungen nur in Bezug auf *eine* Kategorie erfolgen. In den verbleibenden 12 Kategorien liegen zahlreiche Mehrfachzuordnungen vor. Diese führen Mehler et al. auf die Problematik des Polymorphismus zurück, d. h. ein Dokument beinhaltet mehrere Bausteine, die eindeutige Zuordnungen verhindern. Daher wird plädiert für die „exploration of *patterns of page internal structures* in order to disentangle functional equivalents and polymorphic units as a preliminary step to any categorization“ (Hervorhebungen hinzugefügt, G. R.). Diese Beobachtung zeigt, dass die Betrachtung von HTML-Dokumenten als atomare Einheiten von Kategorisierungsprozessen nicht adäquat ist und zu unpräzisen Ergebnissen führt – die Analyse der Binnenstrukturen von Dokumenten ist für die Kategorisierung *zwingend* notwendig. Sie bezieht sich auf die Ermittlung von Hypertextmodulen (vgl. Abschnitt 14.5) und die anschließende Abbildung auf Hypertextsortenmodule. Mehler et al. nehmen keine derartige Ebene an: Die „hypertext building blocks“ der ersten Ebene werden mit exemplarischen HTML-Elementen erläutert; sie ähneln Hypertextmodulen. Die zweite Ebene umfasst „hypertext document types“, die Mehler et al. zufolge durch mindestens ein HTML-Dokument instanziiert werden. In einer erläuternden Grafik wird diese Ebene mit der Instanziierung einer „page“ und die übergeordnete Ebene der „compound hypertext document types“ durch das Beispiel „homepage“ erläutert, wobei jedoch nicht deutlich wird, weshalb eine Einstiegsseite keine Instanz des Strukturtyps „hypertext document“ ist, schließlich handelt es sich ebenfalls um eine „page“.[72]

Das Ergebnis der Studie von Mehler et al. (2004) wird von einem Test bestätigt: In diesem wurden die 727 in der vierten Analyse auf ihre Hypertextknotentypen und -sorten untersuchten Dokumente (vgl. Tabelle 11.10, S. 492) als Trainingsdaten für die Lernverfahren Naive Bayes und kNN eingesetzt, die auf *bag of words*-Vektoren operieren. Als zu kategorisierende Daten dienen 536 HTML-Dokumente. Bei diesen handelt es sich um die Einstiegsseiten und die Dokumente der ersten Hyperlinkebene von 20 weiteren, ebenfalls im Korpus enthaltenen universitären Webauftritten. Die Kategorisierung erfolgt mit einer sehr geringen Präzision von ca. 20% (für die hochfrequenten Kategorien). Für dieses Ergebnis ist einerseits das von Mehler et al. (2004) als Polymorphismus bezeichnete Problem, andererseits die hohe Anzahl von insgesamt 65 Kategorien verantwortlich. Für die meisten dieser Kategorien liegen nur sehr wenige Trainingsdokumente vor.

Zur Identifizierung der Hypertextknotensorte eines HTML-Dokuments ist es also zwingend notwendig, zunächst eine Erkennung der Hypertextsortenmodule vorzunehmen. Falls

[72] Mehler et al. (2004) verwenden den Terminus „compound document“ mit Bezug auf Eiron und McCurley (2003). Diese argumentieren, dass eine abstraktere Ebene als der einzelne Knoten existiert, die sie als „document“ bezeichnen: „This concept is perhaps ambiguous, but we use the term document to mean a coherent body of material on a single topic. [...] Examples include manuals, articles in a newspaper or magazine, or an entire book.“ Das Hypertextsortenmodell (vgl. Kapitel 5) umfasst keine derartige Ebene, weil im Hypertextsystem WWW lediglich einzelne Knoten (d. h. HTML-Dokumente bzw. -Dateien) existieren. Verknüpfte Knoten werden als Hypertext aufgefasst und Hypertexte können eingebettete Hypertexte enthalten. Bei Mehler et al. wird nicht deutlich, weshalb auf das Konzept und die hiermit verbundene mehrdeutige Terminologie der „compound documents“ Bezug genommen wird.

nun angenommen wird, dass ein derartiges Verfahren in einer robusten Implementierung mit einem hohen Abdeckungsgrad vorliegt, ergibt sich die Problematik, dass *mehrere*, z. B. inhaltlich-thematisch markierte Hypertextsortenmodule in einem Dokument existieren. Sofern es die Benutzerschnittstelle einer Suchmaschine dem Anwender erlaubt, nach Hypertextknotensorten zu recherchieren, ist es notwendig, eine Heuristik zu entwickeln, die die individuelle Ausprägung eines Dokuments ermittelt, indem das primäre Hypertextsortenmodul einer Webseite bestimmt wird. Für diesen Zweck könnte z. B. lediglich das initiale Hypertextsortenmodul berücksichtigt werden; zusätzlich könnten der jeweilige Textumfang oder die Abmessungen einbezogen werden (vgl. Haas und Grams, 2000, S. 186). Daneben existiert jedoch auch der Fall, das eine Einstiegsseite vorliegt, die in diesem Sinne kein primäres Hypertextsortenmodul besitzt, das die Hypertextknotensorte determiniert. Die robuste Erkennung von Hypertextsortenmodulen könnte natürlich auch für den Zweck der maschinellen Bestimmung ihrer Vorkommen in großen Dokumentkollektionen eingesetzt werden, so dass Gemeinsamkeiten und Differenzen zwischen den Instanzen unterschiedlicher Hypertextsorten und typische, d. h. konventionalisierte Kombinationen von Hypertextsortenmodulen quantitativ ermittelt werden können (vgl. die auf S. 280 dargestellte Definition von „About this site“ von Reiss, 2000, S. 135).[73] Zudem könnten derartige Informationen zur Ermittlung der typischen Hypertextmodule eingesetzt werden, mit denen spezifische Hypertextsortenmodule in der Regel realisiert werden (z. B. die unterschiedlichen Typen *biografischer Angaben*, vgl. Abbildung 10.2, S. 443) Mit umfangreich annotierten Trainingsdaten könnte auf diese Weise auch der von Ciravegna et al. (2003) beschriebene Zyklus initiiert werden, der aus den Phasen des maschinellen Lernens von Erkennungsregeln und des Annotierens zusätzlicher Trainingsdaten besteht.

Abbildung 14.8 zeigt abschließend ein Beispiel für die innerhalb der Systemarchitektur (vgl. Abbildung 14.1, S. 650) vorgesehene Funktionalität: Ein HTML-Dokument wird von dem Textparser analysiert, wobei die makrostrukturellen Komponenten in der DOM-Struktur durch Elemente und Attribute eines weiteren Namensraumes annotiert werden. Auf der anderen Seite existiert die Hypertextsortenontologie, in der in Bezug auf die drei Ebenen des Hypertextsortenmodells die Konstituenten von Hypertextsorten repräsentiert werden.[74] Mit Hilfe spezieller Werkzeuge kann die Ontologie gefiltert und in Dokumenttyp-Definitionen (z. B. für jede Hypertextsorte eine DTD) überführt werden, die zur Annotierung korrespondierender Instanzen eingesetzt werden können (vgl. hierzu die Abschnitte 13.5.7 und 13.5.8). In dieser DTD werden die Etiketten von Klassen, die Hypertextsortenmodule, Hypertextknotensorten und Hypertextsorten repräsentieren, als Bezeichnungen von XML-Elementen verwendet, d. h. es existiert eine unmittelbare Verbindung zwischen den Klassen der Hypertextsortenontologie und der DTD. Über RDF-Annotationen können externe Ressourcen an einzelne Klassen gebunden werden. Auf diese Weise könnte eine Vielzahl von

[73] Die Existenz eines sehr robusten Mechanismus zur Erkennung von Hypertextsortenmodulen bedeutete, dass Studien, wie sie in Teil III präsentiert werden, mit maschinellen Verfahren durchgeführt werden können.

[74] Der in Abbildung 14.8 dargestellte Ausschnitt der Hypertextsortenontologie beschränkt sich auf diejenigen Konstituenten, die in dem Beispieldokument enthalten sind (HP 42 aus der vierten Analyse, vgl. Tabelle 10.1, S. 427, Abbildung 10.5, S. 456, und Abbildung 14.6, S. 672; es handelt sich um die Hypertextknotensorte *Einstiegsseite der persönlichen Homepage eines Wissenschaftlers*). Zudem wurden verschiedene Bezeichnungen von Klassen und Relationen abgekürzt, um die Lesbarkeit der Darstellung gewährleisten zu können.

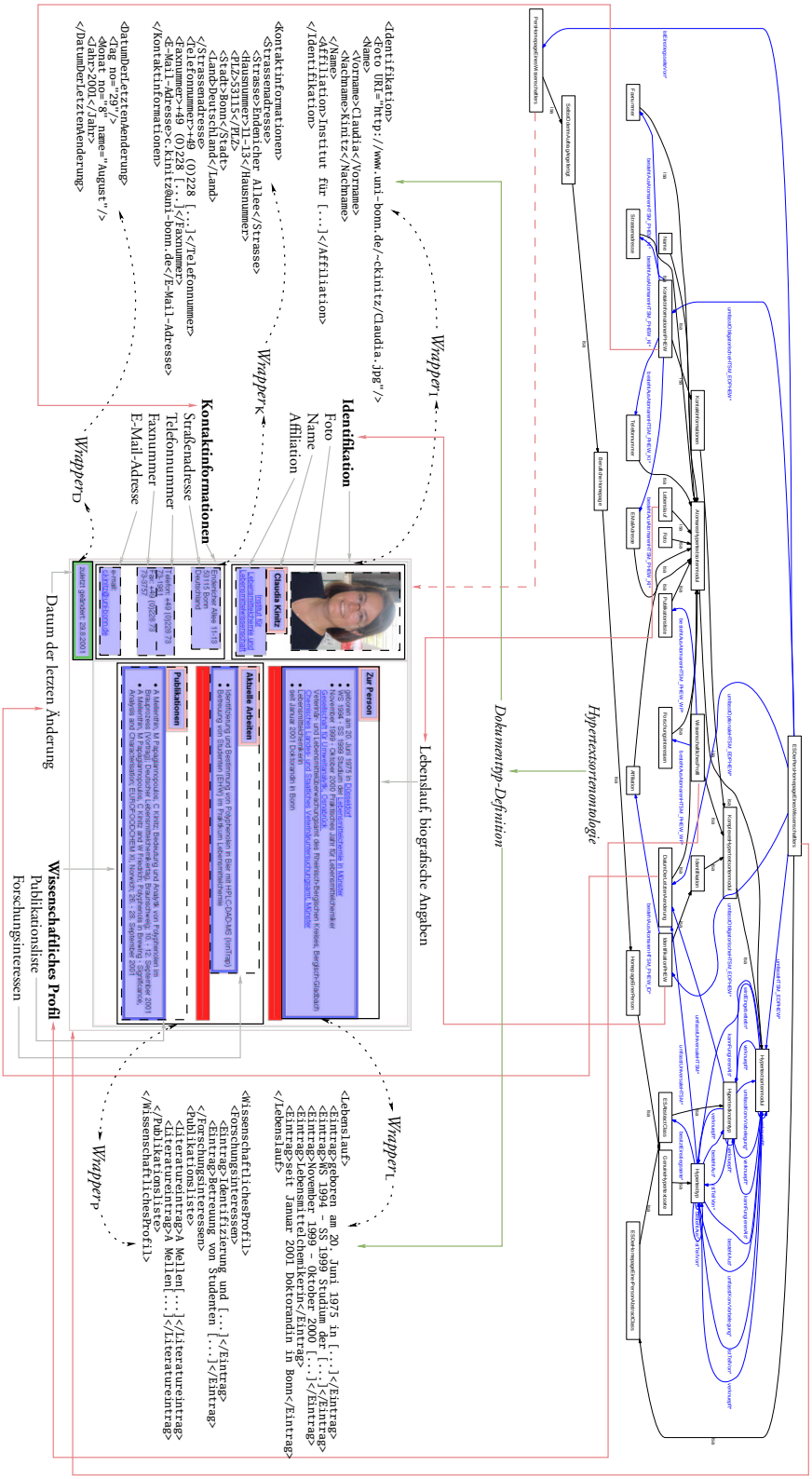


Abbildung 14.8: Abbildung von Hypertextmodulen auf Hypertextsortenmodule auf der Basis der Hypertextsortenontologie

Wrappern, die in der Lage sind, jeweils ein spezifisches Hypertextsortenmodul zu identifizieren und zu analysieren, auf die zugehörigen Hypertextsortenmodule in der Hypertextsortenontologie abgebildet werden.[75] Wenn ein *Wrapper* ein Hypertextmodul z. B. als *Publikationsliste* identifiziert, kann diese Informationen in die Hypertextsortenontologie propagiert werden; die Angaben zu den erkannten Hypertextsortenmodulen eines Hypertextes können zur Ermittlung der korrespondierenden Hypertextsorte eingesetzt werden. Weiterhin kann der *Wrapper* die in der Instanz eines Hypertextsortenmoduls enthaltenen Informationen in ein XML-Format überführen; die auf diese Weise erkannten XML-Fragmente der HTML-Dokumente eines Hypertextes können daraufhin z. B. einzeln gegen die jeweiligen DTD-Fragmente validiert und anschließend zu einer vollständigen Instanz aggregiert werden. Abbildung 14.8 zeigt diesbezüglich fünf Beispiele unterschiedlicher Strukturiertheitsgrade, die sich auf die jeweiligen Definitionen eines Hypertextsortenmoduls beziehen.[76] Ein Vergleich der Hypertextsortenmodule *Datum der letzten Änderung* und *Publikationsliste* verdeutlicht, dass die Identifizierung von Hypertextsortenmodulen mit einem heterogenen Apparat von Methoden durchzuführen ist. Für die *Publikationsliste* bietet es sich an, mit statistischen Methoden oder Lernverfahren zu arbeiten, die unter anderem Eigennamen und Interpunktionszeichen in ihren Merkmalsraum einbeziehen. Beim *Datum der letzten Änderung* hingegen wird in vielen Fällen eine XPath-ähnliche Anfrage ausreichend sein, die dasjenige `hypnotic:TextBlock`-Vorkommen ermittelt, das z. B. aus maximal fünf oder sechs Token bestehet, eine Datumsangabe beinhaltet und am Ende des Dokuments positioniert ist.

14.6.4 Fazit – Eine alternative Architektur

Die vorangegangenen Ausführungen verdeutlichen, dass die in Abschnitt 14.3 vorgeschlagene Architektur mit einer Problematik konfrontiert wird, die die immense Komplexität des Gesamtsystems betrifft: Neben einem robusten Textparser werden Komponenten (unter anderem eine Vielzahl von *Wrappern*) zur Identifizierung von Hypertextsortenmodulen, Hypertextknotensorten und Hypertextsorten benötigt. Daneben erfordert die Architektur einen Mechanismus zur Erkennung der Grenzen von Hypertexten und eingebetteten Hypertexten. Zudem setzt die Identifizierung von Hypertextsorten naturgemäß die Erstellung eines Inventars von Hypertextsorten und der zugehörigen Konstituenten voraus. Diese können in Form einer Hypertextsortenontologie repräsentiert werden, die den Erkennungsprozess mit Hilfe einer Inferenzmaschine unterstützen kann.

[75] In der Hypertextsortenontologie kann eine weitere Ebene vorgesehen werden, die die Spezifizierung von Informationen betrifft, die *Wrapper* zur Identifizierung von Hypertextsortenmodulen einsetzen können. Diese bezieht sich auf die Abbildung von Hypertextsortenmodulen auf ein Inventar derjenigen Hypertextmodule, die typischerweise für ihre Realisierung eingesetzt werden. Auf die Darstellung dieser Zusammenhänge, die in Abschnitt 13.5.8 diskutiert werden, wird in Abbildung 14.8 verzichtet.

[76] Dieser Aspekt bezieht sich auf den in Abschnitt 5.2.1 angesprochenen Umstand, dass eine DTD unterschiedliche Grade der Strukturiertheit besitzen kann (vgl. auch Fußnote 2, S. 262). Das Element `Strassenadresse` in Abbildung 14.8 besitzt einen eher hohen Grad der Strukturierung. Im Gegensatz dazu umfasst `Telefonnummer` lediglich eine fortlaufende Zeichenkette, die durch eine Binnenstrukturierung mit Elementen wie `Landesvorwahl` `Vorwahl` und `Durchwahl` zusätzlich angereichert werden könnte. Es ist zu erwarten, dass *Wrapper* langfristig nicht in der Lage sein werden, für beliebige Hypertextsortenmodule ein hohes Maß der Strukturiertheit mit einer hohen Präzision maschinell ermitteln zu können. Für konventionalisiertere Hypertextsortenmodule können also explizitere Strukturierungen vorgesehen werden.

Für das Ziel der Realisierung einer Suchmaschine, die dem Benutzer eine Filterung nach Hypertextsorten erlaubt, kann diese hochgradig komplexe Architektur in ihrer Funktionalität reduziert werden. Diese alternative Architektur setzt ebenfalls ein Inventar von Hypertextsorten sowie ein präzises Verfahren zur Identifizierung der Grenzen von Hypertexten voraus und basiert vollständig auf maschinellen Lernverfahren: Zunächst wird, z. B. mit Hilfe eines *Crawlers*, eine große Zahl von Hypertexten bezogen, deren Grenzen identifiziert werden. Dieses Verfahren ermitelt also diejenigen HTML-Dokumente, die einen gegebenen Hypertext konstituieren. Daraufhin können diese Knoten maschinell konkateniert werden, d. h. der gesamte Hypertext wird zu einem einzelnen, synthetischen Dokument transformiert.[77] Auf diese Weise entsteht eine Kollektion einzelner HTML-Dokumente, die vollständige Hypertexte umfassen. Über die in Abschnitt 7.3.5 vorgestellte Funktionalität der Web-Oberfläche der Korpusdatenbank können den „Hypertexten“ nun von Experten ihre Hypertextsorten zugewiesen werden, um eine Kollektion von Trainingsdaten aufzubauen. Für diese Kollektion kann ein Merkmalsraum aufgebaut werden, der z. B. auf dem von Lim et al. (2005a,b) vorgeschlagenen Inventar von Merkmalen basiert. Die *tag abuse*-Problematik, die in keiner der verwandten Arbeiten thematisiert wird, kann hierbei durch den Textparser kompensiert werden, indem nicht sämtliche HTML-Elemente, sondern (zusätzlich) die Elemente des `hypnotic:`-Namensraumes in den Merkmalsraum aufgenommen werden. Dabei können Standardmethoden aus dem *Information Retrieval* eingesetzt werden, um unter anderem hochfrequente und somit nicht signifikante Terme zu entfernen. Die individuellen Merkmalsausprägungen der einzelnen Kategorien werden zum Trainieren eines Lernverfahrens eingesetzt, um unbekannten Hypertexten, die ebenfalls zu Einzelknoten transformiert werden, ihre Hypertextsorte zuweisen zu können. Durch die Konkatenation aller Knoten werden sowohl Realisierungsambiguität als auch Polymorphismus (Mehler et al., 2004) ignoriert. Es ist zu erwarten, dass eine derartige Vorgehensweise präzise Resultate liefert, sofern eine umfassende Menge repräsentativer Trainingsbeispiele zur Verfügung steht und eine eher geringe Anzahl von Kategorien eingesetzt wird. Zudem bietet sich diese Architektur zur Anknüpfung an die bislang vorliegenden Ansätze an, um ihr zentrales konzeptionelles Defizit aufzuheben – die Beschränkung auf einzelne HTML-Dokumente (vgl. die Abschnitte 14.2.3 und 14.2.4).

14.7 Hypertextsorten in der Sprach- und Informationstechnologie

Das Hypertextsortenmodell (Kapitel 5), die Hypertextsortenontologie (Kapitel 13) und die Architektur zur maschinellen Identifizierung von Hypertextsorten (Abschnitt 14.3) ermöglichen neuartige informationstechnologische und computerlinguistische Anwendungen. Im Folgenden wird eine Auswahl dieser Perspektiven skizziert. Es ist jedoch anzumerken, dass sie sich auf eine nahezu vollständige Implementierung eines korrespondierenden Systems beziehen und somit lediglich den Charakter eines Ausblicks besitzen.

Hypertextsorten und Suchmaschinen – Die Identifizierung von Hypertextsorten durch Suchmaschinen und die Bereitstellung entsprechender Möglichkeiten der Filterung stellt eine

[77] Dieser Ansatz lehnt sich an die thematische Kategorisierung vollständiger Websites an (vgl. Pierre, 2001, und Kwon und Lee, 2003, sowie Abschnitt 14.2.1).

Anwendung dar, die sich wie ein roter Faden durch die theoretische und angewandte Literatur zieht. Die vorangegangenen Ausführungen haben verdeutlicht, dass mit einer derartigen Anwendung höchstens langfristig zu rechnen ist, da noch eine Vielzahl offener Forschungsfragen existiert, die sich unter anderem auf die Integration von *Semantic Web*-Technologien in Suchmaschinenarchitekturen beziehen (vgl. Finin et al., 2005). Die Entwicklung innovativer Möglichkeiten zur Filterung von Rechercheergebnissen wird jedoch zusehends dringlicher, da das WWW nach wie vor wächst. Die großen Suchmaschinen betreiben zwar einen beträchtlichen technischen Aufwand, um möglichst präzise Ergebnisse für meist sehr kurze Anfragen zu produzieren jedoch steigt mit der Anzahl der indexierten Dokumente zwangsläufig auch die Anzahl der zurückgelieferten Treffer. Eine von Meyer zu Eissen und Stein (2004) durchgeführte Fragebogenuntersuchung unter 286 Studierenden hat ergeben, dass 64% eine Suchmaschine mit integrierter „genre classification" als sehr hilfreich einschätzen, immerhin 29% empfänden eine solche Funktion zumindest gelegentlich als Mehrwert.

Die explizite Filterung nach Hypertextsorten innerhalb von Suchmaschinen kann in einer Benutzerschnittstelle auf unterschiedliche Weise realisiert werden. Der Anwender könnte z. B. einen Informationsbedarf besitzen, der eine Restriktion der Suche auf spezifische Hypertextsorten erlaubt. Somit könnten Recherchen erfolgen, die ausschließlich in Exemplaren von Hypertextsorten münden, die vom Benutzer angegeben wurden. Bei einem weniger eindeutig formulierbaren Informationsbedarf könnten diejenigen Hypertextsorten aus der Suche ausgeschlossen werden, die nach Ansicht des Benutzers keine Relevanz für die Anfrage besitzen. Da die Hypertextsortenontologie bereits eine Vielzahl unterschiedlicher Hypertextsorten umfasst, wäre es auch möglich, sie zur Reduktion der Komplexität auf einer personalisierte Ontologie abzubilden (Chaffee und Gauch, 2000), die die Interesen des Benutzers reflektiert und in der z. B. einzelne Kategorien zu Gruppen gebündelt werden können, um individuelle Konfigurationen von Hypertextsorten für typische Rechercheszenarien zu repräsentieren, die in der Benutzerschnittstelle der Suchmaschine durch eine Auswahlliste aktiviert werden können. Eine weitere Möglichkeit betrifft die Annotierung der Suchergebnisse, d. h. es werden keine Mittel zur Anreicherung der Suchanfrage mit einer oder mehreren Hypertextsorten zur Verfügung gestellt, sondern die Suchergebnisse werden mit einem Etikett versehen, das die korrespondierende Hypertextsorte darstellt (vgl. Amitay et al., 2003, und Rosso, 2005). Die von Roussinov et al. (2001) durchgeführten Interviews deuten an, dass mehrere Hypertextsorten existieren, die für die Mehrzahl der Anwender in Suchmaschinenergebnissen keine Relevanz besitzen (genannt wird die *Homepage einer Person*). Somit kann Wissen über Hypertextsorten auch für das *relevance ranking* eingesetzt werden, indem die Vertreter populärer Hypertextsorten höher gewichtet werden. Derartige Funktionen können jedoch, wie Rosso (2005, S. 189) zu Recht anmerkt, nur dann realisiert werden, „if the web genres could be automatically assigned to the pages. Of course, that is a BIG if."

Falls eine robuste Identifizierung von Hypertextsorten, Hypertextknotensorten und Hypertextsortenmodulen realisiert werden kann, können die ermittelten Informationen dieser drei Ebenen Suchmaschinen zur Verfügung gestellt werden. Die Erkennung von Hypertextsortenmodulen kann als Informationsextraktionsprozess konzeptualisiert werden (vgl. auch Riloff, 1994, sowie Kando, 1999): Da viele Konzepte der Hypertextsortenontologie korrespondierende Konzepte in der Domänenontologie besitzen (vgl. Abschnitt 13.5.6), kann also – die Existenz einer robusten Implementierung vorausgesetzt – eine maschinelle Instanziie-

rung der Klassen erfolgen, die in der Domänenontologie repräsentiert werden (vgl. Missikoff et al., 2003). Eine auf diese Weise erstellte Wissensbasis kann ebenfalls in Suchmaschinen oder „ontological hypertexts“ Verwendung finden (Carr et al., 2004; vgl. Abschnitt 13.2.4).

Hypertextsorten und Metadatenvokabulare – Vokabulare wie *Dublin Core* werden zur Auszeichnung von HTML-Dokumenten mit Metadaten eingesetzt. Neben der expliziten Bereitstellung des Autornamens, des Datums der letzten Änderung oder des Hypertextthemas könnten auf diese Weise auch explizite Angaben zur verwendeten Hypertextsorte in einem Dokument hinterlegt werden. Sobald derartige manuelle Annotierungen in einer Vielzahl von Dokumenten enthalten sind, können Suchmaschinen ihren Anwendern – *ohne* den aufwändigen Prozess der maschinellen Identifizierung – entsprechende Filterkomponenten zur Recherche nach spezifischen Hypertextsorten zur Verfügung stellen. Die Prämisse hierfür stellt jedoch ein standardisiertes Vokabular von Hypertextsortenetiketten dar. Es können zwei Indizien angeführt werden, die die Standardisierung eines derartigen Kategorieninventars unwahrscheinlich erscheinen lassen: Die *Dublin Core Metadata Initiative* hat sich für den „Minimalist Approach“ entschieden, der eine maximale Flexibilität und auch Individualität bezüglich des Elements *Resource Type* garantiert (vgl. Abschnitt 3.6.6). Ein weiteres Problem stellt die Erarbeitung eines übergreifenden Katalogs von Hypertextsorten dar: Die bislang vorgeschlagenen Listen von Web-Genres weisen nur wenige Gemeinsamkeiten auf (vgl. Abschnitt 4.4). Die hochgradige Flexibilität der Kombinierung von Hypertextsortenmodulen und der Umstand, dass oftmals keine eindeutigen Zuordnungen vorgenommen werden können, lassen vermuten, dass die Erstellung eines umfassenden Katalogs mit einem hohen Abdeckungsgrad nur langfristig erreicht werden kann.

Hypertextsorten und Informationsextraktion – Die maschinelle Bestimmung von Hypertextsorten setzt eine präzise Identifizierung von Hypertextsortenmodulen voraus. Für diesen Zweck werden unter anderem Verfahren der Informationsextraktion vorgeschlagen, die auf den ermittelten Hypertextmodulen operieren, um die Bausteine der Textoberfläche auf Hypertextsortenmodule abzubilden. Hypertextsortenmodule stellen innerhalb einer Hypertextsorte konventionalisierte Komponenten dar: Falls es möglich ist, den Hypertexttyp eines gegebenen Hypertextes ohne eine zuvor durchgeführte Ermittlung der Hypertextsortenmodule maschinell zu bestimmen, kann das in der Hypertextsortenontologie enthaltene Wissen über die konventionalisierten Bestandteile eingesetzt werden, um erstens spezielle Erkennungsregeln für diese Hypertextsortenmodule zu aktivieren und ihnen zweitens gezielt Informationen zu entnehmen. Zwischen der Erkennung von Hypertextsortenmodulen und der auf diesem Schritt aufbauenden Informationsextraktion liegt zwar eine fließende Grenze vor, jedoch stellt diese Vorgehensweise dem *Wrapping*-Paradigma neue Perspektiven bereit: Ein *Wrapper* bezieht sich nicht länger auf die Dokumente eines spezifischen Anbieters, sondern auf die Instanzen einer spezifischen Hypertextsorte. Als Beispiel hierfür kann die Suchmaschine *CiteSeer* herangezogen werden (Bollacker et al., 1998), die in der Lage ist, Publikationslisten zu identifizieren, die in ihnen enthaltenen Literatureinträge zu verarbeiten und in einer einheitlichen Datenbasis zu aggregieren.

Hypertextsorten und Webdesign – Hypertextsorten können spezifische Gestaltungsmerkmale aufweisen, die in dem vorliegenden Hypertextsortenmodell von dem Merkmal Dekoration erfasst werden. Obwohl das Thema Webdesign in der vorliegenden Arbeit nur am

Rande diskutiert wurde, können auf Grundlage der maschinellen Erkennung von Hypertextsorten sowie der Hypertextsortenontologie verschiedene Anwendungen konzipiert werden (vgl. auch Haas und Grams, 1998b, S. 99). Hierzu gehört zunächst die Unterstützung der Anfertigung einer Website (vgl. Fußnote 81, S. 203): Ein Produzent verfolgt mit einem Webauftritt ein spezifisches Ziel und eine entsprechende Parametrisierung und Aufbereitung der Hypertextsortenontologie könnte das Gerüst einer Instanz der korrespondierenden Hypertextsorte in einem HTML- oder XML-Editor präsentieren. WWW-Anwender eignen sich sukzessive Wissen über die Komponenten eines Websitetyps (d. h. einer Hypertextsorte) an und erwarten von anderen Vertretern dieses Typs, dass sie diejenigen Komponenten, die als hilfreich, informativ und sinnvoll eingeschätzt werden, ebenfalls aufweisen (vgl. Abschnitt 4.3). Haas und Grams (2000, S. 190) argumentieren, dass eine Website, die einen möglichst typischen Vertreter einer Hypertextsorte darstellt, ihre kommunikative Funktion erfolgreicher erfüllen kann als untypische Vertreter. Durch die implizit in der Hypertextsortenontologie enthaltenen Dokumentgrammatiken, die auf der Ebene der Hypertextsortenmodule von DTD-Fragmenten flankiert werden, wäre es darüber hinaus möglich, Inhalt und Struktur einer Website strikt voneinander zu trennen, indem die korrespondierenden DTDs zur Daten- und Texterfassung innerhalb eines XML-Editors eingesetzt werden. Eine zweite Anwendung bezieht sich auf die Analyse vorhandener Hypertexte: Ivory und Hearst (2002) haben ein Verfahren zur maschinellen Einschätzung der Qualität einer Website entwickelt, das zusätzlich mit Wissen über Hypertextsorten angereichert werden kann (vgl. auch Chan und Yu, 1999). Auf diese Weise wäre es z. B. möglich, extrem untypische Hypertextsortenmodule zu ermitteln und ihre Tilgung vorzuschlagen oder auf fehlende Bausteine hinzuweisen. Da der Benutzer in diesem Szenario die Hypertextsorte auch manuell wählen kann, ist ihre maschinelle Identifizierung für diese Anwendung nicht notwendig. Eine einfachere Anwendung in diesem Bereich betrifft den Einsatz von Informationen über Hypertextsorten, Hypertextknotensorten und Hypertextsortenmodule, um die Benutzbarkeit einer gegebenen Website zu verbessern (ähnlich der Funktion zur Visualisierung der Analyseresultate des Textparsers, vgl. Abschnitt 14.5.5). Auf diese Weise könnten z. B. unterschiedliche Typen von Hyperlinks durch Pop-up-Fenster erläutert werden, die die Funktion einer Verknüpfung verdeutlichen oder es könnten automatisch Sitemaps generiert werden, die die Hypertextknotensorten der HTML-Dokumente einer Website beinhalten (vgl. Haas und Grams, 1998b, S. 100).

Hypertextsorten und linguistische Analysen – Linguistische Analysen können insbesondere im Bereich der Korpuslinguistik von einer maschinellen Identifizierung von Hypertextmodulen und Hypertextsortenmodulen profitieren (vgl. Kilgarriff, 2001). Die automatische Auswertung umfangreicher Bestände von HTML-Dokumenten, die aus heterogenen Quellen stammen, ist bislang nur durch die Implementierung website- oder sogar dokumentspezifischer Vorverarbeitungswerkzeuge möglich, so dass z. B. Wortartfrequenzen oder syntaktische Konstruktionen in Nachrichtenartikeln oder Kolumnen mit POS-Taggern oder Parsern ermittelt, analysiert und mit Daten aus Printpublikationen verglichen werden können. Sobald explizite Informationen über die Makrostrukturbausteine von HTML-Dokumenten vorliegen, können derartige Verfahren auf einer übergreifenden Ebene eingesetzt werden. Hierzu zählt auch die Bestimmung der Ausprägung textueller Dimensionen wie z. B. der Grad der konzeptionellen Mündlichkeit oder Schriftlichkeit (vgl. Abschnitt 14.6.2).

Hypertextsorten und computerlinguistische Anwendungen – Für computerlinguistische Anwendungen gelten zunächst die bereits im vorangegangenen Abschnitt angesprochenen Aspekte: Sobald Informationen über Makrostrukturbausteine vorliegen, können spezielle Werkzeuge, z. B. Parser mit spezialisierten Grammatiken auf den Inhalten der Komponenten operieren (vgl. Kessler et al., 1997). Umgekehrt kann eine solche Anwendung auch ausgeschlossen werden, falls spezifische Kriterien für einem Baustein nicht erfüllt sind (etwa hinsichtlich der Wortfrequenz). Hypertextsorten bieten eine Vielzahl neuartiger Perspektiven, so können etwa auf der Grundlage einer umfangreichen Kollektion unterschiedlicher Hypertextsortenmodule Sammlungen von Testsätzen erstellt werden, um diese zur Evaluation von Parsern mit einem hohen Abdeckungsgrad einzusetzen. Darüber hinaus können auf der Basis der Hypertextsortenmodule einer Hypertextsorte spezifische Verfahren zur maschinellen Textzusammenfassung entwickelt werden (vgl. Crowston und Kwasnik, 2004), die sowohl für die Darstellung auf den kleinformatigen Displays mobiler Endgeräte optimiert sind (vgl. Abschnitt 14.5.2), als auch zur Erzeugung der Kurzzusammenfassungen eingesetzt werden können, die Suchmaschinen bei der Darstellung der Ergebnisse präsentieren. Ein weiterer Aspekt betrifft die Erkennung der Ursprünge neuer Konventionen: Wenn ein extrem umfangreiches Korpus vorliegt und die Instanzen von Hypertextsorten mit einer sehr hohen Präzision identifiziert werden können, ist ein maschineller Vergleich der Instanzen möglich. Da beim Aufbau und der Pflege derartiger Kollektionen Zeitstempel protokolliert werden, können somit spezifische Muster ermittelt werden, die die Entwicklung von Hypertextsorten, die allmähliche Formierung von Hypertextsortenvarianten und die Integration neuer Hypertextsortenmodule reflektieren.

Hypertextsorten und Webstatistiken – Es liegen verschiedene statistische Untersuchungen vor, in denen unterschiedliche Merkmale des *World Wide Web* charakterisiert werden, z. B. der durchschnittliche Umfang eines Dokuments oder die durchschnittliche Anzahl von Hyperlinks (vgl. z. B. Bray, 2004b, Woodruff et al., 1996, Turau, 1998a, sowie ausführlich Anhang A). Eine präzise maschinelle Identifizierung von Hypertextsorten und ihrer beteiligten Konstituenten ermöglichte eine weitere Differenzierungsebene für derartige statistische Analysen. Es könnte z. B. ermittelt werden, wie häufig spezifische Hypertextsorten in bestimmten Ländern instanziiert werden, oder ob sich private Homepages von Studierenden in England und Deutschland in ihrem Durchschnittsumfang und den verwendeten Hypertextsortenmodulen unterscheiden.

14.8 Zusammenfassung

Dieses Kapitel beschäftigt sich mit der maschinellen Identifizierung von Hypertextsorten, Hypertextknotensorten und Hypertextsortenmodulen. Für die „automatic Web genre classification“ liegen erste Arbeiten vor, in denen unterschiedlich komplexe Merkmalsräume verwendet werden, um die Kategorisierung mit Hilfe maschineller Lernverfahren durchzuführen, die auf manuell kategorisierten Trainingsdaten operieren. Allen Ansätzen ist gemein, dass sie sehr eingeschränkte Kategoriensysteme verwenden, die zwischen zwei und 16 Web-Genres umfassen und somit nur einen sehr kleinen Ausschnitt der in Teil III ermittelten Bandbreite von Hypertextsorten reflektieren. Weiterhin teilen sich diese Ansätze

die Gemeinsamkeit, dass lediglich einzelne HTML-Dokumente betrachtet werden, d. h. ein Web-Genre wird durch genau eine Webseite manifestiert, die übergeordneten und untergeordneten Ebenen der Hypertextsorten und Hypertextsortenmodule werden nicht betrachtet. Das in Kapitel 5 präsentierte Hypertextsortenmodell bildet die theoretische Fundierung einer Systemarchitektur, die sowohl diese drei Ebenen als auch die Hypertextsortenontologie als Repräsentationsformat für die unterschiedlichen Konstituenten multipler Hypertextsorten einbezieht. Die Architektur sieht einen Textparser zur initialen Verarbeitung von HTML-Dokumenten vor. Der Einsatz eines derartigen Werkzeugs ist notwendig, weil in *Hypertext Markup Language*-Dokumenten aufgrund der *tag abuse*-Problematik nicht notwendigerweise explizite Annotationen der Dokumentstruktur vorliegen. Der Textparser ermittelt Bausteine der Textoberfläche, die Hypertextmodulen oder einzelnen Bestandteilen von Hypertextmodulen entsprechen; die korrespondierenden Ergebnisse der Analyse des Parsers werden innerhalb des Dokuments durch XML-Elemente und -Attribute eines weiteren Namensraumes annotiert. In einigen Fällen können Hypertextmodule unmittelbar auf Hypertextsortenmodule abgebildet werden (z. B. hinsichtlich der *Navigationshilfe* oder der *Hotlist*). Die innerhalb des Textparsers implementierten Regelsysteme operieren zwar fehlerfrei, müssen jedoch zusätzlich um zahlreiche weitere Regeln ergänzt werden, um eine robuste Identifizierung von Hypertextmodulen in arbiträren HTML-Dokumenten gewährleisten zu können. Eine weitere Komponente der Architektur betrifft die Erkennung der Grenzen von Hypertexten und eingebetteten Hypertexten. Diese Zuordnung ist notwendig, um die an einem Hypertext (und somit an einer Hypertextsorte) beteiligten Knoten zu ermitteln. Zur eigentlichen Erkennung der Hypertextsortenmodule und Hypertextknotensorten werden Verfahren aus der Informationsextraktion vorgeschlagen: *Wrapper* werden üblicherweise dazu eingesetzt, Informationen und Daten aus HTML-Dokumenten zu extrahieren, indem charakteristische Positionen des Elementbaums auf die Existenz von Schlüsselwörtern untersucht werden. Da *Wrapper* nur für diejenigen Websites präzise Resultate liefern, die auf Templates basieren und zu präsentierende Informationen aus Datenbanken beziehen, sind jedoch robustere Erkennungsverfahren notwendig. Die Erkennung kann nach unterschiedlichen Strategien erfolgen. Eine dieser Methoden betrifft z. B. die Verknüpfung von Hypertextsortenmodulen mit zugehörigen *Wrappern* innerhalb der Hypertextsortenontologie, um auf diese Weise die Verarbeitung gegebener HTML-Dokumente mit einer Vielzahl von *Wrappern* durchzuführen. Sobald ein *Wrapper* erfolgreich die von ihm unterstützte Informationseinheit ermittelt hat, liegt die erfolgreich erkannte Instanz eines korrespondierenden Hypertextsortenmoduls vor. Aus der Menge der auf diese Weise identifizierten Hypertextsortemodule kann diejenige Hypertextsorte bestimmt werden, die die meisten dieser Hypertextsortemodule als obligatorische oder optionale Konstituenten enthält. Zusätzlich kann diese Vorgehensweise zur Informationsextraktion eingesetzt werden, indem die *Wrapper* die ermittelten Informationen in ein XML-basiertes Format überführen, das mit den in der Hypertextsortenontologie verwendeten Etiketten oder den von ihr referenzierten DTD-Fragmenten (vgl. die Abschnitte 13.5.7 und 13.5.8) korrespondiert. Da die robuste Implementierung der vorgeschlagenen Architektur, die darüber hinaus hinsichtlich der Untersuchungsdomäne einen sehr großen Abdeckungsgrad besitzen müsste, mit einer immensen Komplexität verbunden ist, wird eine reduzierte Architektur vorgestellt, die an die bislang vorgelegten Ansätze anknüpft, jedoch die zentrale Problematik der Beschränkung auf einzelne HTML-Dokumente durchbricht, indem

alle Knoten eines Hypertextes als ein synthetisches Einzeldokument betrachtet und analysiert werden. Falls eine umfassende Menge von Trainingshypertexten bzw. -dokumenten für jede Hypertextsorte vorliegt, kann durch diese Methodik die eigentliche Anordnung von Hypertextsortenmodulen unbeachtet bleiben, da sie als Bestandteil des synthetischen Einzeldokuments aufgefasst werden, so dass ihre Positionierung (z. B. bezüglich der Realisierung in der Einstiegsseite versus einem untergeordneten Knoten) letztlich keine Rolle spielt.

14.9 Fazit – Zur Identifizierung von Hypertextsorten durch Suchmaschinen

Traditionelle IR-Verfahren werden seit Jahrzehnten in Produktionssystemen angewendet, um Recherchen in großen Textkollektionen zu ermöglichen. Auf das WWW spezialisierte Methoden, die von der Hyperlinkstrukturierung und HTML-Auszeichnung Gebrauch machen, werden seit etwa zehn Jahren mit zunehmender Präzision in Suchmaschinen eingesetzt.

Das Suchen nach Hypertextsorten bzw. Web-Genres ist, wie Rosso (2005, S. 190) anmerkt, zweifelsohne eine sehr reizvolle Idee, deren Implementierung jedoch „a hard [problem]“ darstellt. Die bislang vorgelegten Studien zeigen, dass die maschinelle Identifizierung einer geringen Anzahl von Web-Genres zwar prinzipiell möglich ist (vgl. Abschnitt 14.2.2), jedoch beziehen sich diese Arbeiten nicht auf eine theoretische Grundlage, so dass die realen Gegebenheiten im *World Wide Web* durch die Beschränkung auf einzelne HTML-Dokumente, die als atomare Einheiten der Analyse betrachtet werden, nicht adäquat erfasst werden können. Wie dieses Kapitel gezeigt hat, ist die Operationalisierung des in der vorliegenden Arbeit präsentierten Hypertextsortenmodells ein, um die Charakterisierung von Rosso fortzuführen, *sehr* schwieriges Problem: Die präzise Identifizierung von Hypertextsorten muss zwangsläufig sowohl eine Analyse der Binnenstrukturen der beteiligten HTML-Dokumente als auch die übergeordnete Ebene des Gesamthypertextes sowie der eingebetteten Hypertexte berücksichtigen. Zudem stellt die initiale Identifizierung eines Inventars von Hypertextsorten eine zentrale Prämisse dar, letzten Endes decken die in den bislang vorliegenden Arbeiten eingesetzten Web-Genres nur einen sehr kleinen Teil der existenten Hypertextsorten ab.

Dieser Aspekt betrifft unmittelbar die Frage, weshalb im *World Wide Web* derzeit noch keine Suchmaschine existiert, die es dem Benutzer gestattet, nach Web-Genres bzw. Hypertextsorten zu recherchieren, schließlich wurde die prinzipielle Möglichkeit der Realisierung dieses Ansatzes bereits erfolgreich demonstriert: Die Datenbanken von Anbietern wie z. B. *Google* umfassen Informationen zu mehr als acht Milliarden HTML-Dokumenten. Diese Bestände werden permanent ergänzt, so dass der Berechnungsaufwand für eine einzelne HTML-Datei so gering wie möglich sein muss, um den Index möglichst aktuell halten zu können. Neben dem Aufwand müssen maschinelle Kategorisierungsprozesse im Rahmen einer Produktionsumgebung in der Lage sein, mit *beliebigen* Dokumenten und Inhalten umgehen zu können, um robuste Analysen zu liefern (Stamatatos et al., 2001, S. 471). Selbst wenn zukünftige Ansätze zur maschinellen Identifizierung von Hypertextsorten sehr effiziente Algorithmen zur Verfügung stellen werden, bleibt jedoch die Kernfrage bestehen, auf welche Weise das für diesen Zweck einzusetzende Kategorieninventar bestimmt werden soll. Der in dieser Arbeit analysierten Untersuchungsdomäne der universitären Webangebote ist nur ein kleiner

Teil der mehr als acht Milliarden HTML-Dokumente im WWW zugehörig. Zudem wurden lediglich Stichproben untersucht, so dass die tatsächliche Anzahl von Hypertextsortenmodulen, Hypertextknotensorten und Hypertextsorten für diese Domäne deutlich höher ausfallen dürfte als dies in Teil III dargestellt wurde. Ein Großteil dieser Hypertextsorten ist auf traditionelle Textsorten zurückzuführen. Prinzipiell können beliebige Exemplare von beliebigen schriftlich realisierten Textsorten im WWW publiziert werden, so dass die maschinelle Identifizierung von Hypertextsorten zwangsläufig ein sehr umfangreiches Spektrum traditioneller Textsorten zu berücksichtigen hat. Mit anderen Worten: Die Identifizierung von Hypertextsorten setzt eine sehr robuste Erkennung von Textsorten voraus, die deutlich über die begrenzten Inventare hinausgehen müssen, die in den bislang für diesen Zweck eingesetzten Arbeiten Verwendung finden (vgl. Abschnitt 14.2.2). Bezogen auf die Identifizierung von Hypertextsorten im Rahmen einer domänenunabhängigen Suchmaschine wie z. B. *Google* bedeutet dies, dass ein sprach- und kulturübergreifendes Inventar von Hypertextsorten zu erstellen ist, das sowohl *alle* genuinen Hypertextsorten als auch *alle* im WWW verwendeten traditionellen Textsorten beinhaltet.

Bevor derartige Fragestellungen diskutiert werden können, sind zahlreiche offene Probleme zu bearbeiten; Meyer zu Eissen und Stein (2004, S. 6) weisen zu Recht darauf hin, dass sich die „genre classification of Web pages“ noch in den Kinderschuhen befindet. Hierzu zählt etwa Haas und Grams (2000, S. 186) zufolge die Zuordnung von Dokumenten zu „page types“, die oftmals nicht eindeutig durchgeführt werden kann, weil sich mehrere Bausteine gleichberechtigt in einem einzelnen Dokument befinden. Aus diesem Grund favorisiert Santini (2004a), ähnlich wie Crowston und Kwasnik (2004), einen „multi-faceted approach“, der z. B. ein Dokument als „80% descriptive, 50% instructional, and 30% [...] comment“ bestimmt. Fraglich ist jedoch, ob Anwender eine Benutzerschnittstelle, die auf einer Vielzahl derartiger Parameter beruht, einer Liste intuitiv verständlicher Hypertextsorten vorziehen werden. Das Hypertextsortenmodell bietet bezüglich der Zuordnungsproblematik einen Vorteil: Da Hypertextsortenmodule als Hypertextknotensorten fungieren können, kann die Erkennung einer Hypertextknotensorte als Identifizierung aller Hypertextsortenmodule eines Dokuments konzeptualisiert werden. Somit können z. B. *alle* Bausteine eines Knotens, die eine Relevanz für Suchanfragen besitzen, *parallel* als Hypertextknotensorten dieses Dokuments indexiert werden. Alternativ kann auf Seiten der Benutzerschnittstelle einer Suchmaschine eine Differenzierung zwischen Hypertextknotensorten und Hypertextsortenmodulen implementiert werden, die es den Anwendern gestattet, beide Ebenen in die Recherche einzubeziehen. Bei unklaren Zuordnungen muss für diesen Zweck ein primäres Hypertextsortenmodul bestimmt werden, was z. B. durch seinen Umfang oder die Positionierung innerhalb eines Knotens geschehen kann. Die Berücksichtigung aller drei Ebenen des Hypertextsortenmodells besitzt einen weiteren Vorteil, der sich auf die sehr hohe Anzahl von Hypertextsorten bezieht: Anwendern könnte die Möglichkeit gegeben werden, zunächst einen Hypertexttyp auszuwählen, auf den die Suchanfrage eingeschränkt werden soll. Die Benutzerschnittstelle präsentiert nach dieser Auswahl eine dynamisch aus der Hypertextsortenontologie generierte Liste der zugehörigen Hypertextsorten. Sobald aus dieser Liste einer oder mehrere Einträge selektiert wurden, wird dynamisch eine Aufstellung der zugehörigen Hypertextknotentypen generiert und dargestellt, aus der dann Hypertextknotensorten ausgewählt werden. Eine derartige dynamische Navigation des Inventars von Hypertextsorten

könnte eine intuitive und leicht erlernbare Benutzerschnittstelle für eine Suchmaschine darstellen, die in der Lage ist, Hypertextsorten und ihre beteiligten Komponenten maschinell zu identifizieren. Dieses Kapitel hat jedoch gezeigt, dass die Implementierung eines derartigen Systems, das im Suchmaschinenkontext zwangsläufig einen möglichst umfangreichen Abdeckungsgrad besitzen müsste, mit einer immensen Komplexität verbunden ist, so dass nur langfristig mit einer vollständigen Realisierung gerechnet werden kann.

15 Schlussfolgerungen und Ausblick

15.1 Schlussfolgerungen

Ebenso wie in anderen traditionellen oder digitalen Medien unterschiedliche Textsorten vorliegen, existieren im *World Wide Web* Hypertextsorten. Medieninhärente Spezifika sind dafür verantwortlich, dass eine unmittelbare Übertragung des Konstrukts „Textsorte“ in das WWW nicht möglich ist. Hierzu gehört zunächst die nicht- bzw. multilineare Organisation der Knoten eines Hypertextes: Wenn in der Textlinguistik von einer Textsorte die Rede ist, sind in der Regel Kategorien gemeint, deren Vertreter *eindeutig* als solche identifiziert werden können. Ein im WWW publizierter Hypertext besitzt jedoch zahlreiche Knoten, d. h. HTML-Dokumente, die – ausgehend von einer Einstiegsseite – über Hyperlinks miteinander verknüpft sind. Diese Verknüpfungen können unter anderem in einer linearen, hierarchischen oder auch unsequenzierten Weise realisiert werden. Ein Charakteristikum vieler traditioneller Textsorten bezieht sich auf eine signifikante lineare Reihenfolge der enthaltenen Konstituenten, die in einem WWW-basierten Hypertext aufgrund der unterschiedlichen Sequenziertheitsgrade naturgemäß nur im Falle der selten eingesetzten Monolinearität vorliegen kann. Hypertexte können als Texte aufgefasst werden, deren Inhalte sich in einzelnen Knoten befinden. Knoten können somit als Teiltexte oder auch subordinierte Texte konzeptualisiert werden, die sich auf den übergeordneten Text – den Hypertext – beziehen bzw. diesen erst konstituieren. Der Terminus Hypertextsorte bezieht sich demnach auf konventionalisierte, rekurrent in Erscheinung tretende Eigenschaften des übergeordneten Kommunikats, das durch mindestens einen, in der Regel jedoch mehrere Knoten realisiert wird. Exemplare einer Hypertextsorte entsprechen einander in Bezug auf diese konventionalisierten Eigenschaften.

Das in Kapitel 5 eingeführte Hypertextsortenmodell sieht mehrere Merkmale vor, hinsichtlich derer eine Hypertextsorte markiert sein kann. Hierzu zählt zunächst die *kommunikative Funktion* als wohl zentralste Eigenschaft aller Textsorten. Daneben existieren *kontextuelle Faktoren*, die sich z. B. auf das Verhältnis von Produzent und Rezipient, die allgemeine

oder technisch bedingte Kommunikationssituation und die Produktionsbedingungen beziehen. Weiterhin kann eine Hypertextsorte mit bestimmten *Inhalten bzw. Themen* korrelieren. Während sich diese drei Merkmale von Hypertextsorten an typischen Kriterien traditioneller Textsorten orientieren, beziehen sich die verbleibenden vier Merkmale auf medieninhärente Spezifika. Hypertexte im WWW umfassen oftmals interaktive Komponenten, so dass eine Markierung hinsichtlich der *Interaktion* vorliegen kann, die sich möglicherweise auch auf die *Kommunikation* mit anderen Benutzern oder den Produzenten eines Hypertextes bezieht. Die Eigenschaften einer Hypertextsorte betreffen vollständige Vertreter, d. h. alle Knoten zugehöriger Hypertexte. Diese können charakteristische Formen der *Strukturierung* (z. B. eine lineare oder multilineare Sequenzierung) und eine einheitliche *Dekoration* (d. h. ein konsistentes Webdesign) aufweisen. Spezifische konventionalisierte Eigenschaften einzelner Knoten werden von diesen Merkmalen nicht erfasst, weshalb das Hypertextsortenmodell die zweite konzeptuelle Ebene der Hypertextknotensorte besitzt. Knoten können eine eigenständige oder abweichende *Dekoration* sowie eine spezifische *kommunikative Funktion* aufweisen, die sich unterstützend auf die übergeordnete kommunikative Funktion der Hypertextsorte bezieht. Knoten sind jedoch nicht als monolithische Entitäten aufzufassen. Das Merkmal der *Positionierung* bezieht sich auf die eigentlichen atomaren Bausteine der dritten Ebene: Hypertextsortenmodule können von Autoren prinzipiell frei und sehr flexibel in einem oder mehreren Knoten angeordnet werden. In einer Einstiegsseite existieren meist Hypertextsortenmodule, die primär der Navigation dienen, aber auch solche, die vornehmlich inhaltlich-thematisch oder bezüglich ihrer Dekoration markiert sind. Das für traditionelle Textsorten wichtige Merkmal des Textstrukturmusters kann ebenfalls für Hypertextsortenmodule gelten, denn eine gedruckte *Publikationsliste* entspricht in ihrer Textstruktur einer im WWW veröffentlichten *Publikationsliste*. Ein Knoten kann Instanzen mehrerer Hypertextsortenmodule enthalten – falls spezifische Kombinationen rekurrent in mehreren Hypertexten verwendet werden, können sie als Hypertextknotensorten aufgefasst werden. Darüber hinaus können Hypertextsortenmodule als Hypertextknotensorten fungieren, wenn sie das einzige oder das dominierende Hypertextsortenmodul eines Knotens darstellen. Abschließend sind Hypertextsortenmodule auch in der Lage, als Hypertextsorten zu fungieren, wenn z. B. die einzelnen Bestandteile einer *Publikationsliste* in unterschiedliche Knoten aufgeteilt werden.

Das hier nur in seinen Grundzügen zusammengefasste Hypertextsortenmodell bezieht sich auf beliebige Typen von Gebrauchshypertexten und unterscheidet mit Hypertextsortenmodulen, Hypertextknotensorten und eingebetteten Hypertextsorten drei Ebenen der Konstituenz, die ihrerseits spezifische Merkmale besitzen. Das Modell der Entwicklung von Hypertextsorten (Abschnitt 4.3) betrifft ein Wechselspiel der Rezeption und Produktion von Hypertexten. Im *World Wide Web* existieren Hypertexte mit verwandten kommunikativen Funktionen und ähnlichen Inhalten. Die jeweiligen Autoren rezipieren – neben zahlreichen anderen Websites – auch diese verwandten Hypertexte und modifizieren die eigenen Hypertexte, indem z. B. Bausteine hinzugefügt werden, die aus den anderen Hypertexten stammen oder die von Merkmalen beliebiger anderer Hypertexte inspiriert wurden. Dieser aus den Phasen der Rezeption und der Produktion bzw. Modifikation bestehende Zyklus ist dafür verantwortlich, dass sich allmählich hypertexttyp- oder hypertextsortenspezifische Konventionen etablieren, die in hochfrequenten Fällen den Charakter von Standards besitzen – hierzu zählt z. B. das dreispaltige Layout bei Vertretern der *Online-Zeitung* oder die am lin-

ken Seitenrand positionierte *primäre Navigationshilfe* mit vertikal angeordneten Hyperlinks bei vornehmlich professionell gestalteten und kommerziell ausgerichteten Websites. Derartige Konventionen können in unterschiedlichen Ausprägungen und Detailliertheitsgraden für weitere Hypertextsorten und Hypertextknotensorten ermittelt werden. So umfasst z. B. die *persönliche Homepage eines Wissenschaftlers* eine Vielzahl hochfrequenter Hypertextsortenmodule. Gerade allgemeinere Komponenten der komplexen Hypertextsortenmodule *Identifikation* und *Kontaktinformationen* werden ebenfalls in der Hypertextsorte *private Homepage eines Studierenden* eingesetzt, die somit als übergreifende Hypertextsortenmodule oder auch prototypischer Kern des Hypertexttyps *Homepage einer Person* aufgefasst werden können. Hypertexttypen besitzen nur wenige distinktive Eigenschaften, weisen einen sehr umfassenden Geltungsbereich auf und bündeln mehrere Hypertextsorten, die wiederum ein umfangreicheres Inventar distinktiver Merkmale und einen geringeren Geltungsbereich besitzen.

Da die bislang zum Thema Hypertextsorten bzw. Web-Genres vorliegenden Arbeiten sehr heterogener Natur sind und sich weder eine einheitliche Forschungslinie noch eine homogene Terminologie – die bislang vorgeschlagenen Inventare von Web-Genres eingeschlossen – etabliert hat, setzt die Identifizierung und Analyse von Hypertextsorten eine aufwändige Vorgehensweise voraus: Zunächst ist eine Untersuchungsdomäne zu wählen, auf die sich die Untersuchung beziehen soll, z. B. kommerzielle Websites einer bestimmten Branche oder, wie im Falle der vorliegenden Arbeit, universitäre Webangebote. Für die Untersuchungsdomäne ist ein Korpus anzufertigen, so dass für die Dauer der Studie eine fixierte Dokumentkollektion zur Verfügung steht (Kapitel 7). Daraufhin sind zufällig zusammengestellte Stichproben von Dokumenten zu analysieren. Diese können einerseits mehrere Vertreter einer spezifischen Hypertextsorte oder Hypertextknotensorte oder beliebige Dokumente beinhalten. Im ersten Fall können detaillierte Konventionen ermittelt werden, die sich z. B. auf rekurrente Hypertextsortenmodule oder konventionalisierte Hypertextknotensorten beziehen. Im zweiten Fall geht es primär um die Sammlung und Identifizierung von Hypertextknotensorten und der zugehörigen, übergeordneten Hypertextsorten.

Im Rahmen der fünften Studie, die sich mit eben diesem Ziel beschäftigt (Kapitel 12), konnten insgesamt 112 Hypertextsorten und 159 Hypertextknotensorten in einer Stichprobe von 750 Dokumenten ermittelt werden. Zahlreiche Hypertextknotensorten gehen auf ursprünglich traditionelle Textsorten zurück, z. B. *Lexikoneintrag*, *Pressemitteilung*, *Ankündigung*, *Bibliografie*, *Einladung*, *Kochrezept*, *Reisetagebuch*, *Rundschreiben* und *Wahlergebnis*. Andere Hypertextknotentypen bzw. -sorten sind als WWW- bzw. Hypertext-spezifisch aufzufassen, z. B. *Einstiegsseite*, *Verteiler*, *„under construction"-Hinweis*, *Sitemap*, *Splash-Seite* und *persönliche Homepage eines Wissenschaftlers*. Ohne an dieser Stelle ausführlich auf die Ergebnisse der Analyse einzugehen, zeigen die Beispiele, dass im *World Wide Web* eine immense Bandbreite von Hypertextknotensorten und auch Hypertextsorten eingesetzt wird.

Bezüglich des Ziels der Erstellung einer Typologie von Hypertextsorten für die Untersuchungsdomäne der universitären Webangebote verdeutlichen die genannten Beispiele verschiedene Aspekte. Bereits für traditionelle Textsorten ist diese Aufgabe, wie Heinemann (2000c, S. 538) anmerkt, mit Schwierigkeiten verbunden: „Zum Problem wird dieses – in der praktischen Kommunikation immer [...] funktionierende – Einordnen und Klassifizieren von Texten und Textsorten, wenn es explizit gemacht und in der Form von ›Ordnungen‹ dargestellt werden soll." Hypertextsorten können zunächst in ursprünglich traditionelle Text-

sorten und genuine Hypertextsorten eingeteilt werden, wobei verschiedene Zwischenstufen existieren (vgl. Shepherd und Watters, 1998). Somit können zumindest auf die erste Klasse von Hypertextsorten zahlreiche traditionelle Texttypologien angewendet werden, die z. B. unterschiedliche Funktions-, Situations- und Themenentfaltungstypen differenzieren (vgl. Kapitel 2). Das Hypermediasystem *World Wide Web* legt zudem die Annahme weiterer sehr genereller Typen nahe, die sich z. B. auf unterschiedliche Ausprägungen von Interaktivität, Multimedialität oder Kommunikationspotenziale beziehen. Darüber hinaus existieren, wie die im Hypertextsortenmodell vorgesehenen Stufen der Konstituenz verdeutlichen, weitere Typologisierungsebenen: Neben unterschiedlichen Typen von Hypertextsorten können unterschiedliche Typen von Hypertextknotensorten und auch Hypertextsortenmodulen differenziert werden (vgl. die Kapitel 11 und 12). Eine weitere Schwierigkeit stellen konventionalisierte Hyperlinkstrukturen dar, die z. B. von der *Einstiegsseite eines universitären Webauftritts* zu einem *Verteiler* und von dort zu unterschiedlichen Ausprägungen des Hypertextknotentyps *Webauftritt einer dezentralen Organisationseinheit* führen.

Das im Kontext von Hypertextsorten besonders ausgeprägte Problem der Typologisierung kann nur durch eine integrative Herangehensweise gelöst werden, die alle beteiligten Ebenen berücksichtigt und flexible Zuordnungen einzelner Entitäten zu Kategorien bzw. Typen erlaubt. Für diesen Zweck eignet sich die *Web Ontology Language* als maschinelles Repräsentationsformat. Sowohl die im Hypertextsortenmodell vorgesehenen Ebenen als auch die Beziehungen zwischen diesen Ebenen (Kapitel 5) können unmittelbar als Klassen und Relationen einer OWL-Ontologie repräsentiert werden (Kapitel 13). In die vorliegende Hypertextsortenontologie wurden die in Teil III ermittelten Ergebnisse integriert. Darüber hinaus gestattet diese Vorgehensweise mannigfaltige Visualisierungsmöglichkeiten, die sich z. B. auf die einzelnen Bestandteile einer Hypertextsorte und typologische Aspekte beziehen. Basierend auf dem Hypertextsortenmodell besitzt die Hypertextsortenontologie einen sehr komplexen Aufbau, da sie mehrere Funktionen parallel erfüllt: Unter anderem modelliert sie multiple Typologien, repräsentiert die Konstituenten von Hypertextsorten und umfasst konventionalisierte Hyperlinkstrukturen. Diese Komplexität ist nun keinesfalls als Selbstzweck zu verstehen, vielmehr wird sie benötigt, um die ebenfalls sehr komplexen realen Gegebenheiten im *World Wide Web* adäquat erfassen zu können. Dies gilt insbesondere für den eigentlichen Vorteil der Hypertextsortenontologie, der den Umstand betrifft, dass sie auf einem texttechnologischen Standardformat beruht und somit als Ressource in einem sprachtechnologischen System zur maschinellen Identifizierung von Hypertextsorten fungieren kann.

Für diese Anwendung liegen bereits Arbeiten vor, die jedoch mit verschiedenen Problemen behaftet sind (Abschnitt 14.2.4): Hierzu zählen die begrenzten Kategorieninventare, die lediglich zwischen zwei und 15 Web-Genres beinhalten und somit nur einen sehr kleinen Ausschnitt der existenten Hypertextsorten darstellen. Zudem fokussieren die Studien die technologische Ebene, ohne theoretische Aspekte zu reflektieren, d. h. der Genre-Begriff wird unmittelbar in das WWW übertragen, ohne die eigentlichen Spezifika von *Web*-Genres zu diskutieren – eben dieser Problematik unterliegen auch viele der Studien, die auf die Sammlung unterschiedlicher Web-Genres abzielen (vgl. Abschnitt 4.4). Aus diesem Grund ist die in allen Arbeiten zur maschinellen Identifizierung von Web-Genres vorliegende Annahme, *ein* HTML-Dokument instanziiere genau *ein* Web-Genre, zurückzuweisen. Diese Problemkreise lassen eine Neuausrichtung des noch sehr jungen Forschungsbereiches „automatic Web genre

classification“ notwendig erscheinen und beziehen sich auf seine zentralen Konzepte, schließlich reflektieren die bislang vorliegenden Ansätze nicht die realen Gegegebenheiten im *World Wide Web*, da sie unter anderem weder die übergeordnete Ebene der Hypertextsorten noch die untergeordnete Ebene der Hypertextsortenmodule berücksichtigen.

Die in Kapitel 14 vorgeschlagene Neuausrichtung bezieht sich auf ein mehrstufiges Verfahren: Zunächst sind mit einem Textparser für arbiträre HTML-Dokumente Hypertextmodule zu bestimmen, die die Bausteine der Textoberfläche darstellen. HTML enthält zwar Elemente zur Annotation von Textstrukturen, jedoch liegen derartige Auszeichnungen aufgrund der *tag abuse*-Problematik nur in den seltensten Fällen vollständig und korrekt vor. Der in Abschnitt 14.5 präsentierte Textparser operiert auf beliebigen HTML-Dokumenten, die nach XHTML konvertiert werden, so dass XML-Dokumente zur Verfügung stehen, die mit entsprechendenen Werkzeugen verarbeitet werden können. Über eine Analyse der Element- und Attributstrukturen können nun Hypertextmodule ermittelt und bereits teilweise auf Hypertextsortenmodule abgebildet werden (z. B. *Hotlist* und *primäre Navigationshilfe*). In einer vollständig implementierten Systemarchitektur kann die Hypertextsortenontologie für verschiedene Aufgaben eingesetzt werden. Zunächst dient sie als Format zur Repräsentation des Aufbaus von Hypertextsorten. Da OWL eine XML-basierte Sprache darstellt, können beliebige weitere Ressourcen, die von einzelnen Komponenten des Systems zur Erkennung eingesetzt werden, unmittelbar in der Ontologie referenziert werden. Beispielsweise kann die Klasse, die das Hypertextsortenmodul *Publikationsliste* repräsentiert, mit korrespondierenden Hypertextmodulen, charakteristischen Schlüsselwörtern, Statistiken über die Frequenzen von Interpunktionszeichen und DTD-Fragmenten assoziiert werden, die eine Erkennung dieses Hypertextsortenmoduls ermöglichen. Darüber hinaus kann jeweils ein Hypertextsortenmodul mit einem *Wrapper* verknüpft werden, der auf die Erkennung eines Hypertextsortenmoduls in einem HTML-Dokument abzielt. Für die Gesamtheit von n Hypertextsortenmodulen in einer Untersuchungsdomäne müssten maximal n *Wrapper* implementiert werden, die diese makrostrukturellen Bausteine identifizieren und gegebenenfalls die enthaltenen Informationen extrahieren und auf die in der Ontologie enthaltenen DTD-Fragmente abbilden. Sobald für alle Knoten eines Hypertextes die maschinell ermittelten Hypertextsortenmodule vorliegen, kann eine Inferenzmaschine auf die Hypertextsortenontologie angewendet werden, um basierend auf diesen Informationen die korrespondierende Hypertextsorte zu ermitteln. Da die in der Hypertextsortenontologie enthaltenen Klassen zusammen mit den DTD-Fragmenten in eine Dokumentgrammatik überführt werden können, läge somit eine Transformation von HTML-Dokumenten nach XML vor. Ein derartiges XML-Dokument kann als Sammlung von Individuen aufgefasst werden, die die in der Hypertextsortenontologie repräsentierten Klassen und Relationen instanziieren.

Bereits diese grobe Schilderung verdeutlicht, dass ein derartiges System eine immense Komplexität besäße. Zudem wird vorausgesetzt, dass *Wrapper* in einer Weise generalisiert werden können, so dass beliebige Ausprägungen von Hypertextsortenmodulen identifizierbar sind. Ein weiteres Problem betrifft die Tatsache, dass Verfahren zur Erkennung der Grenzen von Hypertexten und eingebetteten Hypertexten benötigt werden, und auch für die Generierung multipler DTDs aus einer OWL-Ontologie müsste ein spezifisches Werkzeug implementiert werden. Darüber hinaus sind die derzeit verfügbaren Inferenzmaschinen für OWL nicht in der Lage, über Individuen zu schließen. Aus diesen Gründen bietet es sich zum

gegenwärtigen Stand der Forschung an, die Architektur in ihrer Komplexität zu reduzieren (Abschnitt 14.6.4): Wenn *alle* Knoten eines Hypertextes oder eines eingebetten Hypertextes als ein *einzelnes* synthetisches HTML-Dokument aufgefasst werden, kann die Problematik der flexiblen Anordnung von Hypertextsortenmodulen ignoriert werden. Experten können derartigen „Hypertexten" die korrespondierenden Hypertextsorten zuweisen, so dass diese Daten als Trainingsbeispiele in maschinellen Lernverfahren eingesetzt werden können. Diese Vorgehensweise knüpft in methodologischer Hinsicht unmittelbar an die bislang vorliegenden Arbeiten an und behebt eines ihrer Defizite, da nicht länger einzelne HTML-Dokumente, sondern vollständige „Hypertexte" im Zentrum der Kategorisierung stehen.

Ein derartiger Ansatz könnte in einer Suchmaschine eingesetzt werden, um es den Anwendern zu ermöglichen, eine Anfrage auf die Instanzen spezifischer Hypertextsorten einzuschränken und die Vertreter irrelevanter Hypertextsorten aus einer Recherche auszuschließen. Wenn ein solches Verfahren in eine domänenunabhängige Suchmaschine wie z. B. *Google* integriert werden soll, stellt sich eine zentrale Frage: Wieviele und welche Hypertextsorten existieren im *gesamten World Wide Web*? (Und wie können große Mengen repräsentativer Trainingsdaten gesammelt werden?) In Teil III wurden bereits jeweils mehr als 100 Hypertextsorten und Hypertextknotensorten ermittelt, die zum großten Teil zu den traditionellen Textsorten gehören. Ein Erkennungsverfahren in dem genannten Anwendungsszenario müsste zwangsläufig in der Lage sein, nahezu *beliebige* traditionelle Textsorten, spezialisierte Fachtextsorten und genuine Hypertextsorten präzise zu identifizieren. In unterschiedlichen Sprach- und Kulturkreisen existieren oftmals verwandte Textsorten mit einer identischen kommunikativen Funktion, die jedoch sprach- bzw. kulturspezifische Merkmale aufweisen (vgl. Haas und Grams, 1998b, S. 107, und Beghtol, 2001). Die Möglichkeit der Generalisierung maschineller Erkennungsprozesse, so dass derartige verwandte Text- bzw. Hypertextsorten präzise auf ein einheitliches, gleichsam universales Kategorieninventar abgebildet werden können, erscheint – zumindest zum gegenwärtigen Stand der Forschung – fraglich.

15.2 Ausblick

Für die zukünftige Untersuchung von Hypertextsorten ergibt sich aus den Schlussfolgerungen eine Vielzahl von Konsequenzen und weiterführenden Fragestellungen. Bezüglich der maschinellen Identifizierung von Hypertextsorten, Hypertextknotensorten und Hypertextsortenmodulen betrifft dies zunächst, wie bereits in Kapitel 14 angesprochen, eher technologische Aspekte wie z. B. die Ermittlung der Grenzen von Hypertexten und eingebetteten Hypertexten und die Generalisierung von *Wrapping*-Ansätzen auf beliebige Instanzen spezifischer Hypertextsortenmodule. Darüber hinaus sind Verfahren zu entwickeln, die eine semiautomatische Ermittlung und Sammlung von Hypertextsorten erlauben. Für diese Aufgabe bieten sich unüberwachte Lernverfahren an, die einander ähnliche Objekte zu Klassen bündeln. Die zentrale Variable stellen diesbezüglich die Merkmalsräume dar, auf denen derartige *clustering*-Methoden operieren sollen, d. h. es sind zahlreiche unterschiedliche Konfigurationen von Merkmalen zu untersuchen, um schließlich zu einem Inventar zu gelangen, das eine maschinelle Differenzierung unterschiedlicher Hypertextsorten erlaubt. Die Unterstützung durch semiautomatische Verfahren bezieht sich ebenfalls auf den Aufbau der Hypertextsortenontologie, um z. B. die unterschiedlichen Hypertextmodul-Ausprägungen eines Hyper-

textsortenmoduls maschinell in Bezug auf eine Kollektion ermitteln und in der Ontologie repräsentieren zu können.

Dass die maschinelle Identifizierung von Hypertextsorten hochgradig komplexe Prozesse voraussetzt, wird bereits durch eine Betrachtung der Ansätze zur Erkennung von Textsorten bzw. Genres deutlich (vgl. Abschnitt 14.2.2). Fast alle dieser Studien basieren auf vorliegenden Korpora. Die in diesen Kollektionen enthaltenen Texte sind explizit hinsichtlich ihrer Textsorten ausgezeichnet, so dass diese Informationen unmittelbar für maschinelle Prozesse zur Verfügung stehen. Für das *World Wide Web* existieren derartige Korpora noch nicht, weshalb der Aufbau einer Testkollektion von mehreren tausend HTML-Dokumenten aus einer oder mehreren Domänen ein weiteres Desiderat darstellt. Die enthaltenen Dokumente sollten, z. B. mit Hilfe eines interaktiven grafischen Werkzeugs, manuell oder semiautomatisch mit zusätzlichen Informationen angereichert werden, die sich unter anderem auf die verwendeten Hypertextmodule, Hypertextsortenmodule, Hypertextknotensorten und auch auf unterschiedliche Hyperlinktypen beziehen. Für diesen Zweck bietet sich eine Mehrebenen-Annotation auf der Basis von Ontologien und Dokumentgrammatiken in Verbindung mit Informationsextraktionsverfahren an (vgl. Teil IV). Anschließend können die Annotationen z. B. von maschinellen Lernverfahren in ersten prototypischen Systemen (vgl. Abbildung 14.1, S. 650) und Anwendungsszenarien eingesetzt werden.

Die Anreicherung von HTML-Dokumenten mit Metadaten, die mit den Klassen und Relationen einer oder mehrerer Ontologien korrespondieren, stellt einen wichtigen Baustein für das *Semantic Web* dar. Nach der initialen Phase der manuellen Metadatenanreicherung wird der oben geschilderte Ansatz der semiautomatischen und grafisch unterstützten Annotierung mittlerweile auch innerhalb der *Semantic Web*-Initiative verfolgt, um den sehr aufwändigen Prozess der ausschließlich manuell durchgeführten Annotierung – diesbezüglich ist auch vom „knowledge acquisition bottleneck“ die Rede – zu vereinfachen (vgl. z. B. Ciravegna et al., 2002, Vargas-Vera et al., 2002, und Handschuh et al., 2002, 2003). Die genannten Arbeiten beschäftigen sich in der Regel mit einem Beispielszenario: Handschuh et al. (2002) bilden etwa spezifische Inhalte des Webauftritts eines Hotels (Lage, Zimmerpreise etc.) auf eine Domänenontologie ab. Obwohl die Verfasser davon ausgehen, dass es sich bei „hotel homepages“ um „a certain text type“ handelt (ebd., S. 367), der wiederum mit einer „domain ontology“ korrespondiert (hier: „about tourism“, ebd.), wird der Umstand, dass in eben diesem „text type“ spezifische Konventionen existieren, die als Merkmale einer Hypertextsorte konzeptualisiert werden können, *nicht* thematisiert (in keiner der Arbeiten, die dem Verfasser bekannt sind, wird diese Verbindung explizit diskutiert). Die Repräsentation der Konstituenten von Hypertextsorten in einer Hypertextsortenontologie, die auf *Semantic Web*-Standards basiert, kann als eine weitere Ebene in derartigen Vorhaben eingesetzt werden, um den Annotationsprozess zu generalisieren. Statt der unmittelbaren Abbildung von Informationsbestandteilen auf die Domänenontologie kann zunächst eine Verknüpfung mit spezifischen Hypertextsortenmodulen erfolgen, die wiederum Klassen und Relationen der Domänenontologie reflektieren (vgl. die Beispiele in Kapitel 13). Die Hypertextsortenontologie kann also als weitere Repräsentationsschicht innerhalb des *Semantic Web* aufgefasst werden.

Die Existenz eines Systems zur maschinellen Identifizierung von Hypertextsorten vorausgesetzt, ergibt sich die Frage nach einer adäquaten Benutzerschnittstelle: Untersuchungen zum Einsatz von Suchmaschinen zeigen, dass die Benutzer diese meist zurückhaltend einset-

zen, kaum Kenntnisse über die verwendeten Verfahren zur Bestimmung der Relevanz eines Dokuments besitzen und in einer Suchanfrage nur sehr wenige Suchwörter einsetzen (vgl. Kapitel 1). Da mit einem Inventar von deutlich mehr als 100 Hypertextsorten zu rechnen ist (vgl. Teil III), ist eine Schnittstelle zu konzipieren, die einerseits einen intuitiven Zugriff zur Selektion der gewünschten Hypertextsorten ermöglicht und andererseits möglichst viele Hinweise darauf liefert, welchen Abdeckungsgrad eine spezifische Hypertextsorte besitzt. Heinemann (2000c, S. 537) zufolge sollte eine Typologie von Texten „an das konventionelle Alltagswissen der Kommunizierenden über Textsorten anknüpfen, mit ihm kompatibel, zumindest darauf beziehbar sein", was auch für Hypertexte bzw. Hypertextsorten gilt. Ein Großteil der in Teil III ermittelten Hypertextsorten bzw. Hypertextknotensorten bezieht sich unmittelbar auf traditionelle Textsorten, ein weiterer Teil stellt fachspezifische Textsorten dar. Darüber hinaus existieren Kategorien, die WWW-Benutzern untypisch erscheinen könnten (z. B. *Verteiler*). Die Benutzerschnittstelle kann an die Hypertextsortenontologie anknüpfen und zunächst die Auswahl übergeordneter Kategorien (d. h. Hypertexttypen) erlauben, woraufhin der Anwender sukzessive verschiedene Hypertextsorten und die zugehörigen Hypertextknotensorten selektieren kann. Erläuternde Pop-up-Fenster können zusätzliche Informationen zum Geltungsbereich liefern (Abschnitt 14.7). Auf diese Weise kann das in einer Suchmaschine verfügbare Inventar von Hypertextsorten explorativ erkundet und navigiert werden (vgl. auch Crowston und Kwasnik, 2004).

Bezüglich der linguistischen Perspektive ergeben sich ebenfalls zahlreiche weiterführende Fragestellungen. Diese betreffen zunächst die Untersuchung von Webangeboten aus anderen Untersuchungsdomänen, um die verwendeten Hypertextsorten, Hypertextknotensorten und Hypertextsortenmodule zu ermitteln und domänenspezifische von domänenunabhängigen Ebenen bzw. Strategien zu unterscheiden. Zur Modellierung der erhobenen Hypertextsorten kann das Grundgerüst der Hypertextsortenontologie eingesetzt werden (vgl. Kapitel 13), jedoch wird jeweils eine spezifische Domänenontologie benötigt, auf die die Klassen der Hypertextsortenontologie abgebildet werden. Weiterhin zeichnet sich bereits jetzt ein gewisser Standardisierungstrend ab, der sich unter anderem auf generelle und universale Hypertextsortenmodule bezieht: Die meisten Produzenten setzen mittlerweile sehr komfortable HTML-Entwicklungsumgebungen oder CMS-Systeme ein, die vorgefertigte Bausteine anbieten, die in vielen Fällen unmittelbar Hypertextsortenmodulen entsprechen (z. B. Zugriffszähler, Navigationshilfen und Kalender-Applikationen). Die Charakterisierung der Websites von Konferenzen und Tagungen zeigt darüber hinaus, dass sie in übergeordnete Handlungszusammenhänge involviert sind (vgl. Abschnitt 4.6.2). Insbesondere vor dem Hintergrund rezipientenseitiger Fortsetzungserwartungen (Bucher, 2004) erscheint es sinnvoll, die Zusammenhänge zwischen domänenspezifischen Handlungstypen und gegebenenfalls domänenunabhängigen Hypertextknotensorten genauer zu untersuchen.

Des Weiteren konnten in Teil III nur ausgewählte Aspekte des Hypertextsortenmodells exemplarisch diskutiert werden. Weitere Analysen könnten z. B. untersuchen, auf welche Weise spezifische Funktionen der *persönlichen Homepage eines Wissenschaftlers* realisiert werden oder in welchem Zusammenhang die Gestaltung und Dekoration eines Hypertextes zu seinem Inhalt und seiner Funktion stehen. Darüber hinaus liegt mit dem Textparser ein Werkzeug vor (vgl. Abschnitt 14.5), das zumindest rudimentäre, maschinell unterstützte korpuslinguistische Auswertungen ermöglicht, indem z. B. eine syntaktische Annotation und anschließende

Analyse spezifischer Hypertextmodule oder auch Hypertextsortenmodule durchgeführt wird, um z. B. die Unterschiede der Hyperlinkanzeiger in primären Navigationshilfen und Hotlists zu untersuchen (Abschnitt 14.7). In Teil III wurde diesbezüglich eine abstraktere Perspektive eingenommen, um zunächst die eigentlichen Konstituenten von Hypertextsorten zu bestimmen, die nun in detaillierter Weise analysiert werden können.

Bezüglich einer diachronen Untersuchung von Hypertextsorten ist zu ermitteln, wie sich sowohl einzelne Hypertextexemplare als auch Hypertextsorten über einen längeren Zeitraum entwickeln. Für diesen Zweck kann die Korpusdatenbank (vgl. Kapitel 7) um Funktionen zur Konstruktion von Monitorkorpora erweitert werden, so dass in regelmäßigen Abständen aktuelle Versionen vollständiger Websites in das Korpus integriert werden, um Änderungen untersuchen zu können. Im Hinblick auf das zyklische Modell der Entwicklung von Hypertextsorten ist – z. B. durch einen WWW-basierten Fragebogen oder ein Interview (vgl. Thelwall, 2003) – zu ermitteln, weshalb der oder die Produzenten die beobachteten Änderungen durchgeführt haben, um ihre Ursprünge bestimmen zu können (vgl. Abschnitt 4.3). Derartige Änderungsprozesse sollten naturgemäß auch linguistische Befunde einbeziehen, die zum Beispiel die Hypothese belegen könnten, dass die erste Phase einer *privaten Homepage* von einer konzeptionellen Mündlichkeit geprägt ist, wohingegen spätere Entwicklungs- und Modifikationsstufen eher zur Schriftlichkeit tendieren. Derartige Untersuchungen können Aufschluss über die allmähliche Entwicklung von Hypertextsorten liefern, denn es stellt sich die Frage, welche Änderungsprozesse diesbezüglich in einem Zeitraum von beispielsweise zwei, drei oder zehn Jahren existieren. Brandl (2002, S. 160) vergleicht in diesem Zusammenhang ihre 2001 durchgeführte Studie (vgl. Abschnitt 4.4.5) mit den Ergebnissen einer 1998 entstandenen Untersuchung und kommt zu folgendem Schluss: „Man sieht auch, dass offenbar im Zeitraum von drei Jahren [...] die grobe Typeneinteilung von Webangeboten [...] keine bedeutsame Änderung erfahren hat. Dies ist insofern bemerkenswert, als in der Literatur oft das ›rasante Tempo‹ betont wird, mit dem die Entwicklung des WWW voranschreitet. Die Änderungen werden also nicht auf einer übergreifenden Dimensionsebene wahrgenommen, sondern vielmehr auf Detailebene. Dort kommt es konkret auf spezielle Ausprägungen von Funktionalitäten, Formelementen und Inhalten an [...]."

In Bezug auf diese Aspekte können die Beziehungen zwischen der Untersuchung von Hypertextsorten und Usability-Analysen verdeutlicht werden: Letztere gehen von der Perspektive der Rezipienten aus und ermitteln z. B. anhand von Benutzerstudien, welche Navigationspfade in spezifischen Webangeboten verfolgt werden (vgl. z. B. Bucher, 2004). Die Analyse von Hypertextsorten geht von den Produzenten aus und nimmt eine übergeordnete Perspektive ein, d. h. es werden Gemeinsamkeiten und Unterschiede zwischen Hypertextexemplaren untersucht, die die Annahme unterschiedlicher Hypertexttypen oder Hypertextsorten begründen. Dass jedoch enge Beziehungen zwischen diesen beiden Ansätzen existieren, zeigt die vergleichende Usability-Studie von Nielsen und Tahir (2002), in denen Gemeinsamkeiten und Unterschiede zwischen 50 Einstiegsseiten kommerzieller Webangebote ermittelt wurden, um zu einem Kriterienkatalog für die Gestaltung von Exemplaren dieses Hypertextknotentyps zu gelangen. Der zentrale Ratschlag lautet, etablierte Konventionen in das eigene Webangebot zu übernehmen (vgl. Abschnitt 4.6.2). Es bietet sich daher an, die Rezeptionsforschung auf Hypertextsorten auszudehnen, d. h. Benutzern typische und (gegebenenfalls eigens konstruierte, vgl. Dillon und Gushrowski, 2000) untypische Exemplare einer

Hypertextsorte zu präsentieren und z. B. die Auswirkungen des Fehlens obligatorischer Hypertextsortenmodule oder Aspekte der Etikettierung von Hypertextknotensorten zu untersuchen (vgl. Abschnitt 4.3). Diesbezüglich ist es auch notwendig, eine Operationalisierung der „Web-Kompetenz“ eines Rezipienten durchzuführen, die sich unter anderem auch auf die mit der Benutzung des WWW verfolgten Ziele und Merkmale von Hypertextsorten beziehen sollte (vgl. Haas und Grams, 2000, S. 190). Die Ergebnisse großflächig durchgeführter Analysen bezüglich der verfolgten Nutzerziele kann wiederum zur Auswahl spezifischer Hypertextsorten eingesetzt werden, um einen ersten Prototyp zur maschinellen Identifizierung von Hypertextsorten auf diejenigen Kategorien zu beschränken, die einen möglichst großen Teil der Benutzerbedarfe abdecken (vgl. auch Dewe et al., 1998, und Roussinov et al., 2001).

Dass bezüglich des Einsatzes einer Komponente zur maschinellen Identifizierung innerhalb einer Suchmaschine eine derartige Fokussierung auf ausgewählte und mit automatischen und sprachübergreifenden Verfahren präzise detektierbaren Hypertexttypen und Hypertextsorten stattfinden muss, zeigt die Problematik und Komplexität der Aufgabe, eine Liste *aller* im *World Wide Web* verwendeten Hypertextsorten anzufertigen (vgl. auch Greenspun, 1996, Furuta und Marshall, 1996, und Haas und Grams, 2000, S. 182). Der in Zukunft aller Wahrscheinlichkeit nach gerade im kommerziellen Bereich zunehmende Einsatz von *Semantic Web*-Technologien und die hiermit verbundene Zunahme der Nutzung übergreifender Metadatenvokabulare wird sich positiv auf die Realisierung einer solchen Anwendung auswirken, die im Rahmen eines Produktionssystems eingesetzt werden soll.

Anhang

Überblick

Der Anhang umfasst insbesondere weiterführende Informationen über das Korpus und die Korpusdatenbank. Anhang A präsentiert zunächst eine statistische Auswertung des im Korpus enthaltenen Datenbestandes, geht auf verschiedene Merkmale universitärer Webangebote ein und diskutiert zahlreiche Charakteristika der HTML- und XML-Dokumente. Anhang B stellt die Tabellen dar, die in der relationalen Datenbank unter anderem zur Aufnahme von Metadaten und Verwaltungsinformationen Verwendung finden. Anhang C beinhaltet eine Liste der in dieser Arbeit benutzten Abkürzungen.

Die der Arbeit beiliegende CD ROM umfasst für jede der 100 im Korpus verfügbaren universitären Webauftritte verschiedene Kerninformationen, die sich z. B. auf die Anzahl der jeweiligen Webserver und die in das Korpus integrierten Dokumente beziehen. Sie enthält darüber hinaus Aufstellungen der in den Kapiteln 8, 11 und 12 analysierten Stichproben, zwei Bildschirmvideos zur Demonstration der Oberfläche des Textparsers und eine PDF-Version von Kapitel 13, so dass die teils sehr detailliert visualisierten Ausschnitte der Ontologien dieses Kapitels in hochauflösender Form betrachtet werden können.

A

Statistische Charakterisierung

A.1 Einleitung

Dieser Anhang stellt eine statistische Auswertung der im Korpus enthaltenen Dokumente vor. Analysiert werden zahlreiche strukturelle sowie HTML- und HTTP-bezogenen Merkmale, deren individuelle Frequenzen und spezielle Charakteristika unter anderem einen Einfluss auf die Konzeptionierung und Implementierung des Textparsers hatten. Eine weitere Motivation bezieht sich auf verschiedene empirische Studien, die zwischen 1996 und 1998 veröffentlicht wurden, vgl. Bray (1996), Woodruff et al. (1996), O'Neill et al. (1997, 1998) und Turau (1998a,b). Der Vergleich der Auswertung mit den Daten, die in diesen Beiträgen berichtet werden, verfolgt zwei Ziele: Zum einen handelt es sich um ältere Arbeiten, die mit einem Korpus verglichen werden, das in den Jahren 2001 und 2002 erhoben wurde, so dass die generelle Progression des WWW (bezüglich Nutzung, publizierter Inhalte, Erzeugung der Dateien etc.) mit aktuelleren Daten charakterisiert werden kann. Zum anderen zeigt die Kontrastierung domänenspezifische Unterschiede auf, die auf Daten beruhen, die sich einerseits auf das gesamte WWW, andererseits auf eine wohldefinierte Teilmenge – die deutschsprachigen Dokumente der Webserver des deutschen Wissenschaftsnetzes, das WiN-Web – beziehen.

Zunächst geht Abschnitt A.2 auf die technische Realisierung der Datenerhebung ein. Das in Kapitel 7 präsentierte Korpussystem umfasst zwei wesentliche Bestandteile, bei denen es sich um die relationale Datenbank und den Dokumentbestand handelt. Abschnitt A.3 verfolgt eine eher breite Perspektive und basiert auf den Informationen, die mit Hilfe der Datenbank ermittelt werden können (z. B. die Verteilung einzelner Webserver pro Universität oder der Anteil deutschsprachiger Dokumente). Daraufhin widmet sich Abschnitt A.4 den Charakteristika der im Korpus vorhandenen HTML-Dokumente (z. B. Analysen der Dateigrößen oder des Einsatzes von HTML-Elementen und -Attributen). Abschnitt A.5 umfasst eine Untersuchung der ca. 30 000 XML-Dokumentinstanzen, die ebenfalls während der Datensammlung in das Korpus aufgenommen wurden. Abschließend werden in Abschnitt A.6 verwandte Arbeiten und Fragestellungen diskutiert.

A.2 Technische Realisierung

Die in dieser Studie verwendeten Daten wurden mit SQL-Abfragen und *Perl*-Skripten erhoben. Diejenigen Informationen, die sich z. B. auf die Anzahl Webserver, das Alter von Dokumenten oder auf HTTP-Response-Header beziehen (Abschnitt A.3), wurden der Korpusdatenbank unmittelbar mit entsprechenden SQL-Abfragen entnommen, durch *Perl*-Skripte aufbereitet und semiautomatisch in Balkendiagramme überführt. Diejenigen Informationen, die nur durch eine Untersuchung des Dokumentbestandes ermittelt werden können (Abschnitt A.4), wurden durch *Perl*-Skripte gesammelt, die über die ca. 4 000 000 HTML-Dokumente iterieren und jeweils zwischen zehn und 50 verschiedene Merkmale extrahieren, die teilweise aufgrund der entstehenden Datenmenge von weit mehr als einem Gigabyte in Festplatten-gebundenen Hashes abgelegt und anschließend statistisch untersucht wurden. Hierfür wurde das Modul `Statistics::Descriptive` eingesetzt. Die Verarbeitung der HTML-Dokumente erfolgte mit dem Modul `HTML::Parser` (Version 3.28).[1] Die Untersuchung der Hyperlinks basiert auf dem RFC 2396-konformen Modul `URI`.

A.3 Das Web des deutschen Wissenschaftsnetzes

Da bei der Datensammlung verschiedene Beschränkungen eingesetzt wurden (beispielsweise eine Obergrenze bezüglich der Dateigröße, vgl. Abschnitt 7.2.1), ist mit dem Korpus keine exakte Kopie des WiN-Web entstanden. Aus diesem Grund sind die nachfolgend präsentierten Daten zwar nur approximativ interpretierbar, skalieren jedoch durchaus in Bezug auf den gesamten deutschsprachigen Datenbestand der Untersuchungsdomäne.[2]

Tabelle A.1 (S. 723) bietet eine Übersicht über den Umfang des Korpus und die Gesamtgröße der traversierten Webserver. Die Universitäten wurden in fünf Kategorien eingeteilt, wobei *allgemeine Universitäten* und *technische Hochschulen* sowohl den Kern des WiN-Webs als auch den der Kollektion bilden. Diese Gruppen sind mit jeweils 62 bzw. 12 Institutionen vollständig enthalten, wohingegen die Kategorien *Musik- und Kunsthochschulen*, *Wirtschaftshochschulen* und *sonstige Hochschulen* nur teilweise erfasst wurden. Anhang E auf der beiliegenden CD ROM enthält eine universitätsbezogene Aufstellung verschiedener statistischer Informationen. Insgesamt wurden bei der Traversierung der Webauftritte der 100 im Korpus enthaltenen Hochschulen ca. 15 000 Webserver besucht, von denen ca. 93% auf dem HTTP-Standard-Port 80 arbeiten. Die meisten Webserver betreiben die Universität Bremen (638), die Technischen Universitäten Berlin (606) und München (533) sowie die RWTH Aachen (496). Es ist jedoch generell zu beachten, dass zahlreiche experimentelle und nicht mehr gepflegte Webserver – z. B. aufgrund veralteter Hyperlinks – in diese Daten eingeflossen sind.

[1] Eine automatische Konvertierung der HTML-Dateien nach XHTML sowie die anschließende Erhebung der Daten mit Hilfe eines XML-basierten Werkzeugs wurde nicht in Betracht gezogen, da es das Hauptziel dieser Untersuchung war, möglichst *alle* vorhandenen HTML-Dateien zu analysieren und nicht lediglich diejenige Teilmenge, die erfolgreich nach XHTML transformierbar ist (ca. 98,7%, vgl. Abschnitt 14.4).

[2] Sehr umfangreiche HTML-Dokumente kommen in dieser Domäne kaum vor. Die Korpusdatenbank verzeichnet lediglich 2 094 Vorkommen von Dateien, die größer als 500 Kilobyte und somit nicht Teil des Korpus sind. Im Verhältnis zur Gesamtanzahl von 8 465 105 in der Untersuchungsdomäne angebotenen HTML-Dokumenten handelt es sich hierbei um einen Anteil von 0,02%.

Universitäten in der Korpusdatenbank:	100
• Allgemeine Universitäten (*vollständig*)	62
• Technische Hochschulen (*vollständig*)	12
• Musik- und Kunsthochschulen (*partiell*)	5
• Wirtschaftshochschulen (*partiell*)	5
• Sonstige Hochschulen (*partiell*)	16
Dauer der Datensammlung:	16.01.2001 – 07.09.2002
Traversierte Webserver insgesamt:	14 968
Auf Port 80 operierende Webserver:	13 885
Anzahl per HTTP erreichbarer Dateien:	16 196 511
Anzahl HTML-Dokumente:	8 465 105
Gesamtgröße aller entfernten Webserver:	701 464,29 MB
Gesamtgröße der Korpusdatenbank:	40 914,99 MB
Laufende Wortformen (gesamt; bezogen auf Dateien vom Typ `text/html`):	1 138 794 715
Laufende Wortformen (eindeutig; bezogen auf Dateien vom Typ `text/html`):	12 120 162
Dateien im Korpus gesamt:	4 294 417
• Dateien vom Medientyp `text/html`:	3 956 692
• Dateien vom Medientyp `text/plain`:	270 400
• Dateien vom Medientyp `text/css`:	35 651
• Dateien vom Medientyp `text/xml`:	25 871
• Dateien vom Medientyp `text/sgml`:	956
• Dateien vom Medientyp `message/news`:	490
• Dateien vom Medientyp `message/rfc822`:	436

Tabelle A.1: Der Umfang des Korpus sowie Angaben über die Häufungen von Dateitypen

Des Weiteren werden in Tabelle A.1 die Anzahl der laufenden Wortformen sowie eine Übersicht der Datenbestände hinsichtlich ihrer Medientypen dargestellt.[3] Die wortformbezogenen Informationen wurden mit einem *Perl*-Skript ermittelt, das eine Tokenisierung aller HTML-Dokumente vornimmt und die jeweilige Token-Anzahl notiert. Besonders problematisch waren Vorkommen zahlreicher Sonderzeichen aus unterschiedlichen Zeichensatzkodierungen. Das Skript definiert ein Token als eine Zeichenfolge, die verschiedene Whitespace-Zeichen nicht enthalten darf (in *Perl*-Notation: `␣\t\n\r\f\240\013`). Zeichenketten, die Kommata oder Schrägstriche enthalten, wurden in ihre Bestandteile aufgebrochen und erneut tokenisiert, und Zeichenketten, die einen Hyperlink oder eine E-Mail-Adresse darstellen, wurden ebenso wie der extrem häufig vorkommende TEX- bzw. LATEX-Quellcode ignoriert. Wortinitiale oder -finale Interpunktions- und Sonderzeichen wurden getilgt, dies gilt auch für isolierte Zahlenangaben, Auslassungspunkte und einige andere Sonderfälle. Weiterhin musste eine Heuristik implementiert werden, um zu ermitteln, ob es sich tatsächlich um ein Wort im eigentlichen Sinne handelt, um binäre Zeichenfolgen auszuschließen.

Vor der Datensammlung wurden verschiedene Medientypen spezifiziert, die in das Korpus aufgenommen werden sollten (vgl. Abschnitt 7.2.1). Erwartungsgemäß sind 92,1% der im Korpus verfügbaren Dateien HTML-Dokumente, die von einem Webserver mit dem Me-

[3] Die ca. 1,1 Milliarden laufenden und ca. 12 Millionen eindeutigen Wortformen können mit Angaben für (i) das „Australian academic Web“, (ii) das „New Zealand academic Web“ und das (iii) „United Kingdom academic Web“ verglichen werden (Thelwall, 2005, S. 612): (i) 656 653 108 bzw. 1 542 589 Wörter (205 513 HTML-Dokumente), (ii) 60 998 403 bzw. 385 093 Wörter (2 152 386 HTML-Dokumente), (iii) 1 349 418 614 bzw. 3 522 664 Wörter (4 864 271 HTML-Dokumente).

dientyp `text/html` versendet werden. An zweiter Stelle folgen ASCII-Dateien (6,3%). Diese große Zahl von mehr als 270 000 Dateien erklärt sich durch den Umstand, dass viele Programmquelltexte und Konfigurationsdateien als `text/plain` verschickt werden: Nur 26,5% dieser Dateien besitzen die Endung `.txt`, 16,4% umfassen den Suffix `.d`, der unter anderem für Definitionsdateien der Sprache *Modula-2* benutzt wird, und 2% (5 646 Dateien) sind BibTeX-Dateien mit der Endung `.bib`. Die restlichen Dateien enthalten mehr als 3 700 unterschiedliche Endungen, unter denen auch Suffixe wie `.gz` oder `.Z` zu finden sind, die Binärdateien markieren. CSS-Dateien machen ca. 0,8% des Korpusbestandes aus. Die wenigen SGML-Dokumente (0,02%) umfassen Software-Dokumentationen (z. B. der SGML- bzw. XML-basierten *DocBook*-DTD und verschiedener UNIX-Programmpakete), d. h. nur bei wenigen Dutzend der SGML-Instanzen handelt es sich um Dokumente, die innerhalb von Forschungsprojekten oder Lehrveranstaltungen annotiert wurden. Die ebenfalls nur sporadisch im WiN-Web vorkommenden Medientypen `message/news` und `message/rfc822` umfassen Usenet-Postings und einzelne Mitteilungen oder Digests von Mailing-Listen. Die geringe Anzahl dieser Dokumente dürfte zwei Gründe haben: Einerseits sind die Quellen derartiger Informationen mit zahlreichen Werkzeugen direkt in HTML-basierte Formate überführbar, weshalb der für den Endbenutzer unkomfortable Weg der Veröffentlichung der rein textbasierten Originalquelle selten realisiert wird, andererseits bedeutet es einen zusätzlichen Mehraufwand, die unterschiedlichen Dateiendungen innerhalb der Konfiguration des Webservers auf die genannten Medientypen abzubilden, weshalb sich vermutlich weitere Postings und E-Mails in der Gruppe der Dateien des Typs `text/plain` befinden.

A.3.1 Anzahl universitärer Webserver

Die 74 Universitäten der beiden Gruppen *allgemeine Universitäten* und *technische Hochschulen* verfügen über insgesamt 14 628 Webserver (97,7% der 14 968 im Korpus enthaltenen Server), was im Durchschnitt ca. 198 Webservern pro Universität des Kernbereiches entspricht. Von einer Kernuniversität werden im Schnitt ca. 212 296 Dateien per HTTP angeboten (Min.: 6 130, Hochschule Vechta; Max.: 664 168, Universität Karlsruhe), auf `text/html` beschränkt ergibt sich ein Durchschnitt von ca. 110 895 Dokumenten (Min.: 2 903, Hochschule Vechta, Max.: 304 231, Universität Hannover). Wird diese Aufstellung zusätzlich eingeschränkt auf die im Korpus verfügbaren HTML-Dokumente, resultiert dies in dem durchschnittlichen Wert von 51 715 HTML-Dateien pro Hochschule (Min.: 2 035, Universität Vechta, Max.: 150 048, Universität Karlsruhe). Die verbleibenden 26 Universitäten der drei Gruppen *Musik- und Kunsthochschulen*, *Wirtschaftshochschulen* und *sonstige Hochschulen* subsumieren insgesamt 340 Webserver, was im Durchschnitt ca. 13 Webservern und 17 928 per HTTP angebotenen Dateien entspricht. Bei letzterem Wert liegt eine sehr große Streuung vor, so bieten etwa die GISMA Business School (Hannover), die Deutsche Sporthochschule (Köln) oder die Hochschule für Bildende Künste (Braunschweig) jeweils weniger als 100 Dateien an, wohingegen die Universität Lübeck, die Fernuniversität Hagen und die Universität der Bundeswehr München jeweils mehr als 50 000 Dateien zur Verfügung stellen.[4]

[4] Die extrem geringen Werte sind möglicherweise auf den Umstand zurück zu führen, dass die Administratoren der jeweiligen Webserver große Teile ihrer Informationsangebote mit Hilfe des *Robot Exlusion Standard* gegen einen maschinellen Zugriff gesperrt haben (vgl. Fußnote 8 auf S. 325).

Abschnitt A.3.5 geht in detaillierter Form auf die Verteilung aller Webserver bezüglich der per HTTP angebotenen Dokumentmenge ein.

A.3.2 Anteil deutschsprachiger Dokumente

Basierend auf den in Abschnitt A.3.1 genannten Werten sowie der Annahme, dass der Sprachenidentifizierer mit einer Präzision von 100% arbeitet, können ca. 46% der im WiN-Web angebotenen HTML-Dokumente als deutschsprachig aufgefasst werden. Der Algorithmus zur Sprachenidentifizierung verzerrt das Ergebnis nur minimal, da er mit einer Präzision von ca. 96,64% arbeitet (vgl. Abschnitt 7.2.2); fehlerhafte Klassifikationen treten lediglich bei sehr kurzen fachsprachlichen Dokumenten auf. Die laut `germanp.pl` anderssprachigen Dokumente umfassen nach einer stichprobenartigen Inspektion der Dateien, die nicht in das Korpus aufgenommen wurden, z. B. zahlreiche gespiegelte Instanzen verschiedener englischsprachiger Dokumentationen, z. B. zur Programmiersprache *Java* oder als HTML-Dokumente verfügbare *Linux*-Handbuchseiten, sowie englisch- oder französischsprachige Versionen der Informationsangebote der jeweiligen Institutionen. Es kann festgehalten werden, dass der Anteil anderssprachiger Dokumente größer ist als erwartet.

A.3.3 Verwendete Webserver-Typen

Ein HTTP-Response-Header enthält in dem Feld `Server` oftmals eine Angabe über den verwendeten Webserver-Typ, d. h. die eingesetzte Webserver-Software, die in der Korpusdatenbank in `server_info.servertype` abgelegt wird (vgl. Abschnitt 7.2.3). Tabelle A.2 enthält eine aggregierte Auswertung dieser Daten. Der prozentuale Anteil bezieht sich dabei auf die 10 840 Server, die eine Identifizierung innerhalb des HTTP-Response-Headers versendet haben. In einigen der nachfolgenden Analysen wird diese Anzahl als diejenige Menge von Webservern betrachtet, die während der Datensammlung tatsächlich verfügbar war.

Verschiedene Gründe können dafür verantwortlich sein, dass ein Webserver, auf den ein Hyperlink zeigt, keine Rückmeldung sendet (und somit im Client eine HTTP-Zeitüberschreitung erzeugt), so könnte die Maschine kurz- oder langfristig abgeschaltet sein, die Webserver-Software selbst könnte nicht aktiv sein, es könnte ein Netzwerkdefekt vorliegen oder der referenzierende Hyperlink ist nicht mehr aktuell. Da prinzipiell alle Webserver das `Server`-Feld belegen, wird davon ausgegangen, dass die Abwesenheit dieses Feldes in einem „Response-Header“ die Inaktivität dieses Rechners signalisiert. Aus diesem Grund können 4 128 der 14 968 in der Datenbank notierten Webserver als „nicht verfügbar“ gelten.[5]

Die Aufstellung wird angeführt von der Open-Source-Software *Apache*, die auf etwa 75% aller Webserver zum Einsatz kommt, wobei insgesamt ca. 67% auf UNIX-Derivaten (insbesondere *Linux*) operieren. Verschiedene Versionen von *Microsoft Windows* kommen auf 13% der 10 840 Webserver zum Einsatz, wobei nur 260 *Apache*-basierte Webserver diese Betriebssystemfamilie einsetzen, die restlichen 11% teilen sich Versionen des *Microsoft*-Produkts *Internet Information Server* (*IIS*). Varianten der kostenpflichtigen Software *Netscape-Enterprise*

[5] Von diesen 4 128 Webservern waren 178 Webserver zum Zeitpunkt der Datensammlung entgegen dieser Annahme doch verfügbar, wie anhand einer Überprüfung der Dateien, die von diesen Servern im Korpus enthalten sind, gezeigt werden kann. Da es sich hierbei jedoch um lediglich ca. 2 000 Dateien handelt, wird durch die genannte These das Ergebnis der Studie nicht verfälscht.

Webserver-Software		Anzahl	Prozent	Weltweiter Marktanteil
Apache		8 107	74,79	56,89
	UNIX-Versionen	7 252	66,90	
	Ohne Angabe des Betriebssystems	595	5,49	
	Microsoft Windows-Versionen	260	2,40	
Microsoft Internet Information Server (IIS)		1 170	10,793	28,99
	Version 1.0	8	0,07	
	Version 2.0	22	0,20	
	Version 3.0	209	1,93	
	Version 4.0	507	4,68	
	Version 5.0	423	3,90	
	Version 5.1	1	0,01	
Netscape	(verschiedene Versionen)	270	2,49	3,79
NCSA		118	1,09	0,02
	Version 1.3	22	0,20	
	Version 1.4.*	40	0,37	
	Version 1.5.*	55	0,51	
CERN	(fast aussschließlich Version 3.0)	160	1,48	0,00
Roxen	(verschiedene Versionen)	158	1,46	0,02
Verschiedene	(etwa 130 verschiedene Typen)	846	7,80	6,70

Tabelle A.2: Die unterschiedlichen Typen von Webservern im WiN-Web

bzw. *Netscape-FastTrack* sind mit einem Anteil von lediglich 2,5% sehr selten. Dies gilt auch für *NCSA*, *CERN* und *Roxen*. Der *NCSA-HTTPd*, dessen Entwicklung und Pflege bereits 1996 eingestellt wurde, ist der Vorgänger des *Apache*.[6] Der *CERN*-Server ist der ursprünglich von der Gruppe um Tim Berners-Lee entwickelte Daemon.[7] Die Software *Roxen* erfreut sich aufgrund mehrerer Spezialfunktionen bei einigen Anwendern großer Beliebtheit, wird jedoch nur in Einzelfällen eingesetzt.[8] Die letzte Gruppe umfasst ca. 130 weitere Typen, mit denen etwa 7,8% aller Webserver bestückt sind. Zum einen wird also viel Open-Source-Software eingesetzt, was aufgrund der meist nur eingeschränkt zur Verfügung stehenden finanziellen Mittel in der Untersuchungsdomäne zu erwarten ist. Weiterhin arbeiten zahlreiche Webserver mit veralteten Software-Paketen (dies betrifft insbesondere die *NCSA*- und *CERN*-Server), die Sicherheitslücken aufweisen und somit für Angriffe anfällig sind.

Die Tabelle enthält Angaben zum weltweiten Marktanteil, die der *Netcraft*-Studie vom Oktober 2001 entnommen wurden.[9] Die Daten zeigen, dass im WiN-Web gerade der *Apache* überdurchschnittlich häufig eingesetzt wird, wohingegen der *IIS* weniger Marktanteile hat als im gesamten Web. Turau (1998a) stellt ebenfalls verschiedene Daten zur Verteilung der

[6] Informationen zum *NCSA-HTTPd* sind unter `http://hoohoo.ncsa.uiuc.edu` verfügbar. Die letzte stabile Version 1.5.2a dieses Servers stammt vom 29.09.1996.

[7] Die am 15.07.1996 veröffentlichte letzte stabile Version 3.0A des *CERN httpd* wird noch immer durch das *World Wide Web Consortium* vertrieben, hat aber nur noch historischen Wert: `http://www.w3.org/Daemon/`.

[8] *Roxen* selbst ist ein Open-Source-Webserver, der unter `http://www.roxen.com/products/webserver/` erhältlich ist. Um diesen Server herum gruppieren sich einige kommerzielle Produkte wie z. B. *Roxen CMS*.

[9] Das Korpus wurde zwischen dem 16.01.2001 und dem 07.09.2002 gesammelt. Aus diesem Grund wurden für den Vergleich die *Netcraft*-Daten ausgewählt, die nach der Hälfte der Datensammlung verfügbar waren (`http://www.netcraft.com/survey/Reports/200110/byserver/index.html`).

Webserver-Software dar, gibt jedoch keine Informationen zum Bereich „deedu“ an. Seine Datengrundlage (aus dem Jahr 1997) zeigt ein noch breiteres Spektrum: *Apache* (35,6%), *Netscape* (17,0%), *IIS* (9,8%), *NCSA* (10,1%), *CERN* (3,3%) und *WebStar* (2,3%). Boldi et al. (2002) berichten für das afrikanische Web die Verteilung *IIS* (56,10%), *Apache* (37,95%), *Netscape Enterprise* (1,50%) und *Lotus Domino* (1,04%).

A.3.4 Die häufigsten Medientypen

Tabelle A.3 zeigt die häufigsten 30 Medientypen, die innerhalb der 10 840 verfügbaren Webserver (vgl. Abschnitt A.3.3) gefunden wurden. Sie werden mit den Angaben von Heydon und Najork (1999) verglichen, die den *Crawler Mercator* darstellen (vgl. Abschnitt 7.2.1) und Statistiken präsentieren, die aus einem *Crawl* ermittelt wurden, der im Mai 1999 durchgeführt wurde und 77 357 381 HTTP-Response-Header umfasst. Heydon und Najork stellen auch die häufigsten acht Medientypen dar, von denen zwei (`audio/x-pn-realaudio` und `application/zip`, jeweils 0,4%) nicht unter den ersten 30 Medientypen im WiN-Web enthalten sind. Die beiden letzten Spalten der Tabelle beziehen sich auf diejenigen Webserver, die den entsprechenden Medientyp anbieten.

Erstaunlich gering – insbesondere im Vergleich zu den Angaben, die *Mercator* erhoben hat – ist der Anteil von HTML-Dokumenten im WiN-Web: Nur etwa die Hälfte aller per HTTP angebotenen Dateien sind tatsächlich Webseiten, die jedoch auf fast allen erreichbaren Webservern vorhanden sind. Die Unterschiede bei den dominierenden Grafikformaten `image/gif` und `image/jpeg` im Verhältnis zu den von Heydon und Najork (1999) ermittelten Daten sind nur gering. Die Verteilung forschungsbezogener Informationen manifestiert sich im WiN-Web insbesondere in der Publikation von Artikeln und technischen Berichten in Form von Postscript- und PDF-Dateien, wie die prozentual zwar sehr niedrigen, absolut jedoch hohen Werte für die Medientypen `application/pdf` und `application/postscript` belegen; in den von *Mercator* erhobenen Daten sind diese beiden Medientypen deutlich seltener vertreten, was auf eine geringere Priorität dieser Formate für den Datenaustausch schließen lässt.[10] Die hohen Angaben zur Anzahl anbietender Webserver für die Typen `text/css` (28,5%) und `application/javascript` (12,5%) fallen im Vergleich zu den sehr geringen Werten ihrer direkten Nachbarn auf. Dies kann durch den Umstand erklärt werden, dass CSS- und *JavaScript*-Dateien – ähnlich wie Logo-Grafiken – in sehr vielen HTML-Dokumenten referenziert werden, um Layout-Informationen in einer einzelnen Datei zentral pflegen zu können. Auf der gegenüberliegenden Seite des Spektrums befinden sich die jeweiligen Werte zur Anzahl anbietender Webserver für die mit 43 673 bzw. 26 380 HTTP-Response-Headern zahlreich vertretenen Medientypen `application/x-dvi` und `audio/basic`, die nur auf lediglich 588 bzw. 288 Webservern zum Download bereitgestellt werden. Hierbei handelt es sich vermutlich zu großen Teilen um umfangreiche Angebote zu spezialisierten Themen.

A.3.5 Datenumfang der Webserver

Der Umfang der Datenbestände, die von einzelnen Webservern angeboten werden, wurde bereits in Abschnitt A.3.1) betrachtet. Tabelle A.4 enthält eine Aufstellung dieser Datenbe-

[10] Heydon und Najork (1999) machen keine Angaben zur Verteilung der ca. 77 Millionen HTTP-Response-Header auf bestimmte inhaltliche Bereiche oder *top-level*-Domänen.

	Medientyp	Anzahl	Prozent	Mercator	Anzahl Webserver	Prozent
1.	text/html	8 465 201	52,2656	69,2	10 333	95,323
2.	image/gif	3 646 789	22,5159	17,9	8 249	76,098
3.	image/jpeg	2 478 821	15,3047	8,1	6 939	64,013
4.	application/pdf	437 151	2,6990	0,9	4 944	45,609
5.	text/plain	291 960	1,8026	1,5	5 035	46,448
6.	application/postscript	195 773	1,2087	0,3	2 348	21,661
7.	image/png	137 960	0,8518	—	1 459	13,459
8.	application/x-dvi	43 673	0,2696	—	588	5,424
9.	application/octet-stream	36 953	0,2281	—	687	6,338
10.	text/css	36 257	0,2239	—	3 089	28,496
11.	audio/basic	26 380	0,1629	—	288	2,657
12.	text/xml	26 129	0,1613	—	771	7,113
13.	image/x-xbitmap	17 624	0,1088	—	544	5,018
14.	application/x-tex	16 718	0,1032	—	601	5,544
15.	text/rtf	15 161	0,0936	—	390	3,598
16.	audio/x-wav	13 181	0,0813	—	650	5,996
17.	video/mpeg	12 477	0,0770	—	862	7,952
18.	image/tiff	8 640	0,0533	—	509	4,696
19.	application/x-javascript	8 512	0,0526	—	1 356	12,509
20.	audio/midi	7 569	0,0467	—	314	2,897
21.	video/quicktime	5 005	0,0309	—	515	4,751
22.	video/x-msvideo	4 174	0,0258	—	542	5,000
23.	audio/x-aiff	3 993	0,0246	—	97	0,895
24.	image/x-portable-pixmap	3 377	0,0209	—	38	0,351
25.	application/msword	2 294	0,0141	—	79	0,729
26.	chemical/x-pdb	1 480	0,0091	—	86	0,793
27.	text/sgml	1 202	0,0074	—	37	0,341
28.	message/rfc822	1 106	0,0068	—	64	0,590
29.	audio/mpeg	1 022	0,0063	—	55	0,507
30.	application/x-perl	1 007	0,0062	—	82	0,756

Tabelle A.3: Die 30 häufigsten Medientypen

stände, die sich einerseits auf beliebige Dateien (*/*), andererseits auf die vier häufigsten Medientypen bezieht. Die in der letzten Tabellenzeile angegebene Summe der jeweiligen Webserver ist bezüglich der vier Medientypen nicht identisch mit den Angaben aus Tabelle A.3, weil lediglich diejenigen Webserver in die Aufstellung eingegangen sind, die mindestens zwei Dateien eines Typs anbieten, um experimentelle oder noch nicht mit Dokumenten bestückte Webserver auszuschließen. Die in Tabelle A.4 dargestellten Daten reflektieren unmittelbar die Frequenzen der vier Medientypen. Sowohl für alle per HTTP erreichbaren Dateien als auch für die Medientypen gilt, dass Webserver, die mehr als 50 000 Dateien anbieten, im WiN-Web äußerst selten sind. Die Durchschnittskategorien fallen deutlich nach rechts ab, wobei die Mediane zeigen, dass in allen Datenblöcken eine extreme Streuung vorliegt. Sehr markant ist die hohe Anzahl von 4 880 Webservern, die zwischen zwei und 10 Dateien anbieten. Hierbei könnte es sich etwa um Testinstallationen oder um Informationsangebote handeln, die zum größten Teil zugangsgeschützt sind.

A.3.6 Zur Aktualität der angebotenen Informationen

Viele Informationsangebote im WWW werden mehrmals täglich aktualisiert, z. B. Online-Zeitungen oder Nachrichtenportale. Dieser Abschnitt geht auf die Aktualität, d. h. das Alter,

Anzahl	*/*	text/html	image/gif	image/jpeg	application/pdf
⩾ 200 000	3	—	—	—	—
⩾ 150 000	1	1	—	—	—
⩾ 100 000	7	2	—	—	—
⩾ 90 000	5	1	—	—	—
⩾ 80 000	6	2	—	1	—
⩾ 70 000	2	1	—	—	—
⩾ 60 000	5	3	—	—	—
⩾ 50 000	12	4	—	1	—
⩾ 40 000	11	7	1	2	—
⩾ 30 000	40	12	6	1	—
⩾ 20 000	61	37	10	6	—
⩾ 10 000	149	98	37	20	—
⩾ 9 000	45	16	9	9	1
⩾ 8 000	58	21	14	11	—
⩾ 7 000	52	33	9	13	—
⩾ 6 000	75	40	20	12	1
⩾ 5 000	109	45	31	15	—
⩾ 4 000	129	87	37	18	—
⩾ 3 000	205	132	71	39	6
⩾ 2 000	313	202	129	70	14
⩾ 1 000	Ø 783	488	317	214	50
⩾ 900	148	100	61	36	10
⩾ 800	167	Ø 113	93	50	14
⩾ 700	182	149	90	72	15
⩾ 600	198	190	134	94	18
⩾ 500	270	192	142	104	28
⩾ 400	342	270	Ø 181	157	50
⩾ 300	449	396	267	Ø 186	95
⩾ 200	678	592	428	343	150
⩾ 100	1 076	1 127	842	690	430
⩾ 90	160	182	165	124	74
⩾ 80	176	188	173	126	Ø 90
⩾ 70	199	201	198	148	103
⩾ 60	207	(M = 66) 260	211	181	122
⩾ 50	241	296	259	213	160
⩾ 40	312	351	(M = 51) 353	273	216
⩾ 30	343	405	456	(M = 42) 361	306
⩾ 20	(M = 45) 508	528	601	519	361
⩾ 10	769	883	885	699	(M = 23) 641
⩾ 2	4 880	2 405	1 527	1 497	1 391
Σ	13 376	10 059	7 801	6 304	4 345

Tabelle A.4: Aufstellung des Umfangs der Webserver (angeordnet nach der jeweiligen Anzahl der Dateien, die während der Datensammlung auf den erfassten Webservern per HTTP verfügbar waren; Ø: Durchschnittskategorie; M: Median)

derjenigen Dateien ein, deren HTTP-Response-Header das Feld `Last-Modified` enthält. Die Darstellung bezieht sich wie schon im vorherigen Abschnitt, einerseits auf alle Dateien, andererseits auf die vier häufigsten Medientypen. Viele Webserver versenden fehlerhafte Zeitstempel (vgl. Abschnitt 7.2.4), die sich sowohl auf das Datum und die Zeitangabe des Moments beziehen, in dem eine Datei versendet wurde (HTTP-Response-Header-Feld `Date`), als auch den Inhalt des `Last-Modified`-Feldes betreffen, das den Zeitstempel der letzten Änderung einer Datei markiert. Daher hat sich die Erhebung der Daten zur durchschnittlichen Akualität aufwändig gestaltet, weil in den zugehörigen SQL-Abfragen die entsprechenden Fehler antizipiert und ausgeschlossen werden mussten. Es wurden lediglich diejenigen Datensätze berücksichtigt, die den folgenden Beschränkungen genügen:

1. Es liegen Werte für `http_header.date` und `http_header.last_modified` vor.
2. `http_header.date` liegt nach dem Zeitpunkt des Crawl-Beginns.
3. `http_header.date` liegt vor dem Zeitpunkt des Crawl-Endes.
4. Die beiden Angaben besitzen nicht den Wert `0000-00-00 00:00:00`.
5. Das Jahr in `http_header.date` ist größer als `2000` und kleiner als `2003`.
6. Das Jahr in `http_header.last_modified` ist größer als `1990` und kleiner als `2003`.

Die Darstellung der Daten erfolgt mit Hilfe der Abbildungen A.1 bis A.3 sowie Tabelle A.5, die grundlegende statistische Informationen sowie den prozentualen Anteil derjenigen Dateien enthält, die jünger als einen bzw. größer als 10 Tage sind. Die Balkendiagramme decken unterschiedliche Granularitäten ab: Abbildung A.1 enthält die Dateien, die zwischen 11 und 210 Tagen alt sind, wobei die einzelnen Kategorien jeweils in Schritten von 20 Tagen realisiert sind. Abbildung A.2 besitzt ebenfalls diese Darstellungsform und benutzt Intervalle von 50 Tagen. Abbildung A.3 schließlich zeigt den Anteil derjenigen Dateien, die zwischen 700 und 1700 Tagen alt sind und geht in Intervallen von 100 Tagen vor. Tabelle A.5 zeigt bereits die wesentlichen Charakteristika bezüglich des Alters der vier wichtigsten Medientypen: Ein HTML-Dokument ist im Durchschnitt 594,11 Tage alt, was ca. 20 Monaten entspricht.[11] Mehr als 95% aller Webseiten sind älter als 10 Tage, beinahe 70% sind sogar älter als 210 und fast 35% älter als 700 Tage. Es wird davon ausgegangen, dass alle Webserver in dem HTTP-Response-Header-Feld `Date` einen korrekten Zeitstempel liefern, der tatsächlich das zum Zeitpunkt der Datensammlung aktuelle Datum und die Uhrzeit beinhaltet.

Die Verteilung bezüglich der Medientypen ist gleichmäßig: GIF-Dateien sind mit durchschnittlich 772 Tagen deutlich älter als HTML-Dokumente. Da Grafikdateien vermutlich weniger häufig modifiziert oder aktualisiert werden müssen, erscheint dieser Wert plausibel. Unklar ist hingegen, weshalb JPEG-Dateien mit einem Alter von durchschnittlich 488 Tagen deutlich jünger sind als die durchschnittliche Webseite. Einen noch geringeren Wert besitzen PDF-Dateien, die im Schnitt etwa ein Jahr alt sind. Insgesamt widersprechen diese Daten der Annahme, das Durchschnittsalter einer Webseite betrage ca. 50 Tage (vgl. etwa Pitkow, 1998) – nur ein Bruchteil der im WiN-Web verfügbaren HTML-Dokumente kann als „tagesaktuell" gelten (vgl. hierzu auch Koehler, 2002).

[11] Douglis et al. (1997) berichten ein Durchschnittsalter von ca. 1,8 Monaten. Dieser Wert bezieht sich auf HTTP-Anfragen, die in den Intranets der Firmen *AT&T* und *Digital* protokolliert wurden.

	/	text/html	image/gif	image/jpeg	application/pdf
Datensätze	12 536 237	5 617 932	3 375 430	2 260 644	413 724
Minimum	0	0	0	0	0
Maximum	4 199	3 832	4 022	3 610	2 688
Durchschnitt	623,62	594,11	772,37	487,77	374,83
Median	442	415	624	351	250
Modus	0	0	294	222	616
$<$ 1 Tag	0,682	1,029	0,368	0,386	0,266
1 Tag	0,436	0,609	0,233	0,334	0,490
2 Tage	0,346	0,421	0,383	0,177	0,313
3 Tage	0,279	0,398	0,173	0,198	0,309
4 Tage	0,255	0,344	0,120	0,267	0,279
5 Tage	0,190	0,247	0,097	0,171	0,247
6 Tage	0,255	0,385	0,108	0,184	0,321
7 Tage	0,256	0,375	0,124	0,191	0,343
8 Tage	0,294	0,384	0,210	0,235	0,399
9 Tage	0,295	0,334	0,314	0,228	0,241
10 Tage	0,255	0,272	0,230	0,284	0,329
$\geqslant$ 10 Tage	96,458	95,201	97,643	97,344	96,462

Tabelle A.5: Statistische Informationen über das Alter aller per HTTP verfügbaren Dateien sowie der häufigsten vier Medientypen (vgl. Abbildungen A.1 bis A.3)

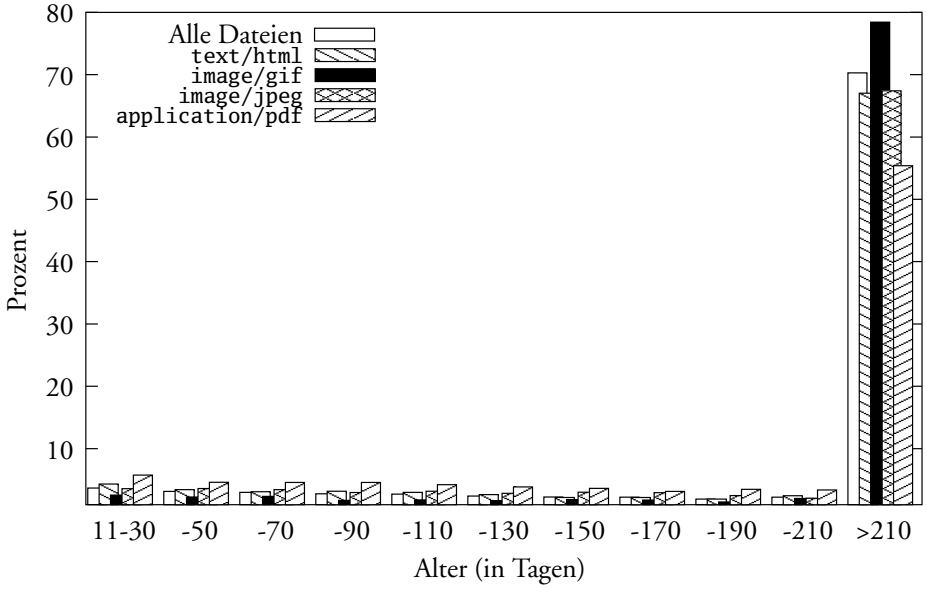


Abbildung A.1: Alter aller per HTTP verfügbaren Dateien sowie der vier häufigsten Medientypen (vgl. Abbildungen A.2 und A.3 sowie Tabelle A.5)

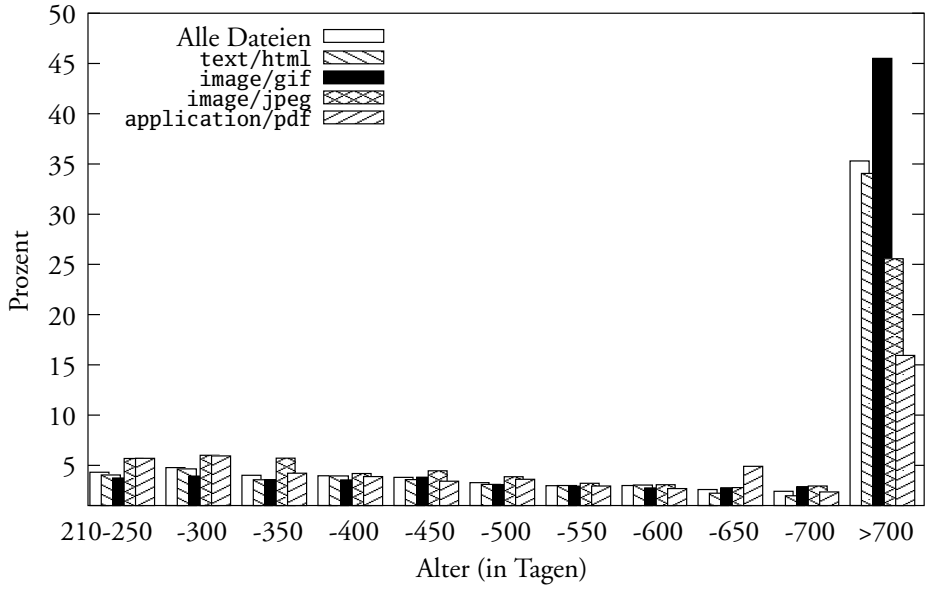


Abbildung A.2: Alter aller per HTTP verfügbaren Dateien sowie der vier häufigsten Medientypen (Fortsetzung, vgl. Abbildung A.1 und A.3 sowie Tabelle A.5)

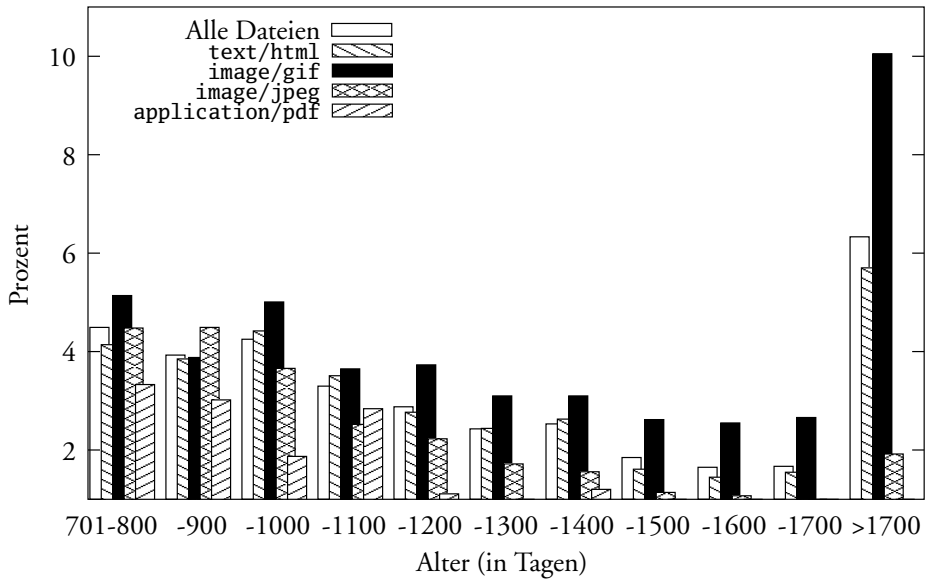


Abbildung A.3: Alter aller per HTTP verfügbaren Dateien sowie der vier häufigsten Medientypen (Fortsetzung, vgl. Abbildung A.1 und A.2 sowie Tabelle A.5)

A.3.7 Verwendung von Cookies

HTTP ist ein zustandsloses Protokoll, weshalb gerade in E-Commerce-Anwendungen häufig Cookies eingesetzt werden. Diese erlauben es, Daten innerhalb des Browsers zu speichern, was üblicherweise in einer speziellen Datei geschieht, die der Benutzer in aktuellen Browsern mit integrierten Editoren manipulieren kann. Ein Webserver kann einen Cookie mit Hilfe des HTTP-Response-Header-Feldes `Set-Cookie` setzen, das dem Cookie einen Namen gibt und ein Ablaufdatum enthält (RFC 2964, RFC 2965). Cookies können z. B. durch CGI-Skripte, *JavaScript*-Anwendungen oder *PHP*-Applikationen erzeugt werden. In der Korpusdatenbank sind insgesamt 81 136 Vorkommen verzeichnet, was einem Anteil von 0,5% entspricht. Sie stammen von insgesamt 593 Webservern aus 85 verschiedenen Universitäten.

Turau (1998a) untersucht die in seiner Stichprobe enthaltenen `Set-Cookie`-Header und berichtet einen Wert von 17,9%, der jedoch zu einem Großteil auf Webserver der *top-level*-Domäne `.com` zurückzuführen ist: in 40,4% aller HTTP-Rückmeldungen aus dieser Domäne war ein `Set-Cookie`-Header enthalten. Für den Bereich „deedu“ (wissenschaftliche Webserver innerhalb der Domäne `.de`) gibt Turau einen Wert von 0,2% an, d. h. Cookies werden im WiN-Web in den Jahren 2002/2003 in der Tat etwas häufiger eingesetzt. Dieser Wert ist durch die Tatsache zu erklären, dass Cookies innerhalb wissenschaftlicher Informationsangebote nur ein sehr begrenztes Anwendungsspektrum besitzen.

A.3.8 Analyse der HTTP-Response-Header

Die meisten Bestandteile des HTTP-Response-Headers wurden bereits in den vorherigen Abschnitten dargestellt, weshalb im Folgenden die verbleibenden Informationen zusammengefasst werden. Eine für HTTP entscheidende Information stellt der Status-Code dar (vgl. Abschnitt 7.2.4). Tabelle A.6 zeigt die 21 verschiedenen Status-Codes, die während der Datensammlung gefunden wurden. Auffällig ist zunächst, dass nicht alle in RFC 2616 definierten Status-Codes in der Liste enthalten sind, was einerseits dadurch zu erklären ist, dass einige Status-Codes den Client nur temporär erreichen und somit per definitionem nicht in der Liste enthalten sein können, andererseits einige Status-Codes zwar definiert, jedoch „reserved for future use“ sind (z. B. 402, „Payment Required“). Weiterhin sind in 214 820 Response-Headern keine Status-Codes verfügbar, was ein Indiz dafür ist, dass der *Crawler* zwar einen HTTP-Request an einen (vermeintlichen) Server abgesetzt hat, dieser jedoch nicht geantwortet hat; in 1,33% aller Anfragen lag also ein Time-Out vor. Heydon und Najork (1999) berichten einen prozentualen Anteil von 2,3%.

Die erfolgreiche Übertragung einer Datei per HTTP wird durch den Status-Code 200 markiert (91,1%). Im Vergleich zu den Daten, die von Heydon und Najork (1999) berichtet werden (87,03%), legt dieser Wert den Schluss nahe, dass Hyperlinks im WiN-Web etwas sorgfältiger gepflegt werden als dies im Durchschnitt der Fall ist. Der Wert 404 wird vom Webserver geliefert, wenn zwar eine gültige HTTP-Anfrage empfangen wurde, an der spezifizierten URL jedoch kein Dokument vorhanden ist. Der Vergleich mit dem Anteil des *Mercator-Crawlers* – 7,4% – belegt die genannte These, denn im WiN-Web münden lediglich 5,5% aller Hyperlinks in einem „Not Found“-Fehler. Turau (1998a) gibt ebenfalls einen prozentualen Anteil von 9,25% an. Der Zustand „Moved Permanently“ wird durch den Status-Code 301 signalisiert, der üblicherweise durch einen Redirect, also eine Weiterleitung erzeugt

Status-Code	Beschreibung (*reason phrase*)	Anzahl	Prozent	Mercator
2xx – Success				
200	OK	14 754 739	91,098250	87,03
204	No Content	2	0,000012	—
206	Partial Content	88	0,000543	—
3xx – Redirection				
300	Multiple Choices	3 114	0,019226	—
301	Moved Permanently	139 778	0,863013	1,12
302	Found	82 114	0,506986	3,33
303	See Other	49	0,000303	—
4xx – Client Error				
400	Bad Request	3 993	0,024653	0,09
401	Unauthorized	36 062	0,222653	0,30
403	Forbidden	66 771	0,412255	0,43
404	Not Found	891 585	5,504796	7,43
406	Not Acceptable	—	—	0,11
407	Proxy Authentication Required	3	0,000019	—
410	Gone	14	0,000086	—
414	Request-URI Too Large	4	0,000025	—
416	Requested Range not Satisfiable	22	0,000136	—
5xx – Server Error				
500	Internal Server Error	2 234	0,013793	0,11
501	Not Implemented	2	0,000012	—
502	Bad Gateway	16	0,000099	—
503	Service Unavailable	1 101	0,006798	—
504	Gateway Time-out	1	0,000006	—
k. A.	*Status-Code nicht verfügbar*	214 820	1,326335	—

Tabelle A.6: Die in der Korpusdatenbank enthaltenen HTTP-Status-Codes

HTTP-Response-Header-Feld	Vorkommen	Prozent	Eindeutige Vorkommen
Content-Length	13 696 993	84,57	439 355
Content-Type	15 975 557	98,64	315
Content-Encoding	6 430	0,04	11
Content-Language	241 203	1,49	26
Content-Location	50 972	0,31	11 509
Location	224 424	1,39	184 011
Date	15 766 620	97,35	3 765 602
Expires	203 947	1,26	81 712
Last-Modified	13 581 691	83,86	5 102 824
WWW-Authenticate	36 078	0,22	—
Cache-Control	144 167	0,89	—
Content-MD5	43 368	0,27	—
Pragma	8 440	0,05	—
Set-Cookie	81 136	0,50	—

Tabelle A.7: Die Anzahl und eindeutige Vorkommen der in der Korpusdatenbank gespeicherten HTTP-Response-Header-Felder (vgl. Tabelle 7.2, S. 334)

wird, die unterschiedliche technische Ursachen haben kann. Alle weiteren Status-Codes kommen im Vergleich zu den bereits erläuterten relativ selten vor.

Tabelle A.7 zeigt abschließend die Anzahl, den Anteil sowie die Menge eindeutiger Werte der in der Datenbank enthaltenen HTTP-Response-Header (vgl. Tabelle 7.2, S. 334).[12] Es fällt auf, dass die Header `Content-Type`, `Date`, `Content-Length` und `Last-Modified` in fast allen Headern enthalten sind, wobei die verbleibenden Felder nur sehr selten in Erscheinung treten. Das Feld `Date` enthält einen Zeitstempel, der Tag und Uhrzeit einer HTTP-Anfrage umfasst. Innerhalb der 15,77 Millionen HTTP-Anfragen sind lediglich 3,77 Millionen eindeutige Zeitstempel enthalten, weil „Sekunden“ als kleinste Zeiteinheit notiert wird und der *Crawler* üblicherweise deutlich mehr als eine Datei pro Sekunde verarbeitet. Im Vergleich zu dieser Anzahl eindeutiger Werte liegt die Angabe beim `Last-Modified`-Feld bei geringerer Anzahl Vorkommen deutlich höher, was ein Indiz für eine sehr große Varianz bezüglich des Datums der letzten Änderung ist (vgl. Abschnitt A.3.6). Bei den wenigen eindeutigen Vorkommen des Feldes `Content-Encoding` handelt es sich zu einem großen Teil um Dateien, die – vermutlich *on the fly* – mittels *gzip* komprimiert wurden (`application/x-gzip`, 39,8%), woraufhin der `Content-Encoding`-Medientyp `text/html` mit 39,4% folgt.[13] Die 26 unterschiedlichen Werte des Feldes `Content-Language` werden angeführt von „de“, das einen prozentualen Anteil von 83,4% hat, gefolgt von „en“ mit 15,6%.

A.4 Charakteristika der HTML-Dokumente

Dieser Abschnitt geht auf Merkmale der HTML-Dokumente ein, die sich im Dateisystem des Korpusservers befinden.[14] Zunächst analysiert Abschnitt A.4.1 die Dateigrößen. Daraufhin stellt Abschnitt A.4.2 dar, welche HTML-Elemente und -Attribute verwendet werden. Anschließend thematisiert Abschnitt A.4.3 zahlreiche Eigenschaften von Hyperlinks, die z. B. die Vernetzung des WiN-Web mit anderen *top-level*-Domänen dokumentieren. In Abschnitt A.4.4 wird der Einsatz und die Verbreitung genereller Metadaten sowie spezifischer Metadatenschemata diskutiert. Abschnitt A.4.5 beschäftigt sich mit der Verwendung multimedialer Dateitypen, d. h. in HTML-Dokumente eingebettete Grafiken, Audio- und Video-Dateien. Abschnitt A.4.6 behandelt Vorkommen interaktiver Elemente und Clientseitiger Anwendungen, die über das Funktionsspektrum von HTML hinausgehen. Daraufhin thematisiert Abschnitt A.4.7, wie hoch der Anteil von Dokumenten ist, die automatisch bzw. semiautomatisch mit Hilfe von Konvertierungswerkzeugen und HTML-Editoren erstellt worden sind. Der vorletzte Abschnitt geht auf unterschiedliche HTML-Versionen ein, und Abschnitt A.4.9 diskutiert abschließend die Verwendung von *Cascading Style Sheets*.

[12] Die letzten fünf Felder besitzen keine eindeutigen Werte, weil ein Vorkommen dieser Felder in einem Response-Header nur mit einem Boole'schen Wert markiert wird.

[13] Dieser hohe Anteil an `text/html`-Encodings ist auf fehlerhaft konfigurierte Webserver zurückzuführen.

[14] Von den 3 956 692 Dateien, die aufgrund ihres Medientyps `text/html` heruntergeladen und in das Korpus aufgenommen wurden, fließen 3 899 341 Dateien in die im Folgenden dargestellten Analysen ein. Die verbleibenden ca. 60 000 Dateien sind entweder leer (Dateigröße von 0 Byte), oder es handelt sich um Verzeichnisse.

A.4.1 Zum durchschnittlichen Umfang der Dokumente

Die Größe von HTML-Dokumenten ist – bezogen auf die in Abschnitt A.6 dargestellten Arbeiten – relativ konstant: Bray (1996) gibt einen Wert von 6 518 Bytes (Median: 2 021 Bytes) an, Woodruff et al. (1996) berichten eine Größe von 4,4 Kilobyte (Median: 2,0 Kilobyte) und laut Turau (1998a) umfasst eine Webseite eine Größe von 7,5 Kilobyte. Die im Korpus enthaltenen Dokumente weichen von diesen Angaben nicht ab: Die durchschnittliche Webseite hat eine Dateigröße von 7 024 Bytes (Median: 3 408 Bytes). Überraschend hoch sind die Werte im afrikanischen Web (Boldi et al., 2002), in dem ein HTML-Dokument eine Größe von durchschnittlich 12 920 Bytes besitzt (Median: 6 935).

Abbildung A.4 stellt die Verteilung der Dateigrößen bezüglich des Gesamtbestandes dar. Die Grafik zeigt, dass ca. 91% aller HTML-Dokumente eine Größe zwischen einem Byte und 16 Kilobyte besitzen, ca. 26% aller Dokumente besitzen eine Größe zwischen zwei und vier Kilobyte. Auf der anderen Seite des Spektrums befinden sich nur sehr wenige Dateien: Lediglich 7 190 Dateien (0,18%) besitzen eine Größe zwischen 128 und 256 Kilobyte, und nur 1 757 Dokumente (0,05%) haben einen Umfang zwischen 256 und 512 Kilobyte. Dieser offensichtliche Mangel an sehr großen Dateien im WiN-Web überrascht, da man in dieser Domäne sehr viele umfangreiche wissenschaftliche Veröffentlichungen als HTML-Dokumente erwarten könnte. Die Daten lassen zwei Schlüsse zu: Entweder wird eine Publikation der Forschungsarbeiten in Formaten wie PDF und Postscript präferiert, oder diese Arbeiten werden tatsächlich in HTML-Versionen veröffentlicht, dabei jedoch pro Abschnitt oder Kapitel in separate Dokumente aufgetrennt – z. B. nimmt das Konvertierungswerkzeug *LaTeX2HTML* in der Grundeinstellung eine derartige Aufteilung vor (vgl. Abschnitt A.4.7) – die Aufstellung der häufigsten Medientypen (vgl. Tabelle A.3) deutet den erstgenannten Schluss an. Bray (1996), Turau (1998a) und Heydon und Najork (1999) geben ebenfalls Diagramme in der Art von Abbildung A.4 an: Bray zeigt, dass fast 60% aller Dokumente eine Größe zwischen einem und zehn Kilobyte besitzen, Dateien über 100 Kilobyte besitzen einen Anteil von ca. 1%. Die von Turau berichtete Verteilung ist bezüglich des prozentualen Anteils nahezu identisch mit den in Abbildung A.4 dargestellten Werten; die Unterschiede betreffen minimal höhere Angaben für die Größenintervalle 1–512 Bytes, 512–1024 Bytes und 1024–2048 Bytes bei Turau (1998a). Gleiches gilt für die in Heydon und Najork (1999) präsentierten Daten, wobei dort die Kategorien um zwei bzw. acht Kilobyte etwas häufiger vorkommen, die durchschnittliche Größe bleibt jedoch konstant bei einem Wert um vier Kilobyte. Interessanterweise bezieht die Darstellung von Heydon und Najork auch andere Dokumenttypen mit ein, z. B. GIF- und JPEG-Dateien (17,9% bzw. 8,1%).

Bray (1996) untersucht die Dokumente seiner Stichprobe bezüglich der durchschnittlichen Anzahl enthaltener Wörter und gibt einen Wert von 1 050 an, wobei ein Wort definiert ist als eine durchgehende Zeichenkette, die mit einem Buchstaben beginnt. Diese Information wurde für die HTML-Dokumente des Korpus ebenfalls erhoben und fällt mit einer durchschnittlichen Anzahl von 310 Wörtern (Median: 95, Modus: 5) deutlich geringer aus. Für diesen Zweck wurden die gleichen Tokenisierungsfunktionen eingesetzt, die auch zur Berechnung der Gesamtanzahl laufender Wortformen benutzt wurden (vgl. Abschnitt A.3). Der Unterschied des arithmetischen Mittels von mehr als 700 Wörtern könnte einerseits auf die komplexeren Tokenisierungsfunktionen zurückzuführen sein, da Bray auf der Grundla-

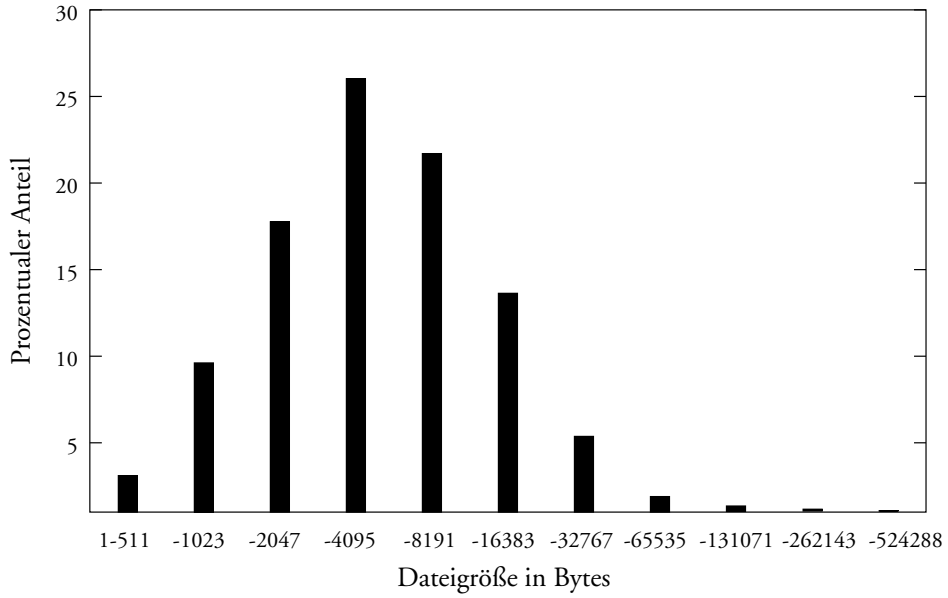


Abbildung A.4: Verteilung der Dateigrößen der HTML-Dokumente

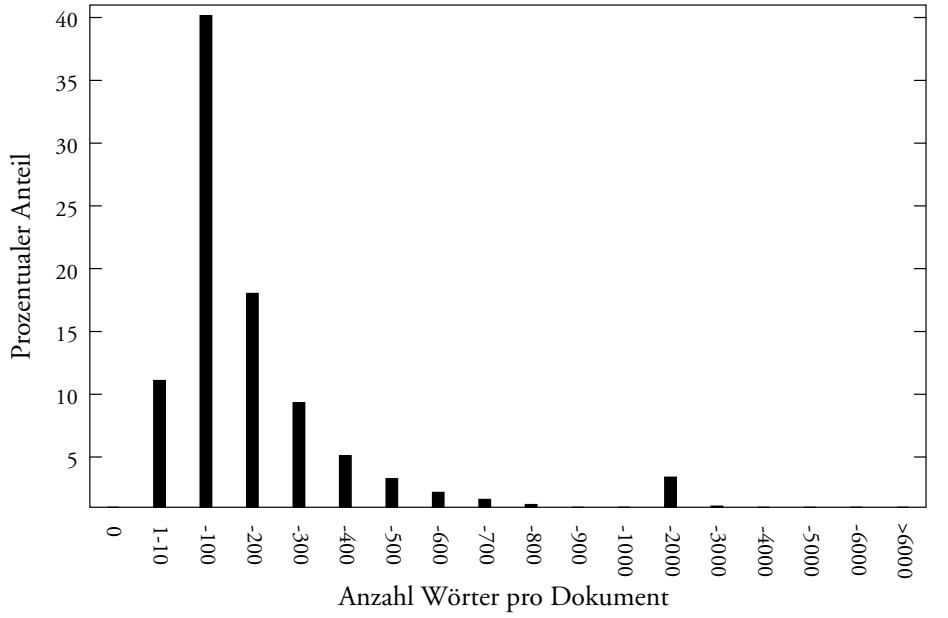


Abbildung A.5: Verteilung der in den HTML-Dokumenten enthaltenen Wörter

ge seiner Definition deutlich mehr Zeichenketten als gültiges Wort interpretiert, andererseits könnte er auch durch den Umstand begründet sein, dass die deutschsprachigen HTML-Dokumente des WiN-Webs schlicht weniger umfangreich sind und z. B. viele Informationen durch binäre Dateiformate wie eingebettete Grafiken transportiert werden.[15] Abbildung A.5 zeigt schließlich die Verteilung der Anzahl Wörter pro Dokument in Bezug zum gesamten Datenbestand. Mehr als 40% aller Dokumente enthalten zwischen 11 und 100 Wörtern, fast 70% aller Dokumente enthalten bis zu 200 Wörter. Auffällig ist der minimale Anstieg in der Kategorie 1001–2000 Wörter (132 651 Dokumente, 3,4%), der möglicherweise durch HTML-Versionen wissenschaftlicher Publikationen verursacht wird.

A.4.2 Benutzung von HTML-Elementen und -Attributen

Woodruff et al. (1996) und Turau (1998a) diskutieren den Einsatz unterschiedlicher HTML-Elemente und -Attribute, um die Verbreitung der zwischen 1995 und 1998 populären, proprietären Browser-Erweiterungen darzustellen. Woodruff et al. berichten, dass eine Webseite im Durchschnitt 71 Elemente und 29 Attribute enthält. Diesen Angaben stehen bezüglich des Korpus 120,57 Elemente (Median: 60, Max.: 29 532) und 236,04 Attribute (Median: 100, Max.: 53 260) gegenüber. Diese deutlich höheren Werte sind – insbesondere bei ähnlichen Durchschnittsangaben bezüglich der Dokumentgröße – auf den Umstand zurück zu führen, dass HTML mittlerweile ein ausgeprägteres Vokabular besitzt und viele Dokumente durch HTML-Editoren angefertigt werden, die umfangreiches Markup erzeugen.

HTML-Elemente

Zur Erhebung der standardisierten und unbekannten HTML-Elemente wurde ein *Perl*-Skript mit den in HTML 4.01 definierten 91 Element- und 119 Attributnamen ausgestattet. Woodruff et al. berichten, dass z. B. für `bdo`, `colgroup` und `noembed` keine Vorkommen gefunden werden konnten. Im Korpus werden *alle* in HTML 4.01 definierten Elemente und Attribute verwendet. Insgesamt wurden 465 830 647 standardisierte Elemente und 448 763 854 standardisierte Attribute gefunden, was einem Anteil von 99,11% bzw. 98,05% entspricht, d. h. proprietäre Elemente bzw. Attribute werden nur noch sehr selten eingesetzt.[16]

Die Tabellen A.8 und A.9 enthalten statistische Angaben für die 91 in HTML 4.01 definierten Elemente. Die Daten wurden nach dem prozentualen Anteil ihrer Vorkommen sortiert. Erwartungsgemäß finden sich an den ersten vier Positionen die Elemente `body`, `html`, `head` und `title`, jedoch fällt bezüglich des prozentualen Anteils auf, dass diese Elemente zwar mit durchschnittlich einem Vorkommen, aber nur in lediglich ca. 97% aller Dokumente enthalten sind. Es kann gefolgert werden, dass es sich bei etwa 3% aller Dokumente,

[15] Thelwall (2005, S. 613) berichtet für die in Fußnote 3 (S. 723) genannten Bereiche, dass die Prozentsätze von Dokumenten, die *keine* Wörter enthalten, (i) 28%, (ii) 27% und (iii) 26% betragen. Bezüglich der Datensammlung erläutert der Verfasser: „Words were extracted from the page title and body only. Words inside tags were ignored." (ebd., S. 612). Falls tatsächlich nur diejenigen Wörter gezählt wurden, die *unmittelbare* Textknoten von `title` und `body` sind, überraschen diese hohen Angaben nicht.

[16] Diese Angaben beziehen sich nur auf die Namen der Elemente und Attribute und nicht auf erlaubte oder unerlaubte Verwendungen von Attributen.

	Element	Vorkommen (absolut)	Prozent	Vorkommen (Anz. Dok.)	Prozent	Vorkommen (Ø pro Dok.)
1.	body	3 918 806	0,8413	3 818 995	97,94	1,03
2.	html	3 830 207	0,8222	3 794 403	97,31	1,01
3.	head	3 825 761	0,8213	3 785 097	97,07	1,01
4.	title	3 799 458	0,8156	3 759 217	96,41	1,01
5.	a	51 679 884	11,0941	3 403 402	87,28	15,18
6.	p	45 444 260	9,7555	3 162 107	81,09	14,37
7.	br	44 126 723	9,4727	2 840 764	72,85	15,53
8.	img	24 327 871	5,2225	2 603 103	66,76	9,35
9.	meta	8 105 345	1,7400	2 586 845	66,34	3,13
10.	table	8 663 097	1,8597	2 497 477	64,05	3,47
11.	td	79 263 995	17,0156	2 480 431	63,61	31,95
12.	tr	29 779 413	6,3928	2 343 908	60,11	12,70
13.	font	53 681 459	11,5238	2 130 568	54,64	25,20
14.	b	22 432 819	4,8157	2 072 720	53,16	10,82
15.	hr	4 948 029	1,0622	1 682 406	43,15	2,94
16.	center	4 595 846	0,9866	1 286 358	32,99	3,57
17.	h1	1 311 328	0,2815	1 088 436	27,91	1,20
18.	h2	2 151 384	0,4618	1 031 791	26,46	2,09
19.	ul	4 249 475	0,9122	1 009 614	25,89	4,21
20.	i	7 095 700	1,5232	992 238	25,45	7,15
21.	li	13 327 877	2,8611	989 574	25,38	13,47
22.	link	1 378 285	0,2959	867 060	22,24	1,59
23.	div	3 386 721	0,7270	756 531	19,40	4,48
24.	h3	2 096 445	0,4500	700 816	17,97	2,99
25.	script	953 725	0,2047	497 809	12,77	1,92
26.	strong	3 202 034	0,6874	482 033	12,36	6,64
27.	address	562 518	0,1208	481 772	12,36	1,17
28.	em	1 420 720	0,3050	279 287	7,16	5,09
29.	style	285 290	0,0612	272 667	6,99	1,05
30.	h4	952 923	0,2046	269 193	6,90	3,54
31.	u	1 298 962	0,2788	260 627	6,68	4,98
32.	pre	849 282	0,1823	260 253	6,67	3,26
33.	blockquote	928 535	0,1993	242 240	6,21	3,83
34.	small	1 528 991	0,3282	201 558	5,17	7,59
35.	span	6 705 197	1,4394	194 098	4,98	34,55
36.	th	1 478 920	0,3175	192 493	4,94	7,68
37.	dl	841 898	0,1807	182 494	4,68	4,61
38.	dd	2 349 605	0,5044	170 767	4,38	13,76
39.	form	211 479	0,0454	166 279	4,26	1,27
40.	map	269 190	0,0578	164 029	4,21	1,64
41.	area	916 831	0,1968	163 251	4,19	5,62
42.	dt	2 126 713	0,4565	162 109	4,16	13,12
43.	tt	2 675 247	0,5743	156 579	4,02	17,09
44.	input	1 099 767	0,2361	141 551	3,63	7,77
45.	base	142 266	0,0305	141 522	3,63	1,01
46.	ol	360 633	0,0774	135 705	3,48	2,66
47.	frameset	199 425	0,0428	132 151	3,39	1,51
48.	frame	346 630	0,0744	130 551	3,35	2,66
49.	big	494 360	0,1061	95 105	2,44	5,20
50.	sup	694 302	0,1490	94 940	2,43	7,31
51.	h5	483 825	0,1039	82 371	2,11	5,87
52.	noframes	79 276	0,0170	78 045	2,00	1,02
53.	code	776 489	0,1667	64 517	1,65	12,04
54.	sub	658 407	0,1413	61 108	1,57	10,77
55.	tbody	148 201	0,0318	50 951	1,31	2,91
56.	dir	366 019	0,0786	39 989	1,03	9,15
57.	caption	74 438	0,0160	38 702	0,99	1,92
58.	textarea	48 285	0,0104	37 413	0,96	1,29
59.	select	59 919	0,0129	32 759	0,84	1,83
60.	option	1 731 204	0,3716	32 470	0,83	53,32
61.	col	172 694	0,0371	30 406	0,78	5,68
62.	h6	62 110	0,0133	29 187	0,75	2,13
63.	basefont	38 130	0,0082	26 174	0,67	1,46
64.	applet	43 746	0,0094	24 751	0,63	1,77
65.	param	431 066	0,0925	24 339	0,62	17,71
66.	noscript	23 757	0,0051	19 871	0,51	1,20
67.	colgroup	38 042	0,0082	17 183	0,44	2,21
68.	menu	37 670	0,0081	14 629	0,38	2,58
69.	cite	56 747	0,0122	8 912	0,23	6,37
70.	thead	14 841	0,0032	8 873	0,23	1,67
71.	object	10 011	0,0021	7 173	0,18	1,40
72.	var	72 619	0,0156	5 077	0,13	14,30
73.	kbd	41 170	0,0088	4 630	0,12	8,89
74.	strike	8 860	0,0019	1 936	0,05	4,58
75.	samp	11 956	0,0026	1 340	0,03	8,92

Tabelle A.8: Verwendung der in HTML 4.01 definierten Elemente im Korpus (Teil 1)

	Element	Vorkommen (absolut)	Prozent	Vorkommen (Anz. Dok.)	Prozent	Vorkommen (Ø pro Dok.)
76.	iframe	1 207	0,0003	945	0,02	1,28
77.	dfn	4 015	0,0009	874	0,02	4,59
78.	tfoot	960	0,0002	845	0,02	1,14
79.	s	5 068	0,0011	841	0,02	6,03
80.	ins	3 691	0,0008	518	0,01	7,13
81.	acronym	2 974	0,0006	354	0,01	8,40
82.	label	1 501	0,0003	320	0,01	4,69
83.	q	637	0,0001	245	0,01	2,60
84.	button	383	0,0001	243	0,01	1,58
85.	abbr	1 509	0,0003	229	0,01	6,59
86.	del	1 888	0,0004	219	0,01	8,62
87.	fieldset	545	0,0001	216	0,01	2,52
88.	legend	510	0,0001	208	0,01	2,45
89.	optgroup	582	0,0001	148	0,00	3,93
90.	isindex	49	0,0000	49	0,00	1,00
91.	bdo	15	0,0000	15	0,00	1,00

Tabelle A.9: Verwendung der in HTML 4.01 definierten Elemente im Korpus (Teil 2)

die von Webservern mit dem Medientyp `text/html` verschickt wurden und auf diese Weise in das Korpus aufgenommen wurden, nicht um HTML-Dokumente handelt, denn diese vier Elemente sind in Dokumentinstanzen obligatorisch. An der fünften Position befindet sich `a` als erstes Inline-Element (vgl. Abschnitt A.4.3). Das Element `p` an Position sechs mit durchschnittlich 15 Vorkommen pro Dokument ist das erste Block-Element der Aufstellung, d. h. es wird global offenbar eine Strukturierung auf der Abschnittsebene präferiert.[17]

Da eine detaillierte Analyse den Rahmen dieses Anhangs sprengte, werden lediglich einige wesentliche Merkmale aufgezeigt: An den Positionen 10 bis 12 befinden sich `table`, `td` (*table data cell*) und `tr` (*table row*), die zur Annotation von Tabellen und auch zur Realisierung Pixel-genauer Layouts eingesetzt werden. Demnach existieren in etwa zwei Drittel aller HTML-Dokumente durchschnittlich drei bis vier Tabellen mit jeweils etwa vier Zeilen und drei Spalten. Da es sich bei dieser Menge von `table`-Vorkommen kaum um Tabellen im eigentlichen Sinne handeln wird, legen diese Daten den Schluss nahe, dass Tabellen auch im WiN-Web sehr häufig für Layout-Zwecke verwendet werden.

Die Elemente `b` (*bold text style*), und `i` (*italic text style*) gehören zu einer Gruppe von Elementen, deren Funktion die explizite Schriftumschaltung ist (die weiteren Elemente lauten `tt`, `big`, `small`, `strike`, `s` und `u`). Da bei diesen Elementen nicht die Annotation logischer oder semantischer Einheiten im Vordergrund steht, sondern lediglich das Aussehen eines Dokumentteils beeinflusst werden soll, werden sie als physikalisches Markup bezeichnet. Kursiv- und Fettdruck können auch mit Hilfe von `em` (*emphasis*) und `strong` (*strong emphasis*) erzielt werden – bei diesen Elementen handelt es sich um logisches Markup, weil eine Aussage über den Inhalt eines Textfragments getroffen wird. Der Standard schreibt zwar nicht vor, wie diese Elemente umgesetzt werden sollen, erwähnt jedoch, dass die meisten grafischem Browser sie in einem kursiven oder fetten Schriftschnitt darstellen. Der Vergleich von einerseits `b` und

[17] Boldi et al. (2002) geben die 15 häufigsten Elemente jeweils nach ihren (a) durchschnittlichen Vorkommen pro Dokument und ihrem (b) Anteil am gesamten Schnappschuss an. (a): `tr` (22,51), `a` (22,49), `font` (20,47), `br` (17,58), `img` (14,07), `p` (13,56), `b` (9,53), `span` (8,40), `table` (6,56), `option` (4,53), `div` (2,62), `o` (2,54), `input` (2,44), `li` (2,36) und `strong` (1,95). (b): `html` (99,41%), `head` (99,39%), `body` (99,38%), `title` (93,37%), `a` (85,47%), `table` (79,03%), `tr` (79,01%), `td` (78,94%), `img` (77,97%), `p` (75,46%), `br` (74,84%), `font` (73,70%), `b` (67,70%), `meta` (63,07%) und `div` (43,09%).

strong (53,16%, 12,36%), andererseits i und em (25,45%, 7,16%) zeigt, dass die Autoren von HTML-Dokumenten im WiN-Web offenbar physikalisches Markup bevorzugen.

Werden die Elemente zur Auszeichnung unterschiedlicher Ebenen von Überschriften betrachtet (h1 bis h6), wird deutlich, dass sie in exakt dieser Reihenfolge im Korpus enthalten sind (h1: 27,9%; h2: 26,5%; h3: 17,8%; h4: 6,9%; h5: 2,1%; h6: 0,8%). In diesem Zusammenhang sind auch die durchschnittlichen Werte der Vorkommen pro Dokument interessant (h1: 1,2; h2: 2,1; h3: 3,0; h4: 3,5; h5: 5,9; h6: 2,1), die – mit Ausnahme von h6 – konstant ansteigende Werte darstellen. Diese Daten deuten an, dass durchschnittlich ein einziger Dokumenttitel (h1) vorliegt, der von mehreren Unterabschnitten begleitet wird.

Woodruff et al. (1996) stellen die zehn meistbenutzten HTML-Elemente in Form zweier Balkendiagramme dar, die über den prozentualen Anteil von Vorkommen und die durchschnittliche Anzahl pro Dokument Auskunft geben: title (91%, ca. 1), a (86%, ca. 15), p (75%, ca. 9,5), hr (71%, ca. 2), body (69%, ca. 0,5), img (61%, ca. 4), head (59%, ca. 0,4), html (59%, ca. 0,4), h1 (57%, ca. 0,5) und br (52%, ca. 7). Die Unterschiede zu Tabelle A.8 sind sehr deutlich und liegen an der derzeit aktuellen Version 4.01 von HTML, die die Elemente html, head, title und body zwingend vorsieht. Im Jahr 1995 wurden die meisten Dokumente manuell erstellt, und da Dokumente auch ohne die genannten obligatorischen Elemente im Browser darstellbar waren, hat es sich bei vielen Nutzern eingebürgert, diese Elemente nicht zu verwenden. Heutzutage werden die meisten Webseiten mit Hilfe von HTML-Editoren erzeugt und gepflegt, die das zugrunde liegende HTML-Markup eines Dokuments häufig vollständig verkapseln und in den meisten Fällen standardkonform halten. Zudem hatten zum Zeitpunkt der Erhebung der von Woodruff et al. (1996) präsentierten Daten einige Elemente noch proprietären Charakter, z. B. table, das erstmals in HTML 3.2 (Raggett, 1997) standardisiert wurde. Auffällig ist weiterhin, dass hr mittlerweile seltener zur Strukturierung von Dokumenten eingesetzt wird, da – gerade durch die Benutzung von Tabellen – mittlerweile funktionale Alternativen existieren.

Es kann festgehalten werden, dass nur 27 Elemente in mehr als 10% aller Dokumente benutzt werden; das Vokabular von 91 Elementen wird nicht vollständig ausgeschöpft. Punktuelle Analysen belegen einen allgegenwärtigen *tag abuse*, weshalb sich maschinelle Strukturanalysen keinesfalls auf das bloße Vorkommen eines Elements verlassen können – der jeweilige Kontext muss ebenfalls einfließen, um ein Elementvorkommen präzise interpretieren zu können. Die Kontrastierung mit den von Woodruff et al. dargestellten Daten zeigt, dass in den Jahren 2001 und 2002 HTML-Dokumente – zumindest bezüglich der Elementfrequenzen – offenbar eher dem Standard entsprechen, als dies noch 1995 der Fall war.

HTML-Attribute

HTML 4.01 definiert 119 Attribute, die größtenteils für mehrere der 91 Elemente gelten; insgesamt existieren 1 529 Kombinationen, von denen 775 im Korpus benutzt werden (vgl. die Tabellen A.10 bis A.20, Woodruff et al., 1996, und Boldi et al., 2002). In ca. 87% aller Dokumente befinden sich durchschnittlich 13,5 Hyperlinks (vgl. Abschnitt A.4.3). Etwa 67% aller Dokumente enthalten im Schnitt 9,3 Bilder, die mittels <img src="..."> referenziert werden (vgl. Abschnitt A.4.5). Überraschend hoch, gerade im Vergleich zu Tabelle A.8, ist die Verwendung von Metadaten (vgl. Abschnitt A.4.4). In mehr als der Hälfte aller Dokumente

wird explizit eine Hintergrundfarbe über das Attribut bgcolor des Elements body gesetzt. Dies ist ein Indiz dafür, dass die Annotation von Dokumenten vornehmlich gestalterische und nicht strukturelle Aspekte betrifft.

Nicht erlaubte Element-Attribut-Kombinationen

Tabelle A.21 (S. 754) stellt diejenigen 40 Element-Attribut-Paare dar, deren Namen zwar dem Standard entsprechen, das Attribut darf jedoch *nicht* für das jeweilige Element benutzt werden. Insgesamt wurden 1 101 dieser fehlerhaften Kombinationen gefunden, die zu einem Großteil auf manuell annotierte HTML-Dokumente zurück zu führen sind, da standardkonforme Editoren keine fehlerhaften Annotationen erzeugen. Die Tabelle bestätigt diese These: An erster Stelle befindet sich das Attribut height, das in 7,7% aller Dokumente für table eingesetzt wurde. Dieses Element verfügt zwar über das Attribut width, mit dem für grafische Browser die Breite einer Tabelle voreingestellt werden kann, ihre Höhe kann jedoch nicht beeinflusst werden. Das Vorhandensein des Attributs width ist offenbar für viele Autoren ein Anlass, auch von der Existenz eines height-Attributs auszugehen. Ähnlich verhält es sich mit dem Attribut cols, das in 4,3% aller Dokumente ebenfalls innerhalb von table benutzt wurde, obwohl es lediglich für frameset und textarea eingesetzt werden kann, um die Struktur eines Framesets bzw. eines Eingabefeldes zu definieren. Die Benennung dieses Attributs (*column*, Spalte) lässt viele Autoren annehmen, dass hierdurch auch die Anzahl der Spalten einer Tabelle voreingestellt werden kann, obwohl der HTML-Standard diese Funktion nicht vorsieht. Die Attribute marginwidth und marginheight, die jeweils hohe Werte für body besitzen, werden ebenfalls nur im Kontext von Framesets benutzt, um in frame- bzw. iframe-Elementen Abstände zu definieren. In body-Elementen sollten derartige Angaben über *Cascading Style Sheets* realisiert werden. Die Attribute border und frameborder werden häufig für das Element frameset gefunden, weil der *Internet Explorer* und die *Netscape*-Browser jeweils proprietäre Mechanismen vorsehen, um die Optik von Framesets zu beeinflussen. Diese nicht dem Standard entsprechenden Attribute wurden in fast allen frameset-Vorkommen festgestellt – insgesamt enthalten 132 151 Dokumente ein solches Element.

Proprietäre Elemente und Attribute

Um den Einsatz proprietärer HTML-Elemente und -Attribute zu erfassen, vergleichen Woodruff et al. (1996) die jeweiligen Vorkommen von table, blink, background, bgcolor, app, applet, dynsrc und embed in zwei Stichproben, die im Sommer bzw. Winter 1995 vom *Inktomi*-Projekt gesammelt wurden. Die ersten vier Elemente bzw. Attribute werden in bis zu 12% aller Dokumente gefunden, wohingegen die letzten vier Elemente nur selten vorkommen (ca. 0,001–0,07%). Für alle propietären Mechanismen ist ein Anstieg in der Benutzung zu verzeichnen, insbesondere bei der Verwendung von Hintergrundgrafiken und Tabellen, die zuerst vom *Netscape Navigator* unterstützt wurden. Turau (1998a) geht auf die Benutzung der Elemente server, bgsound, ilayer, layer, multicol, spacer und marquee ein. Diese Elemente kommen lediglich in maximal 0,49% der HTML-Dokumente vor.

	Element	Attribut	Vorkommen (absolut)	Prozent	Vorkommen (Anz. Dok.)	Prozent	Vorkommen (Ø pro Dok.)
1.	a	href	45 673 301	10,21	3 375 313	86,56	13,532
2.	img	src	24 325 433	5,44	2 602 963	66,75	9,345
3.	meta	content	7 944 199	1,78	2 567 178	65,84	3,095
4.	body	bgcolor	2 330 413	0,52	2 301 657	59,03	1,012
5.	meta	name	6 061 545	1,35	2 294 810	58,85	2,641
6.	img	width	17 373 416	3,88	2 040 611	52,33	8,514
7.	img	height	17 312 188	3,87	2 006 120	51,45	8,630
8.	table	width	6 098 323	1,36	1 927 995	49,44	3,163
9.	font	size	30 292 396	6,77	1 820 104	46,68	16,643
10.	table	border	6 617 584	1,48	1 815 579	46,56	3,645
11.	img	border	14 555 333	3,25	1 805 791	46,31	8,060
12.	meta	http-equiv	2 033 566	0,46	1 761 190	45,17	1,155
13.	td	width	29 725 837	6,64	1 751 432	44,92	16,972
14.	img	alt	13 372 246	2,99	1 731 568	44,41	7,723
15.	body	link	1 595 669	0,36	1 577 442	40,45	1,012
16.	body	vlink	1 554 462	0,35	1 536 562	39,41	1,012
17.	table	cellpadding	5 563 974	1,24	1 503 562	38,56	3,701
18.	td	align	20 236 467	4,52	1 428 460	36,63	14,167
19.	table	cellspacing	5 246 281	1,17	1 412 442	36,22	3,714
20.	body	text	1 400 809	0,31	1 381 794	35,44	1,014
21.	td	valign	20 469 206	4,58	1 235 082	31,67	16,573
22.	font	face	27 534 049	6,16	1 174 426	30,12	23,445
23.	font	color	13 750 374	3,07	1 143 207	29,32	12,028
24.	body	alink	1 144 237	0,26	1 127 922	28,93	1,014
25.	a	name	8 180 175	1,83	946 279	24,27	8,645
26.	img	align	4 635 300	1,04	943 500	24,20	4,913
27.	link	href	1 371 927	0,31	862 147	22,11	1,591
28.	body	background	862 577	0,19	847 672	21,74	1,018
29.	p	align	8 219 029	1,84	816 037	20,93	10,072
30.	td	colspan	5 491 317	1,23	780 252	20,01	7,038
31.	link	rel	1 223 910	0,27	769 921	19,75	1,590
32.	a	target	4 224 135	0,94	727 527	18,66	5,806
33.	td	height	8 646 433	1,93	705 243	18,09	12,260
34.	td	bgcolor	8 648 195	1,93	664 606	17,04	13,013
35.	div	align	2 313 751	0,52	593 574	15,22	3,898
36.	link	type	499 020	0,11	474 649	12,17	1,051
37.	table	bgcolor	965 372	0,22	458 513	11,76	2,105
38.	script	language	804 188	0,18	450 248	11,55	1,786
39.	tr	valign	3 036 163	0,68	399 773	10,25	7,595
40.	hr	width	960 567	0,21	379 258	9,73	2,533
41.	hr	size	991 333	0,22	364 632	9,35	2,719
42.	td	rowspan	1 090 852	0,24	344 837	8,84	3,163
43.	img	hspace	1 404 244	0,31	336 627	8,63	4,172
44.	img	usemap	431 785	0,10	316 108	8,11	1,366
45.	table	align	560 727	0,12	280 848	7,20	1,997
46.	a	onmouseover	2 132 363	0,48	275 330	7,06	7,745
47.	hr	noshade	697 409	0,16	263 233	6,75	2,649
48.	td	nowrap	1 931 866	0,43	241 577	6,20	7,997
49.	a	onmouseout	1 709 728	0,38	227 905	5,84	7,502
50.	tr	align	1 458 793	0,33	224 475	5,76	6,499
51.	img	name	1 697 926	0,38	212 534	5,45	7,989
52.	img	vspace	912 020	0,20	195 186	5,01	4,673
53.	style	type	193 646	0,04	188 583	4,84	1,027
54.	p	class	3 642 283	0,81	177 251	4,55	20,549
55.	tr	bgcolor	902 719	0,20	171 585	4,40	5,261
56.	map	name	269 172	0,06	164 017	4,21	1,641
57.	td	class	3 221 585	0,72	163 615	4,20	19,690
58.	area	coords	888 489	0,20	163 168	4,18	5,445
59.	area	href	903 408	0,20	163 074	4,18	5,540
60.	area	shape	874 124	0,20	159 786	4,10	5,471
61.	hr	align	367 563	0,08	152 526	3,91	2,410
62.	a	class	1 490 905	0,33	141 890	3,64	10,507
63.	input	type	1 049 719	0,23	140 901	3,61	7,450
64.	base	target	140 698	0,03	140 004	3,59	1,005
65.	input	value	825 556	0,18	136 829	3,51	6,033
66.	div	class	627 566	0,14	136 781	3,51	4,588
67.	frame	src	346 151	0,08	130 424	3,35	2,654
68.	frame	name	332 021	0,07	128 917	3,31	2,575
69.	link	rev	152 974	0,03	127 065	3,26	1,204
70.	script	src	166 569	0,04	126 299	3,24	1,319
71.	h1	align	151 217	0,03	125 415	3,22	1,206
72.	p	style	2 757 130	0,62	123 822	3,17	22,267
73.	body	onload	114 524	0,03	114 294	2,93	1,002
74.	input	name	904 334	0,20	113 812	2,92	7,946
75.	span	style	4 548 111	1,02	113 416	2,91	40,101

Tabelle A.10: Verwendung der in HTML 4.01 definierten Attribute im Korpus (Teil 1)

	Element	Attribut	Vorkommen (absolut)	Prozent	Vorkommen (Anz. Dok.)	Prozent	Vorkommen (Ø pro Dok.)
76.	form	action	136 884	0,03	106 339	2,73	1,287
77.	h2	align	154 492	0,04	106 096	2,72	1,456
78.	form	method	124 744	0,03	101 992	2,62	1,223
79.	frameset	cols	106 771	0,02	97 360	2,50	1,097
80.	frame	scrolling	214 922	0,05	94 867	2,43	2,266
81.	br	clear	317 257	0,07	92 836	2,38	3,417
82.	frameset	rows	105 947	0,02	92 206	2,37	1,149
83.	body	lang	89 131	0,02	89 057	2,28	1,001
84.	script	type	126 069	0,03	86 189	2,21	1,463
85.	input	size	307 023	0,07	85 090	2,18	3,608
86.	th	align	478 266	0,11	82 327	2,11	5,809
87.	span	class	1 120 537	0,25	81 286	2,08	13,785
88.	h3	align	127 220	0,03	75 631	1,94	1,682
89.	div	style	381 504	0,09	72 303	1,85	5,276
90.	a	style	552 286	0,12	70 838	1,82	7,796
91.	table	class	184 109	0,04	61 374	1,57	3,000
92.	th	colspan	190 885	0,04	59 798	1,53	3,192
93.	a	onclick	234 030	0,05	59 336	1,52	3,944
94.	frame	noresize	113 060	0,03	58 054	1,49	1,947
95.	link	title	119 150	0,03	57 881	1,48	2,059
96.	frame	marginwidth	134 549	0,03	54 431	1,40	2,472
97.	td	style	1 754 827	0,39	54 179	1,39	32,389
98.	table	style	101 810	0,02	53 876	1,38	1,890
99.	frame	marginheight	128 776	0,03	53 413	1,37	2,411
100.	div	id	311 021	0,07	53 113	1,36	5,856
101.	body	style	52 842	0,01	52 740	1,35	1,002
102.	area	alt	306 795	0,07	51 299	1,32	5,981
103.	a	title	372 280	0,08	50 372	1,29	7,391
104.	img	id	380 946	0,09	47 379	1,22	8,040
105.	form	name	57 513	0,01	47 043	1,21	1,223
106.	h2	class	144 675	0,03	45 836	1,18	3,156
107.	img	ismap	106 013	0,02	43 205	1,11	2,454
108.	th	width	243 203	0,05	42 489	1,09	5,724
109.	ul	type	113 983	0,03	38 230	0,98	2,982
110.	input	onclick	64 636	0,01	37 349	0,96	1,731
111.	textarea	cols	48 092	0,01	37 298	0,96	1,289
112.	textarea	rows	47 939	0,01	37 283	0,96	1,286
113.	textarea	name	47 940	0,01	37 183	0,95	1,289
114.	table	id	66 391	0,01	36 457	0,94	1,821
115.	body	class	36 470	0,01	36 417	0,93	1,001
116.	b	style	350 708	0,08	32 472	0,83	10,800
117.	select	name	58 665	0,01	32 014	0,82	1,832
118.	html	lang	31 898	0,01	31 892	0,82	1,000
119.	th	valign	141 559	0,03	29 135	0,75	4,859
120.	th	bgcolor	282 804	0,06	28 182	0,72	10,035
121.	area	target	165 775	0,04	26 049	0,67	6,36
122.	option	value	1 586 325	0,36	25 632	0,66	61,89
123.	font	style	309 326	0,07	25 060	0,64	12,34
124.	applet	code	43 740	0,01	24 750	0,64	1,77
125.	applet	width	43 694	0,01	24 712	0,63	1,77
126.	applet	height	43 678	0,01	24 703	0,63	1,77
127.	param	value	430 827	0,10	24 326	0,62	17,71
128.	param	name	431 031	0,10	24 332	0,62	17,72
129.	meta	lang	49 656	0,01	23 398	0,60	2,12
130.	input	maxlength	82 374	0,02	22 594	0,58	3,65
131.	option	selected	39 494	0,01	21 083	0,54	1,87
132.	link	media	24 143	0,01	20 891	0,54	1,16
133.	ol	type	68 784	0,02	20 552	0,53	3,35
134.	b	class	89 772	0,02	20 199	0,52	4,44
135.	table	summary	62 088	0,01	19 132	0,49	3,25
136.	tr	class	152 423	0,03	19 099	0,49	7,98
137.	dl	compact	51 797	0,01	18 331	0,47	2,83
138.	basefont	size	29 194	0,01	17 887	0,46	1,63
139.	select	onchange	22 302	0,01	16 729	0,43	1,33
140.	applet	codebase	29 792	0,01	15 200	0,39	1,96
141.	area	title	77 645	0,02	15 065	0,39	5,15
142.	td	id	286 008	0,06	13 967	0,36	20,48
143.	a	id	1 908 338	0,43	13 352	0,34	142,93
144.	th	height	81 114	0,02	12 402	0,32	6,54
145.	basefont	face	12 445	0,00	12 066	0,31	1,03
146.	table	frame	20 904	0,01	11 932	0,31	1,75
147.	input	src	21 667	0,01	11 919	0,31	1,82
148.	th	nowrap	33 105	0,01	11 573	0,30	2,86
149.	select	size	19 689	0,00	11 356	0,29	1,73
150.	i	style	176 785	0,04	11 242	0,29	15,73

Tabelle A.11: Verwendung der in HTML 4.01 definierten Attribute im Korpus (Teil 2)

	Element	Attribut	Vorkommen (absolut)	Prozent	Vorkommen (Anz. Dok.)	Prozent	Vorkommen (Ø pro Dok.)
151.	ol	start	31 476	0,01	10 745	0,27	2,93
152.	table	rules	16 529	0,00	10 170	0,26	1,63
153.	input	style	23 880	0,01	10 123	0,26	2,36
154.	caption	align	20 395	0,01	9 779	0,25	2,09
155.	area	onmouseover	91 094	0,02	9 609	0,25	9,48
156.	th	rowspan	29 519	0,01	9 528	0,24	3,10
157.	p	id	37 109	0,01	9 524	0,24	3,90
158.	span	id	945 340	0,21	9 435	0,24	100,20
159.	tt	class	48 066	0,01	8 943	0,23	5,38
160.	code	class	17 454	0,00	8 846	0,23	1,97
161.	table	rules	16 529	0,00	10 170	0,26	1,625
162.	input	style	23 880	0,01	10 123	0,26	2,359
163.	caption	align	20 395	0,01	9 779	0,25	2,086
164.	area	onmouseover	91 094	0,02	9 609	0,25	9,480
165.	th	rowspan	29 519	0,01	9 528	0,24	3,098
166.	p	id	37 109	0,01	9 524	0,24	3,896
167.	span	id	945 340	0,21	9 435	0,24	100,195
168.	tt	class	48 066	0,01	8 943	0,23	5,375
169.	code	class	17 454	0,00	8 846	0,23	1,973
170.	img	class	30 640	0,01	8 787	0,23	3,487
171.	form	onsubmit	9 717	0,00	8 690	0,22	1,118
172.	li	class	120 287	0,03	8 315	0,21	14,466
173.	th	class	48 616	0,01	8 078	0,21	6,018
174.	input	checked	18 886	0,00	8 033	0,21	2,351
175.	h2	style	19 107	0,00	7 998	0,20	2,389
176.	tr	style	179 392	0,04	7 828	0,20	22,917
177.	a	accesskey	45 881	0,01	7 801	0,20	5,881
178.	form	enctype	8 063	0,00	7 748	0,20	1,041
179.	font	id	101 056	0,02	7 408	0,19	13,641
180.	img	title	37 033	0,01	7 305	0,19	5,070
181.	ul	class	45 355	0,01	7 017	0,18	6,464
182.	br	style	38 898	0,01	7 017	0,18	5,543
183.	img	style	17 651	0,00	6 995	0,18	2,523
184.	area	onmouseout	52 835	0,01	6 702	0,17	7,883
185.	object	height	8 684	0,00	6 610	0,17	1,314
186.	object	width	8 682	0,00	6 606	0,17	1,314
187.	applet	align	7 704	0,00	6 592	0,17	1,169
188.	input	align	9 551	0,00	6 354	0,16	1,503
189.	h3	style	19 474	0,00	6 305	0,16	3,089
190.	object	classid	8 531	0,00	6 203	0,16	1,375
191.	a	rel	29 956	0,01	6 017	0,15	4,979
192.	meta	scheme	22 293	0,01	5 874	0,15	3,795
193.	link	id	5 562	0,00	5 495	0,14	1,012
194.	form	target	7 643	0,00	5 472	0,14	1,397
195.	li	style	68 099	0,01	5 419	0,14	12,567
196.	style	media	5 312	0,00	5 274	0,14	1,007
197.	applet	archive	6 465	0,00	4 941	0,13	1,308
198.	area	nohref	6 545	0,00	4 836	0,12	1,353
199.	th	style	7 799	0,00	4 808	0,12	1,622
200.	input	alt	6 162	0,00	4 678	0,12	1,317
201.	h4	class	13 400	0,00	4 632	0,12	2,893
202.	input	class	13 841	0,00	4 501	0,12	3,075
203.	object	codebase	5 806	0,00	4 403	0,11	1,319
204.	pre	class	10 096	0,00	4 313	0,11	2,341
205.	li	type	42 268	0,01	4 178	0,11	10,117
206.	img	onmouseover	12 904	0,00	4 116	0,11	3,135
207.	col	align	22 376	0,01	3 935	0,10	5,686
208.	tr	id	29 290	0,01	3 816	0,10	7,676
209.	h6	align	7 107	0,00	3 753	0,10	1,894
210.	small	class	27 507	0,01	3 712	0,10	7,410
211.	hr	id	6 381	0,00	3 722	0,10	1,714
212.	h4	style	9 717	0,00	3 592	0,09	2,705
213.	input	id	20 990	0,01	3 553	0,09	5,908
214.	html	version	3 495	0,00	3 494	0,09	1,000
215.	ul	style	10 971	0,00	3 399	0,09	3,228
216.	ul	compact	10 290	0,00	3 198	0,08	3,218
217.	td	onmouseover	20 991	0,01	3 174	0,08	6,613
218.	table	dir	7 343	0,00	3 172	0,08	2,315
219.	hr	class	8 677	0,00	3 122	0,08	2,779
220.	script	defer	5 955	0,00	3 069	0,08	1,940
221.	pre	width	17 820	0,00	3 056	0,08	5,831
222.	frameset	onload	2 994	0,00	2 985	0,08	1,003
223.	body	onclick	2 883	0,00	2 879	0,07	1,001
224.	object	id	3 837	0,00	2 797	0,07	1,372
225.	form	id	9 742	0,00	2 799	0,07	3,481

Tabelle A.12: Verwendung der in HTML 4.01 definierten Attribute im Korpus (Teil 3)

	Element	Attribut	Vorkommen (absolut)	Prozent	Vorkommen (Anz. Dok.)	Prozent	Vorkommen (Ø pro Dok.)
226.	select	id	4 515	0,00	2 771	0,07	1,629
227.	body	onunload	2 748	0,00	2 748	0,07	1,000
228.	i	class	25 042	0,01	2 693	0,07	9,299
229.	blockquote	class	4 350	0,00	2 681	0,07	1,623
230.	param	valuetype	16 361	0,00	2 575	0,07	6,354
231.	td	onmouseout	18 542	0,00	2 447	0,06	7,577
232.	input	onfocus	4 935	0,00	2 476	0,06	1,993
233.	body	onkeypress	2 474	0,00	2 474	0,06	1,000
234.	li	value	13 282	0,00	1 973	0,05	6,732
235.	a	onfocus	21 391	0,01	2 004	0,05	10,674
236.	select	style	2 413	0,00	1 946	0,05	1,240
237.	a	onmousedown	15 643	0,00	1 933	0,05	8,093
238.	body	id	1 916	0,00	1 915	0,05	1,001
239.	img	onmouseout	9 373	0,00	1 816	0,05	5,161
240.	ol	style	5 037	0,00	1 777	0,05	2,835
241.	colgroup	span	2 573	0,00	1 807	0,05	1,424
242.	applet	name	2 365	0,00	1 786	0,05	1,324
243.	col	style	9 364	0,00	1 740	0,04	5,382
244.	h5	class	4 521	0,00	1 695	0,04	2,667
245.	colgroup	width	2 374	0,00	1 689	0,04	1,406
246.	h2	id	1 784	0,00	1 604	0,04	1,112
247.	form	class	1 607	0,00	1 598	0,04	1,006
248.	hr	style	2 775	0,00	1 512	0,04	1,835
249.	div	onmouseout	6 530	0,00	1 405	0,04	4,648
250.	font	lang	2 712	0,00	1 346	0,04	2,015
251.	big	class	4 653	0,00	1 360	0,04	3,421
252.	div	onmouseover	6 133	0,00	1 341	0,03	4,573
253.	h1	id	1 343	0,00	1 284	0,03	1,046
254.	a	onmouseup	12 189	0,00	1 304	0,03	9,347
255.	thead	valign	1 459	0,00	1 229	0,03	1,187
256.	select	onfocus	1 696	0,00	1 246	0,03	1,361
257.	object	type	1 677	0,00	1 214	0,03	1,381
258.	select	class	1 581	0,00	1 161	0,03	1,362
259.	input	onblur	2 297	0,00	1 182	0,03	1,943
260.	h5	style	3 302	0,00	1 163	0,03	2,839
261.	em	class	15 912	0,00	1 162	0,03	13,694
262.	img	onmousemove	1 071	0,00	1 066	0,03	1,005
263.	applet	alt	1 245	0,00	1 038	0,03	1,199
264.	p	lang	17 389	0,00	1 003	0,03	17,337
265.	col	span	1 815	0,00	1 030	0,03	1,762
266.	a	onblur	13 607	0,00	1 028	0,03	13,236
267.	address	class	1 239	0,00	996	0,03	1,244
268.	dt	style	5 530	0,00	982	0,03	5,631
269.	a	type	6 281	0,00	960	0,03	6,543
270.	pre	id	936	0,00	933	0,02	1,003
271.	select	multiple	1 171	0,00	880	0,02	1,331
272.	iframe	width	1 125	0,00	883	0,02	1,274
273.	iframe	src	1 169	0,00	909	0,02	1,286
274.	iframe	height	1 117	0,00	879	0,02	1,271
275.	td	onclick	9 123	0,00	858	0,02	10,633
276.	object	data	1 276	0,00	870	0,02	1,467
277.	object	align	1 324	0,00	857	0,02	1,545
278.	select	onblur	854	0,00	834	0,02	1,024
279.	ol	class	2 005	0,00	827	0,02	2,424
280.	frame	style	1 267	0,00	811	0,02	1,562
281.	em	id	2 700	0,00	827	0,02	3,265
282.	area	onclick	3 414	0,00	810	0,02	4,215
283.	applet	vspace	946	0,00	836	0,02	1,132
284.	address	style	2 628	0,00	825	0,02	3,185
285.	textarea	style	919	0,00	783	0,02	1,174
286.	applet	id	890	0,00	787	0,02	1,131
287.	a	rev	1 790	0,00	762	0,02	2,349
288.	head	lang	750	0,00	750	0,02	1,000
289.	iframe	name	886	0,00	684	0,02	1,295
290.	h6	style	1 460	0,00	703	0,02	2,077
291.	strong	class	3 130	0,00	677	0,02	4,623
292.	h3	id	1 944	0,00	650	0,02	2,991
293.	applet	hspace	846	0,00	665	0,02	1,272
294.	pre	style	6 149	0,00	615	0,02	9,998
295.	iframe	scrolling	806	0,00	607	0,02	1,328
296.	h4	id	874	0,00	628	0,02	1,392
297.	col	class	3 016	0,00	643	0,02	4,691
298.	img	onclick	2 831	0,00	584	0,01	4,848
299.	ol	compact	1 233	0,00	539	0,01	2,288
300.	link	charset	620	0,00	561	0,01	1,105

Tabelle A.13: Verwendung der in HTML 4.01 definierten Attribute im Korpus (Teil 4)

	Element	Attribut	Vorkommen (absolut)	Prozent	Vorkommen (Anz. Dok.)	Prozent	Vorkommen (Ø pro Dok.)
301.	basefont	color	666	0,00	563	0,01	1,183
302.	a	shape	1 463	0,00	535	0,01	2,735
303.	u	class	1 154	0,00	515	0,01	2,241
304.	input	onchange	4 053	0,00	468	0,01	8,660
305.	input	accept	481	0,00	475	0,01	1,013
306.	iframe	frameborder	693	0,00	484	0,01	1,432
307.	div	title	1 492	0,00	487	0,01	3,064
308.	div	onclick	1 780	0,00	450	0,01	3,956
309.	a	hreflang	1 942	0,00	449	0,01	4,325
310.	p	onmouseover	3 311	0,00	443	0,01	7,474
311.	p	onmouseout	3 299	0,00	431	0,01	7,654
312.	input	tabindex	3 050	0,00	411	0,01	7,421
313.	img	onmousedown	1 740	0,00	445	0,01	3,910
314.	iframe	marginwidth	599	0,00	422	0,01	1,419
315.	iframe	marginheight	593	0,00	421	0,01	1,409
316.	iframe	align	437	0,00	414	0,01	1,056
317.	h6	class	3 033	0,00	416	0,01	7,291
318.	center	class	959	0,00	426	0,01	2,251
319.	frame	id	920	0,00	399	0,01	2,306
320.	dt	class	3 463	0,00	381	0,01	9,089
321.	dl	class	1 159	0,00	383	0,01	3,026
322.	col	valign	760	0,00	396	0,01	1,919
323.	br	class	3 326	0,00	393	0,01	8,463
324.	link	style	340	0,00	340	0,01	1,000
325.	link	hreflang	487	0,00	350	0,01	1,391
326.	input	title	689	0,00	337	0,01	2,045
327.	input	readonly	1 106	0,00	338	0,01	3,272
328.	dd	class	3 924	0,00	337	0,01	11,644
329.	blockquote	style	1 145	0,00	332	0,01	3,449
330.	tr	onmouseover	4 463	0,00	313	0,01	14,259
331.	tr	onmouseout	4 443	0,00	301	0,01	14,761
332.	table	lang	1 011	0,00	300	0,01	3,370
333.	p	title	519	0,00	294	0,01	1,765
334.	p	onclick	1 524	0,00	300	0,01	5,080
335.	map	id	398	0,00	321	0,01	1,240
336.	input	accesskey	572	0,00	297	0,01	1,926
337.	hr	title	581	0,00	331	0,01	1,755
338.	form	style	353	0,00	307	0,01	1,150
339.	b	id	1 540	0,00	320	0,01	4,812
340.	tr	onclick	3 757	0,00	255	0,01	14,733
341.	td	scope	747	0,00	270	0,01	2,767
342.	span	dir	46 420	0,01	258	0,01	179,922
343.	label	for	1 242	0,00	279	0,01	4,452
344.	html	dir	284	0,00	284	0,01	1,000
345.	dd	style	1 961	0,00	284	0,01	6,905
346.	textarea	id	339	0,00	216	0,01	1,569
347.	style	title	230	0,00	230	0,01	1,000
348.	span	title	2 537	0,00	252	0,01	10,067
349.	link	lang	325	0,00	250	0,01	1,300
350.	head	profile	233	0,00	233	0,01	1,000
351.	div	onmousedown	439	0,00	234	0,01	1,876
352.	a	lang	734	0,00	224	0,01	3,277
353.	th	id	821	0,00	202	0,01	4,064
354.	textarea	readonly	219	0,00	203	0,01	1,079
355.	td	dir	18 308	0,00	180	0,01	101,711
356.	small	style	481	0,00	197	0,01	2,442
357.	option	style	488	0,00	184	0,01	2,652
358.	object	standby	192	0,00	184	0,01	1,043
359.	img	longdesc	402	0,00	190	0,01	2,116
360.	div	onmouseup	454	0,00	211	0,01	2,152
361.	div	onkeypress	398	0,00	199	0,01	2,000
362.	col	id	2 093	0,00	188	0,01	11,133
363.	button	type	219	0,00	181	0,01	1,210
364.	body	title	198	0,00	197	0,01	1,005
365.	a	tabindex	1 555	0,00	200	0,01	7,775
366.	acronym	title	2 716	0,00	207	0,01	13,121
367.	textarea	tabindex	239	0,00	137	0,00	1,745
368.	table	title	620	0,00	138	0,00	4,493
369.	strong	style	302	0,00	167	0,00	1,808
370.	p	dir	1 669	0,00	157	0,00	10,631
371.	optgroup	label	579	0,00	147	0,00	3,939
372.	object	border	188	0,00	171	0,00	1,099
373.	ins	datetime	3 080	0,00	141	0,00	21,844
374.	ins	cite	3 066	0,00	137	0,00	22,380
375.	center	style	162	0,00	156	0,00	1,038

Tabelle A.14: Verwendung der in HTML 4.01 definierten Attribute im Korpus (Teil 5)

	Element	Attribut	Vorkommen (absolut)	Prozent	Vorkommen (Anz. Dok.)	Prozent	Vorkommen (Ø pro Dok.)
376.	button	value	159	0,00	149	0,00	1,067
377.	button	onclick	185	0,00	162	0,00	1,142
378.	button	name	267	0,00	167	0,00	1,599
379.	b	lang	3 037	0,00	171	0,00	17,760
380.	big	id	586	0,00	154	0,00	3,805
381.	a	onmousemove	2 640	0,00	156	0,00	16,923
382.	abbr	title	838	0,00	142	0,00	5,901
383.	u	style	508	0,00	130	0,00	3,908
384.	textarea	class	185	0,00	121	0,00	1,529
385.	tbody	class	126	0,00	113	0,00	1,115
386.	sup	class	482	0,00	102	0,00	4,725
387.	span	onmouseover	1 227	0,00	132	0,00	9,295
388.	span	onmouseout	1 205	0,00	124	0,00	9,718
389.	select	tabindex	249	0,00	128	0,00	1,945
390.	object	vspace	189	0,00	134	0,00	1,410
391.	object	style	134	0,00	124	0,00	1,081
392.	object	name	123	0,00	108	0,00	1,139
393.	object	hspace	191	0,00	136	0,00	1,404
394.	link	class	112	0,00	100	0,00	1,120
395.	img	onmouseup	365	0,00	104	0,00	3,510
396.	img	lang	119	0,00	110	0,00	1,082
397.	iframe	style	101	0,00	101	0,00	1,000
398.	h1	dir	144	0,00	122	0,00	1,180
399.	frame	title	365	0,00	136	0,00	2,684
400.	form	onreset	107	0,00	105	0,00	1,019
401.	del	datetime	1 650	0,00	109	0,00	15,138
402.	del	cite	1 648	0,00	108	0,00	15,259
403.	colgroup	align	346	0,00	108	0,00	3,204
404.	base	href	130	0,00	130	0,00	1,000
405.	a	charset	163	0,00	102	0,00	1,598
406.	ul	id	159	0,00	82	0,00	1,939
407.	tt	style	130	0,00	91	0,00	1,429
408.	tt	id	598	0,00	66	0,00	9,061
409.	textarea	onfocus	137	0,00	89	0,00	1,539
410.	td	title	302	0,00	82	0,00	3,683
411.	tbody	valign	70	0,00	62	0,00	1,129
412.	tbody	align	132	0,00	62	0,00	2,129
413.	sup	id	178	0,00	60	0,00	2,967
414.	sub	id	292	0,00	68	0,00	4,294
415.	q	cite	81	0,00	70	0,00	1,157
416.	option	label	1 339	0,00	75	0,00	17,853
417.	option	disabled	131	0,00	61	0,00	2,148
418.	object	declare	61	0,00	61	0,00	1,000
419.	menu	compact	105	0,00	60	0,00	1,750
420.	link	target	326	0,00	78	0,00	4,179
421.	li	id	515	0,00	90	0,00	5,722
422.	input	onmouseover	171	0,00	93	0,00	1,839
423.	input	disabled	95	0,00	72	0,00	1,319
424.	i	id	22 705	0,01	73	0,00	311,027
425.	head	dir	87	0,00	87	0,00	1,000
426.	form	title	73	0,00	73	0,00	1,000
427.	em	style	169	0,00	65	0,00	2,600
428.	dl	style	147	0,00	76	0,00	1,934
429.	col	char	59	0,00	59	0,00	1,000
430.	cite	lang	463	0,00	85	0,00	5,447
431.	cite	class	277	0,00	82	0,00	3,378
432.	caption	style	76	0,00	69	0,00	1,101
433.	body	onmouseover	62	0,00	62	0,00	1,000
434.	body	onmouseout	60	0,00	60	0,00	1,000
435.	body	ondblclick	92	0,00	92	0,00	1,000
436.	blockquote	cite	92	0,00	78	0,00	1,179
437.	a	coords	378	0,00	81	0,00	4,667
438.	acronym	lang	648	0,00	84	0,00	7,714
439.	ul	title	83	0,00	33	0,00	2,515
440.	ul	dir	182	0,00	21	0,00	8,667
441.	th	scope	329	0,00	37	0,00	8,892
442.	textarea	title	54	0,00	25	0,00	2,160
443.	textarea	onchange	46	0,00	44	0,00	1,045
444.	textarea	onblur	23	0,00	20	0,00	1,150
445.	td	lang	172	0,00	21	0,00	8,190
446.	td	charoff	598	0,00	57	0,00	10,491
447.	table	onmouseover	204	0,00	34	0,00	6,000
448.	table	onmouseout	190	0,00	23	0,00	8,261
449.	table	onmousemove	22	0,00	22	0,00	1,000
450.	sup	style	162	0,00	22	0,00	7,364

Tabelle A.15: Verwendung der in HTML 4.01 definierten Attribute im Korpus (Teil 6)

	Element	Attribut	Vorkommen (absolut)	Prozent	Vorkommen (Anz. Dok.)	Prozent	Vorkommen (Ø pro Dok.)
451.	sub	class	579	0,00	31	0,00	18,677
452.	span	onmousedown	77	0,00	20	0,00	3,850
453.	span	onclick	155	0,00	57	0,00	2,719
454.	s	style	230	0,00	24	0,00	9,583
455.	select	title	62	0,00	34	0,00	1,824
456.	select	onclick	41	0,00	25	0,00	1,640
457.	option	class	838	0,00	47	0,00	17,830
458.	ol	id	53	0,00	29	0,00	1,828
459.	noframes	title	36	0,00	36	0,00	1,000
460.	noframes	id	32	0,00	32	0,00	1,000
461.	legend	style	23	0,00	23	0,00	1,000
462.	legend	align	76	0,00	34	0,00	2,235
463.	label	accesskey	51	0,00	22	0,00	2,318
464.	input	onmouseout	102	0,00	50	0,00	2,040
465.	i	lang	90	0,00	33	0,00	2,727
466.	iframe	id	67	0,00	42	0,00	1,595
467.	h5	id	71	0,00	35	0,00	2,029
468.	h4	lang	294	0,00	20	0,00	14,700
469.	h2	lang	249	0,00	31	0,00	8,032
470.	h1	onmouseover	77	0,00	51	0,00	1,510
471.	h1	onmouseout	77	0,00	51	0,00	1,510
472.	h1	lang	123	0,00	35	0,00	3,514
473.	frameset	title	84	0,00	52	0,00	1,615
474.	font	title	56	0,00	33	0,00	1,697
475.	fieldset	style	33	0,00	27	0,00	1,222
476.	div	onmousemove	27	0,00	27	0,00	1,000
477.	div	dir	88	0,00	26	0,00	3,385
478.	dir	compact	46	0,00	24	0,00	1,917
479.	cite	style	26	0,00	20	0,00	1,300
480.	caption	class	240	0,00	47	0,00	5,106
481.	br	id	27	0,00	27	0,00	1,000
482.	body	onmousemove	21	0,00	21	0,00	1,000
483.	body	onkeydown	52	0,00	52	0,00	1,000
484.	body	dir	23	0,00	23	0,00	1,000
485.	b	onmouseover	52	0,00	37	0,00	1,405
486.	blockquote	dir	82	0,00	44	0,00	1,864
487.	big	style	420	0,00	45	0,00	9,333
488.	area	tabindex	156	0,00	32	0,00	4,875
489.	area	onfocus	236	0,00	39	0,00	6,051
490.	area	class	77	0,00	50	0,00	1,540
491.	applet	style	44	0,00	37	0,00	1,189
492.	a	ondblclick	69	0,00	23	0,00	3,000
493.	address	lang	27	0,00	27	0,00	1,000
494.	abbr	lang	76	0,00	27	0,00	2,815
495.	var	title	4	0,00	4	0,00	1,000
496.	var	style	5	0,00	5	0,00	1,000
497.	var	id	4	0,00	4	0,00	1,000
498.	var	class	23	0,00	9	0,00	2,556
499.	u	title	1	0,00	1	0,00	1,000
500.	u	onmouseover	78	0,00	6	0,00	13,000
501.	u	onclick	9	0,00	9	0,00	1,000
502.	u	id	9	0,00	9	0,00	1,000
503.	ul	onmouseover	2	0,00	2	0,00	1,000
504.	ul	onmouseout	2	0,00	2	0,00	1,000
505.	ul	onclick	1	0,00	1	0,00	1,000
506.	ul	lang	13	0,00	3	0,00	4,333
507.	tt	title	12	0,00	5	0,00	2,400
508.	tt	onmouseover	4	0,00	4	0,00	1,000
509.	tt	lang	12	0,00	2	0,00	6,000
510.	tr	title	34	0,00	16	0,00	2,125
511.	tr	ondblclick	3	0,00	1	0,00	3,000
512.	tr	lang	34	0,00	12	0,00	2,833
513.	tr	dir	8	0,00	2	0,00	4,000
514.	tr	char	2	0,00	2	0,00	1,000
515.	title	lang	11	0,00	11	0,00	1,000
516.	th	title	54	0,00	17	0,00	3,176
517.	th	onmouseover	69	0,00	6	0,00	11,500
518.	th	onmouseout	69	0,00	6	0,00	11,500
519.	th	onmousedown	2	0,00	2	0,00	1,000
520.	th	onclick	2	0,00	1	0,00	2,000
521.	th	lang	14	0,00	8	0,00	1,750
522.	th	dir	17	0,00	1	0,00	17,000
523.	th	charoff	2	0,00	2	0,00	1,000
524.	th	char	9	0,00	4	0,00	2,250
525.	th	axis	1	0,00	1	0,00	1,000

Tabelle A.16: Verwendung der in HTML 4.01 definierten Attribute im Korpus (Teil 7)

	Element	Attribut	Vorkommen (absolut)	Prozent	Vorkommen (Anz. Dok.)	Prozent	Vorkommen (Ø pro Dok.)
526.	th	abbr	91	0,00	11	0,00	8,273
527.	thead	style	11	0,00	5	0,00	2,200
528.	thead	id	4	0,00	4	0,00	1,000
529.	thead	class	3	0,00	2	0,00	1,500
530.	thead	align	23	0,00	19	0,00	1,211
531.	tfoot	valign	4	0,00	4	0,00	1,000
532.	tfoot	id	4	0,00	4	0,00	1,000
533.	tfoot	align	3	0,00	3	0,00	1,000
534.	textarea	onselect	3	0,00	3	0,00	1,000
535.	textarea	onkeyup	11	0,00	11	0,00	1,000
536.	textarea	onkeypress	9	0,00	9	0,00	1,000
537.	textarea	onkeydown	19	0,00	17	0,00	1,118
538.	textarea	onclick	2	0,00	2	0,00	1,000
539.	textarea	lang	4	0,00	4	0,00	1,000
540.	textarea	disabled	11	0,00	9	0,00	1,222
541.	textarea	dir	1	0,00	1	0,00	1,000
542.	textarea	accesskey	8	0,00	6	0,00	1,333
543.	td	onmouseup	7	0,00	1	0,00	7,000
544.	td	onmousemove	93	0,00	4	0,00	23,250
545.	td	onmousedown	35	0,00	9	0,00	3,889
546.	td	headers	16	0,00	1	0,00	16,000
547.	td	char	63	0,00	16	0,00	3,938
548.	td	abbr	69	0,00	9	0,00	7,667
549.	tbody	title	47	0,00	7	0,00	6,714
550.	tbody	style	17	0,00	17	0,00	1,000
551.	tbody	id	5	0,00	5	0,00	1,000
552.	table	onclick	8	0,00	6	0,00	1,333
553.	sup	title	35	0,00	5	0,00	7,000
554.	sup	lang	22	0,00	2	0,00	11,000
555.	sub	title	4	0,00	4	0,00	1,000
556.	sub	style	21	0,00	18	0,00	1,167
557.	strong	title	1	0,00	1	0,00	1,000
558.	strong	onmouseover	5	0,00	5	0,00	1,000
559.	strong	onmouseout	5	0,00	5	0,00	1,000
560.	strong	lang	30	0,00	14	0,00	2,143
561.	strong	id	4	0,00	4	0,00	1,000
562.	strong	dir	2	0,00	2	0,00	1,000
563.	strike	title	4	0,00	4	0,00	1,000
564.	strike	onmouseover	4	0,00	4	0,00	1,000
565.	strike	id	4	0,00	4	0,00	1,000
566.	s	id	4	0,00	4	0,00	1,000
567.	small	id	1	0,00	1	0,00	1,000
568.	select	onmouseover	25	0,00	5	0,00	5,000
569.	select	onmouseout	23	0,00	3	0,00	7,667
570.	select	disabled	4	0,00	4	0,00	1,000
571.	script	charset	19	0,00	17	0,00	1,118
572.	samp	style	28	0,00	12	0,00	2,333
573.	samp	onmouseover	4	0,00	4	0,00	1,000
574.	samp	id	4	0,00	4	0,00	1,000
575.	samp	class	38	0,00	4	0,00	9,500
576.	q	style	4	0,00	4	0,00	1,000
577.	q	onmouseover	4	0,00	4	0,00	1,000
578.	q	lang	24	0,00	12	0,00	2,000
579.	q	id	4	0,00	4	0,00	1,000
580.	q	dir	2	0,00	1	0,00	2,000
581.	q	class	14	0,00	2	0,00	7,000
582.	pre	title	2	0,00	2	0,00	1,000
583.	pre	onclick	6	0,00	1	0,00	6,000
584.	pre	dir	1	0,00	1	0,00	1,000
585.	p	onmouseup	1	0,00	1	0,00	1,000
586.	p	onmousemove	8	0,00	8	0,00	1,000
587.	p	onmousedown	3	0,00	3	0,00	1,000
588.	p	ondblclick	2	0,00	2	0,00	1,000
589.	param	type	4	0,00	1	0,00	4,000
590.	param	id	48	0,00	6	0,00	8,000
591.	option	onclick	2	0,00	2	0,00	1,000
592.	ol	dir	2	0,00	2	0,00	1,000
593.	object	usemap	4	0,00	4	0,00	1,000
594.	object	title	3	0,00	3	0,00	1,000
595.	object	tabindex	3	0,00	3	0,00	1,000
596.	object	codetype	16	0,00	7	0,00	2,286
597.	object	class	2	0,00	2	0,00	1,000
598.	object	archive	12	0,00	3	0,00	4,000
599.	noscript	title	4	0,00	4	0,00	1,000
600.	noscript	id	4	0,00	4	0,00	1,000

Tabelle A.17: Verwendung der in HTML 4.01 definierten Attribute im Korpus (Teil 8)

	Element	Attribut	Vorkommen (absolut)	Prozent	Vorkommen (Anz. Dok.)	Prozent	Vorkommen (Ø pro Dok.)
601.	noscript	class	3	0,00	3	0,00	1,000
602.	menu	style	9	0,00	5	0,00	1,800
603.	menu	onmouseover	4	0,00	4	0,00	1,000
604.	menu	class	4	0,00	4	0,00	1,000
605.	map	title	4	0,00	4	0,00	1,000
606.	map	onmouseover	16	0,00	2	0,00	8,000
607.	map	onmouseout	16	0,00	2	0,00	8,000
608.	li	title	48	0,00	1	0,00	48,000
609.	li	onmouseup	5	0,00	1	0,00	5,000
610.	li	onmouseover	13	0,00	1	0,00	13,000
611.	li	onmousedown	5	0,00	1	0,00	5,000
612.	li	onclick	3	0,00	1	0,00	3,000
613.	li	lang	42	0,00	18	0,00	2,333
614.	li	dir	4	0,00	3	0,00	1,333
615.	legend	id	8	0,00	4	0,00	2,000
616.	legend	accesskey	2	0,00	1	0,00	2,000
617.	label	id	8	0,00	4	0,00	2,000
618.	label	class	2	0,00	2	0,00	1,000
619.	kbd	style	4	0,00	4	0,00	1,000
620.	kbd	id	4	0,00	4	0,00	1,000
621.	kbd	class	42	0,00	6	0,00	7,000
622.	isindex	prompt	6	0,00	6	0,00	1,000
623.	ins	style	2	0,00	2	0,00	1,000
624.	ins	onmouseover	4	0,00	4	0,00	1,000
625.	ins	id	4	0,00	4	0,00	1,000
626.	input	onselect	5	0,00	5	0,00	1,000
627.	input	onmouseup	2	0,00	2	0,00	1,000
628.	input	onmousemove	2	0,00	1	0,00	2,000
629.	input	onmousedown	24	0,00	15	0,00	1,600
630.	input	onkeyup	11	0,00	10	0,00	1,100
631.	input	onkeypress	34	0,00	18	0,00	1,889
632.	input	onkeydown	12	0,00	11	0,00	1,091
633.	input	ondblclick	10	0,00	10	0,00	1,000
634.	input	lang	18	0,00	2	0,00	9,000
635.	input	ismap	2	0,00	2	0,00	1,000
636.	i	title	10	0,00	6	0,00	1,667
637.	i	onmouseover	11	0,00	11	0,00	1,000
638.	i	onmouseout	2	0,00	2	0,00	1,000
639.	i	onclick	3	0,00	3	0,00	1,000
640.	img	onkeydown	4	0,00	4	0,00	1,000
641.	img	ondblclick	1	0,00	1	0,00	1,000
642.	img	dir	1	0,00	1	0,00	1,000
643.	iframe	class	4	0,00	1	0,00	4,000
644.	hr	dir	1	0,00	1	0,00	1,000
645.	h6	title	2	0,00	2	0,00	1,000
646.	h6	lang	7	0,00	3	0,00	2,333
647.	h6	id	2	0,00	2	0,00	1,000
648.	h5	title	83	0,00	13	0,00	6,385
649.	h5	onmouseover	4	0,00	4	0,00	1,000
650.	h5	onmouseout	4	0,00	4	0,00	1,000
651.	h5	lang	25	0,00	13	0,00	1,923
652.	h4	title	2	0,00	1	0,00	2,000
653.	h4	onmouseover	56	0,00	8	0,00	7,000
654.	h4	onmouseout	56	0,00	8	0,00	7,000
655.	h4	dir	10	0,00	4	0,00	2,500
656.	h3	onmouseover	30	0,00	11	0,00	2,727
657.	h3	onmouseout	30	0,00	11	0,00	2,727
658.	h3	lang	247	0,00	17	0,00	14,529
659.	h3	dir	41	0,00	19	0,00	2,158
660.	h2	title	9	0,00	9	0,00	1,000
661.	h2	onmouseover	12	0,00	12	0,00	1,000
662.	h2	onmouseout	12	0,00	12	0,00	1,000
663.	h2	ondblclick	1	0,00	1	0,00	1,000
664.	h2	dir	30	0,00	19	0,00	1,579
665.	h1	title	18	0,00	18	0,00	1,000
666.	h1	onclick	1	0,00	1	0,00	1,000
667.	frameset	style	20	0,00	18	0,00	1,111
668.	frameset	onunload	19	0,00	19	0,00	1,000
669.	frameset	id	5	0,00	4	0,00	1,250
670.	frameset	class	2	0,00	2	0,00	1,000
671.	frame	longdesc	2	0,00	1	0,00	2,000
672.	frame	class	4	0,00	4	0,00	1,000
673.	form	onmouseover	1	0,00	1	0,00	1,000
674.	form	onmouseout	1	0,00	1	0,00	1,000
675.	form	onclick	4	0,00	4	0,00	1,000

Tabelle A.18: Verwendung der in HTML 4.01 definierten Attribute im Korpus (Teil 9)

	Element	Attribut	Vorkommen (absolut)	Prozent	Vorkommen (Anz. Dok.)	Prozent	Vorkommen (Ø pro Dok.)
676.	form	lang	3	0,00	3	0,00	1,000
677.	form	accept-charset	16	0,00	16	0,00	1,000
678.	form	accept	2	0,00	2	0,00	1,000
679.	font	dir	11	0,00	10	0,00	1,100
680.	fieldset	onmouseover	4	0,00	4	0,00	1,000
681.	fieldset	id	4	0,00	4	0,00	1,000
682.	em	title	12	0,00	8	0,00	1,500
683.	em	onmouseover	4	0,00	4	0,00	1,000
684.	em	onmouseout	4	0,00	4	0,00	1,000
685.	em	lang	63	0,00	13	0,00	4,846
686.	em	dir	1	0,00	1	0,00	1,000
687.	dt	title	28	0,00	4	0,00	7,000
688.	dt	onmouseover	1	0,00	1	0,00	1,000
689.	dt	onclick	8	0,00	4	0,00	2,000
690.	dt	lang	28	0,00	9	0,00	3,111
691.	dt	id	134	0,00	5	0,00	26,800
692.	dl	onmouseover	4	0,00	4	0,00	1,000
693.	dl	id	13	0,00	13	0,00	1,000
694.	div	lang	16	0,00	10	0,00	1,600
695.	dir	style	4	0,00	4	0,00	1,000
696.	dir	onmouseover	4	0,00	4	0,00	1,000
697.	dir	dir	4	0,00	1	0,00	4,000
698.	dir	class	24	0,00	13	0,00	1,846
699.	dfn	title	7	0,00	7	0,00	1,000
700.	dfn	style	10	0,00	8	0,00	1,250
701.	dfn	onmouseover	4	0,00	4	0,00	1,000
702.	dfn	lang	24	0,00	6	0,00	4,000
703.	dfn	id	4	0,00	4	0,00	1,000
704.	dfn	class	8	0,00	2	0,00	4,000
705.	del	onmouseover	4	0,00	4	0,00	1,000
706.	del	id	4	0,00	4	0,00	1,000
707.	dd	title	48	0,00	8	0,00	6,000
708.	dd	lang	30	0,00	3	0,00	10,000
709.	dd	id	12	0,00	1	0,00	12,000
710.	col	title	2	0,00	2	0,00	1,000
711.	col	lang	51	0,00	11	0,00	4,636
712.	col	charoff	2	0,00	1	0,00	2,000
713.	colgroup	valign	10	0,00	9	0,00	1,111
714.	colgroup	style	3	0,00	1	0,00	3,000
715.	colgroup	id	19	0,00	6	0,00	3,167
716.	colgroup	class	7	0,00	3	0,00	2,333
717.	colgroup	char	5	0,00	5	0,00	1,000
718.	code	style	74	0,00	9	0,00	8,222
719.	code	onclick	4	0,00	4	0,00	1,000
720.	code	lang	38	0,00	9	0,00	4,222
721.	code	id	62	0,00	12	0,00	5,167
722.	cite	title	22	0,00	6	0,00	3,667
723.	cite	onmouseover	4	0,00	4	0,00	1,000
724.	cite	id	7	0,00	5	0,00	1,400
725.	center	title	4	0,00	4	0,00	1,000
726.	center	onmouseup	2	0,00	2	0,00	1,000
727.	center	onmouseover	8	0,00	8	0,00	1,000
728.	center	onmouseout	3	0,00	3	0,00	1,000
729.	center	onmousemove	2	0,00	2	0,00	1,000
730.	center	onmousedown	2	0,00	2	0,00	1,000
731.	center	id	7	0,00	7	0,00	1,000
732.	caption	title	4	0,00	4	0,00	1,000
733.	caption	onclick	4	0,00	4	0,00	1,000
734.	caption	id	20	0,00	16	0,00	1,250
735.	button	title	1	0,00	1	0,00	1,000
736.	button	tabindex	1	0,00	1	0,00	1,000
737.	button	style	9	0,00	3	0,00	3,000
738.	button	onmouseover	1	0,00	1	0,00	1,000
739.	button	onmouseout	1	0,00	1	0,00	1,000
740.	button	disabled	1	0,00	1	0,00	1,000
741.	body	onmouseup	3	0,00	3	0,00	1,000
742.	body	onmousedown	11	0,00	11	0,00	1,000
743.	body	onkeyup	6	0,00	6	0,00	1,000
744.	b	title	116	0,00	7	0,00	16,571
745.	b	onmouseout	18	0,00	3	0,00	6,000
746.	b	onclick	17	0,00	8	0,00	2,125
747.	blockquote	title	5	0,00	3	0,00	1,667
748.	blockquote	lang	2	0,00	2	0,00	1,000
749.	blockquote	id	20	0,00	16	0,00	1,250
750.	big	onclick	4	0,00	4	0,00	1,000

Tabelle A.19: Verwendung der in HTML 4.01 definierten Attribute im Korpus (Teil 10)

	Element	Attribut	Vorkommen (absolut)	Prozent	Vorkommen (Anz. Dok.)	Prozent	Vorkommen (Ø pro Dok.)
751.	bdo	dir	10	0,00	10	0,00	1,000
752.	basefont	id	4	0,00	4	0,00	1,000
753.	area	style	3	0,00	1	0,00	3,000
754.	area	onmouseup	61	0,00	5	0,00	12,200
755.	area	onmousemove	11	0,00	3	0,00	3,667
756.	area	onmousedown	43	0,00	11	0,00	3,909
757.	area	ondblclick	6	0,00	1	0,00	6,000
758.	area	id	83	0,00	16	0,00	5,188
759.	area	accesskey	56	0,00	14	0,00	4,000
760.	applet	title	3	0,00	3	0,00	1,000
761.	applet	class	5	0,00	5	0,00	1,000
762.	a	dir	1	0,00	1	0,00	1,000
763.	address	onmouseover	32	0,00	6	0,00	5,333
764.	address	onmouseout	32	0,00	6	0,00	5,333
765.	address	id	25	0,00	7	0,00	3,571
766.	address	dir	9	0,00	6	0,00	1,500
767.	acronym	style	4	0,00	4	0,00	1,000
768.	acronym	onmouseover	4	0,00	4	0,00	1,000
769.	acronym	id	4	0,00	4	0,00	1,000
770.	acronym	class	12	0,00	6	0,00	2,000
771.	abbr	style	4	0,00	4	0,00	1,000
772.	abbr	onmouseover	4	0,00	4	0,00	1,000
773.	abbr	onmouseout	4	0,00	4	0,00	1,000
774.	abbr	id	4	0,00	4	0,00	1,000
775.	abbr	class	6	0,00	6	0,00	1,000

Tabelle A.20: Verwendung der in HTML 4.01 definierten Attribute im Korpus (Teil 11)

Im Korpus existieren, wie die Anteile im Standard definierter Elemente bereits andeuten, kaum proprietäre HTML-Elemente. Die Elemente `nobr` und `o:p` werden mit Vorkommen in 2,34% bzw. 1,14% aller Dokumente am häufigsten eingesetzt.[18] Weitere proprietäre Elemente kommen jeweils in weniger als einem Prozent aller HTML-Dokumente vor, z. B. `spacer` (0,96%), `blink` (0,64%), `embed` (0,27%) oder `wbr` (0,09%). In der sehr umfangreichen Liste nicht standardisierter HTML-Elemente befinden sich auch falsch geschriebene Tags (z. B. `adress`, 0,16%, oder `titel`, 0,10%) und sogar Elemente, deren Bezeichnungen von Autoren fehlerhaft konzeptualisiert wurden, z. B. `h7` (0,03%), `header` (0,09%) statt `head`, `it` (0,02%) statt `em` bzw. `i`, `listing` (0,01%) statt `code` oder `pre`, `item` (103 Dokumente) statt `li`, `comment` (0,02%) statt korrekter SGML-Kommentare[19] (`<!-- ... -->`), `bf` (0,01%) statt `strong` bzw. `b` oder `headline` (0,01%) statt `h1` bis `h6`.[20] Bezüglich der Verwendung proprietärer Attribute können identische Aussagen gemacht werden: Die häufigsten Attribute sind `topmargin` (6,50%), `leftmargin` (6,22%), `bordercolor` (2,13%) und `framespacing` (1,87%). Auch hier wurden erwartungsgemäß falsch geschriebene Attributbezeichnungen gefunden, z. B. `hight` (0,13%) oder `nowarp` (0,02%).[21]

[18] Das Element `o:p` wird von den HTML-Export-Filtern verschiedener *Microsoft Office*-Anwendungen zur Auszeichnung von Abschnitten eingesetzt und nur vom *Internet Explorer* interpretiert. Frühe Versionen des *Netscape Navigator* haben das Element `nobr` eingeführt, das zur Kennzeichnung von Text bestimmt war, der vom Browser nicht umbrochen werden soll. Es gab Überlegungen, dieses Element auch in HTML 3.0 aufzunehmen, diese Version wurde jedoch nie offiziell standardisiert (vgl. `http://www.w3.org/MarkUp/html3/html3.txt`).

[19] Im Korpus wurden insgesamt 13 637 810 derartige Kommentare gefunden.

[20] Die ca. 1 000 Vorkommen von `bf` und `it` sind vermutlich auf Übertragungsfehler seitens der Autoren zurück zu führen: Das Textsatzsystem LaTeX verwendet die Befehle `\textbf{...}` und `\textit{...}`, um auf Fettdruck (*bold face*) bzw. Kursivschrift (*italics*) umzuschalten.

[21] Chidlovskii (2003) berichtet von einem im November 2002 durchgeführten Vergleich von 32 Websites mit den jeweiligen Versionen aus dem Jahr 1998: Es existiert eine Tendenz zu fehlerfreien HTML-Dokumenten und der Verwendung umfangreicherer Bestände von Elementen und Attributen.

	Element	Attribut	Vorkommen (absolut)	Prozent	Vorkommen (Anz. Dok.)	Prozent	Vorkommen (Ø pro Dok.)
1.	table	height	484 059	0,11	299 225	7,67	1,618
2.	body	marginwidth	197 750	0,04	196 705	5,04	1,005
3.	body	marginheight	188 440	0,04	187 395	4,81	1,006
4.	table	cols	336 758	0,07	166 634	4,27	2,021
5.	frameset	border	137 095	0,03	100 014	2,56	1,371
6.	td	background	372 755	0,08	89 927	2,31	4,145
7.	frameset	frameborder	116 785	0,03	85 758	2,20	1,362
8.	hr	color	134 977	0,03	64 127	1,65	2,105
9.	tr	height	485 645	0,11	57 012	1,46	8,518
10.	table	background	56 351	0,01	41 760	1,07	1,349
11.	table	valign	96 154	0,02	38 985	1,00	2,466
12.	table	hspace	52 736	0,01	28 050	0,72	1,880
13.	table	vspace	55 148	0,01	26 908	0,69	2,050
14.	img	valign	85 116	0,02	25 790	0,66	3,300
15.	frame	target	48 309	0,01	24 914	0,64	1,939
16.	meta	value	80 370	0,02	19 814	0,51	4,056
17.	frame	border	32 748	0,01	15 158	0,39	2,160
18.	img	onload	95 299	0,02	13 756	0,35	6,928
19.	table	frameborder	26 903	0,01	12 561	0,32	2,142
20.	hr	valign	12 351	0,00	12 107	0,31	1,020
21.	table	name	15 868	0,00	11 868	0,30	1,337
22.	input	border	13 292	0,00	10 903	0,28	1,219
23.	a	language	54 746	0,01	10 728	0,28	5,103
24.	tr	width	24 165	0,01	7 962	0,20	3,035
25.	font	text	7 422	0,00	7 385	0,19	1,005
26.	td	content	26 491	0,01	7 146	0,18	3,707
27.	a	border	17 088	0,00	7 142	0,18	2,393
28.	table	alt	6 972	0,00	6 933	0,18	1,006
29.	tr	border	7 499	0,00	6 342	0,16	1,182
30.	tr	cellspacing	43 303	0,01	6 274	0,16	6,902
31.	tr	cellpadding	6 269	0,00	5 985	0,15	1,047
32.	meta	type	56 699	0,01	5 881	0,15	9,641
33.	a	alt	18 582	0,00	5 296	0,14	3,509
34.	td	border	18 818	0,00	5 133	0,13	3,666
35.	font	align	8 104	0,00	5 052	0,13	1,604
36.	input	width	7 166	0,00	5 024	0,13	1,426
37.	style	id	4 710	0,00	4 667	0,12	1,009
38.	tr	nowrap	9 681	0,00	4 482	0,12	2,160
39.	input	height	5 570	0,00	4 293	0,11	1,297
40.	a	align	7 912	0,00	4 209	0,11	1,880

Tabelle A.21: Verwendung laut HTML 4.01 nicht erlaubter Element-Attribut-Kombinationen im Korpus (Gesamtanzahl: 1 101)

A.4.3 Hyperlinkbezogene Eigenschaften

Im Folgenden werden verschiedene hyperlinkbezogene Charakteristika der im Korpus enthaltenen Dokumente analysiert. Verknüpfungen werden in HTML mit Hilfe des Elements a realisiert, die Analysen basieren ausschließlich auf den 51 679 884 Vorkommen dieses Elements, die insgesamt ca. 67 Millionen Attribute aufweisen.

Zur Verteilung von Hyperlinks

Von den 3 899 341 untersuchten Dokumenten enthalten insgesamt 3 375 313 (86,56%) mindestens ein a-Element mit dem Attribut href, dessen Wert das Linkziel darstellt. Die durchschnittliche Anzahl pro Dokument beträgt 13,53 mit einem Median von 7 (Modus: 1, Max.: 4 629).[22] Woodruff et al. (1996) berichten einen Schnitt von 17 Hyperlinks. Abbildung A.6 zeigt eine Aufstellung der Anzahl Hyperlinks pro Dokument bezüglich des gesamten Datenbestandes. Aus dem Diagramm geht hervor, dass die Mehrzahl aller Dokumente

[22] Diese Werte beziehen sich ebenfalls nur auf diejenigen Vorkommen des a-Elements, die das Attribut href enthalten, weshalb die in Tabelle A.8 dargestellten Werte leichte Abweichungen aufweisen.

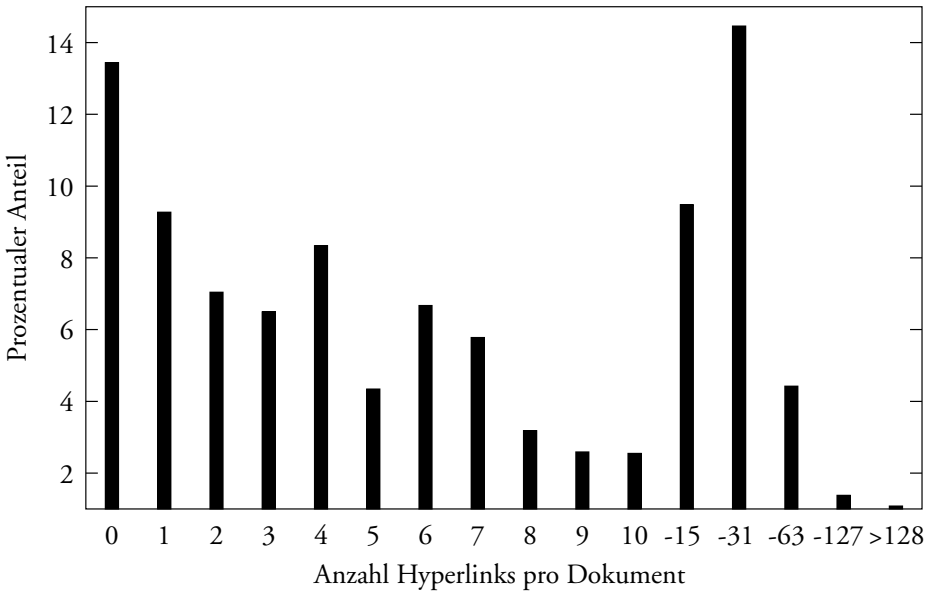


Abbildung A.6: Verteilung der Anzahl Hyperlinks pro Dokument

(55,9%) bis zu 10 Hyperlinks enthält. Zudem beträgt die Anzahl der Dokumente mit mehr als 64 Links lediglich 1,93%, d. h. extensiv verknüpfende Dokumente sind im WiN-Web sehr selten. Bray (1996) berichtet, dass ca. 75% aller analysierten Dokumente mindestens einen Verweis enthalten, wobei die Verteilung, die ebenfalls anhand eines Balkendiagramms illustriert wird, ähnliche Tendenzen wie Abbildung A.6 aufweist. Turau (1998a) gibt ebenfalls Daten zur Hyperlinkverteilung an, diese weisen jedoch mehrere Spezifika auf: In dieser Stichprobe sind pro Dokument durchschnittlich 25,8 Hyperlinks enthalten (Median: 22, Max.: 9 965). Es wird ebenfalls ein Balkendiagramm, wobei auffällig ist, dass lediglich ca. 0,4% der Knoten keinerlei Hyperlinks enthalten. Diese Diskrepanz bezüglich den genannten Daten von 13,4% bzw. 25% bei Bray ist möglicherweise durch die Art der Datensammlung zu erklären.[23] Weiterhin fällt auf, dass in der Stichprobe eine sehr viel größere Zahl von Dokumenten mit mehr als 60 Hyperlinks enthalten ist (vgl. Abbildung A.6). Abgesehen von diesen Unterschieden wird auch in der von Turau (1998a) präsentierten Studie der Trend deutlich, dass die meisten HTML-Dokumente bis zu ca. 30 Hyperlinks enthalten.

Referenzierung unterschiedlicher Internet-Dienste

Um innerhalb eines Hyperlinks nicht auf einen Webserver, sondern auf einen anderen Dienst zu verweisen, werden standardisierte Protokollbezeichner eingesetzt (vgl. Tabelle A.22).[24] Mehr als 90% aller Hyperlinks verwenden HTTP, 4,8% aller Links bieten eine Möglich-

[23] Turau (1998a) benutzt eine reine Breitensuche mit einem vorgegebenen Schwellwert von 500 Dokumenten, so dass sich in dieser Stichprobe vermutlich nur wenige Blattknoten befinden.

[24] Das Schema `rtsp` bezeichnet das *Real Time Streaming Protocol*, das für die strombasierte Übertragung von Audio- und/oder Video-Daten in Verbindung mit dem *Real Player* entworfen wurde. Die Bezeichnung `mocha` kann in einigen Browsern der Firma *Netscape* als alternative Bezeichnung zu `javascript` benutzt werden.

Protokoll	Vorkommen	Prozent	Protokoll	Vorkommen	Prozent
http	41 143 294	90,0821	maito	274	0,0006
mailto	2 224 586	4,8707	rtsp	240	0,0005
javascript	2 159 408	4,7280	ttp	156	0,0003
ftp	86 582	0,1896	hhttp	106	0,0002
news	25 696	0,0563	c	103	0,0002
https	18 029	0,0395	htttp	95	0,0002
gopher	4 495	0,0098	livescript	75	0,0002
telnet	3 607	0,0079	mocha	75	0,0002
about	3 188	0,0070	mailt	70	0,0002
datei	857	0,0019	email	63	0,0001

Tabelle A.22: Die 20 häufigsten in Hyperlinks eingesetzten Protokolle

keit der Kontaktaufnahme per elektronischer Post, und 4,7% aller Verweise referenzieren *JavaScript*-Dateien.[25] Kurz nach der Einführung des WWW waren insbesondere ftp- und news-Verweise beliebt, weil diese etablierten Dienste viele Informationen und Ressourcen umfassen.[26] Die korrespondierenden Clients waren jedoch sehr komplex, weshalb der Benutzerkreis dieser Dienste nur eingeschränkt war. Tabelle A.22 zeigt, dass die Internet-Dienste Telnet, Gopher, FTP und Usenet im WiN-Web nur noch eine untergeordnete Rolle spielen, weil viele ihrer Funktionen auch mit Hilfe von Webservern und HTTP emuliert bzw. abgewickelt werden können. Fast alle weiteren Vorkommen – auch die 290 nicht in der Tabelle präsentierten – sind auf Tippfehler zurückzuführen, z. B. maito, ttp, hhttp oder mailt. Dies ist ein deutlicher Indikator für den Umstand, dass viele Autoren ihre Webangebote nicht testen und Hyperlinks manuell in die Webseiten eintragen, anstatt *Copy & Paste*-Mechanismen zu benutzen (vgl. auch Woodruff et al., 1996, und Turau, 1998a).

Verknüpfung mit anderen top-level-Domänen

Tabelle A.23 stellt die 30 häufigsten *top-level*-Domänen dar, auf die in 11 451 936 Hyperlinks verwiesen wird, die vollständig qualifizierte Namen von Webservern enthalten. Es war zwar zu erwarten, dass die Länderkennung für Deutschland im Rahmen einer Untersuchung eines großen Teils des WWW innerhalb der Domäne .de die meisten Vorkommen hat, ein Anteil von 86,7% ist jedoch überraschend hoch. Es ist zu beachten, dass diese etwa 10 Millionen Links auch die Verweise auf Dokumente beinhalten, die aufgrund der implementierten Datensammlungsbeschränkungen nicht in das Korpus aufgenommen wurden. Tabelle A.1 (S. 723) zeigt, dass es sich hierbei um etwa 4,5 Millionen Hyperlinks handelt.

Auch ohne vorliegende Vergleichsdaten kann aus diesem hohen Wert nur der Schluss gezogen werden, dass innerhalb bestimmter *top-level*-Domänen, die sich vermutlich nur auf länderspezifische Kennungen (im Gegensatz zu den mittlerweile generischen Domänen .com, .org und .net) beschränken, vornehmlich auf Webangebote der gleichen Domäne verwiesen wird. Bezüglich der Verteilung der übrigen Domänen fällt auf, dass .com mit 5,1% Vor-

[25] Von den 2 224 586 mailto-Verweisen enthalten 89 808 Links (4,04%) einen per ?subject=... angegebenen Parameter, der eine vom Autoren des HTML-Dokuments vordefinierte Betreffzeile umfasst.

[26] Boldi et al. (2002) berichten Werte von 0,38% für ftp, 0,02% für gopher und 0,07% für news. Auffällig sind die hochfrequenten Protokolle mailto (29,97%) und javascript (14,94%) innerhalb des afrikanischen Webs.

top-level-Domäne	Vorkommen	Prozent	*top-level*-Domäne	Vorkommen	Prozent
de	9 931 126	86,720	fr	25 788	0,225
com	594 502	5,191	int	24 613	0,215
org	220 115	1,922	ca	17 843	0,156
edu	157 954	1,379	it	14 461	0,126
net	77 314	0,675	se	13 415	0,117
uk	69 639	0,608	au	12 418	0,108
ch	45 142	0,394	jp	9 194	0,080
at	43 112	0,377	es	9 136	0,080
gov	40 060	0,350	be	7 533	0,066
nl	35 731	0,312	dk	7 172	0,063

Tabelle A.23: Die 20 häufigsten *top-level*-Domänen, auf die in Hyperlinks verwiesen wird

kommen die zweite Position einnimmt, denn in dieser *top-level*-Domäne befinden sich nur wenige Angebote aus thematischen Bereichen, die für Forschung oder Lehre von Interesse sind. Turau (1998a) führt in diesem Kontext an, dass im Zuge der Datensammlung seiner Stichprobe extrem viele Verweise auf sehr populäre Websites gefunden wurden, die vermutlich auch für einen Großteil dieser fast 600 000 Hyperlinks nach `.com` verantwortlich sind (z. B. `www.yahoo.com`, `www.google.com`, `www.microsoft.com` und `www.adobe.com`). An vierter Stelle befindet sich `.edu`, die unter anderem alle US-amerikanischen Universitäten umfasst, was – bei aller Abstraktion – ein Indiz dafür sein könnte, dass die dortige Forschung, neben der einheimischen, bei deutschen Wissenschaftlern am deutlichsten wahrgenommen wird. Die verbleibenden Plätze werden fast vollständig vom europäischen Ausland belegt, die Ausnahmen bilden Kanada (`.ca`), Australien (`.au`) und Japan (`.jp`).

Einsatz zusätzlicher Attribute

Das Element `a` umfasst neben `href` weitere Attribute, z. B. `name` (zur Benennung eines Linkziels), `hreflang` (zur Auszeichnung der Sprache eines Ziels) oder `type` (zur Auszeichnung des Medientyps). Tabelle A.24 zeigt die 15 häufigsten Attribute sowie deren prozentuale Anteile an allen 67 379 034 beobachteten Attributen des `a`-Elements. Das Attribut `href` nimmt mit einem Anteil von 67,8% erwartungsgemäß und mit großem Abstand die erste Position ein, gefolgt von `name` mit 12,1%. Insgesamt werden in 946 279 HTML-Dokumenten (24,27%) durchschnittlich 8,65 `a`-Elemente mit dem Attribut `name` benutzt (Median: 2, Max.: 3 259), d. h. dieser Mechanismus zur intrahypertextuellen Navigation wird in fast einem Viertel aller Dokumente eingesetzt und resultiert vermutlich – das Maximum deutet dies an – aus automatischen Konvertierungsprozessen, im Zuge derer häufig auch Inhaltsverzeichnisse erzeugt werden, die umfangreichen Gebrauch von `<a name="...">` machen.[27] An dritter Stelle befindet sich `target` mit ca. 4,2 Millionen Vorkommen, das ausschließlich innerhalb von Framesets eingesetzt wird, so dass sich ein Hyperlink auf einen anderen Frame auswirken

[27] Auf Sprungziele, die durch `<a name="...">` definiert werden, verweisen Hyperlinks, deren abschließendes Segment ein *fragment* ist (RFC 2396), das durch das Zeichen # vom URI abgetrennt wird. Diesem Zeichen folgt das Sprungziel, d. h. der Wert des entsprechenden `name`-Attributs im Zielknoten. Im Korpus befinden sich insgesamt 7 446 641 Hyperlinks (16,30% aller Links), die eine solche Angabe enthalten.

Attribut	Vorkommen	Prozent	Attribut	Vorkommen	Prozent
href	45 673 301	67,79	title	372 280	0,55
name	8 180 175	12,14	onclick	234 030	0,35
target	4 224 135	6,27	add_date	117 695	0,17
onmouseover	2 132 363	3,16	last_visit	117 294	0,17
id	1 908 338	2,83	last_modified	115 628	0,17
onmouseout	1 709 728	2,54	language	54 746	0,08
class	1 490 905	2,21	accesskey	45 881	0,07
style	552 286	0,82			

Tabelle A.24: Die 15 häufigsten Attribute des HTML-Elements a

kann. Dieser absolut betrachtet sehr hohe Wert ist ein Indiz dafür, dass Framesets im WiN-Web häufig eingesetzt werden, doch kann dieses Attribut auch zweckentfremdet werden, um ein neues Browserfenster zu öffnen. Das Element frameset besitzt 199 425 Vorkommen, frame liegt mit 346 630 Vorkommen erwartungsgemäß deutlich höher, weil pro Frameset meist zwei bis drei Frames definiert werden. Bezüglich der Anzahl Dokumente liegen die Werte etwas geringer: frameset kommt in 132 151 Dokumenten (3,39%, bei Turau insgesamt 3,7%) mit durchschnittlich 1,51 Elementen vor (Modus und Median: 1, Max.: 1 023), frame verzeichnet durchschnittlich 2,66 Vorkommen (Modus und Median: 2, Max.: 1 024) in 130 551 Dokumenten (3,35%). Interessanterweise existieren also *nicht* in allen Dokumenten, in denen ein frameset definiert wird, korrespondiere frame-Spezifikationen. Das Element noframes, das Inhalte für Browser aufnehmen soll, die Framesets nicht interpretieren können, wird in nur 78 045 Dokumenten (2,00%) eingesetzt.

Die Attribute onmouseover, onmouseout und onclick werden im Zusammenhang mit *JavaScript* benutzt, um Ereignisse – z. B. das Bewegen des Mauszeigers auf einen Hyperlink – abzufangen und etwa durch eine visuelle Rückmeldung zu bestätigen. Die ebenfalls hohen Werte dieser drei Attribute deuten einen umfangreichen Einsatz von Sprachen wie *JavaScript* im Korpus an (vgl. Abschnitt A.4.6). Auffällig sind drei im Standard nicht definierte Attribute. Hierbei handelt es sich um add_date, last_visit und last_modified mit jeweils nahezu identischen Frequenzen.[28] Alle Browser bieten dem Benutzer die Möglichkeit, Bookmarks bzw. Lesezeichen zu setzen, die oftmals in einer Datei namens bookmarks.html bzw. bookmarks.htm gespeichert werden, d. h. der Benutzer manipuliert durch Hinzufügen oder Löschen eines Bookmarks indirekt ein HTML-Dokument. Bereits der Browser *Mosaic* (Andreessen und Bina, 1994), der Vorläufer des *Netscape Navigator*, enthielt die Funktion, häufig genutzte URLs in Form einer „Hotlist“ zu speichern, die zudem per elektronischer Post als HTML-Dokument verschicket werden konnte.[29] Im Nachfolger *Netscape Navigator* wurden Bookmarks erstmals unmittelbar als HTML-Dateien gepflegt, wobei die Attribute

[28] Streng genommen gehört auch language in diese Gruppe, weil es für das Element a nicht gültig ist und nur innerhalb von script eingesetzt werden kann, um die benutzte Skriptsprache zu identifizieren. Nach HTML 4.01 sollte language jedoch nicht mehr benutzt werden, weil die Angabe eines Medientyps mit Hilfe des Attributs type die präferierte Methode zur Markierung von Dateitypen darstellt.

[29] Siehe die Dokumentation zu *Mosaic*, z. B. http://archive.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/UserGuide/XMosaic.5.4.html: „Use the Mail To. . .button at the bottom of the Hotlist View window to send your hotlist via electronic mail to a colleague. The electronic mail is an HTML document [. . .].“

add_date, last_visit und last_modified die entsprechenden Zeitstempel der Aufnahme in die Liste von Bookmarks, des letzten Besuchs und der letzten Modifikation enthalten. Derartige bookmarks.html-Dateien wurden Mitte bis Ende der neunziger Jahre sehr häufig auf persönlichen Homepages veröffentlicht, um auf effizientem Wege interessante Informationsangebote aufzuzeigen und die eigenen Interessen darzustellen.[30] Die hohe Anzahl von Vorkommen deutet an, dass sich sehr viele Dateien mit Bookmarks im Korpus befinden, die vermutlich zum größten Teil von persönlichen Homepages stammen.

A.4.4 Einsatz und Verbreitung von Metadaten

O'Neill et al. (1998) untersuchen meta-Elemente in einer Stichprobe von ca. 166 000 HTML-Dokumenten, die im Juni 1998 erzeugt wurde (vgl. Abschnitt A.6): In ca. 70% der 1 457 Einstiegsseiten und in ca. 45% der internen Seiten sind durchschnittlich 2,75 bzw. 2,27 Metadaten-Beschreibungen enthalten (Median und Modus jeweils: 2). Die für das Korpus ermittelten Werte liegen etwas höher: Von den ca. vier Millionen HTML-Dokumenten enthalten insgesamt 2 586 845 (66,34%) mindestens ein meta-Element, die durchschnittliche Frequenz beträgt 3,13 (Median und Modus: 2, Max.: 961; vgl. Abbildung A.7).

Attribut	Vorkommen	Prozent	Attribut	Vorkommen	Prozent
content	7 944 199	48,182	charset	2 554	0,016
name	6 061 545	36,756	http-eqiv	2 262	0,014
http-equiv	2 033 566	12,326	contents	1 345	0,008
value	80 370	0,487	author	1 002	0,006
context	64 492	0,391	http_equiv	923	0,006
type	56 699	0,344	http-request	813	0,005
lang	49 656	0,304	language	633	0,004
scheme	22 293	0,135			

Tabelle A.25: Die 15 häufigsten Attribute des HTML-Elements meta

Die Spezifikation von HTML 4.01 sieht für meta die Attribute name (der Wert dieses Attributs benennt eine spezielle Dokumenteigenschaft), http-equiv (kann alternativ zu name, aber insbesondere zur Spezifizierung eines HTTP-Headers benutzt werden), content (der Wert einer Eigenschaft) und scheme vor (zur Referenzierung eines formalen Metadaten-Schemas, auf das sich die Werte der Attribute name und content beziehen). Im Korpus existieren 13 464 verschiedene Attributwerte für meta, wobei ein Großteil dieser Angaben auf nicht valide HTML-Dokumente zurückzuführen ist, die inkorrekte Ereignisse in HTML::Parser verursachen und somit in fehlerhaften Angaben resultieren.[31] Tabelle A.25 stellt die 15 hochfrequenten Attributbezeichnungen dar. Die Aufstellung zeigt, dass die Kombination aus content- und name-Attributen am häufigsten eingesetzt wird, http-equiv wird deutlich seltener benutzt. Das Attribut scheme wird kaum eingesetzt. Wie bei allen untersuchten HTML-Elementen finden sich auch hier zahlreiche Attribute, die nicht standardkonform

[30] Bookmark-Dateien werden auch in zahlreichen Arbeiten zur Benutzermodellierung verwendet, um – basierend auf den vorhandenen Bookmarks, die das Interessenprofil eines spezifischen Benutzers konstituieren – initiale Präferenzen festzulegen (vgl. Goldman et al., 1997).

[31] Beispiele für fehlerhafte Werte sind "keywords",), </head, <title, iso639") und name"keywords".

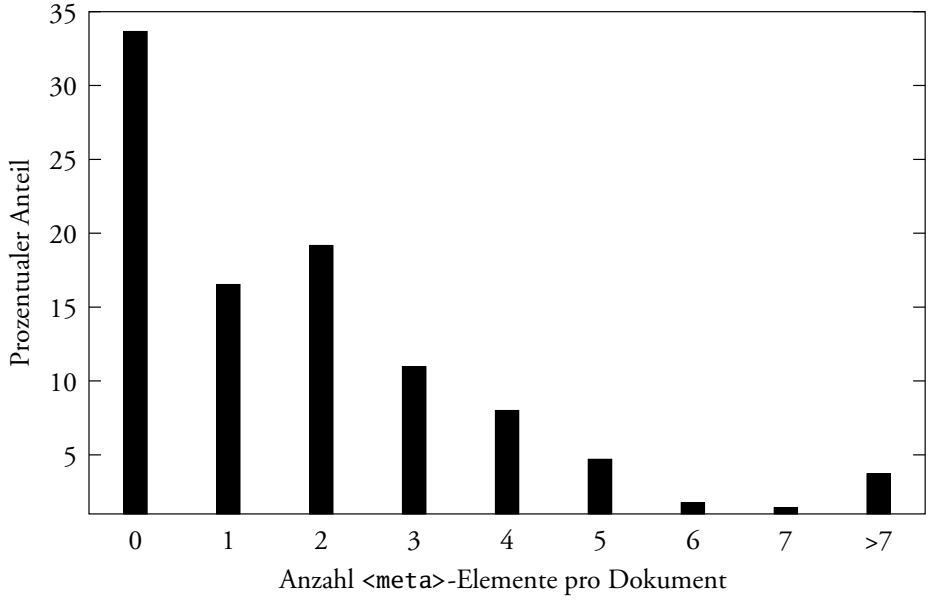


Abbildung A.7: Verteilung der Anzahl von meta-Elementen pro Dokument

sind (z. B. value und context) oder falsch geschrieben wurden wie http-eqiv, http_equiv oder contents.[32] Attribute wie author oder language sollten eigentlich als Attribut*wert* von name eingesetzt werden, diese Vorkommen lassen auf mangelnde Kenntnis der HTML-Spezifikation bzw. des Metadatenkonzepts in HTML schließen.

Tabelle A.26 stellt die Werte des Attributs name dar. Die Angaben werden mit den 20 häufigsten Werten verglichen, die O'Neill et al. ermittelt haben. Es zeigt sich, dass die Attributwerte generator, author, keywords und description sowohl bei O'Neill et al. als auch im Korpus die häufigste Gruppe darstellen. Interessant ist, dass generator (vgl. Abschnitt A.4.7) beide Listen dominieren, und die eigentlich zentralen, den Dokumentinhalt beschreibenden Attributwerte author sowie insbesondere keywords und description im Vergleich zu 1998 nur geringfügig häufiger eingesetzt werden. Die zentrale Schlussfolgerung von O'Neill et al. (1998) lautet, dass sich Metadaten in HTML-Dokumenten in den meisten Fällen nur auf automatisch von HTML-Editoren hinzugefügte Informationen beschränken. Dieser Trend wird durch Tabelle A.26 bestätigt. Auffällig sind auch die hohen Werte für dc.publisher und dc.language. Hierbei handelt es sich um DC-konforme Metadatenauszeichnungen.

Die *Dublin Core Metadata Initiative* (http://purl.org/dc/) hat ein Metadatenschema erarbeitet, das insbesondere für den Einsatz im WWW gedacht ist (vgl. RFC 2413, RFC 2731, und Caplan, 1995). Version 1.1 des Schemas (DCES, 2003) legt 15 Elemente und 12 Typen von Informationsressourcen fest (DCTV, 2004, vgl. auch Schmidt, 2004, und Abschnitt 3.6.6). O'Neill et al. (1998) finden in lediglich sieben der 1 024 untersuchten Einstiegsseiten DC-basierte Metadatenauszeichnungen, im Korpus enthalten hingegen 37 754

[32] Das Attribut value ist zwar in HTML 4.01 spezifiziert, es gilt jedoch nicht für meta, sondern nur für option, param, button und li. Die häufigen Vorkommen des ungültigen Attributs context sind vermutlich auf die orthografische Ähnlichkeit mit content zurück zu führen.

Attributwert	Vorkommen	Prozent	O'Neill et al. (1998): Nur Einstiegsseiten	O'Neill et al. (1998): Nur interne Dokumente
generator	1 684 337	27,78	23	34
author	764 731	12,61	6	7
keywords	672 850	11,10	17	7
description	594 770	9,81	15	6
robots	196 120	3,23	1	1
progid	156 883	2,59	—	—
resource-type	154 494	2,55	1	1
distribution	154 241	2,54	1	1
microsoft border	94 369	1,56	3	1
date	84 368	1,39	—	—
copyright	77 290	1,28	1	1
microsoft theme	75 662	1,25	2	1
template	60 439	1,00	1	2
revisit-after	46 894	0,77	<1	—
publisher	44 829	0,74	—	—
dc.publisher	32 973	0.55	—	—
originator	29 834	0.49	—	—
page-topic	28 205	0.47	—	—
language	28 091	0.46	—	—
dc.language	28 080	0.46	—	—

Tabelle A.26: Die 20 häufigsten Werte des Attributs <meta name="...">

Dokumente (0,97%) DC-Metadaten, wobei im Durchschnitt 8,75 Elemente benutzt werden (Median: 8, Modus: 6, Max.: 88). Falls sich also ein Autor im WiN-Web für DC entscheidet, geschieht die Realisierung meist sehr umfassend. Insgesamt wurden 197 unterschiedliche Elementbezeichnungen gefunden, die mit dc. beginnen (zur Notation siehe RFC 2731 und zur Validierung Kelly et al., 2003), Tabelle A.27 stellt die 20 häufigsten dar.[33] Zum einen existiert eine große Spannbreite bezüglich der Verteilung der DC-Elemente; dies ist naturgemäß ein zentrales Merkmal von Metadaten, da parallel Angaben zum Erstellungsdatum, Autoren, Titel etc. gemacht werden. Zum anderen werden DC-Elemente häufig detaillierter unterteilt, als dies der Standard vorsieht. RFC 2731 führt diese Möglichkeit zwar explizit an („In actual resource description [sic] it is often necessary to qualify *Dublin Core* elements to add nuances of meaning"), doch gilt „qualified DC" mittlerweile als obsolet. Tatsächlich enthalten die meisten der 197 gefundenen DC-Elemente „qualifizierte" Beschreibungen (z. B. dc.creator.personalname.address.keywords oder dc.creator.corporatename.address.email), die die DC-inhärente Vermischung von standardisierten und ad hoc erzeugten Metadaten-Bezeichnungen verdeutlichen, die insbesondere für die Informationserschließung im heterogenen WWW eher kontraproduktiv ist.

Neben dem meta-Attribut scheme kann auch das Attribut profile innerhalb von head eingesetzt werden, um ein Metadatenschema zu referenzieren. O'Neill et al. (1998) berichten für scheme drei Vorkommen (vgl. die 22 293 Vorkommen in Tabelle A.25) und für profile lediglich eines. Im Korpus sind insgesamt 233 profile-Attribute enthalten, die 16 verschie-

[33] Der Name dc.contributer ist ein weiteres Beispiel für ein falsch geschriebenes Element. Die hohe Anzahl von 5 784 Vorkommen deutet an, dass viele DC-Elemente aus einer einzigen Quelle stammen.

Element	Vorkommen	Prozent	O'Neill et al. (1998): Vorkommen
dc.publisher	32 973	9,98	5
dc.language	28 080	8,50	5
dc.subject	27 782	8,40	6
dc.identifier	26 347	7,97	1
dc.creator	26 047	7,88	4
dc.title	22 600	6,84	6
dc.format	18 679	5,65	1
dc.date	17 493	5,30	1
dc.type	17 304	5,24	2
dc.rights	17 231	5,21	—
dc.description	16 562	5,01	6
dc.creator.personalname	5 985	1,81	—
dc.contributer	5 784	1,75	—
dc.date.creation	5 660	1,71	—
dc.date.lastmodified	5 359	1,62	—
dc.source	4 516	1,37	—
dc.creator.personalname.address	3 799	1,15	—
dc.relation	3 469	1,05	—
dc.subject.topic	2 699	0,82	—
dc.publisher.corporatename.address	2 345	0,71	—

Tabelle A.27: Die 20 häufigsten *Dublin Core*-Elemente innerhalb von meta-Elementen

dene Werte aufweisen. Hierfür sieht Raggett et al. (1999) eine URL vor, an der weitere Informationen zu einem Schema oder eine formale Spezifikation hinterlegt sind. 143 profile-Attribute (61,63%) enthalten die URI http://purl.org/metadata/dublin_core, von der auch Alternativen ermittelt wurden (z. B. http://purl.org/dc/elements/1.1/). Die verbleibenden Werte von profile enthalten zu etwa gleichen Teilen einfache Dateinamen und korrekte URIs, die vornehmlich auf studentische Homepages verweisen, auf denen einzelne Gruppen von meta-Elementen in Dateien abgelegt sind.

Bezüglich der Nutzung von Metadaten im WiN-Web kann festgehalten werden, dass etwa zwei Drittel aller HTML-Dokumente durchschnittlich drei meta-Elemente besitzen. Dieser überraschend hohe Anteil von Metadaten lässt sich jedoch, dies zeigen bereits O'Neill et al. (1998), vor allem durch automatisch von HTML-Editoren hinzugefügte Metadaten erklären: Der Anteil inhaltsbeschreibender Informationen ist – bezüglich der Gesamtanzahl von Attributen – mit etwa 23% sehr gering.

A.4.5 Multimedia

Der Begriff Multimedia subsumiert Bilder, Grafiken, Audio- und Videodateien. Das Element img wird zur Referenzierung binärer Bilddateien in 2 603 103 Dokumenten (66,76%) mit durchschnittlich 9,35 Vorkommen eingesetzt (Median: 5, Modus: 1, Max.: 4 097, Summe: 24 327 871), vgl. Abbildung A.8. Bray (1996) berichtet Vorkommen in ca. 50% der untersuchten Dokumente. Die Verteilung wird ebenfalls als Balkendiagramm visualisiert, dessen genereller Trend mit Abbildung A.8 vergleichbar ist. Die beiden Abweichungen betreffen einerseits die Vorkommen von Dokumenten ohne Bilder (ca. 50% vs. 33,24%) und andererseits den generellen Durchschnitt der Anzahl eingebetteter Bilddateien, der bei Bray etwas

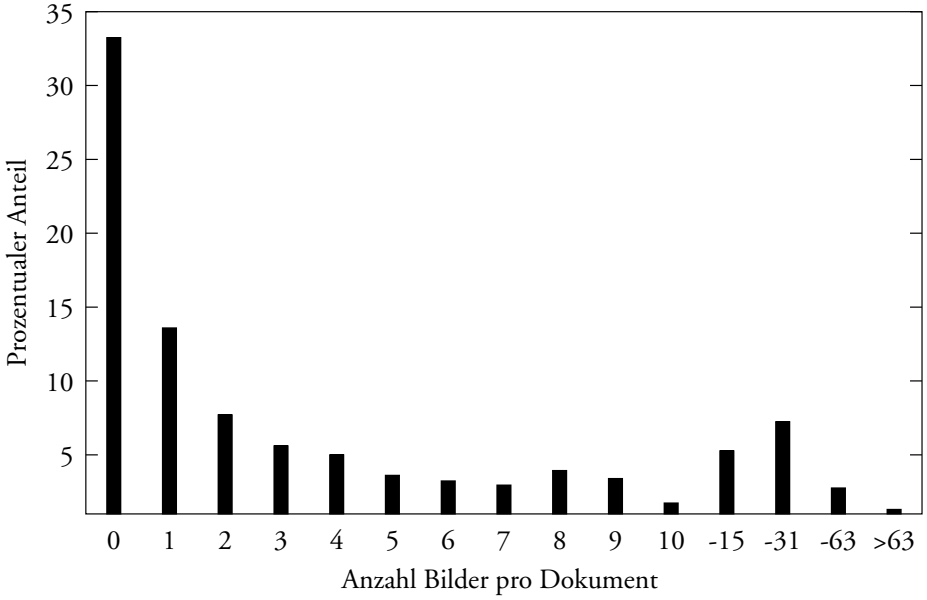


Abbildung A.8: Verteilung der Anzahl eingebetteter Bilder pro Dokument

geringer ist. In der von Woodruff et al. untersuchten Stichprobe existieren durchschnittlich vier `img`-Elemente in ca. 62% aller Dokumente. Woodruff et al. geben auch die Anteile der Formate an: In 61,7% aller Dokumente werden GIF-Dateien eingebunden, das primär zur Komprimierung von Fotos entwickelte JPEG-Format wird in 7,8% eingesetzt. Turau gibt an, dass in ca. 75% aller Dokumente durchschnittlich 10,38 Bilder enthalten sind (Median: 4, Max.: 2 176). Eingeschränkt auf den Bereich „edu“ verringert sich das arithmetische Mittel auf 6,7, wobei „ähnliche Werte“ auch für den Bereich „deedu“ erhoben wurden. Bezüglich der Benutzung unterschiedlicher Formate gibt Turau an, dass 81,6% aller Bilder GIF-Dateien sind, JPEG macht lediglich 4% aus (in „deedu“ jedoch 8,8%). Die in der Korpusdatenbank enthaltenen HTTP-Response-Header zeigen, dass sich die Verteilung im WiN-Web verschoben hat (vgl. Tabelle A.3, S. 728): Auf GIF-Dateien entfällt nur noch ein Anteil von 57,9%, wohingegen JPEG mit 39,3% zugenommen hat. Das Format PNG nimmt mit 2,2% die dritte Stelle ein. Turau (1998a) geht auch auf *clickable maps* ein, d. h. Grafiken, für die mit Hilfe des `map`-Elements separate Bereiche mit unterschiedlichen Hyperlinks verbunden werden können: In der gesamten Stichprobe weisen 21,2% aller Dokumente ein `map`-Element auf, in „deedu“ jedoch nur 3,0%. Im Korpus existieren insgesamt 269 190 dieser Elemente in 164 029 Dokumenten (4,21%; Modus und Median: 1, Max.: 167, Durchschnitt: 1,64).

Zur Referenzierung von Dateitypen, die Browser üblicherweise nicht eigenständig darstellen können – hierzu gehören insbesondere Audio- und Video-Dateien – definiert HTML 4.01 das über zahlreiche Attribute verfügende Element `object`, sehr häufig wird auch das Element `embed` benutzt. Dieses besitzt im Korpus insgesamt 18 126 Vorkommen in 10 591 Dokumenten (0,27%, Modus und Median: 1, Durchschnitt: 1,71, Max.: 71), wohingegen `object` mit lediglich 7 173 Dokumenten (0,18%) durchschnittlich 1,40 mal und somit seltener benutzt wird (Modus und Median: 1, Summe: 10 011). Innerhalb der `embed`-Elemente wur-

den die mittels src angegebenen Dateien auf ihre Endungen untersucht. Demnach sind 31,7% aller eingebetteten Dateien *Shockwave Flash*-Animationen (swf, 4 130 Vorkommen), MIDI-Dateien kommen in 10,6% aller Elemente vor (mid, 1 381 Vorkommen), woraufhin *Quicktime*-Videodateien (mov, 8,4%, 1 091 Vorkommen), Audio-Dateien nach dem WAVE-Standard (wav, 8,1%, 1 053 Vorkommen) und *Powerpoint*-Präsentationen (ppz, 6,7%, 870 Vorkommen) folgen. Weitere häufig gefundene Dateiformate sind rpm (5,0%), mp3 (4,7%), pdb (3,8%), wrl (3,0%), dcr (2,1%), avi (1,7%), cdx (1,4%) und mpg (0,9%). Die eindeutigste Methode zur Untersuchung der in object-Elementen enthaltenen Dateitypen besteht in einer Analyse der Werte des Attributs classid. Diese referenzieren in Form einer hexadezimalen Zeichenkette eindeutige Bezeichner, die vom Browser benutzt werden, um korrespondierende *Plug-ins* zu starten. In den 10 011 object-Elementen wurden insgesamt 8 531 Werte dieses Attributs gefunden, die 3 637 *Macromedia Flash*-Dateien (42,6%) referenzieren. Es folgen *ActiveX-OLE*-Objekte (975 Vorkommen, 11,4%), *ActiveX-Powerpoint* (869 Vorkommen, 10,2%), der *RealPlayer* (638 Vorkommen, 7,5%) und die *Java Virtual Machine* (619 Vorkommen, 7,3%). In der Liste der nachfolgenden Dateitypen dominieren weitere *Microsoft ActiveX*-Objekte, z. B. *ActiveX-Tabular-Data-Control* (1,5%), *ActiveX-Media-Player-2* (0,9%), *ActiveX-ActiveMovie* (0,8%) oder *ActiveX-Excel* (0,5%). Die genannten Häufigkeiten bestätigen Turau (1998a), der in seiner Stichprobe kaum Beispiele für den unmittelbaren Einsatz von Audio- (ca. 1% aller Dokumente) und Video-Dateien (ca. 0,25% aller Dokumente) in Webseiten findet. Dem steht die relativ hohe Anzahl Audio- und Video-basierter Medientypen gegenüber, die anhand der Korpusdatenbank aus der Sammlung der HTTP-Response-Header ermittelt wurden (vgl. Tabelle A.3 auf S. 728). Es kann gefolgert werden, dass derartige multimediale Dateien im WiN-Web eher über Hyperlinks eingebunden werden, anstatt den technisch komplexeren und weniger plattformunabhängigen Weg zu wählen, multimediale Inhalte unmittelbar in ein HTML-Dokument einzubetten.

A.4.6 Interaktive Elemente – Client-seitige Anwendungen

Eine Erweiterung des Funktionsumfangs von HTML kann durch den Einsatz von HTML-Formularen, CGI-Skripten, Client-seitigen Skriptsprachen und *Java*-Applets realisiert werden. Turau (1998a) stellt hierzu fest, dass selbst im „experimentierfreudigerem Umfeld der Universitäten“ keinesfalls überdurchschnittlich viele *Java*-Anwendungen eingesetzt werden: Im Bereich „deedu“ enthalten lediglich 0,68% aller Dokumente ein Applet, 2,7% enthalten *JavaScript*-Code (insgesamt: 0,86% bzw. 9,1%). Die mit Hilfe des Korpus ermittelten Werte zeigen, dass *JavaScript* an Bedeutung gewonnen hat: Diese Sprache wird – vemutlich größtenteils für optische Effekte während der Navigation eines Dokuments – in 529 230 Dokumenten (13,57%) benutzt, wohingegen die Verbreitung von *Java* mit Vorkommen in 22 985 Dokumenten (0,59%) im Vergleich zu Turau (1998a) als stagnierend bezeichnet werden kann.[34] Andere Skriptsprachen weisen nur minimale Anteile auf: *JScript* wird in 2 950 Webseiten (0,08%), *LiveScript* in 699 (0,02%) und *VBScript* nur in 445 Dokumenten (0,01%) verwendet. *JScript* ist die *Microsoft*-eigene Implementierung des ECMAScript-Standards in der Version 3, *LiveScript* ist die ehemalige, im *Netscape Navigator* 2.0 eingesetzte Bezeich-

[34] Dieser Wert bezieht sich auf applet-Elemente, in denen der Wert von code auf class, java oder jar endet.

nung für *JavaScript* und *VBScript* ist schließlich eine Erweiterung von *Visual Basic*, die in *MS Office*-Anwendungen und im *Internet Explorer* verwendet werden kann.[35]

HTML-Formulare werden unter anderem für Suchfunktionen eingesetzt und durch das Element `form` realisiert, das wiederum Elemente wie `input`, `button` oder `select` enthalten kann, um Eingabemasken, Knöpfe oder Menüs darzustellen. Eingegebene Daten werden entweder Server- (z. B. mittels *PHP*- oder CGI) oder Client-seitig (etwa *JavaScript*) ausgewertet. In den Dokumenten des Korpus sind insgesamt 211 479 `form`-Elemente enthalten, die sich auf 166 279 Webseiten (4,26%) verteilen. Durchschnittlich umfasst ein HTML-Dokument 1,27 dieser Elemente (Modus und Median: 1, Max.: 111). Turau (1998a) berichtet, dass im Bereich „deedu“ 5,5% aller Dokumente ein `form`-Element beinhalten, in dieser Domäne scheint in Bezug auf interaktive Angebote also überraschenderweise ein rückläufiger Trend vorzuliegen.[36] Komplementär zu diesen Angaben wurde auch die Anzahl von Hyperlinks erhoben, deren URLs die Zeichenketten `/cgi-bin/` oder `/cgi/` enthalten – diese Verzeichnisse umfassen meist CGI-Skripte. Bezüglich der 45 673 131 insgesamt beobachteten Hyperlinks treffen die beiden obigen Muster in 367 674 Fällen zu, d. h. CGI-Skripte haben einen Anteil von ca. 0,81% an der gesamten Hyperlinkbasis im WiN-Webs. RFC 2396 sieht vor, dass arbiträre Anfragen (z. B. in dynamisch erzeugten Seiten, zur permanenten Weitergabe einer Session-ID etc.) innerhalb einer URI/URL in einer „query component“ hinterlegt werden können, die mittels `?` von der Pfadkomponente abgetrennt wird. Im Anfrageteil sind z. B. die Zeichen `/`, `+`, `&` und `=` reserviert, mit deren Hilfe Attribut-Wert-Paare und Abfolgen von Suchwörtern enkodiert werden. Derart aufgebaute Adressen werden unter anderem von Skriptsprachen erzeugt, um dynamische Seiten zu realisieren, d. h. in dieser Form aufgebaute Hyperlinks sind unabhängig von `form`-Elementen zu betrachten. Dies wird von den absoluten Angaben belegt, denn derartige Hyperlinks können in insgesamt 1 101 021 Fällen beobachtet werden, was 2,41% aller Verweise entspricht.

A.4.7 Zur ursprünglichen Erstellung der HTML-Dokumente

Viele HTML-Editoren und Konvertoren (vgl. Abschnitt 3.3.6) hinterlassen in `<meta name="generator" content="...">` einen Fingerabdruck, der die verwendete Software identifiziert. In 1 684 337 Dokumenten des Korpus (43,195%) befinden sich diesbezüglich insgesamt 8 076 unterschiedliche Identifikationen. Turau findet in 10,5% aller Dateien 893 unterschiedliche Informationen (13,1% im Bereich „deedu“). Tatsächlich hinterlässt aber nicht jede Software zur Be- oder Verarbeitung von HTML-Dokumenten eine solche Angabe, häufig ist diese Funktion auch deaktivierbar, weshalb der Anteil manuell mittels ASCII-Editoren gepflegter Webseiten deutlich geringer sein dürfte als 56,81%. Bereits diese absoluten Zahlen zeigen jedoch, dass mittlerweile sehr viel häufiger HTML-Editoren und Konverter eingesetzt werden als noch 1998. Dieser gestiegene Bedarf ist unter anderem auf zahlreiche technische

[35] Die Terminologie ist verwirrend: *LiveScript* hat nach der Veröffentlichung des *Netscape Navigator* 2.0 schnell eine große Akzeptanz gefunden. Im November 1996 hat die *European Computer Manufacturers Association* (ECMA) mit der Standardisierung dieser Sprache (die nun von *Netscape* als *JavaScript* bezeichnet wurde) begonnen, deren erste Fassung im Juni 1997 verabschiedet wurde. ECMAScript wurde ab Version 1 ein ISO-Standard und liegt derzeit in Version 3 vor (ECMAScript, 1999). Da ECMAScript ein internationaler Standard ist, gelten *JavaScript* und *JScript* als die beiden wesentlichen Implementierungen der Standarddefinition.

[36] Zum Vergleich die Angaben zu „com“: 49,0%, „edu“: 9,3%, „de“: 8,6%, „decom“: 10,1% und „jp“: 29,7%.

Editor/Konverter	Vorkommen	Versionen	Prozent	Turau (1998a)
Netscape / Mozilla	437 346	3 730	25,97	24,6
Microsoft FrontPage	384 704	34	22,84	35,9
Microsoft Internet Assistant	312 951	25	18,58	4,0
Microsoft Word	101 582	39	6,03	—
GoLive CyberStudio	49 436	18	2,94	2,6
NetObjects	42 151	45	2,50	2,4
LaTeX2HTML	28 253	16	1,68	—
Phase 5	25 918	27	1,54	—
Adobe Page Mill	25 117	7	1,49	10,0
CatalogBuilder	19 732	1	1,17	—
Microsoft Powerpoint	17 466	4	1,04	—
Adobe Photoshop	15 148	5	0,90	—
StarOffice	14 131	35	0,84	—
UDO6	11 586	10	0,69	—
Modular DocBook Stylesheets	11 186	21	0,66	—
Claris Home Page	9 390	6	0,56	5,6
Visual Page	9 059	14	0,54	—
SGML-Tools	7 850	3	0,47	—
Tidy	7 463	14	0,44	—
Corel WordPerfect	3 565	17	0,21	—

Tabelle A.28: Die 20 meistbenutzten HTML-Editoren bzw. Konvertierungsprogramme

Neuerungen und einen erweiterten Funktionsumfang der Browser zurückzuführen, der eine manuelle Erstellung anspruchsvoller Dokumente immer komplexer werden lässt.

Tabelle A.28 listet die 20 meistbenutzten Werkzeuge auf, die innerhalb von `meta`-Elementen gefunden wurden. Die Spalte „Versionen“ gibt an, wie viele unterschiedliche Versionen einer speziellen Software in die enthaltenen Angaben eingeflossen sind.[37] Der in der mittleren Spalte angegebene prozentuale Anteil bezieht sich auf die 1 684 337 Dokumente, in denen eine Generator-Angabe vorliegt. Die rechte Spalte zeigt zum Vergleich sieben der acht Angaben von Turau (1998a), die Software *InfoOffice* (bei Turau: 3,2%) wurde im Korpus in lediglich 33 Dokumenten eingesetzt und ist daher nicht in der Tabelle vertreten. Der Vergleich mit den von Turau präsentierten Daten verdeutlicht zahlreiche interessante Aspekte: Die unterschiedlichen *Netscape Communicator*- bzw. *Mozilla*-basierten Editoren haben einen praktisch unveränderten Anteil. Der Grund für die große Verbreitung liegt vermutlich in der Verfügbarkeit dieser Software, die kostenlos zum Download bereitsteht bzw. unter einer Open-Source-Lizenz erhältlich ist. *Microsoft FrontPage* wird im Vergleich zu 1998 deutlich seltener eingesetzt, liegt mit ca. 23% jedoch noch immer an zweiter Stelle. Da in den Hochschulen viele Arbeitsplatzrechner das *Microsoft Office*-Paket enthalten, ist häufig „automatisch“ *MS FrontPage* verfügbar. Die Vorkommen des *Microsoft Internet Assistant* stammen aus älteren *Office*- bzw. Anwendungs-Paketen, in denen der HTML-Export-Filter separat heruntergeladen werden musste; etwa 300 000 der 312 951 Vorkommen gehen auf ältere

[37] Die 3 730 unterschiedlichen Versionen der Software-Gruppe „Netscape / Mozilla“ sind in der Tatsache begründet, dass diese Software die Version, eine Lokalisierungsangabe, das verwendete Betriebssystem und deren Prozessorarchitektur notiert. Beispielsweise wurden allein für das UNIX-Derivat IRIX 184 unterschiedliche Angaben gefunden, z. B. `Mozilla/4.75C-SGI [en] (X11; I; IRIX64 6.5 IP27) [Netscape]`, `Mozilla/3.0b6Gold (X11; I; IRIX 5.3 IP22) [Netscape]` und `Mozilla/4.03j1 [en] (X11; I; IRIX 6.3 IP32) [Netscape]`.

Powerpoint-Versionen zurück, die restlichen auf *MS Word*, d. h. der eigentliche Anteil von *Powerpoint*-Präsentationen, die zur Veröffentlichung im Web nach HTML konvertiert wurden, ist deutlich höher, als dies die lediglich 17 466 expliziten Vorkommen von *Microsoft Powerpoint* zunächst suggerieren. Bemerkenswert sind auch die beinahe 30 000 Vorkommen von HTML-Dokumenten, die mit Hilfe des Konverters *LaTeX2HTML* erzeugt wurden und die in den meisten Fällen auf wissenschaftliche Artikel und technische Berichte zurückzuführen sind. Da ältere Versionen dieses Werkzeugs keine Identifikation in einem `meta`-Element hinterlassen, sondern SGML-Kommentare für diesen Zweck einsetzen, dürfte deren tatsächlicher Anteil noch höher sein. Die *SGML-Tools* sind eine Sammlung Open-Source-basierter Werkzeuge zur Verarbeitung und Konvertierung von SGML-Dokumentinstanzen, die für *Linux*-Dokumentationen und Handbücher zusammengestellt wurde. *Tidy* (vgl. Abschnitt 14.4) beinhaltet die Funktion, HTML-Dokumente, die von *Microsoft Word* generiert wurden, zu bereinigen, d. h. das proprietäre Markup zu entfernen, das nur in Verbindung mit dem *Internet Explorer* optische Ergebnisse erzielt. Dass die Software einen Fingerabdruck hinterlässt, muss allerdings explizit konfiguriert werden, weshalb der Anteil ursprünglicher *MS Word*-Dokumente, die mittels *Tidy* nachbearbeitet wurden, erheblich höher sein dürfte.

A.4.8 Einsatz unterschiedlicher HTML-Versionen

HTML ist eine SGML-Applikation, weshalb streng genommen jede Datei über eine Dokumenttyp-Deklaration verfügen müsste, die die verwendete DTD über einen *Formal Public Identifier* (FPI) referenziert (ISO 8879). Tatsächlich enthalten nur lediglich 1 442 628 der Dokumente (36,99%) eine HTML-Deklaration. Keiner der verbreiteten Browser ist auf eine formale Deklaration angewiesen, stattdessen wird die Strategie verfolgt, auch fehlerhafte Markup-Strukturen verarbeiten und darstellen zu können.

Innerhalb des *Perl*-Moduls `HTML::Parser` wird ein Ereignis ausgelöst, falls eine HTML-Deklaration vorliegt. Aufgrund dieser Information wurden insgesamt 6 239 unterschiedliche FPIs, von denen jedoch etwa 4 100 eine fehlerhafte Syntax oder ungültige Angaben enthalten. Es existieren nur wenige standardisierte HTML-DTDs, mehrere diskutierte Versionen wurden nie offiziell vom W3C unterstützt, z. B. HTML+ und HTML 3.0 (vgl. `http://www.w3.org/MarkUp/historical`); die standardisierten Versionen lauten 2.0 (RFC 1866), 3.2 (Raggett, 1997), 4.0 (Raggett et al., 1997) und 4.01 (Raggett et al., 1999).[38] Darüber hinaus existiert mit ISO 15445 (2000) auch ein ISO-Standard, der eine Teilmenge von HTML 4.0 darstellt und weitere Einschränkungen implementiert, so darf z. B. ein `h3`-Element nicht auf `h1` folgen, wenn zwischen ihnen kein `h2`-Element existiert. Neben den SGML-basierten HTML-Standards existiert mit XHTML (Pemberton, 2002) auch eine XML-Variante (vgl. Abschnitt 14.4). XHTML 1.0 wird flankiert von Version 1.1 (Altheim und McCarron, 2001), die insbesondere die Modularisierung funktionsbezogener Element- und Attributgruppen verfolgt (Baker et al., 2000, Altheim et al., 2001).

[38] HTML 2.0 wurde von der *HTML Working Group* der IETF entwickelt. Die Gruppe hat ihre Arbeit im September 1996 eingestellt; die zukünftige Standardisierung wurde vom W3C übernommen. Die Veröffentlichung von RFC 2854, das `text/html` definiert, hatte den Zweck, alle bis dato von der IETF bezüglich HTML publizierten Standards aufzuheben: „This document [...] defines the 'text/html' MIME type *by pointing to the relevant W3C recommendations*" [Hervorhebung hinzugefügt, G. R.]. Die auf diese Weise obsolet gewordenen Standards sind z. B. RFC 1866, RFC 1867, RFC 1942, RFC 1980 und RFC 2070.

	Formal Public Identifier	Standard	Anzahl	Prozent
1.	`-//w3c//dtd html 4.0 transitional//en`	HTML 4.0	491 706	34,08
2.	`-//w3c//dtd html 3.2//en`	HTML 3.2	206 297	14,30
3.	`-//ietf//dtd html//en`	HTML 2.0	176 575	12,24
4.	`-//w3c//dtd html 3.2 final//en`	HTML 3.2	123 279	8,55
5.	`-//w3c//dtd html 4.0//en`	HTML 4.0	38 810	2,69
6.	`-//w3c//dtd html 3.2//de`	—	32 864	2,28
7.	`-//w3c//dtd html 4.01 transitional//en`	HTML 4.01	49 675	3,44
8.	`-//ietf//dtd html 2.0//de`	—	19 897	1,38
9.	`-//ietf//dtd html 3.0//de`	—	17 517	1,21
10.	`-//ietf//dtd html 3.2//en`	—	16 289	1,13
11.	`-//w3c//dtd html//en`	—	15 909	1,10
12.	`-//w3c//dtd html 3.2 final//de`	—	15 206	1,05
13.	`-//w3o//dtd w3 html 2.0//en`	—	14 754	1,02
14.	`-//w3c//dtd html 4.0 //en`	HTML 4.0	14 736	1,02
15.	`-//ietf//dtd html//en//3.2`	—	7 899	0,55
16.	`-//w3c//dtd html 4.0 frameset//en`	HTML 4.0	7 550	0,52
17.	`-//ietf//dtd html 2.0//en`	HTML 2.0	7 527	0,52
18.	`-//w3c//dtd xhtml 1.0 transitional//en`	XHTML 1.0	6 865	0,48
19.	`-//advasoft//dtd html 3.2 extended 961018//en`	—	5 718	0,40
20.	`-//sq//dtd html 2.0 hotmetal + extensions//en`	—	5 520	0,38
21.	`-//sq//dtd html 2.0 + alle erweiterungen//en`	—	5 387	0,37
22.	`-//w3c//dtd html 3.2 //en`	HTML 3.2	5 112	0,35
23.	`-//ietf//dtd html 3.0//en`	—	4 874	0,34
24.	`-//w3c//dtd html 4.01 frameset//en`	HTML 4.01	4 473	0,31
25.	`-//ietf//dtd html//de`	—	4 256	0,30

Tabelle A.29: In HTML-Deklarationen gefundene *Formal Public Identifier*

Tabelle A.29 zeigt die 25 häufigsten FPIs, die im Korpus gefunden wurden.[39] Die gelegentlich auch als Loose-DTD bezeichnete, in HTML 4.0 definierte Transitional-DTD wird in ca. einer halben Million Dokumente eingesetzt.[40] HTML 4.0 und 4.01 spezifizieren jeweils drei DTDs: Die Strict-DTD enthält alle Elemente und Attribute, die nicht den Status *deprecated* (abgelehnt, veraltet) besitzen oder nicht in Frameset-basierten Dokumenten enthalten sein dürfen. Die Transitional-DTD enthält im Unterschied zur Strict-DTD alle Präsentationselemente und umfasst auch veraltete Konstrukte. Die Frameset-DTD unterstützt Framesets. Die Strict-DTD stellt die Grundeinstellung dar, d. h. der an der fünften Position aufgeführte FPI referenziert die Strict-DTD. Aus der Tabelle geht hervor, dass etwa die Hälfte aller FPIs keine Entsprechung in den IETF- bzw. W3C-Standards besitzen, d. h. nicht korrekt sind. Der FPI an Position 19 geht auf den HTML-Editor *ASWedit* zurück, der erstmals im Juni 1995 von der Firma *AdvaSoft* für UNIX-Plattformen angeboten wurde. Die FPIs an den Positionen 20 und 21 stammen von dem Editor *HoTMetaL* der Firma *SoftQuad*.

Viele der ungültigen FPIs enthalten an letzter Stelle die Zeichenkette `de` statt `en`. Um diesen Umstand erklären zu können, ist zunächst eine Darstellung des generischen Formats einer HTML-Deklaration notwendig. Innerhalb einer Dokumenttyp-Deklaration wird die verwendete DTD anhand eines symbolischen FPIs oder eines Dateinamens referenziert, der

[39] Die Zeichenketten wurden auf konsistente Kleinschreibung normalisiert.

[40] Boldi et al. (2002) treffen eine derartige Unterscheidung nicht. Sie berichten für etwa 2 000 000 Dokumente aus dem afrikanischen Web, dass 78,81% keine HTML-Deklaration enthalten. Die Verteilung der Versionen lautet: HTML 4: 7,71%, HTML 3: 4,16%, HTML 2: 2,07%, HTML 1: 0,03%, Sonstige: 7,11%.

als *Formal System Identifier* (FSI) bezeichnet wird.[41] Sobald DTDs benutzt werden, die eine standardisierte Bezeichnung besitzen und die Dokumentinstanzen ausgetauscht werden sollen, sollte ein FPI verwendet werden, weil dieser eine gewisse Plattformunabhängigkeit garantiert, da die Dateien, die DTDs enthalten, auf unterschiedlichen Rechnern meist unterschiedliche Dateinamen besitzen, wohingegen der *Formal Public Identifier* als symbolischer Name standardisiert ist (vgl. Grosso, 1997, und RFC 3151). Der generelle Aufbau des FPI wird durch den SGML-Standard (ISO 8879) festgelegt: Durch – bzw. + wird zu Beginn repräsentiert, ob es sich um einen registrierten oder unregistrierten Besitzer handelt („registered" bzw. „unregistered owner identifier"); als Besitzer wird die Person, Gruppe von Personen, Firma oder Organisation angesehen, die für die Pflege der markierten Entität verantwortlich ist. Darauf folgt die Kennung des Besitzers, woraufhin – wiederum abgetrennt durch // – die „public text class" folgt, im Falle von DTDs ist dies immer `DTD`.[42] Es folgt die „public text description", in der erneut das Wurzelelement bzw. die Kurzbezeichnung der DTD sowie Versionsinformationen enthalten sein können. Nach dem letzten Trennelement (//) folgt die Angabe einer Länderkennung nach ISO 639, die spezifiziert, in welcher Sprache der „public text" (in diesem Fall die referenzierte DTD) verfasst wurde. Abschließend kann optional die „public text display version" folgen, die Angaben über unterstützte Geräte oder verwendete Kodierungsschemata macht. Es wird deutlich, dass die fehlerhaften Angaben bezüglich der Sprache (`//de` und `//en`) auf einem Missverständnis seitens der Autoren der betroffenen HTML-Dokumente beruhen. Offenbar vermuten diese, dass innerhalb des FPI die Sprache, in der ein HTML-Dokument verfasst wurde, hinterlegt werden kann, tatsächlich handelt es sich aber um eine standardisierte Angabe, die nicht modifiziert werden darf. Weiterhin fällt auf, dass weder das W3C noch die IETF als „registrierte Besitzer" gelten, da zu Beginn der standardisierten FPIs ein – statt eines + einzuesetzen ist. ISO 9070 zufolge kann sich eine Organisation bei der zuständigen Registrierungsstelle offiziell als „public text owner" registrieren lassen und einen standardisierten „owner identifier" erhalten. Es ist unklar, weshalb weder das W3C noch die IETF diesen Weg verfolgt haben. Laut ISO 9070 wird jedoch lediglich der „owner identifier" registriert, d. h. der Besitzer ist selbst für die „public text description" verantwortlich, weshalb die Angaben in den HTML-Standards verbindlich sind.

A.4.9 Verwendung von Cascading Style Sheets

Cascading Style Sheets (Lie und Bos, 1997, Bos et al., 1998) werden eingesetzt, um globale Layout-Informationen mittels einer zentral gepflegten Datei an HTML-Elemente zu binden. Neben der Möglichkeit, diese Angaben in einer separaten Datei zu hinterlegen existieren auch die beiden Alternativen, CSS-Informationen mittels des `style`-Elements unmittelbar in einem HTML-Dokument zu notieren (und hierdurch eine eventuell referenzierte CSS-Datei

[41] Ein Beispiel für einen FPI, der in einer Dokumenttyp-Deklaration benutzt wird, lautet `<!doctype html public "-//w3c//dtd html 4.0 transitional//en">`. Zunächst wird mit Hilfe des Schlüsselwortes `doctype` die nachfolgende Zeichenkette als Dokumenttyp-Deklaration identifiziert. Es folgt der Name der DTD, der zugleich ihr Wurzelelement markiert. Anschließend wird zwischen FPI und FSI unterschieden. Der letzte Bestandteil der Deklaration referenziert entweder einen Dateinamen oder einen abstrakten symbolischen Namen.

[42] Die weiteren Klassenbezeichner lauten `CAPACITY`, `CHARSET`, `DOCUMENT`, `ELEMENTS`, `ENTITIES`, `LPD`, `NONSGML`, `NOTATION`, `SHORTREF`, `SUBDOC`, `SYNTAX` und `TEXT`. Falls das Zeichen – als Bezeichner für die „public text class" vorliegt, gilt die Bezeichnung als nicht verfügbar.

zu überschreiben) oder sie innerhalb des `style`-Attributs in den inhaltsorientierten HTML-Elementen aufzuführen. Von den drei unterschiedlichen Möglichkeiten der Einbindung von CSS-Informationen machen insgesamt 846 443 im Korpus enthaltene HTML-Dokumente Gebrauch (21,71%). Dabei werden durchschnittlich 14,39 CSS-Referenzierungen pro Dokument vorgnommen (Modus und Median: 1, Max.: 6 126). Die Werte für Modus und Median deuten an, dass der Einsatz separater CSS-Dateien präferiert wird, wohingegen der Durchschnittswert von 14,39 zeigt, dass auch das Element bzw. Attribut `style` häufig eingesetzt wird. Diese Vermutungen werden durch eine weitere Erhebung bestätigt: Am häufigsten werden externe CSS-Dateien benutzt, diese kommen in 474 404 Dokumenten vor (12,17%; Modus und Median: 1, Durchschnitt: 1,04, Max.: 35). Das `style`-Attribut wird in 367 416 Webseiten (9,42%) mit durchschnittlich 31,27 Vorkommen verwendet (Modus: 1, Median: 4, Max.: 6 126). Das `style`-Element wird eher selten eingesetzt, es konnte – in Verbindung mit dem Attribut-Wert-Paar `type="text/css"` – in lediglich 188 058 Dokumenten (4,82%) festgestellt werden (Modus und Median: 1, Durchschnitt: 1,03, Max.: 15). Die Summe dieser Gruppen zeigt, dass in einigen Dokumenten mehr als eine Alternative verwendet wird.

A.5 Charakteristika der XML-Dokumente

Mit Mignet et al. (2003) liegt eine Studie über die Verbreitung von XML-Dokumenten im WWW vor. Die Autoren untersuchen 190 417 öffentlich zugängliche Dateien, die dem Datenbestand eines Anbieters für Suchmaschinentechnologien entnommen wurden und von 19 254 Webservern gesammelt wurden. Etwa 53% der Dokumente stammen aus den *top-level*-Domänen `.com` und `.net`. In 48% der Dokumente wird eine DTD referenziert, und nur 0,09% (179 Dokumente) enthalten einen Verweis auf eine XML Schema-Beschreibung. Die Stichprobe enthält Verweise auf 75 DTDs, wobei sich 92% dieser Referenzierungen auf die *Wireless Markup Language* beziehen. Die von Mignet et al. untersuchten XML-Dateien sind mit einer durchschnittlichen Größe von 4 641 Byte relativ klein, wobei der Anteil des Markups ausgesprochen hoch ausfällt. Für diese Untersuchungen wurden die XML-Dokumente bezüglich ihrer Dateigröße in vier Cluster aufgeteilt, wobei der prozentuale Anteil des Markups im Verhältnis zur eigentlich enkodierten Information zwischen 90% und 52% beträgt. Weiterhin betrachten Mignet et al. die Verteilung von Dateisuffixen.

Die Korpusdatenbank enthält insgesamt 25 871 Dateien, die von 758 Webservern mit dem Medientyp `text/xml` an den *Crawler* geschickt wurden. Zahlreiche XML-Dateien, die korrekterweise diesen Medientyp besitzen sollten, wurden jedoch (unter anderem) als `text/html` oder `text/plain` versendet, was insbesondere für Dateien mit den Endungen `.wml`, `.rdf` und sogar `.xml` gilt. Diese Dateien wurden, ebenso wie Dateien, die auf `.xsl`, `.xslt`, `.svg` und `.rss` enden, aus dem Korpus extrahiert und in die Gruppe der XML-Dateien aufgenommen, die insgesamt 29 551 Dateien mit einem Umfang von ca. 200 Megabyte enthält. Mignet et al. (2003) geben nicht an, auf welcher Basis während des Aufbaus ihrer Stichprobe entschieden wurde, ob es sich um eine XML-annotierte Datei handelt. Die im Folgenden dargestellten Daten wurden mit einem *Perl*-Skript erhoben, das das Modul `XML::Parser` einsetzt, das wiederum den nicht validierenden Parser *expat* kapselt (vgl. Abschnitt 14.4).

A.5.1 Überprüfung der Dokumentinstanzen auf Wohlgeformtheit

Insgesamt 28 462 der 29 551 Dateien (96,3%) stellen wohlgeformten XML-Code dar, 1 028 (3,5%) Dateien wurden von *expat* mit einer oder mehreren Fehlermeldungen zurückgewiesen; bei 61 Dateien (0,2%) handelt es sich um DTDs. Tabelle A.30 zeigt die Anzahl der von *expat* gemeldeten Fehler, wobei die zahlreichen Vorkommen des Fehlers *mismatched tag* auffallen, die aus einer unerlaubten Elementschachtelung resultieren. Da die meisten XML-Dateien mit maschineller Unterstützung angefertigt werden – entweder mit Hilfe eines validierenden Editors oder durch eine vollständige maschinelle Erzeugung –, verwundert gerade dieser Fehler, der durch den Einsatz von XML-Software eigentlich nicht auftreten dürfte. Seine Ursache liegt jedoch in fast allen Fällen in nicht deklarierten Entitäten oder falsch deklarierten Zeichensatzangaben: *expat* findet z. B. im Inhalt eines Elements `title` einen in ISO Latin 1 kodierten Umlaut, der einen Konflikt mit der Zeichensatzgrundeinstellung Unicode hervorruft (gemeldet wird ein `bad token`), woraufhin das schließende `title`-Element ignoriert wird. Das nächste gefundene Element ist in diesem Beispiel ein schließendes Tag, so dass die Fehlermeldung *mismatched tag* vorliegt.

Fehlermeldung	Vorkommen	Fehlermeldung	Vorkommen
mismatched tag	536	no element found	19
not well-formed (invalid token)	273	undefined entity	15
syntax error	117	junk after document element	12
xml declaration not at start of external entity	26	unknown encoding	2
Handler couldn't resolve external entity	20	duplicate attribute	1
error in processing external entity reference	20		

Tabelle A.30: Fehlermeldungen des XML-Parsers *expat*

A.5.2 Dokumenttyp-Definitionen und XML Schema-Beschreibungen

Von den XML-Dateien besitzen 16 892 (59,3%) eine XML-Deklaration, die 1 892 `encoding`- und 831 `standalone`-Attribute umfassen (6,6% bzw. 2,9%). Unicode scheint sich aufgrund der überraschend geringen Anzahl expliziter Zeichensatzangaben als Grundeinstellung durchgesetzt zu haben. In 56,1% aller Dateien wurden 15 967 Referenzen auf DTDs gefunden (bei Mignet et al.: 48%). 15 833 der Referenzen (99,1%) verweisen auf einen FSI (Angabe eines Dateinamens) und nur 778 Referenzen (4,9%) auf einen FPI (Referenzierung durch einen symbolischen Namen, vgl. Abschnitt A.4.8). In einigen Dateien waren beide Angaben enthalten, weshalb sich in der Summe nicht 15 967 Vorkommen ergeben.

Mignet et al. (2003) berichten, dass in den 190 417 untersuchten XML-Dateien nur 75 DTDs referenziert werden, wobei sie keine Angaben zur Unterscheidung zwischen FSI und FPI machen. 92% aller DTD-Verweise beziehen sich dabei auf die „norms 1.1 or 1.2 of the WAP protocol“, womit vermutlich die korrespondierenden Versionen der *Wireless Markup Language* gemeint sind. Im XML-Web des deutschen Wissenschaftsnetzes sind andere Typen vorherrschend – insgesamt wurden 33 unterschiedliche DTDs innerhalb von FPIs gefunden (vgl. Tabelle A.31). Die Abwesenheit von WML-Dateien (es sind insgesamt lediglich 12 derartige Dateien enthalten, die sich auf drei DTDs beziehen) lässt sich dadurch erklären, dass

Formal Public Identifier	Anzahl	Prozent
-//W3C//DTD HTML 4.0//EN	393	50,5141
-//W3C//DTD SVG 20000802//EN	132	16,9666
-//W3C//DTD SVG 20000303 Stylable//EN	59	7,5835
-//W3C//DTD HTML 4.0 Transitional//EN	47	6,0411
-//W3C//DTD SVG 1.0//EN	33	4,2416
-//W3C//DTD SVG 20001102//EN	22	2,8278
-//Sun Microsystems, Inc.//DTD Web Application 2.2//EN	15	1,9280
-//BEA Systems, Inc.//DTD WebLogic 5.1.0 EJB//EN	10	1,2853
-//Sun Microsystems, Inc.//DTD Enterprise JavaBeans 1.1//EN	10	1,2853
-//Uppaal Team//DTD Flat System 1.0//EN	9	1,1568
-//W3C//DTD HTML 4.01 Transitional//EN	7	0,8997
-//WAPFORUM//DTD WML 1.1//EN	7	0,8997
-//UNIBW Inf//DTD UniBw Projekt Studium//EN	6	0,7712
-//W3C//DTD XHTML 1.0 Strict//EN	6	0,7712
+//IDN python.org//DTD XML Bookmark Exchange Language 1.0//EN//XML	5	0,6427
-//W3C//DTD XHTML 1.0 Transitional//EN	5	0,6427
-//BEA Systems, Inc.//DTD WebLogic 5.1.0 EJB RDBMS Persistence//EN	5	0,6427
-//Sun Microsystems Inc.//DTD JavaHelp TOC Version 1.0//EN	4	0,5141
http://www.web3D.org/TaskGroups/x3d/translation/x3d-compromise.dtd	4	0,5141
-//Norman Walsh//DTD Simplified DocBk XML V3.1.7.1//EN	3	0,3856
-//TEI//DTD TEI Lite XML ver. 1.3//EN	3	0,3856
-//WAPFORUM//DTD WML 1.1//EN	3	0,3856
-//W3C//DTD SVG April 1999//EN	2	0,2571
-//Netscape Communications//DTD RSS 0.91//EN	2	0,2571
-//Norman Walsh//DTD DocBk XML V3.1.7//EN	2	0,2571
-//St. Andrew's Hospital//DTD for Patient Administration V2.1//EN	2	0,2571
-//Sun Microsystems Inc.//DTD JavaHelp Index Version 1.0//EN	2	0,2571
-//Sun Microsystems, Inc.//DTD Enterprise JavaBeans 2.0//EN	2	0,2571
-//Norman Walsh//DTD Image Library 1.0//EN	2	0,2571
-//WAPFORUM//DTD WML 1.1//EN	2	0,2571
-//TUD//KOM//DTD lipi's WORKLOG 1.0 //EN	2	0,2571
-//Sun Microsystems, Inc.//DTD Enterprise JavaBeans 1.2//EN	2	0,2571
-//BEA Systems, Inc.//DTD Enterprise JavaBeans 1.1//EN	2	0,2571

Tabelle A.31: Die als *Formal Public Identifier* referenzierten Dokumenttyp-Definitionen

die Ermöglichung des Zugriffes auf hochschulbezogene Informationen mit mobilen Endgeräten zum Zeitpunkt des Korpusaufbaus innerhalb der Universitäten keine hohe Priorität besaß.[43] Mit dem zunehmenden Einsatz von E-Learning-Anwendungen hat sich dieser Umstand mittlerweile geändert, da hierbei der ortsunabhängige Zugriff auf Informationen eine zentrale Rolle spielt; da WAP und WML aber nur eine Art Übergangstechnologie darstellten, werden diese nur noch selten eingesetzt.

In Tabelle A.31 sind insbesondere die HTML-DTDs auffällig, deren Vorkommen sich prinzipiell nicht in XML-, sondern in SGML-Instanzen befinden, da nicht auf die XHTML-DTD Bezug genommen wird. Einige Server der TU Dresden enthalten HTML-Dokumente mit dem Dateisuffix `.xml`, weshalb der Webserver diese dem Client gegenüber als `text/xml` identifiziert. Neben der hohen Anzahl von SVG-Dateien sind vor allem mehrere Versionen der *Enterprise Java Beans* (EJB) enthalten. Hervorzuheben ist weiterhin die sehr geringe Anzahl

[43] In der Korpusdatenbank sind lediglich drei Webserver enthalten, deren Namen mit `wap` beginnen: `wap.physik.uni-dortmund.de`, `wap.uni-mannheim.de` sowie `wapserv.chemie.uni-halle.de` und nur bei einem dieser Rechner handelt es sich um den offiziellen WAP-Server einer Universität.

TEI-basierter XML-Dokumente. Ihre Abwesenheit lässt sich eventuell ebenfalls durch Verzerrungen erklären, die von fehlerhaft fehlerhaft gesetzten Medientypen verursacht werden.

In 99,1% aller Deklarationen werden FSIs benutzt, d. h. die DTD wird als voll qualifizierter Dateiname angegeben; insgesamt weden in den 29 551 XML-Dateien 265 unterschiedliche DTDs durch einen FSI referenziert. Den mit Abstand größten Anteil an diesen 15 833 Referenzierungen hat `vokoxml.dtd` (14 176 Vorkommen, 89,53%), die auf mehreren Webservern der Universität Leipzig zur Annotation eines umfangreichen Esperanto-Lexikons eingesetzt wird (`http://www.uni-leipzig.de/esperanto/voko/revo/`). Eine weitere Gruppe häufig vorkommender XML-Dateien enthält Verweise auf lokal vorhandene W3C-Standards, z. B. HTML und SVG in unterschiedlichen Versionen (etwa 800 Vorkommen, ca. 4%). Die dritte Gruppe von XML-Instanzen, die lokale DTDs referenzieren, stammt aus dem DFG-Projekt „Digitale Dissertationen im Internet“ (`http://www.educat.hu-berlin.de/diss_online/`). In diesem wird eine an HTML angelehnte `artikel.dtd` verwendet, um projektbezogene Inhalte im WWW mit einem einheitlichen Layout zu publizieren (150 Vorkommen, 0,95%). Es folgen 70 Vorkommen (0,44%) der `jvx.dtd` zur Repräsentation geometrischer Modelle (vgl. `http://www.eg-models.de`), 53 Referenzierungen (0,33%) einer Metadaten-DTD des *IMS Global Learning Consortium* (`http://www.imsproject.org/xml/`), die jedoch nicht aus E-Learning-Materialien, sondern aus Beispieldateien stammen, sowie 29 Vorkommen (0,18%) der `wbqs.dtd` (*Web Based Quiz System*, `http://wwwagse.informatik.uni-kl.de/teaching/swp/ss2002/`), die an der Universität Kaiserslautern zur Strukturierung von Multiple-Choice-Tests in einem E-Learning-Szenario entwickelt wurde.

Lediglich 35 Dokumente (0,12%) referenzieren eine XML Schema-Beschreibung. Mignet et al. (2003) geben für ihre Stichprobe einen Wert von 0,09% an (179 Dokumente). Zur Bestimmung dieses Wertes wurde mit der Überprüfung der Vorkommen von Attributen namens `SchemaLocation` bzw. `noNameSpaceSchemaLocation` derjenige Test durchgeführt, der auch von Mignet et al. eingesetzt wurde. Die wenigen referenzierten Schema-Beschreibungen umfassen Metadaten-Schemata des IMS, ein Schema zur Korpusannotation des Instituts für maschinelle Sprachverarbeitung der Universität Stuttgart (`http://www.ims.uni-stuttgart.de/projekte/TIGER/`), Beispieldateien aus einem Lehrbuch zum Thema XML sowie aus einer Lehrveranstaltung und abschließend eine exemplarische Dokumentinstanz zur strukturierten Erfassung des Lehrangebots der Universität München.

A.5.3 Dateigröße – Verhältnis von Markup zu Information

Die 190 417 von Mignet et al. untersuchten Instanzen besitzen Dateigrößen zwischen 10 und 500 608 Bytes, wobei sich ein Durchschnitt von 4 641 Bytes ergibt. Die im Korpus enthaltenen Dateien sind mit diesen Angaben vergleichbar: Die Dateigrößen bewegen sich zwischen 29 und 475 989 Bytes mit einem Durchschnitt von 2 814 Bytes (Median: 734, Modus: 164). Mignet et al. teilen die Dateien in vier Cluster auf:

- C_1: $<$ 512 Bytes – 48 671 Dokumente, 60% (29 556 Dateien) mit dem Suffix `.xml`, 12% (5 871 Dateien) mit der Endung `.wml`
- C_2: 512–1024 Bytes – 39 449 Dokumente, 62% (24 500 Dateien) mit dem Suffix `.wml`, 20% (8 035 Dateien) mit der Endung `.xml`

- C_3: 1024–4096 Bytes – 69 846 Dokumente, 30% (21 403 Dateien) mit dem Suffix `.wml`, 36% (25 115 Dateien) mit der Endung `.xml` und 16% (11 765 Dateien) mit den Suffixen `.rdf` oder `.rss`
- C_4: > 4096 Bytes – 32 361 Dokumente, 58% (18 733 Dateien) mit den Suffixen `.rdf` oder `.rss`, 37% (12 156 Dateien) mit der Endung `.xml` und 1% (356 Dateien) mit dem Suffix `.wml`

WML-Dateien sind meist relativ klein – ca. 88% befinden sich in C_1 oder C_2, wohingegen RDF- bzw. RSS-Dokumente umfangreicher sind, da ca. 96% aus C_3 oder C_4 stammen. Für andere Dokumenttypen ergeben sich nach Mignet et al. keine derartigen offensichtlichen Zuordnungen. Die Cluster-Verteilungen zeigen, dass etwa 46% der Dateien kleiner sind als 1 024 Bytes. Im Falle der XML-Instanzen des Korpus sind dies etwa 57% (16 972 Dateien), wobei 36% (10 741 Dateien) kleiner sind als 512 Bytes. Umfangreiche Instanzen sind kaum im Korpus enthalten: Lediglich 111 Dateien (0,38%) sind größer als 64 Kilobyte.

Mignet et al. untersuchen auch das Verhältnis von Markup zu enkodierter Information, d. h. den prozentualen Anteil von Elementen und Attributen an der Größe einer Instanz. Für C_1 bis C_4 geben die Autoren 90%, 72%, 68% und 52% an (ein Durchschnittswert wird nicht genannt). Diese Werte decken sich mit den Daten, die bei der Auswertung der im Korpus enthaltenen Instanzen erhoben wurden: Das in den XML-Dateien enthaltene Markup hat einen durchschnittlichen Anteil von 80,42% mit einem Median von 82%. Das Minimum beträgt 0%, das Maximum 99% mit einem Modus von 95%.[44] Die von Mignet et al. gezogene Schlussfolgerung, dass die strukturellen Informationen die textuellen Daten eindeutig dominieren, kann also bestätigt werden.

A.5.4 Verteilung unterschiedlicher Dateisuffixe

Mignet et al. (2003) untersuchen die Verteilung von Dateisuffixen, wobei erwartungsgemäß `.xml` (29,70%) und `.wml` (26,55%) am häufigsten vorkommen. Mit 21,08% wird die Suffix-Klasse „Form“ angegeben, womit „form-accessible documents“ gemeint sind.[45] Die Endungen `.rdf` und `.rss` besitzen zusammen 16,69% Vorkommen, wohingegen `.xsl` und `.xslt` lediglich mit einem Anteil von 0,60% vertreten sind. Andere Endungen umfassen 5,38%.

Diese Angaben decken sich nur teilweise mit den Daten, die mit Hilfe der Korpusdatenbank erhoben wurden; die enthaltenen Suffixe, die Anzahl der Vorkommen sowie ihr prozentualer Anteil sind in Tabelle A.32 dargestellt. Die Liste der Endungen wird von `.xml` mit einem prozentualen Anteil von fast 97% dominiert, wohingegen die bei Mignet et al. mit 26,55% vertretene Endung `.wml` lediglich 0,06% einnimmt. XSLT-Stylesheets (`.xsl` und `.xslt`) sind bezüglich ihrer Vorkommen vergleichbar (0,60% bei Mignet et al. vs. 1,36%). Unklar bleibt, was bei Mignet et al. (2003) mit „form-accessible documents“ gemeint ist. Auch die in Tabelle A.32 aufgeführte Liste von Dateiendungen enthält keinen Hinweis auf diese Klasse von Dateiendungen.

[44] Das Minimum von 0% erklärt sich durch einige Dateien, die zwar den Dateisuffix `.rdf` besitzen, jedoch kein Markup enthalten, obgleich RDF-bezogene Inhalte vorhanden sind. Vermutlich handelt es sich um RDF-Dateien, aus denen das Markup entfernt wurde.

[45] Für diese Klasse von Dateiendungen werden keine Suffixe aufgeführt.

Dateisuffix	Vorkommen	Prozent	Dateisuffix	Vorkommen	Prozent
xml	28 566	96,67	wml	19	0,06
xsl	391	1,32	xslt	12	0,04
svg	271	0,91	php	3	0,01
rdf	126	0,43	html	2	0,01
jxv	105	0,36	xsd	2	0,01
dtd	62	0,21			

Tabelle A.32: Die Dateisuffixe der im Korpus enthaltenen XML-Instanzen

Dateiname	Vorkommen	Prozent	Dateiname	Vorkommen	Prozent
filelist.xml	8 877	30,04	dwsitecolumnsme.xml	19	0,06
pres.xml	2 008	6,78	index_filelist.xml	19	0,06
001_filelist.xml	126	0,43	web.xml	17	0,06
index.xml	100	0,34	foo.xml	12	0,04
master03.xml	97	0,33	links.xml	12	0,04
master04.xml	87	0,29	ejb-jar.xml	11	0,04
master05.xml	31	0,10	profil.rdf	10	0,03
001_pres.xml	24	0,08			

Tabelle A.33: Die 15 häufigsten Dateinamen der im Korpus enthaltenen XML-Instanzen

A.5.5 Verteilung unterschiedlicher Dateinamen

Das bislang gezeichnete Bild entspricht weitestgehend den von Mignet et al. berichteten Daten. Die Ausnahmen stellen die im Korpus hochfrequente Endung `.xml` sowie die wenigen Vorkommen von WML-Dokumenten dar. Auf einen zentralen Aspekt gehen Mignet et al. nicht ein: Die Verteilung unterschiedlicher Dateinamen, die Hinweise auf die Erzeugung von XML-Dateien liefern können. Mit 9 826 unterschiedlichen Dateinamen ist eine sehr große Varietät zu verzeichnen (vgl. Tabelle A.33). Die Aufstellung wird mit 8 877 Vorkommen von `filelist.xml` angeführt, von dem auch Varianten in der Liste enthalten sind. Diese Dateien werden von *Microsoft Office* angelegt, wenn eine Datei nach HTML exportiert wird. In `filelist.xml` befindet sich nach dem Export eine Liste der generierten Dateien, z. B. im Falle einer *Powerpoint*-Datei pro Folie eine Zeile der Form `<o:File HRef="slide0010_image099.gif"/>`. Die etwa 2 000 Dateien namens `pres.xml` sind auf *Powerpoint* zurückzuführen: Sie enthalten Informationen zur Darstellung von nach HTML exportierten *Powerpoint*-Präsentationen, die für die Anzeige im *Internet Explorer* optimiert sind. Es zeigt sich, dass mehr als ein Drittel der XML-Instanzen zwar wohlgeformtes XML-Markup enthalten, jedoch nur (proprietäre) Nebenprodukte der Konvertierung bzw. des Exports binärer Dateiformate nach HTML sind und somit streng genommen keine strukturierte Informationen im eigentlichen Sinn darstellen.

A.5.6 Einsatz von Namespaces

Ein ähnliches Bild zeichnet sich bezüglich des Einsatzes von Namespaces ab: In 12 120 Dateien (41,01%) werden 46 Namensräume in durchschnittlich 57,7 Elementen (Median: 16,

Namespace	Vorkommen	Prozent	Namespace	Vorkommen	Prozent
o	520 105	75,69	xlink	490	0,07
p	111 649	16,25	dc	476	0,07
v	34 705	5,05	xslt	450	0,07
xsl	13 369	1,95	html	364	0,05
rdf	1 130	0,16	fo	330	0,05
dieper	1 106	0,16	a	279	0,04
uml	997	0,14	rdfs	241	0,04
dcq	602	0,09			

Tabelle A.34: Die 15 häufigsten Namespaces der im Korpus enthaltenen XML-Dateien

Modus: 4) eingesetzt. Tabelle A.34 zeigt die 15 häufigsten Namespaces sowie deren Vorkommen in den Instanzen. Auffällig sind die drei ersten Namensräume o, p und v, die zusammen für ca. 97% aller Vorkommen verantwortlich sind. Die Deklarationen dieser Namensräume in den Wurzelelementen der referenzierenden Instanzen lauten:

```
xmlns:o="urn:schemas-microsoft-com:office:office"
xmlns:p="urn:schemas-microsoft-com:office:powerpoint"
xmlns:v="urn:schemas-microsoft-com:vml"
```

Auch hier dominieren *Microsoft*-Formate, die insbesondere beim Export von *Powerpoint*-Präsentationen zur Erzeugung von Instanzen eingesetzt werden, die lediglich vom *Internet Explorer* unterstützt werden. Dabei ist o der allgemeine *Office*-Namespace, p wird für *Powerpoint*-bezogene Objekte und Effekte verwendet, und v bezieht sich auf *Vector Markup Language*, die 1998 von *Microsoft* und einigen anderen Firmen als Alternative zur damals ebenfalls in der Planung befindlichen XML-Applikation *Scalable Vector Graphics* (SVG) entwickelt wurde.[46] Bei dieper handelt es sich um einen Namensraum, der im Rahmen des Projekts *Digitised European Periodicals*, http://gdz.sub.uni-goettingen.de/dieper/) gemeinsam mit RDF und *Dublin Core* (dc bzw. dcq) zur Auszeichnung von Metadaten verwendet wird. Der Präfix uml wird in mehreren Instanzen der Universität Würzburg als XML-Variante der *Unified Modelling Language* eingesetzt.

A.5.7 Events des XML-Parser-Moduls

Das für die Datenerhebung verwendete *Perl*-Modul XML::Parser gestattet die Ereignis-basierte Verarbeitung XML-annotierter Daten (vgl. Ray und McIntosh, 2002, S. 47 ff.). Dieses auch als Event-basierte Parsing bezeichnete Paradigma eignet sich insbesondere zur effizienten Verarbeitung großer Datenmengen, da die Instanz nicht vollständig in den Speicher eingelesen werden musss (vgl. Rehm, 2004e). Stattdessen wird sie als Datenstrom betrachtet, dessen spezielle Charakteristika (öffnende und schließende Elemente, CDATA-Abschnitte etc.) Ereignisse innerhalb auslösen. Diese Ereignisse können über Handler abgefangen werden, die an Funktionen gebunden werden, denen – abhängig vom auslösenden Ereignis – die in der Eingabe gefundenen Daten als Argumente übergeben werden (vgl. Abschnitt 14.4.2).

[46] Unter http://www.w3.org/TR/NOTE-VML ist die ursprüngliche „Note“ zu VML verfügbar, die im *World Wide Web Consortium* zwar diskutiert, aber nicht weiter verfolgt wurde.

Event	Auslöser	Vorkommen
`Char`	Non-Markup/Text	4 169 944
`Start`	Öffnendes Element	1 728 253
`End`	Schließendes Element	1 723 095
`XMLDecl`	XML-Deklaration	16 892
`Doctype`	DOCTYPE-Deklaration	15 967
`Comment`	Kommentar	11 712
`CdataStart`	Beginn eines CDATA-Abschnitts	1 118
`CdataEnd`	Ende eines CDATA-Abschnitts	1 118
`Element`	Deklaration eines Elements	809
`Proc`	Processing Instruction	598
`Attlist`	Deklaration einer Attributliste	399
`Entity`	Deklaration einer Entität	309
`ExternEnt`	Referenzierung einer externen Entität	21
`Notation`	Deklaration einer Notation	8
`Unparsed`	Deklaration einer nicht verarbeiteten Entität	2

Tabelle A.35: Die *expat*-Events während der Verarbeitung der 29 551 XML-Dateien

Das Modul `XML::Parser` erlaubt den Zugriff auf 20 Ereignisse, von denen die wichtigsten mit ihren ursprünglichen Namen in Tabelle A.35 aufgeführt sind, wobei die rechte Spalte die absolute Anzahl aller Vorkommen eines Ereignisses nach der Verarbeitung der ca. 30 000 XML-Dateien enthält. An oberster Stelle rangiert das `Char`-Ereignis mit mehr als 4,1 Millionen Vorkommen. Dies erscheint auf den ersten Blick ungewöhnlich, da der durchschnittliche Anteil des Markups einer XML-Instanz ca. 80% beträgt, jedoch wird das `Char`-Ereignis für fortlaufende Zeichenketten aus Gründen der Performanzmaximierung mittels eines Caching-Verfahrens wiederholt ausgelöst. An zweiter und dritter Stelle folgen die Ereignisse `Start` und `End`, die bei öffnenden bzw. schließenden Tags ausgelöst werden. Obwohl XML im Gegensatz zu SGML keine Tag-Minimierung gestattet, werden etwa 5 000 mehr öffnende Elemente registriert, was an den ca. 1 000 nicht wohlgeformten XML-Dokumenten liegt, die ebenfalls in die hier dargestellte Aufstellung einfließen (vgl. Abschnitt A.5.1).[47]

A.6 Verwandte Arbeiten

Im Folgenden werden die Hintergründe der im Laufe dieses Anhangs referenzierten Studien zusammengefasst, deren Ausrichtung häufig unter dem Schlagwort *Web Characterization* subsumiert wird. Es existieren vielfältige Zielsetzungen, die nur einen kleinen Ausschnitt der Arbeiten zur Charakterisierung des WWW darstellen. Mit den Ergebnissen statistischer Analysen werden z. B. Load-Balancing- und Caching-Algorithmen von Proxy-Servern optimiert (Douglis et al., 1997), weiterhin wurden sie – gemeinsam mit Analysen von Zugriffsprotokollen, Benutzerstudien und Messungen von Datenvolumina – im W3C zur Konzeptionierung des Protokolls *HTTP Next Generation* (HTTP-NG) eingesetzt (vgl. Pitkow, 1998).[48]

[47] Für ein leeres Element wie `</br>`, wird zunächst ein `Start`- und unmittelbar darauf ein `End`-Ereignis ausgelöst.

[48] Zum Themenkomplex *Web Characterization* existierte 1998 und 1999 eine Arbeitsgruppe im W3C, die jedoch ebenso wie die HTTP-NG-Initiative mittlerweile ihre Arbeit eingestellt hat (vgl. `http://www.w3.org/WCA/`, `http://www.w3.org/Protocols/HTTP-NG/` und Gourley et al., 2002, S. 247 ff.).

Bray (1996) untersucht die Größe des WWW und Charakteristika der „durchschnittlichen Seite“. Als Datenbasis werden ca. 1,5 Millionen Dokumente benutzt, die im November 1995 vom *Crawler* des *Open Text Index* (`http://www.opentext.com`) gesammelt wurden; insgesamt sind 11 366 121 eindeutige und mit `http`, `ftp` oder `gopher` beginnende Hyperlinks auf 223 851 unterschiedliche Server enthalten. Bray geht auch auf die „größten und sichtbarsten Websites“ ein. Da die Ergebnisse aus heutiger Sicht veraltet sind, wird auf eine Darstellung verzichtet. Die dreidimensionale Visualisierung der Resultate stellt jedoch nach wie vor eine interessante Alternative zur rein Text- bzw. Link-basierten Navigation im WWW dar.

Woodruff et al. (1996) benutzen eine 2,6 Millionen Dokumente umfassende Stichprobe, die das *Inktomi*-Projekt im November 1995 an der Universität Berkeley angefertigt hat. Es werden ausschließlich Charakteristika von Dokumenten thematisiert, die mit Hilfe verschiedener Werkzeuge analysiert werden. Die Dokumente der Stichprobe setzen sich zu etwa gleichen Teilen aus den *top-level*-Domänen `.edu` (27%), `.com` (20%), `.gov` (4%), `.net` (4%), `.org` (3%) und `.mil` (1%) sowie anderen *top-level*-Domänen wie z. B. `.de` zusammen (41%). Eine syntaktische Überprüfung einer ca. 92 000 Dateien umfassenden Teilmenge mittels des Werkzeugs *weblint* zeigt, dass etwa 40% aller HTML-Dokumente mindestens einen Fehler enthalten; die häufigsten Ursachen sind fehlende `html`- und `head`-Elemente. Diese Ursache dürfte heutzutage nicht mehr häufig in Erscheinung treten, da die meisten HTML-Dokumente mit maschineller Unterstützung durch spezielle HTML-Editoren angefertigt werden und daher meist standardkonformes Markup enthalten.

Im Gegensatz zu Bray und Woodruff et al. benutzen O’Neill et al. (1997) keine vorhandene Stichprobe, sondern stellen ein Verfahren vor, das eine zufällige Generierung von IP-Nummern beinhaltet, um auf diese Weise eine zufällige[49] Stichprobe von Webservern zu erhalten (vgl. O’Neill, 1997). O’Neill et al. (1998) stellen die Resultate einer Studie zur Benutzung von Metadaten im WWW vor, die auf einer Stichprobe von ca. 165 000 Dokumenten basiert, die im Juni 1998 erzeugt wurde. O’Neill et al. unterscheiden zwischen „home pages“, d. h. den Einstiegsseiten von Webservern, und „internal pages“, da Metadaten häufig in den Einstiegsseiten notiert werden und für das gesamte verlinkte Webangebot gültig sind. Die Verfasser untersuchen ausschließlich Metadaten, die mittels `meta` notiert werden, wobei in ca. 70% aller 1 457 Einstiegsseiten durchschnittlich 2,75 Elemente enthalten sind (vgl. Abschnitt A.4.4). O’Neill et al. (1998) kommen zu dem Schluss, dass Metadaten zwar sehr häufig im WWW vorkommen, jedoch in den meisten Fällen vom HTML-Editor automatisch hinzugefügt werden, wobei diejenigen Metadaten, die von Menschen eingetragen werden, meist nicht auf standardisierten Schemata wie *Dublin Core* basieren und häufig eine zu grobe Granularität bezüglich der verwendeten Verschlagwortungsmethodik besitzen, so dass sie letztlich für Suchmaschinen nur von begrenztem Wert sind.

Turau (1998a,b) untersucht 147 384 HTML-Dokumente, die aus verschiedenen Bereichen (z. B. Hochschule, Industrie, Computer, Automobilbau) und unterschiedlichen geografischen Regionen stammen (z. B. Europa, USA, Asien), wobei es insbesondere um die Untersuchung multimedialer Elemente, Client-seitiger Anwendungen und verschiedener Tech-

[49] Die Autoren merken an, dass eine derartige Stichprobe streng genommen nicht zufällig erhoben werden kann. Hierfür sind verschiedene technische Gründe verantwortlich, die sich z. B. auf Weiterleitungen, Duplikate, multiple IP-Nummern für einen Rechnernamen oder mehrere Webangebote für eine IP-Nummer beziehen, die unter verschiedenen virtuellen Rechnernamen erreichbar sind (vgl. Turau, 1998a).

niken wie Cookies oder dynamische Dokumente sowie den nach Herkunftsländern aufgeschlüsselten Vergleich dieser Eigenschaften geht. Turau benutzt einen in *Java* implementierten *Crawler*, der Webserver mit Hilfe einer Breitensuche traversiert und für jedes HTML-Dokument einen Datensatz mit etwa 60 Feldern anlegt, der anschließend ausgewertet wird. Für die meisten Merkmale gibt Turau Werte für den Bereich „deedu“ an, der sich auf die Webserver des deutschen Wissenschaftsnetzes bezieht, die mit einfachen Mustervergleichen bestimmt wurden.[50] Turau (1998a) stellt meines Wissens die einzige Studie dar, die sich mit statistischen Analysen in diesem Bereich des Web beschäftigt.

Neben diesen verwandten Arbeiten existiert ein weiterer Themenkomplex, der häufig als „Informetrics“, „Cybermetrics“, „Cyberinformetrics“, „Webography“ oder „Webometrics“ (Almind und Ingwersen, 1997) bezeichnet wird, innerhalb dessen Korpora von Webseiten – insbesondere deren Hyperlinkstrukturen – statistisch untersucht werden. Generell beschäftigen sich diese Analysen, die traditionell meist auf bibliometrischen Untersuchungen aufbauen, mit quantitativen Aspekten informationsverarbeitender Prozesse. Im WWW stehen insbesondere die Graphenstrukturen, die in sehr großen Dokumentkollektionen von Hyperlinks aufgespannt werden, im Zentrum des Interesses (Broder et al., 2000), weshalb z. B. Methoden der Zitationsanalyse eingesetzt werden (vgl. Lawrence et al., 1999, Chakrabarti et al., 1999). Untersucht werden Fragestellungen, die z. B. die Größe des Webs, den Abdeckungsgrad von Suchmaschinen und digitalen Bibliotheken, Benutzerstudien in spezifischen Suchkontexten und die Entdeckung von Communitys betreffen (vgl. Bar-Ilan, 2001).

Im Kontext der Analyse länderspezifischer Schnappschüsse der Webserver wissenschaftlicher Netzwerke sind insbesondere die Arbeiten von Thelwall von Interesse, (z. B. Thelwall, 2001a, Thelwall und Harries, 2003, Park und Thelwall, 2003, Thelwall und Wilkinson, 2003, und Thelwall, 2005). Es wurde ein *Crawler* konzipiert (Thelwall, 2001b), der für Domänen wie z. B. `.ac.uk`, also die wissenschaftlichen Webserver Großbritaniens, die in den entsprechenden HTML-Dokumenten enthaltenen Hyperlinkstrukturen protokolliert, die daraufhin statistisch analysiert werden können. Mehrere dieser Datensammlungen stehen – z. B. für die wissenschaftlichen Netze Australiens, Neuseelands, Spaniens und Chinas – unter `http://cybermetrics.wlv.ac.uk` zur Verfügung. Thelwall und Harries (2003) zeigen mit derartigen Analysen beispielsweise, dass Korrelationen zwischen der in den HTML-Dokumenten einer bestimmten Universität enthaltenen Hyperlinkstruktur und ihrer Forschungsproduktivität existieren. Mit Sanguanpong et al. (2000) liegt eine Analyse des thailändischen Webs vor, dass im März 2000 etwa 700 000 Dokumente auf 8 000 Webservern umfasst. Neben der Analyse HTTP- und HTML-bezogener Eigenschaften stellen die Autoren auch den *Crawler* vor, der für diese Untersuchung entwickelt wurde. Boldi et al. (2002) gehen auf das afrikanische Web ein, wobei der Sprachenidentifizierer *TextCat* (vgl. Fußnote 18, S. 329) eingesetzt wird. Die Autoren zeigen, dass die ermittelten Sprachen keineswegs mit der gegebenen Verteilung von Sprachen auf dem afrikanischen Kontinent korreliert: 74,68% der Webdokumente sind in Englisch verfasst, der Anteil der Englisch-Muttersprachler an der Gesamtbevölkerung beträgt jedoch nur etwa 0,07%.

[50] Ein Webserver gehört zum Bereich „deedu“, falls sein Name eines der Muster `uni-`, `tu-` oder `fh-` enthält.

A.7 Zusammenfassung

Die statistischen Analysen zeichnen ein relativ klares und homogenes Bild vom WiN-Web: Eine Hochschule aus dem Bereich der Kernuniversitäten bietet durchschnittlich ca. 110 895 HTML-Dokumente auf 198 Webservern an, nur etwa 51 715 dieser Dokumente (46%) sind deutschsprachig. Die im WiN-Web verfügbaren Webseiten sind durchschnittlich 594 Tage alt. Dieser Wert ist überraschend hoch und erstaunt umso mehr, als dass JPEG- und PDF-Dateien im Durchschnitt deutlich jünger sind (488 bzw. 375 Tage). Das arithmetische Mittel der Dokumentgröße beträgt 7 024 Byte (91% aller HTML-Dokumente besitzen eine Größe bis 16 Kilobyte), d. h. komplexe strukturelle Analyseverfahren werden im Regelfall nicht mit extrem großen Dokumenten konfrontiert. Die durchschnittliche Anzahl von Wörtern pro Dokument fällt mit 310 eher gering aus, mehr als 40% aller Dokumente enthalten nur zwischen 11 und 100 Wörtern, fast 70% aller Dokumente enthalten bis zu 200 Wörter. Insbesondere diese Daten stellen die entscheidende Rolle metasprachlicher struktureller Analyseverfahren heraus, denen im Schnitt 120,57 HTML-Elemente (davon etwa 13,5 Hyperlinks) und 236,04 HTML-Attribute pro Dokument zur Verfügung stehen (vgl. Abschnitt 14.5). Fast alle Element- und Attribut-Bezeichnungen entsprechen dem HTML 4.01-Standard, proprietäre Elemente werden de facto nicht mehr eingesetzt. Bezüglich des Einsatzes unterschiedlicher Elemente kann festgehalten werden, dass das mit 91 Elementen sehr umfangreiche Vokabular von HTML 4.01 nicht annähernd ausgeschöpft wird. Lediglich 27 HTML-Elemente werden in mehr als 10% aller Dokumente eingesetzt. Für den Textparser bedeutet dies, dass zahlreiche sehr niedrigfrequente Elemente unberücksichtigt bleiben können. Im Bereich der Metadaten werden frühere Arbeiten bestätigt, in denen zwar eine hohe Anzahl von `meta`-Elementen (im Korpus durchschnittlich 3 Elemente in zwei Dritteln aller Dokumente) festgestellt wird, deren Inhalte sich jedoch größtenteils auf automatisch hinzugefügte Informationen wie den Medientyp, Zeichensatz oder Namen des Autors beziehen. Großflächige Metadatenauszeichnungen sind im WiN-Web nicht existent. Ähnliches kann für den Einsatz multimedialer Objekte berichtet werden, der im Vergleich zu anderen Domänen als spartanisch bezeichnet werden kann: Zwar werden in etwa zwei Dritteln aller Dokumente durchschnittlich 9,4 eingebettete Bilder eingesetzt, Audio- und/oder Video-Dateien werden hingegen kaum unmittelbar in HTML-Dokumenten angeboten. Auch interaktive Anwendungen, die mittels *Java* oder *JavaScript* realisiert werden, sind eher die Ausnahme als die Regel: Applets werden nur in 0,59% aller Dokumente verwendet, *JavaScript*-Code wird in 13,57% aller Fälle eingesetzt, Direktiven zur Setzung eines Cookies wurden in lediglich 0,5% aller HTTP-Response-Header gefunden. Auch die Benutzung des `form`-Elements in 4,26% aller Dokumente ist verschwindend gering. Etwa 43% aller Dokumente enthalten den Fingerabdruck eines HTML-Editors, d. h. etwa die Hälfte aller im Korpus befindlichen Dokumente wurde mit maschineller Unterstützung angefertigt. Die verwendete Software generiert meist aktuelles Markup, am häufigsten werden die HTML-Standards 4.0, 3.2 und 4.01 benutzt. Der XML-Bereich des WiN-Web ist mit ca. 30 000 Dokumentinstanzen kaum ausgeprägt, das Gros der Instanzen geht auf fehlerhaft übertragene HTML-Dokumente und automatisch von *Microsoft Office*-Anwendungen generierte Dateien zurück.

Insgesamt ergibt sich für das WiN-Web aus Sicht der *Hypertext Markup Language* das Bild, dass zwar eine große Menge Webseiten angeboten wird, diese enthalten jedoch äußerst über-

schaubare Auszeichnungsstrukturen und kaum multimediale oder interaktive Elemente. Hyperlinks werden im Vergleich zu anderen Studien sehr gut gepflegt, was ein Indiz dafür ist, dass es den Autoren eher um inhaltliche als designbezogene Aspekte geht. Für Zwecke der automatisierten Informationsrecherche einsetzbare Metadatenauszeichnungen sind kaum vorhanden, und der Anteil von RDF-Dateien ist mit 126 Vorkommen extrem gering.

Dieser Anhang enthält statistische Analysen, die über die Größe des WiN-Web und zahlreiche Charakteristika der enthaltenen HTML-Dokumente Auskunft geben. In dieser Hinsicht handelt es sich um eine Fortführung von Bray (1996), Woodruff et al. (1996), O'Neill et al. (1998) und insbesondere Turau (1998a). Eine Analyse der Frequenzen aller in HTML 4.01 definierten Elemente und Attribute sowie der erlaubten und nicht erlaubten Kombinationen wurde meines Wissens bislang nicht durchgeführt. Diese Daten wurden der *HTML Working Group* des W3C unter dem Vorsitz von Steven Pemberton zur Verfügung gestellt, so dass sie in zukünftige Revisionen des XHTML-Standards einfließen können. Auf diese Weise könnten z. B. kaum benutzte Elemente wie `isindex` oder `bdo` als „deprecated" eingestuft werden. Abschließend kann die Liste der nicht erlaubten Element-Attribut-Kombinationen benutzt werden, um entsprechende Warnhinweise in den Standard aufzunehmen, z. B. dass das Attribut `height` nicht für das Element `table` verwendet werden darf.

Die Studie besitzt eine Relevanz für IR-Anwendungen, da sie die Bestimmung verschiedener Parameter für IR-Systeme am WiN-Web exemplifiziert. Zwei Parameter, ein Mangel an expliziten und maschinell auswertbaren Metadaten und ein großer Bedarf für die Analyse von Markup-Strukturen, wurden bereits vorgestellt. Die Verteilung der Medientypen zeigt, dass es möglich sein könnte, Referenzierungen externer CSS-Stylesheets zur Bestimmung der Grenzen von Hypertexten einzusetzen (vgl. Abschnitt 14.6.1). Die Untersuchung über die Verteilung von HTML-Elementen zeigt, dass Tabellen zwar sehr häufig verwendet werden, es handelt sich dabei jedoch oftmals nicht um genuine Tabellen (vgl. Abschnitt 14.5.2).

Die Domäne der wissenschaftlichen Webserver deutscher Universitäten könnte bezüglich des gesamten *World Wide Web* zu unabhängig und zu homogen erscheinen, so dass die Ergebnisse dieser Studie keine generelle Anwendbarkeit besäßen. Tatsächlich zeigt jedoch die Menge von 4 000 000 HTML-Dokumenten, die von 100 Universitäten stammen und der Einsatz von mehr als 200 verschiedenen HTML-Editoren und -Konvertoren, dass das *Web des deutschen Wissenschaftsnetzes* in Bezug auf seinen inneren Aufbau extrem heterogen ist. Eine der wesentlichen Fragen zukünftiger Studien wird es sein, universale sowie domänenspezifische Charakteristika derartiger Kollektionen zu bestimmen. Ein Vergleich der vorliegenden Daten mit früheren Arbeiten zeigt bereits einen gewissen Trend: Vielversprechende Kandidaten für universale Merkmale sind etwa die Verteilung von Dateigrößen und Hyperlinks. Deutliche Unterschiede existieren hingegen bezüglich des durchschnittlichen Alters von Dateien, der durchschnittlichen Anzahl von Wörtern, der Verteilung von HTML-Elementen und -Attributen sowie in der Benutzung von Audio- und Video-Dateien.

B

Die Tabellen des Korpusdatenbankservers

Nachfolgend sind die Felder, Datentypen, Indizes etc. der einzelnen Tabellen der Korpusdatenbank aufgeführt (vgl. ausführlich hierzu Abschnitt 7.2.3). Die Strukturen wurden in der Client-Anwendung `mysql` des Datenbanksystems *MySQL* mit dem SQL-Kommando `desc` (*describe*) erzeugt.

Field	Type	Null	Key	Default	Extra
id	int(10) unsigned		PRI	0	auto_increment
uri	text			NULL	
file	text	YES		NULL	
server_info	int(10) unsigned		MUL	0	
status_code	smallint(5) unsigned		MUL	0	
cont_length	int(10) unsigned		MUL	0	
cont_type	varchar(30)		MUL		
cont_encoding	tinytext	YES		NULL	
cont_language	tinytext	YES		NULL	
cont_location	tinytext	YES		NULL	
location	tinytext	YES		NULL	
date	datetime	YES		NULL	
expires	datetime	YES		NULL	
last_modified	datetime	YES		NULL	
www_authen	enum('0','1')			0	
cache_control	enum('0','1')			0	
content_md5	enum('0','1')			0	
pragma	enum('0','1')			0	
set_cookie	enum('0','1')			0	

Tabelle B.1: Struktur der Tabelle `http_header`

Field	Type	Null	Key	Default	Extra
id	int(10) unsigned		PRI	0	
servername	varchar(255)		MUL		
port	mediumint(8) unsigned		MUL	0	
servertype	varchar(255)		MUL		
http_version	varchar(5)		MUL		
city	char(3)		MUL		

Tabelle B.2: Struktur der Tabelle server_info

Field	Type	Null	Key	Default	Extra
id	int(10) unsigned		PRI	0	auto_increment
city	char(3)				
type	varchar(100)				
in_corpus	enum('0','1')			0	
home_url	varchar(100)				
name	varchar(100)				
location	varchar(100)				
diskspace	bigint(20) unsigned	YES		NULL	
crawl_start	datetime	YES		NULL	
crawl_end	datetime	YES		NULL	
totalspace	float(10,2) unsigned	YES		NULL	

Tabelle B.3: Struktur der Tabelle universities

Field	Type	Null	Key	Default	Extra
id	int(10) unsigned		PRI	0	auto_increment
login	varchar(10)				
password	varchar(40)				
status	varchar(15)				
firstname	varchar(20)				
middlename	varchar(20)	YES		NULL	
lastname	varchar(20)				
email	varchar(100)	YES		NULL	
url	varchar(100)	YES		NULL	
city	varchar(30)				
country	varchar(30)				

Tabelle B.4: Struktur der Tabelle user

Field	Type	Null	Key	Default	Extra
sample_id	smallint(5) unsigned		PRI	0	auto_increment
title	tinytext			NULL	
generated	datetime		MUL	0000-00-00 00:00:00	
num_docs	mediumint(8) unsigned		MUL	0	
matching_docs	mediumint(8) unsigned		MUL	0	
cities_available	text			NULL	
user	varchar(10)		MUL		
description	text			NULL	
sql_query	text			NULL	
permission	tinytext			NULL	

Tabelle B.5: Struktur der Tabelle meta_sample

Field	Type	Null	Key	Default	Extra
template_id	smallint(5) unsigned		PRI	0	
name	varchar(20)				
generated	datetime		MUL	0000-00-00 00:00:00	
user	varchar(10)		MUL		
description	text			NULL	

Tabelle B.6: Struktur der Tabelle `meta_template`

Field	Type	Null	Key	Default	Extra
id	int(10) unsigned		PRI	0	auto_increment
sample_id	smallint(5) unsigned		MUL	0	
template_id	int(10) unsigned		MUL	0	
template_id2	int(10) unsigned	YES		NULL	
template_id3	int(10) unsigned	YES		NULL	
template_id4	int(10) unsigned	YES		NULL	
template_id5	int(10) unsigned	YES		NULL	

Tabelle B.7: Struktur der Tabelle `sample_template`

Field	Type	Null	Key	Default	Extra
id	int(10) unsigned		PRI	0	auto_increment
sample_id	smallint(5) unsigned		MUL	0	
doc_id	int(10) unsigned		MUL	0	
server_id	int(10) unsigned		MUL	0	
doc_url	text			NULL	
city	char(3)		MUL		

Tabelle B.8: Struktur der Tabelle `sample`

Field	Type	Null	Key	Default	Extra
id	int(10) unsigned		PRI	0	auto_increment
template_id	smallint(5) unsigned		MUL	0	
sample_id	smallint(5) unsigned		MUL	0	
doc_id	int(10) unsigned		MUL	0	
hts	tinytext			NULL	
comment	text	YES		NULL	
standalone	enum('0','1')			0	
user	varchar(10)				
date	datetime			0000-00-00 00:00:00	

Tabelle B.9: Struktur der Tabelle `template1`

Field	Type	Null	Key	Default	Extra
id	int(10) unsigned		PRI	0	auto_increment
template_id	smallint(5) unsigned		MUL	0	
sample_id	smallint(5) unsigned		MUL	0	
doc_id	int(10) unsigned		MUL	0	
path	tinytext	YES		NULL	
hts	tinytext	YES		NULL	
comment	text	YES		NULL	
feat0	enum('0','1')	YES		NULL	
feat1	enum('0','1')	YES		NULL	
feat2	enum('0','1')	YES		NULL	
feat3	enum('0','1')	YES		NULL	
feat4	enum('0','1')	YES		NULL	
feat5	enum('0','1')	YES		NULL	
feat6	enum('0','1')	YES		NULL	
feat7	enum('0','1')	YES		NULL	
feat8	enum('0','1')	YES		NULL	
feat9	enum('0','1')	YES		NULL	
feat10	enum('0','1')	YES		NULL	
feat11	enum('0','1')	YES		NULL	
feat12	enum('0','1')	YES		NULL	
feat13	enum('0','1')	YES		NULL	
feat14	enum('0','1')	YES		NULL	
feat15	enum('0','1')	YES		NULL	
feat16	enum('0','1')	YES		NULL	
feat17	enum('0','1')	YES		NULL	
feat18	enum('0','1')	YES		NULL	
feat19	enum('0','1')	YES		NULL	
module0	enum('0','1')	YES		NULL	
module1	enum('0','1')	YES		NULL	
module2	enum('0','1')	YES		NULL	
module3	enum('0','1')	YES		NULL	
module4	enum('0','1')	YES		NULL	
module5	enum('0','1')	YES		NULL	
module6	enum('0','1')	YES		NULL	
module7	enum('0','1')	YES		NULL	
module8	enum('0','1')	YES		NULL	
module9	enum('0','1')	YES		NULL	
module10	enum('0','1')	YES		NULL	
module11	enum('0','1')	YES		NULL	
module12	enum('0','1')	YES		NULL	
module13	enum('0','1')	YES		NULL	
module14	enum('0','1')	YES		NULL	
module15	enum('0','1')	YES		NULL	
module16	enum('0','1')	YES		NULL	
module17	enum('0','1')	YES		NULL	
module18	enum('0','1')	YES		NULL	
module19	enum('0','1')	YES		NULL	
text_value	tinytext	YES		NULL	
contact_list	enum('0','1')	YES		NULL	
contact_sort	tinytext	YES		NULL	
contact_search	enum('0','1')	YES		NULL	
contact_name	enum('0','1')	YES		NULL	
contact_address	enum('0','1')	YES		NULL	
contact_phone	enum('0','1')	YES		NULL	
contact_fax	enum('0','1')	YES		NULL	
contact_email	enum('0','1')	YES		NULL	
linklist_sort	tinytext	YES		NULL	
linklist_links	tinytext	YES		NULL	
search_interac	enum('0','1')	YES		NULL	
search_topic	tinytext	YES		NULL	
search_scope	tinytext	YES		NULL	
user	varchar(10)				
date	datetime			0000-00-00 00:00:00	

Tabelle B.10: Struktur der Tabelle `template2`

Field	Type	Null	Key	Default	Extra
id	int(10) unsigned		PRI	0	auto_increment
template_id	smallint(5) unsigned		MUL	0	
sample_id	smallint(5) unsigned		MUL	0	
doc_id	int(10) unsigned		MUL	0	
greeting	enum('0','1')	YES		NULL	
prom_name	enum('0','1')	YES		NULL	
prom_title	enum('0','1')	YES		NULL	
prom_job	enum('0','1')	YES		NULL	
prom_aff	enum('0','1')	YES		NULL	
photo	enum('0','1')	YES		NULL	
cont_street	enum('0','1')	YES		NULL	
cont_room	enum('0','1')	YES		NULL	
cont_postal	enum('0','1')	YES		NULL	
cont_offhours	enum('0','1')	YES		NULL	
cont_phone	enum('0','1')	YES		NULL	
cont_phonesec	enum('0','1')	YES		NULL	
cont_fax	enum('0','1')	YES		NULL	
cont_email	enum('0','1')	YES		NULL	
cont_url	enum('0','1')	YES		NULL	
cont_sms	enum('0','1')	YES		NULL	
cont_pgp	enum('0','1')	YES		NULL	
cont_x500	enum('0','1')	YES		NULL	
cont_route	enum('0','1')	YES		NULL	
cont_priv_street	enum('0','1')	YES		NULL	
cont_priv_phone	enum('0','1')	YES		NULL	
cont_priv_cell	enum('0','1')	YES		NULL	
cont_priv_fax	enum('0','1')	YES		NULL	
cont_priv_email	enum('0','1')	YES		NULL	
cont_priv_url	enum('0','1')	YES		NULL	
cv	enum('0','1')	YES		NULL	
talks	enum('0','1')	YES		NULL	
publications	enum('0','1')	YES		NULL	
prom_books	enum('0','1')	YES		NULL	
interests	enum('0','1')	YES		NULL	
projects	enum('0','1')	YES		NULL	
memberships	enum('0','1')	YES		NULL	
transfer	enum('0','1')	YES		NULL	
functions	enum('0','1')	YES		NULL	
courses	enum('0','1')	YES		NULL	
theses	enum('0','1')	YES		NULL	
hints	enum('0','1')	YES		NULL	
univ_logo	enum('0','1')	YES		NULL	
univ_name	enum('0','1')	YES		NULL	
link_univ	enum('0','1')	YES		NULL	
link_dep	enum('0','1')	YES		NULL	
link_group	enum('0','1')	YES		NULL	
link_hotlinks	enum('0','1')	YES		NULL	
tech_webmaster	enum('0','1')	YES		NULL	
tech_lastmod	enum('0','1')	YES		NULL	
tech_counter	enum('0','1')	YES		NULL	
tech_alternate	enum('0','1')	YES		NULL	
tech_guestbook	enum('0','1')	YES		NULL	
cont_sec_name	enum('0','1')	YES		NULL	
cont_sec_add	enum('0','1')	YES		NULL	
cont_sec_room	enum('0','1')	YES		NULL	
cont_sec_hours	enum('0','1')	YES		NULL	
cont_sec_phone	enum('0','1')	YES		NULL	
cont_sec_fax	enum('0','1')	YES		NULL	
cont_sec_email	enum('0','1')	YES		NULL	
cont_col_name	enum('0','1')	YES		NULL	
cont_col_list	enum('0','1')	YES		NULL	
cont_col_add	enum('0','1')	YES		NULL	
cont_col_room	enum('0','1')	YES		NULL	
cont_col_phone	enum('0','1')	YES		NULL	
cont_col_email	enum('0','1')	YES		NULL	
cont_col_sass	enum('0','1')	YES		NULL	
user	varchar(10)				
date	datetime			0000-00-00 00:00:00	

Tabelle B.11: Struktur der Tabelle `template3`

Field	Type	Null	Key	Default	Extra
id	int(10) unsigned		PRI	NULL	auto_increment
template_id	smallint(5) unsigned		MUL	0	
sample_id	smallint(5) unsigned		MUL	0	
doc_id	int(10) unsigned		MUL	0	
htks	tinytext	YES		NULL	
hts	tinytext	YES		NULL	
path	tinytext	YES		NULL	
gen_feat	tinytext	YES		NULL	
comment	text	YES		NULL	
user	varchar(10)				
date	datetime			0000-00-00 00:00:00	

Tabelle B.12: Struktur der Tabelle `template4`

Abkürzungsverzeichnis

AI *Artificial Intelligence* → KI

API *Application Programming Interface* – Schnittstellenbeschreibung

ASCII *American Standard Code for Information Interchange*

BLOB *Binary Large Object* → Abschnitt 7.2.1

BNC *British National Corpus* → Abschnitt 14.2.2

CD ROM *Compact Disc – Read Only Memory*

CERN *Centre Européen pour la Recherche Nucléaire* → Abschnitt 3.2 und `http://www.cern.ch`

CFP *Call for Papers* → Abschnitt 4.6.2

CGI *Common Gateway Interface* → `http://hoohoo.ncsa.uiuc.edu/cgi/`

CMC *Computer-Mediated Communication* → Abschnitt 4.2.1 und Abschnitt 8.3

CMS *Content Management System* → Abschnitt 3.3.6

CPAN *Comprehensive Perl Archive Network* → `http://www.cpan.org`

CSCW *Computer-Supported Collaborative Work* → Abschnitt 4.2.2

CSS *Cascading Style Sheets* → Bos et al. (1998)

DAML *DARPA Agent Markup Language* → Abschnitt 13.2.3

DC *Dublin Core* – Metadatenschema → Abschnitt 3.6.6 und Abschnitt A.4.4

DCMI *Dublin Core Metadata Initiative* → Abschnitt 3.6.6 und Abschnitt A.4.4

DFG *Deutsche Forschungsgemeinschaft* → `http://www.dfg.de`

DFN *Deutsches Forschungsnetz* → Abschnitt 6.3 und `http://www.dfn.de`

DL *Description Logic* → Abschnitt 13.2.3

DNS *Domain Name System* → RFC 1034 und RFC 1035

DOM *Document Object Model* → Abschnitt 14.5 und Hors et al. (2000)

DSSSL *Document Style Semantics and Specification Language* → ISO 10179

DTD *Document Type Definition* → ISO 8879 und Bray et al. (2004b)

ECMA *European Computer Manufacturers Association* → Abschnitt A.4.6

EDM *Enterprise Document Management* → Abschnitt 4.2.3

FAQ *Frequently Asked Questions* → Abschnitt 3.6.4, Abschnitt 3.6.6, Abschnitt 4.5.3

FPI *Formal Public Identifier* → Abschnitt A.4.8 und Abschnitt A.5.2

FSF *Free Software Foundation* → `http://www.fsf.org` und `http://www.gnu.org`

FSI *Formal System Identifier* → Abschnitt A.4.8 und Abschnitt A.5.2

FTP *File Transfer Protocol* → RFC 0959

GERHARD *German Harvest Automated Retrieval and Directory* → Abschnitt 13.3

GFDL *GNU Free Documentation License* → `http://www.gnu.org/copyleft/fdl.html`

GIF *Graphics Interchange Format*

GNU *GNU's Not UNIX* → `http://www.gnu.org` und `http://www.fsf.org`

GPL *GNU General Public License* → `http://www.gnu.org/copyleft/gpl.html`

HCI *Human-Computer-Interaction* → Abschnitt 6.3.3

HICSS *Hawai'i International Conference on System Sciences* → Abschnitt 4.1

HP *Homepage* → Abschnitt 4.6.1

HTKS *Hypertextknotensorte* → Abschnitt 5.5

HTKT *Hypertextknotentyp* → Abschnitt 5.5

HTM *Hypertextmodul* → Abschnitt 5.6.2

HTML *Hypertext Markup Language* → Abschnitt A.4.8 und Raggett et al. (1999)

HTS *Hypertextsorte* → Abschnitt 5.4

HTSM *Hypertextsortenmodul* → Abschnitt 5.6

HTT *Hypertexttyp* → Abschnitt 5.4

HTTP *Hypertext Transfer Protocol* → RFC 2616

HTTP-NG *HTTP Next Generation* → `http://www.w3.org/Protocols/HTTP-NG/`

IE *Information Extraction* → Abschnitt 14.6.3

IETF *Internet Engineering Task Force* → `http://www.ietf.org`

IMS *IMS Global Learning Consortium* → `http://www.imsproject.org`

IP *Internet Protocol* → RFC 0791

IRC *Internet Relay Chat* → RFC 2810, RFC 2811, RFC 2812, RFC 2813

IR *Information Retrieval*

ISO *International Organization for Standardization* → `http://www.iso.org`

JPEG *Joint Photographic Experts Group*

KI *Künstliche Intelligenz* → AI

k**NN** k *Nearest Neighbour* → Abschnitt 14.2.1

LAMP *Linux Apache MySQL PHP* → Abschnitt 7.3

LCN *least common node* → Abschnitt 14.5.4

LDP *Linux Documentation Project* → `http://www.tldp.org`

LOB *Lancaster-Oslo-Bergen Corpus* → Abschnitt 14.2.2

MD *Message Digest* → Abschnitt 7.3.1 und RFC 1321

MIME *Multipurpose Internet Mail Extensions* → RFC 2045, RFC 2046, RFC 2047, RFC 2048

MOO *MUD Object Oriented*

MUD *Multi User Dungeon*

NAGT *North American Genre Theory* → Abschnitt 2.3.7

NCSA *National Center for Supercomputing Applications* → `http://www.ncsa.uiuc.edu`

NLP *Natural Language Processing*

NLS *oNLine System* → Abschnitt 3.2

NNTP *Network News Transport Protocol* → RFC 0977 und RFC 1036

OCR *Optical Character Recognition*

OIL *Ontology Inference Layer* → Abschnitt 13.2.3

OSI *Open Source Initiative* → `http://www.opensource.org`

OWL *Web Ontology Language* → Abschnitt 13.2.3

PDA *Personal Digital Assistant*

PDF *Portable Document Format* → `http://www.adobe.com/products/acrobat/adobepdf.html`

PGP *Pretty Good Privacy* → RFC 1991

PHP *PHP: Hypertext Preprocessor* → Abschnitt 7.3 und `http://www.php.net`

PNG *Portable Network Graphics* → RFC 2083

POS *Part-of-Speech* – Wortart

PRIDE *Personalized Retrieval Indexing and Documentary Evolution* → Abschnitt 3.2

RACER *Renamed ABox and Concept Expression Reasoner* → Abschnitt 13.2.3

RDF *Resource Description Framework* → Abschnitt 13.2.3

RDFS *RDF Schema* → Abschnitt 13.2.3

RFC *Request for Comments* → `http://www.ietf.org/rfc/`

RSS *Really Simple Syndication*

RST *Rhetorical Structure Theory* → Abschnitt 2.2.5

RTSP *Real Time Streaming Protocol* → RFC 2326
SGML *Standard Generalized Markup Language* → ISO 8879
SMS *Short Message Service*
SOAP *Simple Object Access Protocol* → Box et al. (2000), Mitra et al. (2002)
SOM *Self-Organizing Map* – Selbstorganisierende Karte
SQL *Structured Query Language*
SVG *Scalable Vector Graphics* → Ferraiolo et al. (2003)
SVM *Support Vector Machine* → Abschnitt 14.2.1
TCP/IP *Transmission Control Protocol/Internet Protocol* → RFC 0791 und RFC 0793
TEI *Text Encoding Initiative* → Sperberg-McQueen und Burnard (2002)
TREC *Text Retrieval Conference* → Abschnitt 7.5.4
TR *Text Retrieval* → Abschnitt 7.5.4
UCS *Universal Character Set* → `http://www.unicode.org`
UDK *Universale Dezimalklassifikation* → `http://www.udcc.org` und Abschnitt 13.3.1
UNA *Unique Name Assumption* → Abschnitt 13.2.3
URI *Uniform Resource Identifier* → RFC 2396
URL *Uniform Resource Locator* → RFC 1738 und RFC 1808
UTF *UCS Transformation Formats* → RFC 2279 und `http://www.unicode.org`
VLC *Very Large Corpus* bzw. *Very Large Collection* → Abschnitt 7.5.4
VML *Vector Markup Language* → `http://www.w3.org/TR/NOTE-VML`
VRML *Virtual Reality Markup Language* → `http://www.web3d.org/x3d/vrml/`
W3C *World Wide Web Consortium* → `http://www.w3.org` und Abschnitt 3.2
WAIS *Wide Area Information Service* → RFC 1625
WAP *Wireless Access Protocol* → `http://www.openmobilealliance.org/tech/affiliates/wap/`
WiN *Wissenschaftsnetz* → DFN und Abschnitt 6.3
WiN-Web *Web des deutschen Wissenschaftsnetzes* → WiN, DFN und Abschnitt A.3
WML *Wireless Markup Language* → `http://www.openmobilealliance.org/tech/affiliates/wap/`
WSDL *Web Services Description Language* → Christensen et al. (2001)
WWW *World Wide Web* → Abschnitt 3.2
WYSIWYG *What You See Is What You Get*
XHTML *Extensible Hypertext Markup Language* → Abschnitt 14.4 und Abschnitt A.4.8
XML *Extensible Markup Language* → Bray et al. (2004b)
XPath *XML Path Language* → Clark und DeRose (1999) und Berglund et al. (2002)
XSL *Extensible Stylesheet Language* → Clark (1999) und Adler et al. (2001)
XSLT *XSL Transformations* → Clark (1999)
XTM *XML Topic Maps* → Pepper und Moore (2001)

Danksagungen

An dieser Stelle möchte ich mich bei denjenigen Personen bedanken, die zum Entstehen der vorliegenden Dissertation beigetragen haben.

Zuallererst möchte ich meinem Betreuer, Prof. Dr. Henning Lobin, ganz besonders herzlich für seine langjährige und sehr engagierte Unterstützung sowie für zahlreiche konstruktive Diskussionen und Anregungen danken, die diese Arbeit maßgeblich geprägt haben. Bei Prof. Dr. Gerd Fritz bedanke ich mich ebenfalls sehr herzlich für die vielen wertvollen Hinweise.

Ein besonders großes und herzliches Dankeschön schulde ich Maja Bärenfänger, Sabine Heuser, Alexander Krumeich, Harald Lüngen, Ulrike Naumann, Maik Stührenberg, Janusz Taborek und Stefan Ulrich, die sich freundlicherweise bereit erklärt haben, einzelne Kapitel oder die gesamte Arbeit Korrektur zu lesen und mit denen ich im Zuge dessen viele interessante Diskussionen geführt habe.

Frank H. Müller und Tylman Ule haben mir dankenswerterweise das in einer Analyse verwendete *taz*-Korpus in einer XML-Version zur Verfügung gestellt. Claus-Rainer Rollinger, Klaus Dalinghaus sowie Hans-Joachim Wätjen und Bernd Diekmann danke ich für die Bereitstellung der im Projekt GERHARD eingesetzten Rohdaten und die ausführlichen Anmerkungen zum Datenformat. Stefan Ondrejicka war so freundlich, den von ihm implementierten Crawler *Pavuk* um verschiedene Funktionen zu ergänzen, die für den Korpusaufbau benötigt wurden. Holger Knublauch hat das *protégé OWL Plug-in* entwickelt und stand mir bei dessen Anwendung mit Rat und Tat zur Seite. Markus Kohm, der federführende Entwickler des *LaTeX*-Pakets *KOMA-Script*, mit dem die vorliegende Arbeit erstellt wurde, hat meine Anfragen mit ebenso viel Engagement wie Erfolg beantwortet.

Für anregende Diskussionen und Literaturhinweise möchte ich mich bedanken bei: David Firth, Stephanie Haas, Eva-Maria Jakobs, Thomas Lampe, Alexander Mehler, Gerd Richter, Marina Santini, Mike Shepherd, Angelika Storrer und Volker Turau.

Des Weiteren möchte ich den bislang nicht genannten ehemaligen Arbeitskollegen im Fachgebiet Angewandte Sprachwissenschaft und Computerlinguistik sowie im Zentrum für Medien und Interaktivität für die Zusammenarbeit während meiner Zeit an der Justus-Liebig-Universität Gießen danken. Hierzu zählen insbesondere Silvia Baumgart, Petra Saskia Bayerl, Ralf Grünspahn, Sabine Heymann und Susanne Schneider.

Meinen Eltern und meiner Familie danke ich für ihre fortwährende Unterstützung und das Verständnis für zuletzt immer seltenere Besuche.

Finally I would like to express my deepest gratitude to Katherine for her patience, support and encouragement over the years, especially during the final stages of writing this thesis. Merci vielmals! It's done now!

Literaturverzeichnis

Åkesson, Maria (2003): *Digital Patterns for the Online Newspaper Genre – A Genre Analysis of 85 Swedish Daily Online Newspapers*. Diplomarbeit, Institutionen för informatik, Handelshögskolan vid Göteborgs universitet. Online verfügbar: `http://www.handels.gu.se/epc/archive/00002936/`.

Abels, Eileen G.; White, Marilyn Domas und Hahn, Karla (1997): "Identifying user-based criteria for Web pages". *Internet Research* 7 (4): S. 252–262.

Abels, Eileen G.; White, Marilyn Domas und Hahn, Karla (1998): "A user-based design process for Web sites". *Internet Research* 8 (1): S. 39–48.

Abiteboul, Serge; Buneman, Peter und Suciu, Dan (2000): *Data on the Web – From Relations to Semistructured Data and XML*. San Francisco: Morgan Kaufmann.

Abiteboul, Serge; Cobéna, Gregory; Masanes, Julien und Sedrati, Gerald (2002): "A First Experience in Archiving the French Web". In: *Research and Advanced Technology for Digital Technology – 6th European Conference, ECDL 2002*, herausgegeben von Agosti, M. und Thanos, C. Heidelberg: Springer, Band 2458 von *Lecture Notes in Computer Science*, S. 1–15.

Adamzik, Kirsten (1995): "Aspekte und Perspektiven der Textsortenlinguistik". In: *Textsorten – Texttypologie. Eine kommentierte Bibliographie*, herausgegeben von Adamzik, Kirsten, Münster: Nodus, S. 11–40.

Adamzik, Kirsten (Hrsg.) (2000a): *Textsorten – Reflexionen und Analysen*, Band 1 von *Textsorten*. Tübingen: Stauffenburg.

Adamzik, Kirsten (2000b): "Was ist pragmatisch orientierte Textsortenforschung?" In: Adamzik (2000a), S. 91–112.

Adamzik, Kirsten (2004): *Textlinguistik – Eine einführende Darstellung*, Band 40 von *Germanistische Arbeitshefte*. Tübingen: Niemeyer.

Adler, Sharon; Berglund, Anders; Caruso, Jeff; Deach, Stephen; Graham, Tony; Grosso, Paul; Gutentag, Eduardo; Milowski, Alex; Parnell, Scott; Richman, Jeremy und Zilles, Steve (2001): "Extensible Stylesheet Language (XSL) 1.0". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/xsl/`.

Agosti, Maristella und Smeaton, Alan (Hrsg.) (1996): *Information Retrieval and Hypertext*. Dordrecht: Kluwer.

Aladwani, Adel M. und Palvia, Prashant C. (2002): "Developing and validating an instrument for measuring user-perceived web quality". *Information & Management* 39 (6): S. 467–476.

Alam, Hassan; Rahman, Fuad; Tarnikova, Yuliya und Kumar, Aman (2003): "When is a List is a List?: Web Page Re-authoring for Small Display Devices". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

de Almeida Santos, Hélia Vannucchi (1999): "Thwarting the Web Users' Expectations". In: *Proceedings of the third Conference on Creativity & Cognition*. Loughborough, S. 199–200.

Almind, Tomas und Ingwersen, Peter (1997): "Informetric analyses on the World Wide Web: A methodological approach to »Webometrics«." *Journal of Documentation* 53 (4): S. 404–426.

Altheim, Murray; Boumphrey, Frank; Dooley, Sam; McCarron, Shane; Schnitzenbaumer, Sebastian und Wugofski, Ted (2001): "Modularization of XHTML". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/xhtml-modularization/`.

Altheim, Murray und McCarron, Shane (2001): "XHTML 1.1 – Module-based XHTML". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/xhtml11/`.

Amitay, Einat (1997): *Hypertext: The Importance of being Different*. Master thesis, Centre for Cognitive Science, University of Edinburgh, Edinburgh.

Amitay, Einat (1998): "Using Common Hypertext Links to Identify the Best Phrasal Description of Target Web Documents". In: *Proceedings of the SIGIR '98 Post-Conference Workshop on Hypertext Information Retrieval for the Web*. Melbourne.

Amitay, Einat (2000a): "Anchors in Context: A Corpus Analysis of Web Pages Authoring Conventions". In: Pemberton und Shurville (2000), S. 25–35. Online verfügbar: `http://www.mri.mq.edu.au/~einat/`. Die Printfassung dieses Textes wurde gekürzt; Zitate beziehen sich auf die online erschienene Version.

Amitay, Einat (2000b): "InCommonSense – Rethinking Web Search Results". In: *ICME 2000 – IEEE International Conference on Multimedia and Expo*. New York City.

Amitay, Einat (2001): "Trends, Fashions, Patterns, Norms, Conventions. . .and Hypertext Too". *Journal of the American Society for Information Science* 52: S. 36–43.

Amitay, Einat; Carmel, David; Darlow, Adam; Lempel, Ronny und Soffer, Aya (2003): "The Connectivity Sonar: Detecting Site Functionality by Structural Patterns". In: *Conference on Hypertext and Hypermedia – Proceedings of the 14th ACM Conference on Hypertext and Hypermedia*. New York, S. 38–47.

Amitay, Einat und Oberlander, Jon (1997): "Convention says . . .". In: *Proceedings of the Flexible Hypertext Workshop of the Eighth ACM International Hypertext Conference*. Southampton.

Andreessen, Marc und Bina, Eric (1994): "NCSA Mosaic: A Global Hypermedia System". *Internet Research* 4 (1): S. 7–17.

Androutsopoulos, Jannis und Kraft, Daniel (2003): "Homepage-Design: Gestaltungslösungen in einer neuen Kommunikationsform". Ausarbeitung eines Vortrags auf dem Germanistentag 2001, 30.09.–03.10.2001, Erlangen. Online verfügbar: `http://www.ids-mannheim.de/prag/medienstil/`.

Androutsopoulos, Jannis K. (2000): "Die Textsorte Flyer". In: Adamzik (2000a), S. 175–213.

Angell, David und Heslop, Brent (1994): *The Elements of E-mail Style*. Reading, Menlo Park, New York etc.: Addison-Wesley.

Antoniou, Grigoris und van Harmelen, Frank (2004): "Web Ontology Language: OWL". In: Staab und Studer (2004), S. 67–92.

Antos, Gerd und Tietz, Heike (Hrsg.) (1997a): *Die Zukunft der Textlinguistik – Traditionen, Transformationen, Trends*, Band 188 von *Reihe Germanistische Linguistik*. Tübingen: Niemeyer.

Antos, Gerd und Tietz, Heike (1997b): "Einleitung: Quo vadis, Textlinguistik?" In: Antos und Tietz (1997a), S. vii–x.

Arasu, Arvind; Cho, Junghoo; Garcia-Molina, Hector; Paepcke, Andreas und Raghavan, Sriram (2001): "Searching the Web". *ACM Transactions on Internet Technology* 1 (1): S. 2–43.

Ashish, Naveen und Knoblock, Craig A. (1997): "Semi-Automatic Wrapper Generation for Internet Information Sources". In: *Proceedings of the 2nd International Conference on Cooperative Information Systems (CoopIS '97)*. S. 160–169.

Asirvatham, Arul Prakash und Ravi, Kranthi Kumar (2001): "Web Page Classification Based on Document Structure". Technischer Bericht, International Institute of Information Technology, Hyderabad.

Askehave, Inger und Nielsen, Anne Ellerup (2005): "What are the Characteristics of Digital Genres? – Genre Theory from a Multi-modal Perspective". In: *Proceedings of the 38th Hawaii International Conference on Systems Sciences (HICSS-38)*. Big Island, Hawaii.

Attardi, Giuseppe; Gullí, Antonio und Sebastiani, Fabrizio (1999): "Automatic Web Page Categorization by Link and Context Analysis". In: *Proceedings of THAI-99, 1st European Symposium on Telematics, Hypermedia and Artificial Intelligence*, herausgegeben von Hutchison, Chris und Lanzarone, Gaetano. Varese, S. 105–119.

Austin, John (1962/1979): *Zur Theorie der Sprechakte (How to do things with words)*. Stuttgart: Reclam.

Baader, Franz; Calvanese, Diego; McGuinness, Deborah L.; Nardi, Daniele und Patel-Schneider, Peter F. (Hrsg.) (2003): *The Description Logic Handbook – Theory, Implementation and Applications*. Cambridge: Cambridge University Press.

Baader, Franz; Horrocks, Ian und Sattler, Ulrike (2004): "Description Logics". In: Staab und Studer (2004), S. 3–28.

Baddeley, Alan D. (1990): *Human Memory: Theory and Practice*. Hillsdale: Erlbaum.

Baeza-Yates, Ricardo und Riberiro-Neto, Berthier (1999): *Modern Information Retrieval*. Harlow: Addison Wesley.

Bahl, Anke (1997): *Zwischen On- und Offline – Identität und Selbstdarstellung im Internet*. München: KoPäd.

Baker, Mark; Ishikawa, Masayasu; Matsui, Shinichi; Stark, Peter; Wugofski, Ted und Yamakami, Toshihiko (2000): "XHTML Basic". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/xhtml-basic/`.

Bar-Ilan, Judit (2001): "Data collection methods on the Web for infometric purposes – A review and analysis". *Scientometrics* 50 (1): S. 7–32.

Barnard, David T.; Burnard, Lou; DeRose, Steven J.; Durand, David G. und Sperberg-McQueen, C.M. (1996): "Lessons for the World Wide Web from the Text Encoding Initiative". *The World Wide Web Journal* 1 (1): S. 349–357.

Barroso, Luiz André; Dean, Jeffrey und Hölzle, Urs (2003): "Web Search for a Planet: The Google Cluster Architecture". *IEEE Micro* 23 (2): S. 22–28.

Bartlett, Frederic C. (1932): *Remembering – A Study in Experimental and Social Psychology*. Cambridge: Cambridge University Press.

Bates, Marcia J. und Lu, Shaojun (1997): "An Exploratory Profile of Personal Home Pages: Content, Design, Metaphors". *Online & CD ROM Review* 21 (6): S. 331–340.

Bauer, Christian; Bauer, Dietmar und Scharl, Arno (2000): "Measuring the web: A pilot study for web site classification based on empirical evidence". *Trends in Communication* 4 (6): S. 113–132.

Baumgartner, Robert; Flesca, Sergio und Gottlob, Georg (2001): "Visual Web Information Extraction with Lixto". In: *Proceedings of the 27th VLDB Conference*. Rom, S. 119–128.

Bayerl, Petra Saskia (2002): *Linguistische Analyse studentischer persönlicher Homepages*. Magisterarbeit, Institut für deutsche Sprache und mittelalterliche Literatur, Justus-Liebig-Universität, Gießen.

Bazerman, Charles (1988): *Shaping Written Knowledge – The Genre and Activity of the Experimental Article in Science*. Madison: University of Wisconsin Press.

Bazerman, Charles (1994): "Systems of Genres and the Enactment of Social Intentions". In: *Genre and the New Rhetoric*, herausgegeben von Freedman, Aviva und Medway, Peter, London: Taylor and Francis, S. 79–101.

de Beaugrande, Robert-A. und Dressler, Wolfgang (1981): *Einführung in die Textlinguistik*, Band 28 von *Konzepte der Linguistik*. Tübingen: Niemeyer.

de Beaugrande, Robert (1997): "Textlinguistik: Zu neuen Ufern?" In: Antos und Tietz (1997a), S. 1–11.

Bechhofer, Sean; van Harmelen, Frank; Hendler, Jim; Horrocks, Ian; McGuinness, Deborah L.; Patel-Schneider, Peter F. und Stein, Lynn Andrea (2004): "OWL Web Ontology Language – Reference". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/2004/REC-owl-ref-20040210/`.

Beckett, Dave (2004): "RDF/XML Syntax Specification (revised)". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/`.

Beesley, Kenneth R. (1988): "Language Identifier: A Computer Program for Automatic Natural-Language Identification of On-line Text". In: *Languages at Crossroads: Proceedings of the 29th Annual Conference of the American Translators Association*. S. 47–54. Online verfügbar: `http://www.rxrc.xerox.com/people/beesley/langid.html`.

Beghtol, Clare (2001): "The Concept of Genre and Its Characteristics". *Bulletin of The American Society for Information Science and Technology* 27 (2). Online verfügbar: `http://www.asis.org/Bulletin/`.

Behme, Henning (2000a): "Gute Seiten, schlechte Seiten – Was Webmaster richtig falsch machen können". *iX, Magazin für professionelle Informationstechnik* (8): S. 54–58.

Behme, Henning (2000b): "Sichten teilen – MySQL-Daten mit PHP ins Web bringen". *iX, Magazin für professionelle Informationstechnik* (6): S. 56–58.

Beißwenger, Michael (Hrsg.) (2001): *Chat-Kommunikation. Sprache, Interaktion, Sozialität & Identität in synchroner computervermittelter Kommunikation. Perspektiven auf ein interdisziplinäres Forschungsfeld*. Stuttgart: ibidem.

Beißwenger, Michael; Storrer, Angelika und Runte, Maren (2004): "Modellierung eines Terminologienetzes für das automatische Linking auf der Grundlage von WordNet". *LDV Forum* 19 (1/2): S. 113–125.

Belew, Richard K. (2000): *Finding Out About – A Cognitive Perspective on Search Engine Technology and the WWW*. Cambridge: Cambridge University Press.

Bellamy, Rachel; Boguraev, Branimir und Kennedy, Christopher (1999): "Dynamic Visual Metaphors for News Story Abstractions". In: *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32)*.

Benbow, S. Mary P. (1998): "File not found: the problems of changing URLs for the World Wide Web". *Internet Research* 8 (3): S. 247–250.

Benczúr, András A.; Csalogány, Károly; Fogaras, Dániel; Friedman, Eszter; Sarlós, Tamás; Uher, Máté und Windhager, Eszter (2003): "Searching a small national domain – Preliminary report". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Berglund, Anders; Boag, Scott; Chamberlin, Don; Fernandez, Mary F.; Kay, Michael; Robie, Jonathan und Siméon, Jérôme (2002): "XML Path Language (XPath) – Version 2.0". Technische Spezifikation (Working Draft), W3C. Online verfügbar: `http://www.w3.org/TR/xpath20/`.

Bergman, Michael K. (2000): "The Deep Web: Surfacing Hidden Value". White Paper, BrightPlanet.com LLC.

Bergquist, Magnus und Ljungberg, Jan (1999): "Genres in Action: Negotiating Genres in Practice". In: *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32)*.

van Berkel, Arrie und de Jong, Mariët (1999): "Coherence Phenomena in Hypertextual Environments". In: Jakobs et al. (1999), S. 29–40.

Berker, Thomas (2001): *Internetnutzung in den 90er Jahren. Wie ein junges Medium alltäglich wurde*. Frankfurt, New York: Campus.

Berners-Lee, Tim (1999): *Weaving the Web – The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. San Francisco: Harper San Francisco.

Berners-Lee, Tim; Cailliau, Robert; Groff, Jean-François und Pollermann, Bernd (1992): "World-Wide Web: The Information Universe". *Electronic Networking: Research, Applications and Policy* 1 (2): S. 52–58.

Berners-Lee, Tim; Cailliau, Robert; Luotonen, Ari; Nielsen, Henrik Frystyk und Secret, Arthur (1994): "The World-Wide Web". *Communications of the ACM* 37 (8): S. 76–82.

Berners-Lee, Tim; Hendler, James und Lassila, Ora (2001): "The Semantic Web". *Scientific American* 284 (5): S. 34–43.

Bernstein, Mark (1998): "Patterns of Hypertext". In: *Proceedings of the ninth ACM conference on Hypertext and Hypermedia: Links, Objects, Time and Space – Structure in Hypermedia Systems*. Pittsburgh, S. 21–29.

Biber, Douglas (1988): *Variation across Speech and Writing*. Cambridge, New York, Port Chester etc.: Cambridge University Press.

Bieber, Christoph und Leggewie, Claus (Hrsg.) (2004): *Interaktivität – Ein transdisziplinärer Schlüsselbegriff*. Frankfurt/Main, New York: Campus.

Bieber, Michael; Vitali, Fabio; Ashman, H.; Balasubramanian, V. und Oinas-Kukkonen, H. (1997): "Fourth generation hypermedia: Some missing links for the world wide web". *International Journal of Human-Computer Studies* 47 (1): S. 31–65.

Bins, Elmar K. und Piwinger, Boris-A. (1997): *Newsgroups – Weltweit diskutieren*. Bonn, Albany, Belmont etc.: Thomson.

Bittner, Johannes (2003): *Digitalität, Sprache, Kommunikation – Eine Untersuchung zur Medialität von digitalen Kommunikationsformen und Textsorten und deren varietätenlinguistischer Modellierung*. Berlin: Erich Schmidt.

Black, John B. und Wilensky, Robert (1979): "An Evaluation of Story Grammars". *Cognitive Science* 3: S. 213–229.

Blood, Rebecca (2004): "How Blogging Software Reshapes the Online Community". *Communications of the ACM* 47 (12): S. 53–55.

Boardman, Mark (2005): *The Language of Websites*. London, New York: Routledge.

Boguraev, Branimir; Bellamy, Rachel und Kennedy, Christopher (1999): "Dynamic Presentation of Phrasally-Based Document Abstractions". In: *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32)*.

Boldi, Paolo; Codenotti, Bruno; Santini, Massimo und Vigna, Sebastiano (2002): "Structural Properties of the African Web". In: *Proceedings of the 11th International World Wide Web Conference (WWW 2002)*. Honolulu.

Bollacker, Kurt; Lawrence, Steve und Giles, C. Lee (1998): "CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications". In: *Proceedings of the Second International Conference on Autonomous Agents*, herausgegeben von Sycara, Katia P. und Wooldridge, Michael. New York, S. 116–123.

Bolter, Jay David (1991): *Writing Space – The Computer, Hypertext, and the History of Writing*. Hillsdale: Erlbaum.

Bos, Bert; Lie, Håkon Wium; Lilley, Chris und Jacobs, Ian (1998): "Cascading Style Sheets, Level 2 – CSS2 Specification". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/REC-CSS2/`.

Boudourides, Moses A.; Mavrikakis, Manolis und Vasileiadou, Eleftheria (2002): "E-Mail Threads, Genres & Networks in a Project Mailing List". Online verfügbar: `http://www.math.upatras.gr/~mboudour/publ.html`. Dep. of Mathematics, University of Patras, Greece.

Boudourides, Moses A. und Peticca, S. (2001): "Genres in Organizational Communication – Literature Review". Online verfügbar: `http://www.math.upatras.gr/~mboudour/publ.html`. Dep. of Mathematics, University of Patras, Greece.

Box, Don; Ehnebuske, David; Kakivaya, Gopal; Layman, Andrew; Nielsen, Noah Mendelsohn Henrik Frystyk; Thatte, Satish und Winer, Dave (2000): "Simple Object Access Protocol (SOAP) 1.1". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/SOAP/`.

Brachman, Ronald J. (1979): "On the Epistemological Status of Semantic Networks". In: *Associative Networks: Representation and Use of Knowledge by Computers*, herausgegeben von Findler, N. V., New York: Academic Press, S. 3–50.

Brachman, Ronald J. und Levesque, Hector J. (Hrsg.) (1985): *Readings in Knowledge Representation*. San Mateo: Morgan Kaufmann.

Brandl, Annette (2002): *Webangebote und ihre Klassifikation – Typische Merkmale aus Experten- und Rezipientenperspektive*, Band 21 von *Angewandte Medienforschung – Schriftenreihe des Medien Instituts Ludwigshafen*. München: R. Fischer.

Braun, Herbert (2005): "Vom Web auf die Platte – Informationen aus dem Internet sichern und archivieren". *c't, Magazin für Computertechnik* (5): S. 174–179.

Bray, Tim (1996): "Measuring the Web". *Computer Networks and ISDN Systems* 28 (7-11): S. 993–1005.

Bray, Tim; Hollander, Dave; Layman, Andrew und Tobin, Richard (2004a): "Namespaces in XML 1.1". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/xml-names11/`.

Bray, Tim; Paoli, Jean; Sperberg-McQueen, C. M.; Maler, Eve; Yergeau, François und Cowan, John (2004b): "Extensible Markup Language (XML) 1.1". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/2004/REC-xml11-20040204/`.

Breaux, Travis D. und Reed, Joel W. (2005): "Using Ontology in Hierarchical Information Clustering". In: *Proceedings of the 38th Hawaii International Conference on Systems Sciences (HICSS-38)*. Big Island, Hawaii.

Bretan, Ivan; Dewe, Johan; Hallberg, Anders; Wolkert, Niklas und Karlgren, Jussi (1998): "Web-Specific Genre Visualization". In: *Proceedings of the 3rd WebNet Conference (WebNet '98)*. Orlando.

Breure, Leen (2001): "Development of the Genre Concept". Online verfügbar: `http://www.cs.uu.nl/people/leen/GenreDev/GenreDevelopment.htm`. Information and Computing Sciences, University of Utrecht. Version 1.0.1.

Brickley, Dan und Guha, R.V. (2004): "RDF Vocabulary Description Language 1.0: RDF Schema". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/2004/REC-rdf-schema-20040210/`.

Brin, Sergey und Page, Lawrence (1998): "The Anatomy of a Large-Scale Hypertextual Web Search Engine". *Computer Networks and ISDN Systems* 30 (1–7): S. 107–117.

Brinker, Klaus (1973): "Zum Textbegriff in der heutigen Linguistik". In: *Studien zur Texttheorie und zur deutschen Grammatik*, herausgegeben von Sitta, Horst und Brinker, Klaus, Düsseldorf: Schwann, S. 9–41.

Brinker, Klaus (2001): *Linguistische Textanalyse – Eine Einführung in Grundbegriffe und Methoden*, Band 29 von *Grundlagen der Germanistik*. Berlin: Erich Schmidt, 5. Auflage.

Brinker, Klaus; Antos, Gerd; Heinemann, Wolfgang und Sager, Sven F. (Hrsg.) (2000): *Text- und Gesprächslinguistik*, Band 16.1 von *Handbücher zur Sprach- und Kommunikationswissenschaft (HSK)*. Berlin, New York: de Gruyter.

Brinker, Ulrich und Hoffmann, Horst-Joachim (2004): "Kommunikationskiller – Was Besucher von Websites vertreibt". *c't, Magazin für Computertechnik* (4): S. 70–75.

Broder, Andrei; Kumar, Ravi; Maghoul, Farzin; Raghavan, Prabhakar; Rajagopalan, Sridhar; Stata, Raymie; Tomkins, Andrew und Wiener, Janet (2000): "Graph structure in the Web". *Computer Networks and ISDN Systems* 33: S. 309–320.

Brown, David J. (2004): "Web Search Considered Harmful". *ACM Queue* 2 (2): S. 84–85.

Brown, Peter J. (1990): "Assessing the quality of hypertext documents". In: *Hypertext: Concepts, Systems and Applications*, herausgegeben von Streitz, Norbert; Rizk, Antoine und André, Jacques, Cambridge: Cambridge University Press, S. 1–12.

Bruce, Harry (1999): "Perceptions of the Internet: what people think when they search the internet for information". *Internet Research* 9 (3): S. 187–199.

Brückner, Thomas (2001): "Textklassifikation". In: Klabunde et al. (2001), S. 442–447.

Bucher, Hans-Jürgen (1996): "Textdesign – Zaubermittel der Verständlichkeit? Die Tageszeitung auf dem Weg zum interaktiven Medium". In: Hess-Lüttich et al. (1996), S. 31–59.

Bucher, Hans-Jürgen (1999): "Die Zeitung als Hypertext – Verstehensprobleme und Gestaltungsprinzipien für Online-Zeitungen". In: Lobin (1999b), S. 9–32.

Bucher, Hans-Jürgen (2000): "Formulieren oder Visualisieren? Multimodalität in der Medienkommunikation". In: Richter et al. (2000), S. 661–691.

Bucher, Hans-Jürgen (2001): "Wie interaktiv sind die neuen Medien?" In: Bucher und Püschel (2001), S. 139–171.

Bucher, Hans-Jürgen (2004): "Online-Interaktivität – Ein hybrider Begriff für eine hybride Kommunikationsform". In: Bieber und Leggewie (2004), S. 132–167.

Bucher, Hans-Jürgen; Büffel, Steffen und Wollscheid, Jörg (2004): "Digitale Zeitung als E-Paper – Ein Hybridmedium zwischen Print- und Online-Zeitung". Ifra Special Report 6.32. Ifra Special Reports, Forschungsberichte, Untersuchungsberichte zu Technik und Organisation sowie Dokumente zur Standardisierung der Verlagstechnik.

Bucher, Hans-Jürgen und Püschel, Ulrich (Hrsg.) (2001): *Die Zeitung zwischen Print und Digitalisierung*. Wiesbaden: Westdeutscher Verlag.

Buhofer, Annelies Häcki (2000): "Mediale Voraussetzungen: Bedingungen von Schriftlichkeit allgemein". In: Brinker et al. (2000), S. 251–261.

Burger, Harald (2000): "Textsorten in den Massenmedien". In: Brinker et al. (2000), S. 614–628.

Burke, Sean M. (2002): *Perl & LWP – Fetching Web Pages, Parsing HTML, Writing Spiders & More*. Beijing, Cambridge, Farnham etc.: O'Reilly & Associates.

Burner, Mike (1997): "Crawling towards Eternity: Building an archive of the World Wide Web". *Web Techniques Magazine* 2 (5). Online verfügbar: `http://www.webtechniques.com/archives/1997/05/burner/`.

Bush, Vannevar (1941): "Memorandum Regarding Memex". In: Nyce und Kahn (1991a), S. 81–84. Anhang eines am 10.04.1941 verfassten Briefes von Vannevar Bush an Eric Hodgins (Herausgeber von *Fortune*), Teil der Vannevar-Bush-Sammlung in der Library of Congress.

Bush, Vannevar (1945a): "As We May Think". *Atlantic Monthly* 176 (1): S. 101–108.

Bush, Vannevar (1945b): "As We May Think". In: Nyce und Kahn (1991a), S. 85–110. Ursprünglich erschienen in *Atlantic Monthly* 176 (1), S. 101–108 und *Life* 19 (11), S. 112–114, 116, 121, 123–124.

Bush, Vannevar (1959): "Memex II". In: Nyce und Kahn (1991a), S. 165–184. Das letzte bekannte Manuskript (datiert 27.08.1959), Teil der Vannevar-Bush-Sammlung des MIT Archives; der Entwurf wurde bearbeitet von Vannevar Bush und Emily Flint (*Atlantic Monthly*) und eingereicht bei *Life*.

Bush, Vannevar (1967): "Memex Revisited". In: Nyce und Kahn (1991a), S. 197–216. Ursprünglich erschienen in *Science is Not Enough*, Vannevar Bush, New York, S. 75–101.

Buten, John (1996): "Personal Home Page Survey". Ursprünglich online verfügbar unter `http://www.asc.upenn.edu/usr/sbuten/phpi.htm`. In archivierter Form verfügbar unter `http://www.archive.org`.

Buttler, David; Liu, Ling und Pu, Calton (2001): "A Fully Automated Object Extraction System for the World Wide Web". In: *Proceedings of the 2001 International Conference on Distrubuted Computing Systems (ICDCS '01)*. Phoenix.

Buyukkokten, Orkut; Garcia-Molina, Hector und Paepcke, Andreas (2001a): "Accordion summarization for end-game browsing on PDAs and cellular phones". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. S. 213–220.

Buyukkokten, Orkut; Garcia-Molina, Hector und Paepcke, Andreas (2001b): "Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices". In: *Proceedings of the 10th International World Wide Web Conference (WWW 2001)*. Hong Kong.

Bußmann, Hadumod (2002): *Lexikon der Sprachwissenschaft*. Stuttgart: Kröner, 3. Auflage.

Calishain, Tara und Dornfest, Rael (2003): *Google Hacks – 100 Industrial-Strength Tips & Tools*. Peking, Cambridge, Farnham etc.: O'Reilly & Associates.

Campbell, Brad und Goodman, Joseph M. (1988): "HAM: A general purpose hypertext abstract machine". *Communications of the ACM* 31 (7): S. 856–861.

Campos, João und Silva, Mário J. (2001): "Versus: A Model for a Web Repository". In: *CRC'01 – 4ª Conferência de Redes de Computadores*. Covilhã.

Caplan, Priscilla (1995): "You Call It Corn, We Call It Syntax-Independent Metadata for Document-Like Objects". *The Public-Access Computer Systems Review* 6 (4): S. 19–23.

Capstick, Joanne; Diagne, Abdel Kader; Erbach, Gregor; Uszkoreit, Hans; Leisenberg, Anne und Leisenberg, Manfred (2000): "A System for Supporting Cross-Lingual Information Retrieval". *Information Processing and Management* 36 (2): S. 275–289.

Carberry, Sandra; Chu, Jennifer; Green, Nancy und Lambert, Lynn (1993): "Rhetorical Relations: Necessary but not Sufficient". In: *Proceedings of the ACL Workshop on Intentionality and Structure in Discourse Relations*, herausgegeben von Rambow, Owen. S. 1–4.

Carchiolo, Vincenza; Longheu, Alessandro und Malgeri, Michele (2002): "Extraction of Hidden Semantics from Web Pages". In: *Intelligent Data Engineering and Automated Learning – IDEAL 2002*, herausgegeben von Yin, Hujun; Allinson, Nigel; Freeman, Richard; Keane, John und Hubbard, Simon, Berlin, Heidelberg, New York etc.: Springer, Band 2412 von *Lecture Notes in Computer Science*, S. 117–122.

Carchiolo, Vincenza; Longheu, Alessandro und Malgeri, Michele (2003): "Extracting Logical Schema from the Web". *Applied Intelligence* 18: S. 341–355.

Carr, Leslie; Kampa, Simon; Hall, Wendy; Bechhofer, Sean und Goble, Carole (2004): "Ontologies and Hypertext". In: Staab und Studer (2004), S. 517–531.

Carroll, Jeremy J. und Roo, Jos De (2004): "OWL Web Ontology Language – Test Cases". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/2004/REC-owl-test-20040210/`.

Carstensen, Kai-Uwe; Ebert, Christian; Endriss, Cornelia; Jekat, Susanne; Klabunde, Ralf und Langer, Hagen (Hrsg.) (2004): *Computerlinguistik und Sprachtechnologie – Eine Einführung*. Heidelberg: Spektrum, 2. Auflage.

Cavnar, William B. und Trenkle, John M. (1994): "n-Gram-Based Text Categorization". In: *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR '94)*. Las Vegas, S. 161–175.

Cayzer, Steve (2004): "Semantic Blogging and Decentralized Knowledge Management". *Communications of the ACM* 47 (12): S. 47–52.

Chaffee, Jason und Gauch, Susan (2000): "Personal Ontologies for Web Navigation". In: *Proceedings of the 9th International Conference on Information and Knowledge Management (CIKM 2000)*. McLean, S. 227–237.

Chakrabarti, Soumen (2003): *Mining the Web – Discovering Knowledge from Hypertext Data*. Amsterdam, Boston, London etc.: Morgan Kaufmann.

Chakrabarti, Soumen; Dom, Byron; Kumar, S. Ravi; Raghavan, Prabhakar; Rajagopalan, Sridhar; Tomkins, Andrew; Kleinberg, Jon M. und Gibson, David (1999): "Hypersearching the Web". *Scientific American* 280 (6): S. 54–60.

Chakrabarti, Soumen; Dom, Byron E. und Indyk, Piotr (1998): "Enhanced Hypertext Categorization Using Hyperlinks". In: *Proceedings of SIGMOD'98, ACM International Conference on Management of Data*, herausgegeben von Haas, Laura M. und Tiwary, Ashutosh. Seattle, S. 307–318.

Chan, Michael und Yu, Gin (1999): "Extracting Web Design Knowledge: The Web De-Compiler". In: *IEEE International Conference on Multimedia Computing and Systems (ICMCS 1999)*. Florence, Band 2, S. 547–552.

Chandler, Daniel (1998): "Personal Home Pages and the Construction of Identities on the Web". Online verfügbar: `http://www.aber.ac.uk/media/Documents/short/webident.html`. Aberystwyth Post-International Group: Linking Theory and Practice – Issues in the Politics of Identity (9–11.09.1998), University of Wales, Aberystwyth.

Chandler, Daniel und Roberts-Young, Dilwyn (1999): "The Construction of Identity in Adolescent Personal Home Pages". In: *Internet-Based Teaching and Learning (IN-TELE) '98*, herausgegeben von Marquet, Pascal; Alain Jaillet, Stéphanie Mathey und Nissen, Elke, Frankfurt: Peter Lang, S. 461–466. In einer erweiterten Version online verfügbar unter `http://www.aber.ac.uk/media/Documents/short/strasbourg.html`.

Chen, Francine R.; Bloomberg, Dan S. und Wilcox, Lynn D. (1996): "Detection and Location of Multi-Character Sequences in Lines of Imaged Text". *Journal of Electronic Imaging* 5 (1).

Chen, Hsin-Hsi; Tsai, Shih-Chung und Tsai, Jin-He (2000): "Mining Tables from Large Scale HTML Texts". In: *The 18th International Conference on Computational Linguistics (COLING 2000)*. Saarbrücken.

Chen, Jinlin; Zhou, Baoyao; Shi, Jin; Zhang, Hongjiang und Fengwu, Qui (2001): "Function-Based Object Model Towards Website Adaption". In: *Proceedings of the 10th International World Wide Web Conference (WWW-10)*. Hong Kong, S. 587–596.

Chen, Lihui und Chue, Wai Lian (2005): "Using Web Structure and Summarisation Techniques for Web Content Mining". *Information Processing and Management* 41 (5): S. 1225–1242.

Chen, Li-Qun; Xie, Xing; Ma, Wei-Ying; Zhang, Hong-Jiang; Zhou, Heqin und Feng, Huanqing (2003a): "DRESS: A Slicing Tree Based Web Page Representation for Various Display Sizes". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Chen, Yu; Ma, Wei-Ying und Zhang, Hong-Jiang (2003b): "Detecting Web Page Structure for Adaptive Viewing on Small Form Factor Devices". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Chi, Chi-Hung; Ding, Chen und Lim, Andrew (1999): "Word Segmentation and Recognition for Web Document Framework". In: *Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM '99)*. S. 458–465.

Chidlovskii, Boris (2003): "Information Extraction from Tree Documents by Learning Subtree Delimiters". In: *Proceedings of the IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03)*, herausgegeben von Kambhampati, Subbarao und Knoblock, Craig A. Acapulco, S. 3–8.

Cho, Junghoo; Shivakumar, Narayanan und Garcia-Molina, Hector (2000): "Finding Replicated Web Collections". In: *Proceedings of the ACM International Conference on Management of Data (SIGMOD'2000)*, herausgegeben von Chen, Weidong; Naughton, Jeffrey F. und Bernstein, Philip A. Dallas.

Christensen, Erik; Curbera, Francisco; Meredith, Greg und Weerawarana, Sanjiva (2001): "Web Services Description Language (WSDL) 1.1". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/wsdl`.

Chung, Christina Yip; Gertz, Michael und Sunaresan, Neel (2002): "Reverse Engineering for Web Data: From Visual to Semantic Structures". In: *Proceedings of the 18th International Workshop on Data Engineering (ICDE '02)*. San Jose, S. 53–63.

Chung, Christina Yip; Gertz, Michael und Sundaresan, Neel (2001): "Quixote: Building XML Repositories from Topic Specific Web Documents". In: *Fourth International Workshop on the Web and Databases (WebDB 2001)*, herausgegeben von Giansalvatore Mecca, Jérôme Siméon. Santa Barbara, S. 103–108.

Ciravegna, Fabio; Dingli, Alexiei; Guthrie, David und Wilks, Yorick (2003): "Integrating Information to Bootstrap Information Extraction from Web Sites". In: *Proceedings of the IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03)*, herausgegeben von Kambhampati, Subbarao und Knoblock, Craig A. Acapulco, S. 9–14.

Ciravegna, Fabio; Dingli, Alexiei; Petrelli, Daniela und Wilks, Yorick (2002): "User-System Cooperation in Document Annotation Based on Information Extraction". In: Gómez-Pérez und Benjamins (2002), S. 122–137.

Clark, James (1999): "XSL Transformations (Version 1.0)". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/xslt/`.

Clark, James und DeRose, Steve (1999): "XML Path Language (XPath) – Version 1.0". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/xslt/`.

Cohen, William W.; Hurst, Matthew und Jensen, Lee S. (2002): "A Flexible Learning System for Wrapping Tables and Lists in HTML Documents". In: *Proceedings of the 11th International World Wide Web Conference (WWW 2002)*. Honolulu.

Cölfen, Elisabeth; Cölfen, Hermann und Schmitz, Ulrich (1997): *Linguistik im Internet*. Opladen: Westdeutscher Verlag.

Conklin, Jeff (1987): "Hypertext: An Introduction and Survey". *IEEE Computer* 20 (9): S. 17–41.

Constable, Peter und Simons, Gary (2000): "Language Identification and IT – Addressing Problems of Linguistic Diversity on a Global Scale". In: *Proceedings of the 17th International Unicode Conference*. San José. Erweiterte Fassung online erhältlich unter `http://www.sil.org/silewp/2000/001/`.

Cooley, R.; Mobasher, B. und Srivastava, J. (1997): "Web Mining: Information and Pattern Discovery on the World Wide Web". In: *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*. Newport Beach, S. 558–567.

Cooper, James W.; Coden, Anni R. und Brown, Eric W. (2002): "A Novel Method for Detecting Similar Documents". In: *Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35)*. Big Island, Hawaii.

Cowie, Jim; Ludovik, Evgeny und Zacharski, Ron (1998): "An Autonomous, Web-based, Multilingual Corpus Collection Tool". In: *Proceedings of the International Conference on Natural Language Processing and Industrial Applications*. Moncton, S. 142–148. Online verfügbar: `http://crl.nmsu.edu/~raz/langrec/nlpia.htm`.

Cowie, Jim und Wilks, Yorick (2000): "Information Extraction". In: Dale et al. (2000), S. 241–260.

Craven, Mark; DiPasquo, Dan; Freitag, Dayne; McCallum, Andrew; Mitchell, Tom; Nigam, Kamal und Slattery, Seán (1998): "Learning to Extract Symbolic Knowledge from the World Wide Web". In: *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98)*. American Association for Artificial Intelligence.

Craven, Mark; DiPasquo, Dan; Freitag, Dayne; McCallum, Andrew; Mitchell, Tom; Nigam, Kamal und Slattery, Seán (2000): "Learning to Construct Knowledge Bases from the World Wide Web". *Artificial Intelligence* (118): S. 69–113.

Crijns, Rogier (2001): "Elemente textuellen Appellierens in der digitalen Produktwerbung – Textgestaltung und kulturspezifische Appellformen im Webvertising". In: Handler (2001), S. 277–293.

Cronin, Blaise; Snyder, Herbert W.; Rosenbaum, Howard; Martinson, Anna und Callahan, Ewa (1998): "Invoked on the Web". *Journal of the American Society for Information Science* 49 (14): S. 1319–1328.

Crowston, Kevin und Kwasnik, Barbara H. (2004): "A Framework for Creating a Facetted Classification for Genres: Addressing Issues of Multidimensionality". In: *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS-37)*. Big Island, Hawaii.

Crowston, Kevin und Williams, Marie (1997): "Reproduced and Emergent Genres of Communication on the World-Wide Web". In: *Proceedings of the 30th Hawaii International Conference on Systems Sciences (HICSS-30)*. Band 6, S. 30–39.

Crowston, Kevin und Williams, Marie (1999): "The Effects of Linking on Genres of Web Documents". In: *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32)*.

Crowston, Kevin und Williams, Marie (2000): "Reproduced and Emergent Genres of Communication on the World Wide Web". *The Information Society* 16 (3): S. 201–215.

Crystal, David (2001): *Language and the Internet*. Cambridge: Cambridge University Press.

Cunliffe, Daniel (2000): "Developing usable Web sites – a review and model". *Internet Research* 10 (4): S. 295–308.

Cutler, Michael; Shih, Yungming und Meng, Weiyi (1997): "Using the Structure of HTML Documents to Improve Retrieval". In: *USENIX Symposium on Internet Technologies and Systems (NSITS '97)*. Monterey, S. 241–251.

Cyr, Dianne und Trevor-Smith, Haizley (2004): "Localization of Web design: An empirical comparison of German, Japanese, and United States Web site characteristics". *Journal of the American Society for Information Science and Technology* 55 (13): S. 1199–1208.

Dahlström, Mats (2002): "When is a Webtext?" *Text Technology* 11 (1): S. 139–161.

Dalal, Nikunji P.; Quible, Zane und Wyatt, Katherine (2000): "Cognitive Design of Home Pages: An Experimental Study of Comprehension on the World Wide Web". *Information Processing and Management* 36 (4): S. 607–621.

Dale, Robert; Moisl, Hermann und Somers, Harold (Hrsg.) (2000): *Handbook of Natural Language Processing*. New York, Basel: Marcel Dekker.

Dalgaard, Rune (2001): "Hypertext and the Scholarly Archive: Intertexts, Paratexts and Metatexts at Work". In: *Proceedings of the twelfth ACM Conference on Hypertext and Hypermedia*. Århus, S. 175–184.

Dammann, Günter (2000): "Textsorten und literarische Gattungen". In: Brinker et al. (2000), S. 546–561.

Daneš, František (1974a): "Functional Sentence Perspective and the Organization of the Text". In: Daneš (1974b), S. 106–128.

Daneš, František (Hrsg.) (1974b): *Papers on Functional Sentence Perspective*. Prag: Academia.

Daneš, František (1974c): "Zur Terminologie der funktionalen Satzperspektive". In: Daneš (1974b), S. 217–222.

Davidov, Dmitry; Gabrilovich, Evgeniy und Markovitch, Shaul (2004): "Parameterized Generation of Labeled Datasets for Text Categorization Based on a Hierarchical Directory". In: *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval*. Sheffield, S. 250–257.

Davies, John; Fensel, Dieter und van Harmelen, Frank (2003a): "Introduction". In: Davies et al. (2003b), S. 1–9.

Davies, John; Fensel, Dieter und van Harmelen, Frank (Hrsg.) (2003b): *Towards the Semantic Web – Ontology-Driven Knowledge Management*. Chichester: John Wiley.

Davulcu, Hasan; Koduri, Sukumar und Nagarajan, Saravanakumar (2003): "Datarover: A Taxonomy Based Crawler for Automated Data Extraction from Data-Intensive Websites". In: *Proceedings of the fifth ACM International Workshop on Web Information and Data Management*. New Orleans, S. 9–14.

DCES (2003): "Dublin Core Metadata Element Set, Version 1.1: Reference Description". Dublin Core Metadata Initiative, DCMI. Version vom 02.06.2003. Online verfügbar: `http://dublincore.org/documents/2003/02/04/dces/`.

DCMT (2004): "DCMI Metadata Terms". Dublin Core Metadata Initiative, DCMI. Version vom 14.06.2004. Online verfügbar: `http://dublincore.org/documents/dcmi-terms/`.

DCTV (2004): "DCMI Type Vocabulary". Dublin Core Metadata Initiative, DCMI. Version vom 14.06.2004. Online verfügbar: `http://dublincore.org/documents/dcmi-type-vocabulary/`.

Dennis, Simon; Bruza, Peter und McArthur, Robert (2002): "Web Searching: A Process-Oriented Experimental Study of Three Interactive Search Paradigms". *Journal of the American Society for Information Science and Technology* 53 (2): S. 120–133.

DeRose, Steve; Maler, Eve und Orchard, David (2001): "XML Linking Language (XLink) Version 1.0". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/xlink/`.

Dewdney, Nigel; VanEss-Dykema, Carol und MacMillan, Richard (2001): "The Form is the Substance: Classification of Genres in Text". In: *Proceedings of Workshop on Human Language Technology and Knowledge Management*. Association for Computational Linguistics, Toulouse.

Dewe, Johan; Karlgren, Jussi und Bretan, Ivan (1998): "Assembling a Balanced Corpus from the Internet". In: *Proceedings of the 11th Nordic Conference of Computational Linguistics*. Kopenhagen, S. 100–107.

DiBona, Chris; Ockman, Sam und Stone, Mark (Hrsg.) (1999): *Open Sources: Voices from the Open Source Revolution*. Peking, Cambridge, Köln etc.: O'Reilly & Associates.

Diekmannshenke, Hajo (2000): "Die Spur des Internetflaneurs – Elektronische Gästebücher als neue Kommunikationsform". In: *Soziales im Netz – Sprache, Beziehungen und Kommunikationskulturen im Internet*, herausgegeben von Thimm, Caja, Opladen, Wiesbaden: Westdeutscher Verlag, S. 131–155.

Diekmannshenke, Hajo (2004): "›Gesprächsstrategien‹ in Politik-Chats". *Osnabrücker Beiträge zur Sprachtheorie* (68): S. 123–140.

van Dijk, Teun A. (1972): *Some Aspects of Text Grammars – A Study in Theoretical Linguistics and Poetics*. The Hague, Paris: Mouton.

van Dijk, Teun A. (1980): *Macrostructures – An interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition*. Hillsdale: Erlbaum.

van Dijk, Teun A. und Kintsch, Walter (1983): *Strategies of Discourse Comprehension*. New York, London, Paris etc.: Academic Press.

Dillon, Andrew (1996): "Myths, Misconceptions, and an Alternative Perspective on Information Usage and the Electronic Medium". In: Rouet et al. (1996), S. 25–42.

Dillon, Andrew und Gushrowski, Barbara A. (2000): "Genres and the Web: Is the Personal Home Page the First Uniquely Digital Genre?" *Journal of the American Society for Information Science* 51 (2): S. 202–205.

Dillon, Andrew und Vaughan, Misha (1997): "»It's the journey and the destination«: Shape and the ermergent property of genre in evaluating digital documents". *New Review of Multimedia and Hypermedia* 3: S. 91–106.

Dimitrova, Maya und Kushmerick, Nicholas (2003): "Dimensions of Web Genre". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Dimter, Matthias (1981): *Textklassenkonzepte heutiger Alltagssprache – Kommunikationssituation, Textfunktion und Textinhalt als Kategorien alltagssprachlicher Textklassifikation*, Band 32 von *Reihe Germanistische Linguistik*. Tübingen: Niemeyer.

Dmitriev, Pavel; Lagoze, Carl und Suchkov, Boris (2005): "Finding the Boundaries of Information Resources on the Web". In: *Proceedings of the 14th International World Wide Web Conference (WWW 2005)*. Chiba, S. 1124–1125. Poster Track.

Dollar Consulting (2002): "Archival Preservation of Web Resources: HTML to XHTML Migration Test – Technical Considerations, Evaluation, and Recommendations". Von Dollar Consulting angefertigte Studie für die Smithsonian Institution Archives, 1. Juli 2002. Online verfügbar: `http://www.si.edu/archives/archives/dollarrpt2.html`.

Döring, Nicola (1997): "Lernen mit dem Internet". In: Issing und Klimsa (1997), S. 305–336.

Döring, Nicola (2001a): "Persönliche Homepages im WWW". *Medien & Kommunikationswissenschaft* 49 (3): S. 325–349.

Döring, Nicola (2001b): "Selbstdarstellung mit dem Computer". In: *Neue Medien im Alltag: Die Vielfalt individueller Nutzungsweisen*, herausgegeben von Boehnke, Klaus und Döring, Nicola, Lengerich: Pabst, S. 196–234.

Döring, Nicola (2002): "Personal Home Pages on the Web: A Review of Research". *Journal of Computer-Mediated Communication* 7 (3). Online verfügbar: `http://www.ascusc.org/jcmc/`.

Dörre, Jochen; Gerstl, Peter und Seiffert, Roland (2001): "Volltextsuche und Text Mining". In: Klabunde et al. (2001), S. 425–441.

Douglis, Fred; Feldmann, Anja; Krishnamurthy, Balachander und Mogul, Jeffrey (1997): "Rate of Change and other Metrics: a Live Study of the World Wide Web". In: *Proceedings of the 1st USENIX Symposium on Internet Technologies and Systems (USITS '97)*. Monterey, S. 147–158.

Dressler, Wolfgang (1974): "Funktionelle Satzperspektive und Texttheorie". In: Daneš (1974b), S. 87–105.

DuBois, Paul (1999): *MySQL*. Indianapolis: New Riders.

Duckett, Jon; Griffin, Oliver; Mohr, Stephen; Norton, Francis; Stokes-Rees, Ian; Williams, Kevin; Cagle, Kurt; Ozu, Nikola und Tennison, Jeni (2001): *Professional XML Schemas*. Birmingham: Wrox.

Duden, Band 01 (2004): *Die deutsche Rechtschreibung*, Band 1 von *Der Duden in 12 Bänden*. Mannheim, Leipzig, Wien, Zürich: Dudenverlag, 23. Auflage.

Dumais, Susan; Platt, John; Heckerman, David und Sahami, Mehran (1998): "Inductive Learning Algorithms and Representations for Text Categorization". In: *Proceedings of 7th International Conference on Information and Knowledge Management*. Bethesda, Maryland.

Dumais, Susan T. und Chen, Hao (2000): "Hierarchical Classification of Web Content". In: *Proceedings of the 23rd Conference on Research and Development in Information Retrieval (SIGIR 2000)*, herausgegeben von Belkin, Nicholas J.; Ingwersen, Peter und Leong, Mun-Kew. Athens, S. 256–263.

Dunning, Ted (1994): "Statistical Identification of Language". Technischer Bericht MCCS 94-273, New Mexico State University, New Mexico.

Dürscheid, Christa (1999): "Zwischen Mündlichkeit und Schriftlichkeit". *Papiere zur Linguistik* 60 (1): S. 17–30.

Dürscheid, Christa (2000): "Sprachliche Merkmale von Webseiten". *Deutsche Sprache* 28 (1): S. 60–73.

Dürscheid, Christa (2004): "Netzsprache – ein neuer Mythos". *Osnabrücker Beiträge zur Sprachtheorie* (68): S. 141–157.

ECMAScript (1999): "ECMAScript Language Specification". Standard ECMA-262, 3rd Edition, European Computer Manufacturers Association. Online verfügbar: `http://www.ecma-international.org`.

Eckkrammer, Eva Martha (2001): "Textsortenkonventionen im Medienwechsel". In: Handler (2001), S. 45–66.

Efimova, Lilia und de Moor, Aldo (2005): "Beyond Personal Webpublishing: An Exploratory Study of Conversational Blogging Practices". In: *Proceedings of the 38th Hawaii International Conference on Systems Sciences (HICSS-38)*. Big Island, Hawaii.

Eichhoff-Cyrus, Karin M. (2000): "Vom Briefsteller zur Nettikette: Textsorten gestern und heute". In: *Die deutsche Sprache zur Jahrtausendwende*, herausgegeben von Eichhoff-Cyrus, Karin M. und Hoberg, Rudolf, Mannheim, Leipzig, Wien, Zürich: Dudenverlag, S. 53–62.

Eikvil, Line (1999): "Information Extraction from World Wide Web – A Survey". Technischer Bericht 945, Norwegian Computing Center.

Eiron, Nadav und McCurley, Kevin S. (2003): "Untangling Compound Documents on the Web". In: *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia*. S. 85–94. Nottingham.

El-Bayoumi, Janice G. (1999): "Focus on Service – Departmental Web Site Reorganization". In: *Proceedings of the 27th annual ACM SIGUCCS Conference on User Services: Mile High Expectations*. Denver, S. 56–60.

Embley, David W.; Campbell, Douglas M.; Smith, Randy D. und Liddle, Stephen W. (1998): "Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Documents". In: *Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM '98)*. Bethesda, S. 52–59.

Embley, David W.; Tao, Cui und Liddle, Stephen W. (2002): "Automatically Extracting Ontologically Specified Data from HTML Tables with Unknown Structure". In: *Conceptual Modelling – ER 2002: 21st International Conference on Conceptual Modeling*, herausgegeben von Spaccapietra, Stefano; March, Salvatore T. und Kambayashi, Yahiko, Berlin, Heidelberg, New York etc.: Springer, Band 2503 von *Lecture Notes in Computer Science*, S. 322–337.

Emde, Werner (1991): "Managing Lexical Knowledge in LEU/2". In: *Text Understanding in LILOG*, herausgegeben von Herzog, Otthein und Rollinger, Claus-Rainer, Berlin, Heidelberg: Springer, Band 546 von *Lecture Notes in Artificial Intelligence*, S. 167–179.

Emigh, William und Herring, Susan C. (2005): "Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias". In: *Proceedings of the 38th Hawaii International Conference on Systems Sciences (HICSS-38)*. Big Island, Hawaii.

Endres, Albert und Fellner, Dieter W. (2000): *Digitale Bibliotheken*. Heidelberg: dpunkt.

Endres-Niggemeyer, Brigitte (1998): *Summarizing Information*. Berlin, Heidelberg, New York etc.: Springer.

Endres-Niggemeyer, Brigitte (2004): "Automatisches Textzusammenfassen". In: Lobin und Lemnitzer (2004), S. 407–432.

Engelbart, Douglas C. (1962): "Letter to Vannevar Bush and Program On Human Effectiveness". In: Nyce und Kahn (1991a), S. 235–244.

Engelbart, Douglas C. und English, William K. (1968): "A research center for augmenting human intellect". In: *American Federation of Information Processing Societies Conference Proceedings of the Fall Joint Computer Conference*. Washington D. C.: Thompson, Band 33, S. 395–410.

Erdmann, Michael (2001): *Ontologien zur konzeptuellen Modellierung der Semantik von XML*. Dissertation, Fakultät für Wirtschaftswissenschaften der Universität Fridericiana zu Karlsruhe, Karlsruhe. Erhältlich als Book on Demand.

Erickson, Thomas (1996): "The World Wide Web as Social Hypertext". *Communications of the ACM* 39 (1): S. 15–17.

Erickson, Thomas (1997): "Social Interaction on the Net: Virtual Community as Participatory Genre". In: *Proceedings of the 30th Hawaii International Conference on Systems Sciences (HICSS-30)*. Band 6, S. 13–21.

Erickson, Thomas (1999): "Ryhme and Punishment: The Creation and Enforcement of Conventions in an On-Line Participatory Limerick Genre". In: *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32)*.

Erickson, Thomas (2000): "Making Sense of Computer-Mediated Communication (CMC): Conversations as Genres, CMC Systems as Genre Ecologies". In: *Proceedings of the 33rd Hawaii International Conference on Systems Sciences (HICSS-33)*.

Eriksen, Lars Bo (1997): "Digital Documents, Work and Technology, Three Cases of Internet News Publishing". In: *Proceedings of the 30th Hawaii International Conference on Systems Sciences (HICSS-30)*. Band 6, S. 87–96.

Eriksen, Lars Bo und Ihlström, Carina (1999): "In the Path of the Pioneers – Longitudinal Study of Web News Genre". In: *Proceedings of the 22nd Information Systems Research Seminar in Scandinavia (IRIS 22): "Enterprise Architectures for Virtual Organizations"*, herausgegeben von Käkölä, Timo K. University of Jyväskylä, Keuruu, S. 289–304.

Eriksen, Lars Bo und Ihlström, Carina (2000): "Evolution of the Web News Genre – The Slow Move Beyond the Print Metaphor". In: *Proceedings of the 33rd Hawaii International Conference on Systems Sciences (HICSS-33)*.

Eriksen, Lars Bo und Sørgaard, Pål (1996): "Organisational Implementation of WWW in Scandinavian Newspapers: Tradition Based Approaches Dominate". In: *Proceedings of the 19th Information Systems Research Seminar in Scandinavia (IRIS 19)*, herausgegeben von Dahlbom, Bo; Ljungberg, Fredrik; Nuldén, Urban; Simon, Kai; Stage, Jan und Sørensen, Carsten. University of Götheberg, Lökeberg, S. 333–349.

Essid, Joe (2004): "Film as explicador for hypertext". *Computers and the Humanities* 38: S. 317–333.

Ester, Martin; Kriegel, Hans-Peter und Schubert, Matthias (2002): "Web Site Mining: A new way to spot Competitors, Customers and Suppliers in the World Wide Web". In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. Edmonton, S. 249–258.

Etzioni, Oren (1996): "The World-Wide Web: Quagmire or Gold Mine?" *Communications of the ACM* 39 (11): S. 65–68.

Fairon, Cédrick (2000): "GlossaNet: Parsing a Web Site as a Corpus". *Linguisticæ Investigationes* 22 (2): S. 327–340.

Fallside, David C.; Thompson, Henry S.; Beech, David; Maloney, Murray; Mendelsohn, Noah; Biron, Paul V. und Malhotra, Ashok (2001): "XML Schema". Technische Spezifikation, W3C. Besteht aus Part 0 (Primer), Part 1 (Structures), Part 2 (Datatypes). Online verfügbar: `http://www.w3.org/XML/Schema`.

Feilke, Helmuth (2000): "Die pragmatische Wende in der Textlinguistik". In: Brinker et al. (2000), S. 64–82.

Feldweg, Helmut; Kibiger, Ralf und Thielen, Christine (1995): "Zum Sprachgebrauch in deutschen Newsgruppen". *Osnabrücker Beiträge zur Sprachtheorie* (50): S. 143–154.

Fellbaum, Christiane (Hrsg.) (1998): *Wordnet – An Electronic Lexical Database*. Cambridge: MIT Press.

Fensel, Dieter; van Harmelen, Frank und Horrocks, Ian (2003a): "OIL and DAML+OIL: Ontology Languages for the Semantic Web". In: Davies et al. (2003b), S. 11–31.

Fensel, Dieter; Hendler, James; Lieberman, Henry und Wahlster, Wolfgang (Hrsg.) (2003b): *Spinning the Semantic Web – Bringing the World Wide Web to its Full Potential*. Cambridge, London: MIT Press.

Ferraiolo, Jon; Fujisawa, Jun und Jackson, Dean (2003): "Scalable Vector Graphics (SVG) 1.1 Specification". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/SVG11/`.

Figge, Udo L. (2000): "Die kognitive Wende in der Textlinguistik". In: Brinker et al. (2000), S. 96–104.

Finin, Tim; Mayfield, James; Joshi, Anupam; Cost, R. Scott und Fink, Clay (2005): "Information Retrieval on the Semantic Web". In: *Proceedings of the 38th Hawaii International Conference on Systems Sciences (HICSS-38)*. Big Island, Hawaii.

Finn, Aidan; Kushmerick, Nicholas und Smyth, Barry (2001): "Fact or Fiction: Content Classification for Digital Libraries". In: *Proceedings of the Second Joint DELOS-NSF Workshop on Personalisation and Recommender Systems in Digital Libraries*. Network of Excellence for Digital Libraries, Dublin. Online verfügbar: `http://www.ercim.org/publication/ws-proceedings/DelNoe02/index.html`.

Finn, Aidan; Kushmerick, Nicholas und Smyth, Barry (2002): "Genre Classification and Domain Transfer for Information Filtering". In: *Advances in Information Retrieval – 24th BCS-IRSG European Colloquium on IR Research*, herausgegeben von Crestani, F.; Girolami, M. und van Rijsbergen, C. J., Berlin, Heidelberg, New York etc.: Springer, Band 2291 von *Lecture Notes in Computer Science*, S. 353–362.

Firth, David und Lawrence, Cameron (2003): "Genre Analysis in Information Systems Research". *Journal of Information Technology Theory and Application* 5 (3): S. 63–77.

Flämig, Christian; Kimminich, Otto; Krüger, Hartmut; Meusel, Ernst-Joachim; Rupp, Hans Heinrich; Scheven, Dieter; Schuster, Hermann Josef und Stenbock-Fermor, Friedrich Graf (Hrsg.) (1996): *Handbuch des Wissenschaftsrechts*. Berlin, Heidelberg, New York: Springer, 2. Auflage.

Fleming, Jennifer (1998): *Web Navigation: Designing the User Experience*. Cambridge, Köln, Paris etc.: O'Reilly & Associates.

Flender, Jürgen und Christmann, Ursula (2000): "Hypertext: prototypische Merkmale und deren Realisierung im Hypertext »Visuelle Wahrnehmung«". *Medienpsychologie* 12 (2): S. 95–116.

Fletcher, William H. (2001): "Concordancing the Web with KWICFinder". In: *Proceedings of the 3rd North American Symposium on Corpus Linguistics and Language Teaching*. Boston.

Florescu, Daniela; Levy, Alon und Mendelzon, Alberto (1998): "Database Techniques for the World-Wide Web: A Survey". *SIGMOD Record* 27 (3): S. 59–74.

Fluit, Christiaan; Sabou, Marta und van Harmelen, Frank (2004): "Supporting User Tasks through Visualization of Lightweight Ontologies". In: Staab und Studer (2004), S. 415–432.

Fogg, B. J.; Soohoo, Cathy; Danielson, David R.; Marable, Leslie; Stanford, Julianne und Tauber, Ellen R. (2003): "How do users evaluate the credibility of web sites?: a study with over 2,500 participants". In: *Proceedings of the 2003 Conference on Designing for User Experiences*. San Francisco, S. 1–15.

Fogg, B. J.; Swani, Preeti; Treinen, Marissa; Marshall, Jonathan; Laraki, Othman; Osipovich, Alex; Varma, Chris; Fang, Nicholas; Paul, Jyoti; Rangnekar, Akshay und Shon, John (2001): "What Makes Web Sites Credible?: A Report on A Large Quantitative Study". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Seattle, S. 61–68.

Foltz, Peter W. (1996): "Comprehension, Coherence, and Strategies in Hypertext". In: Rouet et al. (1996), S. 109–136.

Ford, Karen Susan (1995): "Converting from Paper to HTML". In: *Proceedings of the 23rd annual ACM SIGUCCS Conference on User Services: Winning the Networking Game*. St. Louis, S. 115–118.

Ford, Nigel; Miller, David und Moss, Nicola (2001): "The Role of Individual Differences in Internet Searching: An Empirical Study". *Journal of the American Society for Information Science and Technology* 52 (12): S. 1049–1066.

Fortanet, Inmaculada; Palmer, Juan Carlos und Posteguillo, Santiago (1998): "Netvertising: Content-Based Subgeneric Variations in a Digital Genre". In: *Proceedings of the 31st Hawaii International Conference on Systems Sciences (HICSS-31)*. Big Island, Hawaii, Band 2, S. 87–96.

Fortanet, Inmaculada; Palmer, Juan Carlos und Posteguillo, Santiago (1999): "The Emergence of a New Genre: Advertising on the Internet (netvertising)". *Hermes, Journal of Linguistics* (23): S. 93–113.

Foucou, Pierre-Yves und Kübler, Natalie (2000): "A web-based environment for teaching technical English". In: *Rethinking Language Pedagogy from a Corpus Perspective – Papers from the third international conference on Teaching and Language Corpora*, herausgegeben von Burnard, Lou und McEnery, Tony, Frankfurt/Main, Berlin, Bern etc.: Peter Lang, Band 2 von *Łódź Studies in Language*, S. 65–73.

Fox, Edward A.; McMillan, Gail und Eaton, John L. (1999): "The Evolving Genre of Electronic Theses and Dissertations". In: *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32)*.

Franconi, Enrico (2003): "Natural Language Processing". In: *The Description Logic Handbook – Theory, Implementation and Applications*, herausgegeben von Baader, Franz; Calvanese, Diego; McGuinness, Deborah L.; ; Nardi, Daniele und Patel-Schneider, Peter F., Cambridge: Cambridge University Press, S. 450–461.

Freisler, Stefan (1994): "Hypertext – Eine Begriffsbestimmung". *Deutsche Sprache* 22: S. 19–50.

Friedl, Jeffrey E. F. (1997): *Mastering Regular Expressions*. Cambridge, Köln, Paris etc.: O'Reilly & Associates.

Fritz, Gerd (1999): "Coherence in Hypertext". In: *Coherence in Spoken and Written Discourse*, herausgegeben von Bublitz, Wolfram; Lenk, Uta und Ventola, Eija, Amsterdam, Philadelphia: John Benjamins, Band 63 von *Pragmatics And Beyond New Series*, S. 221–232.

Fürnkranz, Johannes (1999): "Exploiting Structural Information for Text Classification on the WWW". In: *Advances in Intelligent Data Analysis: Proceedings of the 3rd Symposium (IDA-99)*, Berlin, Heidelberg, New York etc.: Springer, Band 1642 von *Lecture Notes in Computer Science*, S. 487–497.

Furuta, Richard (1989): "An Object-based Taxonomy for Abstract Structure in Document Models". *The Computer Journal* 32 (6): S. 494–504.

Furuta, Richard und Marshall, Catherine C. (1996): "Genre as Reflection of Technology in the World-Wide Web". In: *Hypermedia Design, Proceedings of the International Workshop on Hypermedia Design (IWHD 1995)*, herausgegeben von Fraïssé, Sylvain; Garzotto, Franca; Isakowitz, Tomás; Nanard, Jocelyne und Nanard, Marc, Berlin, Heidelberg, New York etc.: Springer, Workshops in Computing, S. 182–195.

Gaberell, Roger (2000): "Probleme einer deutschen Textsortengeschichte – die »Anfänge«". In: Adamzik (2000a), S. 155–174.

Gansel, Christina und Jürgens, Frank (2002): *Textlinguistik und Textgrammatik*, Band 6 von *Studienbücher zur Linguistik*. Wiesbaden: Westdeutscher Verlag.

Gansner, Emden R. und North, Stephen C. (2000): "An Open Graph Visualization System and its Applications to Software Engineering". *Software – Practice and Experience* 30 (11): S. 1203–1233.

Gao, Xiaoying und Sterling, Leon (1999): "AutoWrapper: Automatic Wrapper Generation for Multiple Online Services". In: *Proceedings of Asia Pacific Web Conference 1999 (APWeb99)*. S. 61–70.

Garfinkel, Simson und Spafford, Gene (1996): *Practical UNIX & Internet Security*. Cambridge, Köln, Paris etc.: O'Reilly & Associates, 2. Auflage.

Gattermann, Günter (1996): "Wissenschaftliche Einrichtungen". In: Flämig et al. (1996), S. 897–928.

Genette, Gérard (2001): *Paratexte – Das Buch vom Beiwerk des Buches*, Band 1510 von *Suhrkamp Taschenbuch Wissenschaft*. Frankfurt/Main: Suhrkamp.

Geroimenko, Vladimir und Chen, Chaomei (Hrsg.) (2003): *Visualizing the Semantic Web – XML-based Internet and Information Visualization*. London, Berlin, Heidelberg: Springer.

Géry, Mathias (2002a): "Considering HyperDocuments and Context for Indexing the Web". In: *Proceedings of the International Conference on Artificial Intelligence (IC-AI'02)*. Las Vegas.

Géry, Mathias (2002b): "Non-linear reading for a structured Web indexation". In: *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02)*. Tampere.

Géry, Mathias und Chevallet, Jean-Pierre (2001): "Toward a Structured Information Retrieval System on the Web: Automatic Structure Extraction of Web Pages". In: *International Workshop on Web Dynamics*. London.

Ghaoui, Claude; George, Steven M.; Rada, Roy; Beer, Martin D. und Getta, Janus (1990): "Text to Hypertext and Back Again". In: *Computers and Writing: The State of the Art*, herausgegeben von Holt, Patrik O'Brian und Williams, Noel, Oxford: Intellect Books, S. 109–130.

Gibson, David; Punera, Kunal und Tomkins, Andrew (2005): "The Volume and Evolution of Web Page Templates". In: *Proceedings of the 14th International World Wide Web Conference (WWW 2005)*. Chiba, S. 830–839. Refereed Papers Track.

Giguet, Emmanuel (1995): "Multilingual Sentence Categorization According to Language". In: *Proceedings of the European Chapter of the Association for Computational Linguistics SIGDAT Workshop, "From Text to Tags: Issues in Multilingual Language Analysis"*. Dublin, S. 73–76.

Giguet, Emmanuel (1996): "The Stakes of Multilinguality: Multilingual Text Tokenization in Natural Language Diagnosis". In: *Proceedings of the 4th Pacific Rim International Conference on Artificial Intelligence Workshop "Future Issues for Multilingual Text Processing"*. Cairns. Online verfügbar: `http://users.info.unicaen.fr/~giguet/diagnostic.html`.

Gillenson, Mark; Sherrell, Daniel L. und da Chen, Lei (2000): "A Taxonomy of Web Site Traversal Patterns and Structures". *Communications of the AIS* 3 (4).

Glover, Eric J.; Tsioutsiouliklis, Kostas; Lawrence, Steve; Pennock, David M. und Flake, Gary W. (2002): "Using Web Structure for Classifying and Describing Web Pages". In: *Proceedings of the 11th International World Wide Web Conference (WWW 2002)*. Honolulu.

Gobbetti, Enrico und Turner, Russell (1997): "Exploring Annotated 3D Environments on the World Wide Web". In: *Intelligent Hypertext – Advanced Techniques for the World Wide Web*, herausgegeben von Nicholas, Charles und Mayfield, James, Berlin, Heidelberg, New York etc.: Springer, Band 1326 von *Lecture Notes in Computer Science*, S. 31–46.

Goertz, Lutz (1995): "Wie interaktiv sind Medien?" In: Bieber und Leggewie (2004), S. 97–117. Ursprünglich erschienen in: *Rundfunk und Fernsehen* (4), S. 477–493.

Goldfarb, Charles F. (1990): *The SGML Handbook*. Oxford: Oxford University Press.

Goldman, Claudia V.; Langer, Amir und Rosenschein, Jeffrey S. (1997): "Musag: An Agent that Learns what You Mean". *Applied Artificial Intelligence* 11 (5): S. 413–435.

Goller, Christoph; Löning, Joachim; Will, Thilo und Wolff, Werner (2000): "Automatic Document Classification: A thorough Evaluation of Various Methods". In: *7. Internationales Symposium für Informationswissenschaft*. Darmstadt.

Gómez-Pérez, Asunción und Benjamins, V. Richard (Hrsg.) (2002): *Knowledge Engineering and Knowledge Management – Ontologies and the Semantic Web*, Band 2473 von *Lecture Notes in Artificial Intelligence*. Berlin, Heidelberg, New York etc.: Springer.

Goodman, Danny (1998): *Dynamic HTML: The Definitive Guide*. Cambridge, Köln, Paris etc.: O'Reilly & Associates.

Goosens, Michel und Rahtz, Sebastian (1999): *The LaTeX Web Companion – Integrating TeX, HTML, and XML*. Reading, Harlow. Menlo Park etc.: Addison-Wesley.

Göpferich, Susanne (1995): *Textsorten in Naturwissenschaften und Technik: Pragmatische Typologie – Kontrastierung – Translation*, Band 27 von *Forum für Fachsprachen-Forschung*. Tübingen: Gunter Narr.

Görz, Günther (Hrsg.) (1995): *Einführung in die Künstliche Intelligenz*. Bonn, Reading, Menlo Park etc.: Addison-Wesley, 2. Auflage.

Gourley, David; Totty, Brian; Sayer, Marjorie; Reddy, Sailu und Aggarwal, Anshu (2002): *HTTP – The Definitive Guide*. Peking, Cambridge, Farnham etc.: O'Reilly & Associates.

Graefen, Gabriele (1997): *Der Wissenschaftliche Artikel – Textart und Textorganisation*, Band 27 von *Arbeiten zur Sprachanalyse*. Frankfurt/Main, Berlin, Bern etc.: Peter Lang.

Graham, Leah und Metaxas, Panagiotis Takis (2003): "»Of course it's true; I saw it on the Internet!«: Critical Thinking in the Internet Era". *Communications of the ACM* 46 (5): S. 70–75.

Grant, Jan und Beckett, Dave (2004): "RDF Test Cases". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/2004/REC-rdf-testcases-20040210/`.

Greenhill, Stewart und Venkatesh, Svetha (1999): "Constructing and navigating personalised views of the web". *Information Processing and Management* 35 (5): S. 679–689.

Greenspun, Philip (1995): "We Have Chosen Shame and Will Get War". *Electronic Publishing* 1 (1). Online verfügbar: `http://philip.greenspun.com/research/shame-and-war`.

Greenspun, Philip (1996): "Shame and War Revisited – Adding Semantic Markup to HTML". Online verfügbar: `http://philip.greenspun.com/research/shame-and-war-revisited`.

Greenspun, Philip (1999): *Philip and Alex's Guide to Web Publishing*. San Francisco: Morgan Kaufmann.

Grefenstette, Gregory (1999): "The World Wide Web as a resource for example-based machine translation tasks". In: *Proceedings of the ASLIB Conference on Translating and the Computer*. London.

Grefenstette, Gregory und Nioche, Julien (2000): "Estimation of English and non-English Language Use on the WWW". In: *Proceedings of RIAO'2000: Content-Based Multimedia Information Access*. Paris, S. 237–246. Online verfügbar: `http://de.arXiv.org/abs/cs.CL/0006032`.

Grefenstette, Gregory und Tapanainen, Pasi (1994): "What is a Word, What is a Sentence? – Problems of Tokenization". In: *Proceedings of the Third Conference on Computational Lexicography and Text Research (COMPLEX '94)*. Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, S. 79–87. Online verfügbar: `http://www.xerox.fr/grenoble/mltt/articles/home.html`.

Grice, Paul H. (1975): "Logic and Conversation". In: *Syntax and Semantics*, herausgegeben von Cole, Peter und Morgan, Jerry L., New York: Academic Press, Band 3, S. 41–58.

Grigar, Dene (2002): "MOOtextuality". *Text Technology* 11 (1): S. 163–179.

Grigoleit, Uwe (1995): *Internet vertraulich – Highlights, Expertentips, Undokumentierte Features*. Düsseldorf, San Francisco, Paris, Soest: Sybex.

Grimes, Deborah J. und Boening, Carl H. (2001): "Worries with the Web: A Look at Student Use of Web Resources". *College & Research Libraries* 62 (1): S. 11–23.

Grishman, Ralph (2003): "Information Extraction". In: Mitkow (2003), S. 545–559.

Grosso, Paul (1997): "Entity Management: OASIS Technical Resolution 9401 (Amendment 2 to TR 9401)". OASIS – Organization for the Advancement of Structured Information Standards, Entity Management Subcommittee. Online verfügbar: `http://www.oasis-open.org/specs/a401.htm`.

Groth, Kristina (1998): "Personal Home Pages on the World Wide Web – A Simple Version of a Knowledge Net?" *Trends in Communication* 6 (4): S. 47–60.

Groß, Annette (2000): "Prozesse beim Verstehen fremdsprachlicher Hypertexte". *Sprache und Datenverarbeitung* 24 (1): S. 5–22.

Große, Ernst Ulrich (1976): *Text und Kommunikation – Eine linguistische Einführung in die Funktionen der Texte*. Stuttgart, Berlin, Köln, Mainz: Kohlhammer.

Gruber, Helmut (1997): "Themenentwicklung in wissenschaftlichen E-mail-Diskussionslisten. Ein Vergleich zwischen einer moderierten und einer nichtmoderierten Liste". In: Weingarten (1997a), S. 105–128.

Gruber, Helmut (2000): "Scholarly Email Discussion List Postings: a single new genre of academic communication". In: Pemberton und Shurville (2000), S. 36–43.

Gruber, Thomas R. (1993): "A Translation Approach to Portable Ontology Specifications". *Knowledge Acquisition* 5 (2): S. 199–220.

Grzega, Joachim (1999): "Some Observations on E-Mail Style vs. Traditional Style". *Papiere zur Linguistik* 60 (1): S. 3–16.

Gudivada, Venkat N.; Raghavan, Vijay V.; Grosky, William I. und Kasanagottu, Rajesh (1997): "Information Retrieval on the World Wide Web". *IEEE Internet Computing* 1 (5): S. 58–68.

Guha, Ramanathan; McCool, Rob und Miller, Eric (2003): "Semantic Search". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Gülich, Elisabeth und Raible, Wolfgang (Hrsg.) (1972): *Textsorten – Differenzierungskriterien aus linguistischer Sicht*, Band 5 von *Athenäum-Skripten Linguistik*. Frankfurt/Main: Athenäum.

Gülich, Elisabeth und Raible, Wolfgang (1977): *Linguistische Textmodelle – Grundlagen und Möglichkeiten*, Band 130 von *UTB für Wissenschaft*. München: Fink.

Gulli, Antonio und Signorini, Alessio (2005): "The Indexable Web is More than 11.5 Billion Pages". In: *Proceedings of the 14th International World Wide Web Conference (WWW 2005)*. Chiba, S. 902–903. Poster Track.

Gullikson, Shelley; Blades, Ruth; Bragdon, Marc; McKibbon, Shelley; Sparling, Marnie und Toms, Elaine G. (1999): "The impact of information architecture on academic web site usability". *The Electronic Library* 17 (5): S. 293–305.

Günther, Ulla und Wyss, Eva Lia (1996): "E-mail-Briefe – eine neue Textsorte zwischen Mündlichkeit und Schriftlichkeit". In: Hess-Lüttich et al. (1996), S. 61–86.

Gupta, Suhit; Kaiser, Gail; Neistadt, David und Grimm, Peter (2003): "DOM-based Content Extraction of HTML Documents". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Haack, Johannes (1997): "Interaktivität als Kennzeichen von Multimedia und Hypermedia". In: Issing und Klimsa (1997), S. 151–166.

Haas, Stephanie W. und Grams, Erika S. (1998a): "A Link Taxonomy for Web Pages". In: *Proceedings of the 61st Annual Meeting of the American Society for Information Science*, herausgegeben von Preston, C. S. 485–495.

Haas, Stephanie W. und Grams, Erika S. (1998b): "Page and Link Classifications: Connecting Diverse Resources". In: *Proceedings of Digital Libraries '98 – Third ACM Conference on Digital Libraries*, herausgegeben von Witten, I.; Akscyn, R. und Shipman, F. Pittsburgh, S. 99–107.

Haas, Stephanie W. und Grams, Erika S. (2000): "Readers, Authors, and Page Structure – A Discussion of Four Questions Arising from a Content Analysis of Web Pages". *Journal of the American Society for Information Science* 51 (2): S. 181–192.

Haase, Martin; Huber, Michael; Krumeich, Alexander und Rehm, Georg (1997): "Internetkommunikation und Sprachwandel". In: Weingarten (1997a), S. 51–85.

Habel, Christopher (1986): "Stories – An Artificial Intelligence Perspective (?)". *Poetics* 15: S. 111–125.

Halavais, Alexander M. Campbell (2001): *The Slashdot Effect: Analysis of a Large-Scale Public Conversation on the World Wide Web*. Ph. d. thesis, University of Washington. Online verfügbar: `http://alex.halavais.net/research/diss.pdf`.

Hammwöhner, Rainer (1997): *Offene Hypertextsysteme – Das Konstanzer Hypertextsystem (KHS) im wissenschaftlichen und technischen Kontext*, Band 32 von *Schriften zur Informationswissenschaft*. Konstanz: Universitätsverlag Konstanz.

Handler, Peter (1997): "Stileigenschaften elektronisch vermittelter Wissenschaftstexte". In: Jakobs und Knorr (1997), S. 89–108.

Handler, Peter (Hrsg.) (2001): *E-Text: Strategien und Kompetenzen – Elektronische Kommunikation in Wissenschaft, Bildung und Beruf*, Band 7 von *Textproduktion und Medium*. Frankfurt/Main, Berlin, Bern etc.: Peter Lang.

Handschuh, Siegfried; Staab, Steffen und Ciravegna, Fabio (2002): "S-CREAM – Semi-automatic CREAtion of Metadata". In: Gómez-Pérez und Benjamins (2002), S. 358–372.

Handschuh, Siegfried; Staab, Steffen und Volz, Raphael (2003): "On Deep Annotation". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Harrison, Claire (2002): "Hypertext Links: Whither Thou Goest, and Why". *First Monday* 7 (10). Online verfügbar: `http://firstmonday.org/issues/issue7_10/harrison/index.html`.

Hars, Alexander und Ou, Shaosong (2001): "Working for Free? – Motivations of Participating in Open Source Projects". In: *Proceedings of the 34th Hawaii International Conference on Systems Sciences (HICSS-34)*.

Hartmann, Peter (1964): "Text, Texte, Klassen von Texten". *Bogawus* 2: S. 15–25.

Hartmann, Peter (1971): "Texte als linguistisches Objekt". In: *Beiträge zur Textlinguistik*, herausgegeben von Stempel, Wolf-Dieter, München: Fink, S. 9–29.

Hawking, David; Craswell, Nick und Harman, Donna (1999a): "Results and Challenges in Web Search Evaluation". In: *The Eigth International World Wide Web Conference*, herausgegeben von Tang, E. International World Wide Web Conference Committee, Foretec Seminars, NRC Canada, Toronto.

Hawking, David; Craswell, Nick und Thistlewaite, Paul (1999b): "Overview of the TREC-7 Very Large Collection Track". In: *The Seventh Text REtrieval Conference (TREC 7)*. National Institute of Standards and Technology, S. 91–104. NIST Special Publication 500-242.

Hawking, David; Voorhees, Ellen; Craswell, Nick und Bailey, Peter (2000): "Overview of the TREC-8 Web Track". In: *The Eigth Text REtrieval Conference (TREC 8)*. National Institute of Standards and Technology, S. 131–150. NIST Special Publication 500-246.

Hayes, Patrick (2004): "RDF Semantics". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/2004/REC-rdf-mt-20040210/`.

Heflin, Jeff (2004): "OWL Web Ontology Language – Use Cases and Requirements". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/2004/REC-webont-req-20040210/`.

Heiber, Daniela (2001): "Die Textsorte »Lebenslauf« in studentischen Homepages". *Punkt.de – Online Journal für deutsche Sprache, Literatur und Landeskunde* Online verfügbar: `http://punktde.ruhr-uni-bochum.de`.

Heinemann, Margot (2000a): "Textsorten des Alltags". In: Brinker et al. (2000), S. 604–614.

Heinemann, Margot (2000b): "Textsorten des Bereichs Hochschule und Wissenschaft". In: Brinker et al. (2000), S. 702–709.

Heinemann, Margot und Heinemann, Wolfgang (2002): *Grundlagen der Textlinguistik*, Band 230 von *Reihe Germanistische Linguistik*. Tübingen: Niemeyer.

Heinemann, Wolfgang (2000c): "Aspekte der Textsortendifferenzierung". In: Brinker et al. (2000), S. 523–546.

Heinemann, Wolfgang (2000d): "Textsorte – Textmuster – Texttyp". In: Brinker et al. (2000), S. 507–523.

Heinemann, Wolfgang (2000e): "Textsorten. Zur Diskussion um Basisklassen des Kommunizierens. Rückschau und Ausblick". In: Adamzik (2000a), S. 9–29.

Heinemann, Wolfgang und Viehweger, Dieter (Hrsg.) (1991): *Textlinguistik – Eine Einführung*, Band 115 von *Reihe Germanistische Linguistik*. Tübingen: Niemeyer.

Heines, Jesse; Börner, Katy; Ivory, Melody Y. und Gehringer, Edward F. (2003): "Panel on the Development, Maintenance, and Use of Course Web Sites". In: *Proceedings of the 34th SIGCSE Technical Symposium on Computer Science Education*. Reno, S. 94–95.

Heißing, Christian (2000): *Klassifizieren von URLs durch Generalisierung von Pfadnamen*. Magisterarbeit, Institut für Semantische Informationsverarbeitung, Universität Osnabrück.

Helander, Martin G.; Landauer, Thomas K. und Prabhu, Prasad V. (Hrsg.) (1997): *Handbook of Human-Computer Interaction*. Amsterdam, Lausanne, New York etc.: Elsevier, 2. Auflage.

Hemenway, Keven und Calishain, Tara (2003): *Spidering Hacks – 100 Industrial-Strength Tips & Tools*. Peking, Cambridge, Farnham etc.: O'Reilly & Associates.

Herring, Susan C.; Kopuer, Inna; Paolillo, John C.; Scheidt, Lois Ann; Tyworth, Michael; Welsch, Peter; Wright, Elijah und Yu, Ning (2005): "Conversations in the Blogosphere: An Analysis »From the Bottom Up«". In: *Proceedings of the 38th Hawaii International Conference on Systems Sciences (HICSS-38)*. Big Island, Hawaii.

Herring, Susan C.; Scheidt, Lois Ann; Bonus, Sabrina und Wright, Elijah (2004): "Bridging the Gap: A Genre Analysis of Weblogs". In: *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS-37)*. Big Island, Hawaii.

Hess-Lüttich, Ernest W. B. (1997): "Text, Intertext, Hypertext – Zur Texttheorie der Hypertextualität". In: Klein und Fix (1997), S. 125–148.

Hess-Lüttich, Ernest W. B.; Holly, Werner und Püschel, Ulrich (Hrsg.) (1996): *Textstrukturen im Medienwandel*, Band 29 von *Forum Angewandte Linguistik*. Frankfurt/Main, Berlin, Bern etc.: Peter Lang.

Heydon, Allan und Najork, Marc (1999): "Mercator: A Scalable, Extensible Web Crawler". *World Wide Web* 2 (4): S. 219–229.

Heyer, Gerhard; Läuter, Martin; Quasthoff, Uwe und Wolff, Christian (2001): "Wissensextraktion durch linguistisches Postprocessing bei der Corpusanalyse". In: Lobin (2001b), S. 71–83.

Heyer, Gerhard und Wolff, Christian (1999): "Strukturierungshilfen für Hypermediadokumente und ihre Umsetzung". In: Lobin (1999b), S. 89–119.

Hinman, Lawrence M. (2002): "Academic Integrity and the World Wide Web". *ACM SIGCAS Computers and Society* 32 (1): S. 33–42.

Hirai, Jun; Raghavan, Sriram; Garcia-Molina, Hector und Paepcke, Andreas (2000): "WebBase: A Repository of Web Pages". In: *Proceedings of the 9th International World Wide Web Conference*. Amsterdam, S. 277–293.

Hirst, Graeme (2004): "Ontology and the Lexicon". In: Staab und Studer (2004), S. 209–229.

Ho, James (1997): "Evaluating the World Wide Web: A Global Study of Commercial Sites". *Journal of Computer-Mediated Communication* 3 (1). Online verfügbar: `http://www.ascusc.org/jcmc/`.

Hoffmann, Ludger (2004): "Chat und Thema". *Osnabrücker Beiträge zur Sprachtheorie* (68): S. 103–122.

Hofmann, Martin und Simon, Lothar (1995): *Problemlösung Hypertext: Grundlagen – Entwicklung – Anwendung*. München, Wien: Hanser.

Hofstede, Geert H. (1980): *Culture's consequences: International differences in work-related values*. Beverly Hills: Sage.

Holloway, Janet (1987): "The Evolution of a Genre: The Computer Center Newsletter". In: *Proceedings of the ACM SIGUCCS Conference XV on User Services*. Kansas City, S. 217–224.

Honkaranta, Anne (2003): "Evaluating the 'Genre Lens' for Analyzing Requirements for Content Assembly". In: *Proceedings of the 8th CAiSE/IFIP8.1 International Workshop on Evaluation of Modeling Methods in Systems Analysis and Design (EMMSAD '03)*, herausgegeben von Siau, Keng; Krogstie, John und Halpin, Terry. Velden, S. 95–105.

Honkaranta, Anne und Lyytikäinen, Virpi (2003): "Operationalizing a Genre-based Method for Content Analysis: A Case of a Church". In: *Business Information Systems – Proceedings of BIS 2003*, herausgegeben von Abramowicz, Witold und Klein, Gary. Colorado Springs, S. 108–116.

Hopkins, David (2000): "Web Documentation Project at the University of Delaware". In: *Proceedings of the 28th annual ACM SIGUCCS conference on User services: Building the future*. Richmond, S. 102–105.

Horn, Robert E. (1989): *Mapping Hypertext – Analysis, Linkage, and Display of Knowledge for the Next Generation of On-Line Text and Graphics*. Lexington: The Lexington Institute.

Hors, Arnaud Le; Hégaret, Philippe Le; Wood, Lauren; Nicol, Gavin; Robie, Jonathan; Champion, Mike und Byrne, Steve (2000): "Document Object Model (DOM) Level 2 Core Specification". Technische Spezifikation, W3C.

Hovy, Eduard H. (1990): "Parsimonious and Profligate Approaches to the Question of Discourse Structure Relations". In: *Proceedings of 5th International Workshop on Language Generation*. Pittsburgh, S. 128–136.

Hsu, Chun-Nan und Dung, Ming-Tzung (1998): "Generating Finite-State Transducers for Semi-Structured Data Extraction from the Web". *Information Systems* 23 (8): S. 521–538.

Hu, Jianying und Bagga, Amit (2003): "Functionality-Based Web Image Categorization". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Hu, Wen-Chen; Chen, Yining; Schmalz, Mark S. und Ritter, Gerhard X. (2001): "An Overview of World Wide Web Search Technologies". In: *Proceedings of the 5th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2001)*. Orlando.

Huber, Oliver (2002): *Hyper-Text-Linguistik – TAH: ein textlinguistisches Analysemodell für Hypertexte. Theoretisch und praktisch exemplifiziert am Problemfeld der typisierten Links von Hypertexten im World Wide Web.* Dissertation, Fakultät für Sprach- und Literaturwissenschaften, Ludwig-Maximilians-Universität München, München.

Huck, Gerald; Fankhauser, Peter; Aberer, Karl und Neuhold, Erich (1998): "Jedi: Extracting and Synthesizing Information from the Web". In: *Proceedings of the 3rd International Conference on Cooperative Information Systems a (CoopIS '98)*. Online verfügbar: `ftp://ftp.darmstadt.gmd.de/pub/oasys/reports/P-98-11.pdf`.

Huizingh, Eelko K. R. E. (2000): "The content and design of web sites: an empirical study". *Information & Management* 37 (3): S. 123–134.

Hurst, Matthew (2002): "Classifying TABLE Elements in HTML". In: *Proceedings of the 11th International World Wide Web Conference (WWW 2002)*. Honolulu.

Ihlström, Carina und Åkesson, Maria (2004): "Genre Characteristics – A Front Page Analysis of 85 Swedish Online Newspapers". In: *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS-37)*. Big Island, Hawaii.

Ihlström, Carina und Lundberg, Jonas (2003): "The Online News Genre through the User Perspective". In: *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS-36)*. Big Island, Hawaii.

Iorio, Angelo Di und Vitali, Fabio (2005): "Web Authoring: A Closed Case?" In: *Proceedings of the 38th Hawaii International Conference on Systems Sciences (HICSS-38)*. Big Island, Hawaii.

Ipsen, Guido (1999): "Dynamische Verweise in Hypertexten – Zur Verwendbarkeit von Metaphern bei der Hypertextproduktion". In: Jakobs et al. (1999), S. 11–27.

Ipsen, Guido (2001): "Pragmatik des Hypertextes – Linguistische Aspekte WWW-gebundener Informationsmedien als designtechnisches Instrument". In: Handler (2001), S. 67–80.

Isenberg, Horst (1968): "Überlegungen zur Texttheorie". In: *ASG-Bericht*, Berlin: Deutsche Akademie der Wissenschaften zu Berlin, Arbeitsstelle Strukturelle Grammatik, Nummer 2.

Isenberg, Horst (1978): "Probleme der Texttypologie – Variation und Determination von Texttypen". *Wissenschaftliche Zeitschrift der Karl-Marx-Universität Leipzig* 27: S. 565–579.

ISO 10179 (1996): "Information Processing – Processing Languages – Document Style Semantics and Specification Language (DSSSL)". Internationaler Standard, International Organization for Standardization, Genf. Online verfügbar: `http://www.ornl.gov/sgml/wg8/`.

ISO 10744 (1997): "Information Processing – Hypermedia/Time-Based Structuring Language (HyTime) – Second Edition". Internationaler Standard, International Organization for Standardization, Genf. Online verfügbar: `http://www.ornl.gov/sgml/wg8/`.

ISO 15445 (2000): "ISO/IEC 15445: Information Technology – Document Description and Processing Languages – HyperText Markup Language (HTML)". Internationaler Standard, International Organization for Standardization, Genf. Online verfügbar: `http://purl.org/NET/ISO+IEC.15445/15445.html`.

ISO 639 (1998): "ISO/IEC 639: Codes for the representation of names of languages. Part 1 (1988), Part 2 (1998)". Internationaler Standard, International Organization for Standardization, Genf.

ISO 8879 (1986): "Information Processing – Text and Office Information Systems – Standard Generalized Markup Language". Internationaler Standard, International Organization for Standardization, Genf.

ISO 9070 (1991): "ISO/IEC 9070: Information Technology – SGML Support Facilities – Registration Procedures for Public Text Owner Identifiers". Internationaler Standard, International Organization for Standardization, Genf.

ISO/IEC 13250 (2000): "Information Technology – Document Description and Processing Languages – Topic Maps". Internationaler Standard, International Organization for Standardization, Genf. Online verfügbar: `http://www.ornl.gov/sgml/wg4/`.

Issing, Ludwig J. und Klimsa, Paul (Hrsg.) (1997): *Information und Lernen mit Multimedia*. Weinheim: Beltz Psychologie Verlags Union, 2. Auflage.

Ivory, Melody Y. und Hearst, Marti A. (2002): "Statistical Profiles of Highly-Rated Web Sites". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing our World, Changing Ourselves*. Minneapolis, S. 367–374.

Jackson, Peter und Moulinier, Isabelle (2002): *Natural Language Processing for Online Applications – Text Retrieval, Extraction and Categorization*, Band 5 von *Natural Language Processing*. Amsterdam, Philadelphia: John Benjamins.

Jakobs, Eva-Maria (2003): "Hypertextsorten". *Zeitschrift für germanistische Linguistik* 31 (2): S. 232–252.

Jakobs, Eva-Maria und Knorr, Dagmar (Hrsg.) (1997): *Textproduktion in elektronischen Umgebungen*, Band 2 von *Textproduktion und Medium*. Frankfurt/Main, Berlin, Bern etc.: Peter Lang.

Jakobs, Eva-Maria; Knorr, Dagmar und Pogner, Karl-Heinz (Hrsg.) (1999): *Textproduktion: HyperText, Text, KonText*, Band 5 von *Textproduktion und Medium*. Frankfurt/Main, Berlin, Bern etc.: Peter Lang.

Jansen, Bernard J. und Pooch, Udo (2001): "A Review of Web Searching Studies and a Framework for Future Research". *Journal of the American Society for Information Science and Technology* 52 (3): S. 235–246.

Jansen, Bernard J.; Spink, Amanda und Saracevic, Tefko (2000): "Real life, real users, and real needs: A study and analysis of user queries on the web". *Information Processing and Management* 36 (2): S. 207–227.

Janson, Bernd (1996): "Zentrale Wissenschaftliche Einrichtungen". In: Flämig et al. (1996), S. 883–896.

Jensen, Carlos und Potts, Colin (2004): "Privacy Policies as Decision-Making Tools: An Evaluation of Online Privacy Notices". In: *Proceedings of the 2004 Conference on Human Factors in Computing Systems*. Wien, S. 471–478.

Johnson-Laird, P. N. (1983): *Mental Models – Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge: Cambridge University Press.

Jonassen, David H. (1989): *Hypertext/Hypermedia*. Englewood Cliffs: Educational Technology Publications.

Jones, Rosie und Ghani, Rayid (2000): "Automatically Building a Corpus for a Minority Language from the Web". In: *Proceedings of the Student Workshop at the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*. Hong Kong. Online verfügbar: `http://www.cs.cmu.edu/~webkb/`.

Jones, Russ und Nye, Adrian (1995): *HTML und das World Wide Web – Selbst publizieren im WWW*. A Nutshell Handbook. Bonn: O'Reilly/International Thomson. Deutsche Übersetzung, Aktualisierung und Erweiterung von Thomas Merz.

Jucker, Andreas (2004): "Gutenberg und das Internet – Der Einfluss von Informationsmedien auf Sprache und Sprachwissenschaft". Networx, Nr. 40. Online verfügbar: `http://www.mediensprache.net/de/networx/`.

Jucker, Andreas H. (2000): "Multimedia und Hypertext. Neue Formen der Kommunikation oder alter Wein in neuen Schläuchen?" In: *Kommunikationsformen im Wandel der Zeit*, herausgegeben von Fritz, Gerd und Jucker, Andreas H., Tübingen: Niemeyer, S. 7–28.

Kahn, Paul und Nyce, James M. (1991): "The Idea of a Machine: The Later Memex Essays". In: Nyce und Kahn (1991a), S. 113–144.

Kallmeyer, Werner (2000a): "Sprache und neue Medien – zum Diskussionsstand und zu einigen Schlussfolgerungen". In: Kallmeyer (2000b), S. 292–315.

Kallmeyer, Werner (Hrsg.) (2000b): *Sprache und neue Medien. Jahrbuch des Instituts für deutsche Sprache 1999*. Berlin, New York: de Gruyter.

Kamenz, Uwe; Hülsmann, Petra und Heiland, Thomas (1998): *Internet-Studie Hochschulen 1998*. Praxis-Studien zum Internet. Dortmund: ProfNet.

Kan, Min-Yen (2004): "Web Page Categorization without the Web Page". In: *Proceedings of the 13th Conference on World Wide Web (WWW-2004)*. New York, S. 262–263. Poster Track.

Kando, Noriko (1999): "Text Structure Analysis as a Tool to Make Retrieved Documents Usable". In: *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages*. Taipei, S. 126–135.

Karjalainen, Anne; Päivärinta, Tero; Tyrväinen, Pasi und Rajala, Jari (2000): "Genre-Based Metadata for Enterprise Document Management". In: *Proceedings of the 33rd Hawaii International Conference on Systems Sciences (HICSS-33)*.

Karjalainen, Anne und Salminen, Airi (2000): "Bridging the Gap between Hard and Soft Information Genres". In: *Challenges of Information Technology Management in the 21st Century*, herausgegeben von Khosrowpour, Mehdi. Hershey: Idea Group, S. 92–95.

Karlgren, Jussi; Bretan, Ivan; Dewe, Johan; Hallberg, Anders und Wolkert, Niklas (1998): "Iterative Information Retrieval Using Fast Clustering and Usage-Specific Genres". In: *Proceedings of the 8th DELOS Workshop on User Interfaces in Digital Libraries*. Stockholm, S. 85–92.

Karlgren, Jussi und Cutting, Douglass (1994): "Recognizing Text Genres with Simple Metrics Using Discriminant Analysis". In: *COLING 94 – The 15th International Conference on Computational Linguistics*. Association for Computational Linguistics, Kyoto, Band 2, S. 1071–1075.

Kehoe, Andrew und Renouf, Antoinette (2002): "WebCorp: Applying the Web to Linguistics and Linguistics to the Web". In: *Proceedings of the 11th International World Wide Web Conference (WWW 2002)*. Honolulu.

Keitel, Evelyn; Boehnke, Klaus und Wenz, Karin (Hrsg.) (2003): *Neue Medien im Alltag: Nutzung, Vernetzung, Interaktion*, Band 3 von *DFG-Forschergruppe »Neue Medien im Alltag«*. Lengerich, Berlin, Bremen etc.: Pabst.

Kelly, Brian; Johnston, Pete und Powell, Andy (2003): "Approaches To Validation of Dublin Core Metadata Embedded In (X)HTML Documents". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Kendall, Robert (2000): "Hypertext: Foe to Print?" *ACM SIGWEB* 9 (2-3): S. 5–7.

Kengeri, Rekha; Seals, Cheryl D.; Harley, Hope D.; Reddy, Himabindu P. und Fox, Edward A. (1999): "Usability study of digital libraries: ACM, IEEE-CS, NCSTRL, NDLTD". *International Journal of Digital Libraries* 2 (2–3): S. 157–169.

Kennedy, Alistair und Shepherd, Michael (2005): "Automatic Identification of Home Pages on the Web". In: *Proceedings of the 38th Hawaii International Conference on Systems Sciences (HICSS-38)*. Big Island, Hawaii.

Kessler, Brett; Nunberg, Geoffrey und Schütze, Hinrich (1997): "Automatic Detection of Text Genre". In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics*. San Francisco, S. 32–38.

Kilgarriff, Adam (2001): "Web as Corpus". In: *Proceedings of the Corpus Linguistics 2001 Conference*, herausgegeben von Rayson, Paul; Wilson, Andrew; McEnery, Tony; Hardie, Andrew und Khoja, Shereen. Lancaster, S. 342–344.

Killoran, John B. (2002): "Homepage, Homebound; Web Log, We Blog: Web Genres for Personal Civic (Dis-)Engagement". Online verfügbar: `http://communication.cudenver.edu/~jkillora/research/2002rsa.html`. Beitrag für die Konferenz der Rhetoric Society of America, Las Vegas (25. Mai 2002).

Killoran, John B. (2003a): "From Print Objects to Web Subjects: The Reorientation of the Resume". Online verfügbar: `http://communication.cudenver.edu/~jkillora/research/2003nca.html`. Beitrag für die National Communication Association Conference, Miami Beach (22. November 2003).

Killoran, John B. (2003b): "From Print Resume to Web Resume: The Destabilizing of a Genre". Online verfügbar: `http://communication.cudenver.edu/~jkillora/research/2003cattw.html`. Beitrag für die Konferenz der Canadian Association of Teachers of Technical Writing, Halifax (29. Mai 2003).

Killoran, John B. (2004): "Genre, Agency, and Technological Determinism: A Survey of Web Resume Authors". Online verfügbar: `http://communication.cudenver.edu/~jkillora/research/2004rmca.html`. Beitrag für die Rocky Mountain Communication Association convention, Denver (6. März, 2004).

Kim, Sun und Zhang, Byoung-Tak (2003): "Genetic Mining of HTML Structures for Effective Web-Document Retrieval". *Applied Intelligence* 18 (3): S. 243–256.

Kintsch, Walter und van Dijk, Teun A. (1978): "Towards a model of text comprehension and production". *Psychological Review* 85 (5): S. 363–394.

Kistler, Thomas und Marais, Hannes (1998): "WebL – A Programming Language for the Web". *Computer Networks and ISDN Systems* 30: S. 259–270.

Klabunde, Ralf; Carstensen, Kai-Uwe; Ebert, Christian; Endriss, Cornelia; Jekat, Susanne; Langer, Hagen und Schiehlen, Michael (Hrsg.) (2001): *Computerlinguistik und Sprachtechnologie – Eine Einführung*. Heidelberg: Spektrum.

Klein, Josef (2000): "Intertextualität, Geltungsmodus, Texthandlungsmuster. Drei vernachlässigte Kategorien der Textsortenforschung – exemplifiziert an politischen und medialen Textsorten". In: Adamzik (2000a), S. 31–44.

Klein, Josef und Fix, Ulla (Hrsg.) (1997): *Textbeziehungen: linguistische und literaturwissenschaftliche Beiträge zur Intertextualität*. Tübingen: Stauffenburg.

Klein, Michel; Broekstra, Jeen; Fensel, Dieter; van Harmelen, Frank und Horrocks, Ian (2003): "Ontologies and Schema Languages on the Web". In: Fensel et al. (2003b), S. 95–139.

Kleinberg, Jon (1998): "Authoritative Sources in a Hyperlinked Environment". In: *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*. San Francisco, S. 668–677.

Kleinz, Torsten (2004): "Qualitätsoffensive – Freie Online-Enzyklopädie Wikipedia stellt die Weichen für die Zukunft". *c't, Magazin für Computertechnik* (14): S. 38–39.

Klyne, Graham und Carroll, Jeremy J. (2004): "Resource Description Framework (RDF): Concepts and Abstract Syntax". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/`.

Kneece, Nancy S. (1996): "An Electronic Publishing Spectrum: A Framework for Text Modules". In: *Proceedings of the 14th annual International Conference on Systems Documentation: Marshaling new Technological Forces: Building a Corporate, Academic, and User-Oriented Triangle*. Research Triangle Park, S. 195–203.

Knublauch, Holger; Fergerson, Ray W.; Noy, Natalya F. und Musen, Mark A. (2004a): "The protégé OWL Plugin: An Open Development Environment for Semantic Web Applications". In: *Proceedings of the Third International Semantic Web Conference – ISWC 2004*. Hiroshima. Online verfügbar: `http://protege.stanford.edu/plugins/owl/documentation.html`.

Knublauch, Holger; Musen, Mark A. und Rector, Alan L. (2004b): "Editing Description Logic Ontologies with the protégé OWL Plugin". In: *Proceedings of the International Workshop on Description Logics – DL 2004*. Whistler. Online verfügbar: `http://protege.stanford.edu/plugins/owl/documentation.html`.

Kober, Katharina; Krumeich, Alexander; von der Landwehr, Klaus; Langer, Hagen und Rehm, Georg (1998): "Projektbericht pronto: Probleme der Eigennamenerkennung". Institut für Semantische Informationsverarbeitung, Universität Osnabrück. Unveröffentlichtes Manuskript.

Koch, Peter und Oesterreicher, Wulf (1994): "Schriftlichkeit und Sprache". In: *Schrift und Schriftlichkeit*, herausgegeben von Günther, Hartmut und Ludwig, Otto, Berlin, New York: de Gruyter, Band 10.1 von *Handbücher zur Sprach- und Kommunikationswissenschaft (HSK)*, S. 587–604.

Koehler, Wallace (2002): "Web Page Change and Persistence – A Four-Year Longitudinal Study". *Journal of the American Society for Information Science and Technology* 53 (2): S. 162–171.

Kopak, Richard W. (1999): "Functional Link Typing in Hypertext". *ACM Computing Surveys* 31 (4): S. 16–22.

Kosala, Raymond und Blockeel, Hendrik (2000): "Web Mining Research: A Survey". *Newsletter of the Special Interest Group on Knowledge Discovery & Data Mining* 2 (1): S. 1–15.

Kowtha, N. Rao und Choon, Timothy Whai Ip (2001): "Determinants of website development: a study of electronic commerce in Singapore". *Information & Management* 39 (3): S. 227–242.

Kraus, Michael (2000): "Websites von Universitäten". Projektbericht, Institut für Informatik, Universität München. Online verfügbar: `http://www.pms.informatik.uni-muenchen.de/~krausm/`.

Krause, Robin D. (2003): "Managing Higher Ed Web Sites: Balancing the need for timely updates, the requirements of institutional marketing, and the development of content". In: *Proceedings of the 31st annual ACM SIGUCCS conference on User services*. San Antonio, S. 139–141.

Krol, Ed (1994): *The Whole Internet – User's Guide & Catalog*. Sebastopol: O'Reilly & Associates, 2. Auflage.

Kruschwitz, Udo (2001): "Exploiting Structure for Intelligent Web Search". In: *Proceedings of the 34th Hawaii International Conference on Systems Sciences (HICSS-34)*.

Kuhlen, Rainer (1991): *Hypertext – Ein nicht-lineares Medium zwischen Buch und Wissensbank*. Berlin, Heidelberg, New York etc.: Springer.

Kuhlen, Rainer (1997): "Hypertext". In: *Grundlagen der praktischen Information und Dokumentation. Ein Handbuch zur Einführung in die fachliche Informationsarbeit*, herausgegeben von Buder, Marianne; Rehfeld, Werner; Seeger, Thomas und Strauch, Dietmar, München etc.: Saur, S. 355–369. 4. Auflage.

Kuhlen, Rainer (2004): "Kollaboratives Schreiben". In: Bieber und Leggewie (2004), S. 216–239.

Kumar, Ravi; Novak, Jasmine; Raghavan, Prabhakar und Tomkins, Andrew (2004): "Structure and Evolution of Blogspace". *Communications of the ACM* 47 (12): S. 35–39.

Kurzidim, Michael (2004): "Wissenswettstreit – Die kostenlose Wikipedia tritt gegen die Marktführer Encarta und Brockhaus an". *c't, Magazin für Computertechnik* (21): S. 132–139.

Kushmerick, Nicholas; Weld, Daniel S. und Doorenbos, Robert B. (1997): "Wrapper Induction for Information Extraction". In: *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI '97)*. Nagoya, S. 729–737.

Kwasnik, Barbara H.; Crowston, Kevin; Nilan, Michael und Roussinov, Dmitri (2001): "Identifying Document Genre to Improve Web Search Effectiveness". *Bulletin of The American Society for Information Science and Technology* 27 (2). Online verfügbar: `http://www.asis.org/Bulletin/`.

Kwon, Oh-Woog und Lee, Jong-Hyeok (2003): "Text categorization based on k-nearest neighbor approach for Web site classification". *Information Processing & Management* 39 (1): S. 25–44.

Labrou, Yannis und Finin, Tim (1999): "Yahoo! as an Ontology – Using Yahoo! Categories to Describe Documents". In: *Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM '99)*. S. 180–187.

Landow, George P. (1992): *Hypertext: The Convergence of Technology and Contemporary Critical Theory*. Baltimore, London: John Hopkins University Press.

Landow, George P. (1997): *Hypertext 2.0: The Convergence of Technology and Contemporary Critical Theory*. Baltimore, London: John Hopkins University Press.

Lang, Ewald (1973): "Über einige Schwierigkeiten beim Postulieren einer Textgrammatik". In: *Generative Grammar in Europe*, herausgegeben von Kiefer, F. und Ruwet, N., Dordrecht: Reidel, S. 284–314.

Langer, Stefan (2001): "Sprachen auf dem WWW". In: *Sprach- und Texttechnologie in digitalen Medien – Proceedings der Frühjahrstagung der Gesellschaft für Linguistische Datenverarbeitung*, herausgegeben von Lobin, Henning. Gesellschaft für linguistische Datenverarbeitung, Justus-Liebig-Universität Gießen, S. 85–91.

Langer, Stefan (2002): "Grenzen der Sprachenidentifizierung". In: *KONVENS 2002 – 6. Konferenz zur Verarbeitung natürlicher Sprache*, herausgegeben von Busemann, Stefan. Saarbrücken: Deutsches Forschungszentrum für Künstliche Intelligenz, S. 99–106. DFKI Document D-02-01.

Lawrence, Steve (2001): "Online or Invisible?" *Nature* 411 (6837): S. 521.

Lawrence, Steve; Giles, C. Lee und Bollacker, Kurt (1999): "Digital Libraries and Autonomous Citation Indexing". *IEEE Computer* 32 (6): S. 67–71.

Lee, Yong-Bae und Myaeng, Sung Hyon (2002): "Text Genre Classification with Genre-Revealing and Subject-Revealing Features". In: *Proceeding of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR '02)*. Tampere, S. 145–150.

Lee, Yong-Bae und Myaeng, Sung Hyon (2004): "Automatic Identification of Text Genres and Their Roles in Subject-Based Categorization". In: *Proceedings of the 37th Hawaii International Conference on Systems Sciences (HICSS-37)*.

LeFurgy, William G. (2001): "Records and Archival Management of World Wide Web Sites". *Government Record News – Newsletter of the Government Records Section of American Archivists* (2). Online verfügbar: `http://www.archivists.org/saagroups/gov/newsletters.asp`.

Lenat, Douglas B. (1995): "CYC: A Large-Scale Investment in Knowledge Infrastructure". *Communications of the ACM* 38 (11): S. 32–38.

Lenke, Nils und Schmitz, Peter (1995): "Geschwätz im ›Globalen Dorf‹ – Kommunikation im Internet". *Osnabrücker Beiträge zur Sprachtheorie* (50): S. 117–141.

Leutner, Detlev (1997): "Adaptivität und Adaptierbarkeit multimedialer Lehr- und Informationssysteme". In: Issing und Klimsa (1997), S. 139–149.

Leuze, Dieter (1996): "Mitwirkungsrechte der Mitglieder". In: Flämig et al. (1996), S. 859–881.

Li, Fang; Sheng, Huanye und Weisweber, Wilhelm (2001): "World Wide Web – A Multilingual Language Resource". In: *Web Intelligence: Research and Development*, herausgegeben von Zhong, Ning; Yao, Yiyu; Liu, Jiming und Ohsuga, Setsuo. Berlin, Heidelberg, New York etc.: Springer, Band 2198 von *Lecture Notes in Computer Science*, S. 373–378.

Li, Wen-Syan; Kolak, Okan und Vu, Quoc (2000): "Defining Logical Domains in a Web Site". In: *Proceedings of Hypertext 2000*. San Antonio, S. 123–132.

Liang, Ting-Peng und Lai, Hung-Jen (2002): "Effect of store design on consumer purchases: an empirical study of on-line bookstores". *Information & Management* 39 (6): S. 431–444.

Lie, Håkon Wium und Bos, Bert (1997): *Cascading Style Sheets – Layouts für das Web-Publishing*. Bonn, Reading etc.: Addison-Wesley.

Lim, Chul Su; Lee, Kong Joo und Kim, Gil Chang (2005a): "Automatic Genre Detection of Web Documents". In: *Natural Language Processing – IJCNLP 2004*, herausgegeben von Su, Keh-Yih; Tsujii, Jun'ichi; Lee, Jong-Hyeok und Kwong, Oi Yee, Berlin, Heidelberg, New York etc.: Springer, Band 3248 von *Lecture Notes in Artificial Intelligence*, S. 310–319.

Lim, Chul Su; Lee, Kong Joo und Kim, Gil Chang (2005b): "Multiple Sets of Features for Automatic Genre Classification of Web Documents". *Information Processing and Management* 41 (5): S. 1263–1276.

Lim, Seung-Jin und Ng, Yiu-Kai (1999): "An Automated Approach for Retrieving Hierarchical Data from HTML Tables". In: *Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM '99)*. S. 466–474.

Linke, Angelika; Nussbaumer, Markus und Portmann, Paul R. (2001): *Studienbuch Linguistik*, Band 121 von *Reihe Germanistische Linguistik*. Tübingen: Niemeyer, 4. Auflage.

Liu, Bing; Chin, Chee Wee und Ng, Hwee Tou (2003): "Mining Topic-Specific Concepts and Definitions on the Web". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Liu, Chang; Arnett, Kirk P.; Capella, Louis M. und Beatty, Robert C. (1997): "Web sites of the Fortune 500 companies: Facing customers through home pages". *Information & Management* 31 (6): S. 335–345.

Liu, Ling; Pu, Calton und Han, Wei (2000): "XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources". In: *Proceedings of the International Conference on Data Engineering (ICDE)*. S. 611–621.

Lobin, Henning (1999a): "Intelligente Dokumente – Linguistische Repräsentation komplexer Inhalte für die hypermediale Wissensvermittlung". In: Lobin (1999b), S. 155–178.

Lobin, Henning (Hrsg.) (1999b): *Text im digitalen Medium – Linguistische Aspekte von Textdesign, Texttechnologie und Hypertext Engineering*. Wiesbaden: Westdeutscher Verlag.

Lobin, Henning (2000): "Service-Handbücher – Linguistische Aspekte im Document Lifecycle". In: Richter et al. (2000), S. 791–808.

Lobin, Henning (2001a): *Informationsmodellierung in XML und SGML*. Berlin, Heidelberg, New York etc.: Springer. Korrigierter Nachdruck der 1. Auflage.

Lobin, Henning (Hrsg.) (2001b): *Sprach- und Texttechnologie in digitalen Medien – Proceedings der Frühjahrstagung der Gesellschaft für Linguistische Datenverarbeitung*, Justus-Liebig-Universität Gießen.

Lobin, Henning (2004): "Textauszeichnung und Dokumentgrammatiken". In: Lobin und Lemnitzer (2004), S. 51–82.

Lobin, Henning und Lemnitzer, Lothar (Hrsg.) (2004): *Texttechnologie – Anwendungen und Perspektiven*. Stauffenburg Handbücher. Tübingen: Stauffenburg.

Lobin, Henning; Stührenberg, Maik und Rehm, Georg (2003): "eLearning und offene Standards: Zum Einsatz XML-strukturierter Lernobjekte". *Sprache und Datenverarbeitung* 27 (1/2): S. 75–94.

Lucas, Wendy und Topi, Heikki (2002): "Form and Function: The Impact of Query Term and Operator Usage on Web Search Results". *Journal of the American Society for Information Science and Technology* 53 (2): S. 95–108.

Lucas, Wendy und Topi, Heikki (2004): "Training for Web search: Will it get you in shape?" *Journal of the American Society for Information Science and Technology* 55 (13): S. 1183–1198.

Luger, George F. (2001): *Künstliche Intelligenz – Strategien zur Lösung komplexer Probleme*. München: Pearson Studium, Addison-Wesley, 4. Auflage.

Luger, George F. und Stubblefield, William A. (1993): *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. Redwood City: Benjamin/Cummings, 2. Auflage.

Lutz, Benedikt (1995): "Hypertextlinguistik: Erfahrungen aus der Praxis – Anregungen für die linguistische Forschung". *Osnabrücker Beiträge zur Sprachtheorie* 50: S. 155–163.

Machilek, Franz; Schütz, Astrid und Marcus, Bernd (2004): "Selbstdarsteller oder Menschen wie du und ich? Intentionen und Persönlichkeitsmerkmale von Homepagebesitzer/inne/n". *Zeitschrift für Medienpsychologie* 16 (3): S. 88–98.

Makhfi, Jamshid (2002): *Medienkultur – Eine qualitative und quantitative Analyse von Webpages*, Band 3 von *Beiträge zur Computersoziologie*. Münster, Hamburg, London: Lit.

Maler, Eve und Andaloussi, Jeanne El (1996): *Developing SGML DTDs – From Text to Model to Markup*. Upper Saddle River: Prentice Hall.

Malik, Ayesha (2003): "XML, Ontologies, and the Semantic Web". *XML Journal* 4 (2): S. 26–30.

Mancini, Clara und Shum, Simon Buckingham (2004): "Towards ›Cinematic‹ Hypertext". In: *Proceedings of the fifteenth ACM conference on Hypertext & Hypermedia*. Santa Cruz, S. 215–224.

Mandler, Jean M. und Johnson, Nancy S. (1980): "On Throwing out the Baby with the Bathwater: A Reply to Black and Wilensky's Evaluation of Story Grammars". *Cognitive Science* 4: S. 305–312.

Mangasser-Wahl, Martina (Hrsg.) (2000a): *Prototypentheorie in der Linguistik: Anwendungsbeispiele – Methodenreflexion – Perspektiven*. Tübingen: Stauffenburg.

Mangasser-Wahl, Martina (2000b): "Roschs Prototypentheorie – Eine Entwicklung in drei Phasen". In: Mangasser-Wahl (2000a), S. 15–31.

Mann, William C.; Matthiessen, Christian M. I. M. und Thompson, Sandra A. (1989): "Rhetorical Structure Theory and Text Analysis". Information Sciences Institute Research Report ISI/RR-89-242, University of Southern California.

Mann, William C. und Thompson, Sandra A. (1987): "Rhetorical Structure Theory: Description and Construction of Text Structures". In: *New Results in Artificial Intelligence, Psychology and Linguistics*, herausgegeben von Kempen, Gerhard, Dordrecht, Boston, Lancaster: Martinus Nijhoff Publishers, S. 85–96.

Mann, William C. und Thompson, Sandra A. (1988): "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization". *Text* 8: S. 243–281.

Manola, Frank und Miller, Eric (2004): "RDF Primer". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/2004/REC-rdf-primer-20040210/`.

Marcus, Aaron und Gould, Emilie West (2000): "Crosscurrents: Cultural Dimensions and Global Web User-Interface Design". *Interactions* 7 (4): S. 32–46.

Marlow, David M. (2004): "Investigating Technical Trouble Tickets: An analysis of a homely CMC genre". In: *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS-37)*. Big Island, Hawaii.

Matsuda, Katsushi und Fukushima, Toshikazu (1999): "Task-Oriented World Wide Web Retrieval by Document Type Classification". In: *Proceedings of the International Conference on Information and Knowledge Management (CIKM '99)*. Kansas City, S. 109–113.

Maxwell, Christine und Grycz, Czeslaw Jan (1994): *The New Rider's Official Internet Yellow Pages*. Indianapolis: New Riders.

May, Wolfgang und Lausen, Georg (2000): "Information Extraction from the Web". Technischer Bericht 136, Computer Science Institute, Freiburg University. Online verfügbar: `http://www.informatik.uni-freiburg.de/~may/Publics/`.

McGillis, Louise und Toms, Elaine G. (2001): "Usability of the Academic Library Web Site: Implications for Design". *College & Research Libraries* 62 (4): S. 355–367.

McGuinness, Deborah L.; Fikes, Richard; Stein, Lynn Andrea und Hendler, James (2003): "DAML-ONT: An Ontology Language for the Semantic Web". In: Fensel et al. (2003b), S. 65–93.

McGuinness, Deborah L. und van Harmelen, Frank (2004): "OWL Web Ontology Language – Overview". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/2004/REC-owl-features-20040210/`.

McKnight, Cliff; Dillon, Andrew und Richardson, John (1991): *Hypertext in Context*. The Cambridge Series on Electronic Publishing. Cambridge: Cambridge University Press.

McMillan, Sally J. (2001): "Survival of the Fittest Online: A Longitudinal Study of Health-Related Web Sites". *Journal of Computer-Mediated Communication* 6 (3). Online verfügbar: `http://www.ascusc.org/jcmc/`.

Mehler, Alexander (2001): *Textbedeutung – Zur prozeduralen Analyse und Repräsentation struktureller Ähnlichkeiten von Texten*, Band 5 von *Sprache, Sprechen und Computer*. Frankfurt/Main: Peter Lang.

Mehler, Alexander (2004): "Textmining". In: Lobin und Lemnitzer (2004), S. 329–352.

Mehler, Alexander; Dehmer, Matthias und Gleim, Rüdiger (2004): "Towards Logical Hypertext Structure — A Graph-Theoretic Perspective". In: *Proceedings of the Fourth International Workshop on Innovative Internet Computing Systems (I2CS '04)*, herausgegeben von Böhme, Thomas und Heyer, Gerhard. Berlin, New York: Springer, Lecture Notes in Computer Science.

Mehler, Alexander und Lobin, Henning (Hrsg.) (2004): *Automatische Textanalyse – Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte*. Wiesbaden: Verlag für Sozialwissenschaften.

Meyer zu Eissen, Sven und Stein, Benno (2004): "Genre Classification of Web Pages". In: *Proceedings of the 27th German Conference on Artificial Intelligence (KI-2004)*. Ulm.

Meyrowitz, Norman (1989): "Hypertext – Does It Reduce Cholestorol, Too?" In: Nyce und Kahn (1991a), S. 287–318.

Michalak, Susan und Coney, Mary (1993): "Hypertext and the Author/Reader Dialogue". In: *Proceedings of the fifth ACM conference on Hypertext*. Seattle, S. 174–182.

Middleton, Iain; McConnell, Mike und Davidson, Grant (1999): "Presenting a model for the structure and content of a university World Wide Web site". *Journal of Information Science* 25 (3): S. 219–227.

Mignet, Laurent; Barbosa, Denilson und Veltri, Pierangelo (2003): "The XML Web: a First Study". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Mikheev, Andrei; Grover, Claire und Moens, Marc (1999): "XML Tools and Architecture for Named Entity Recognition". *Markup Languages* 1 (3): S. 89–113.

Miles-Board, Timothy; Kampa, Simon; Carr, Leslie und Hall, Wendy (2001): "Hypertext in the Semantic Web". In: *Proceedings of the twelfth ACM Conference on Hypertext and Hypermedia*. Århus, S. 237–238.

Miles-Board, Timothy J.; Bailey, Christopher P.; Hall, Wendy und Carr, Leslie A. (2004): "Building a Companion Website in the Semantic Web". In: *Proceedings of the 13th Conference on World Wide Web (WWW-2004)*. New York, S. 365–373. Refereed Papers Track.

Miller, Carolyn R. (1984): "Genre as Social Action". *Quarterly Journal of Speech* (70): S. 151–167. Ebenfalls in und zitiert nach: Freedman, Aviva und Medway, Peter (1994): *Genre and the New Rhetoric*, London: Taylor and Francis, S. 23–42.

Miller, George A. (1956): "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information". *Psychological Review* 63 (2): S. 81–97.

Miller, Hugh (1995): "The Presentation of Self in Electronic Life: Goffman on the Internet". Online verfügbar: `http://ess.ntu.ac.uk/miller/cyberpsych/goffman.htm`. Beitrag für die Embodied Knowledge and Virtual Space Conference, Goldsmiths' College, University of London, Juni.

Miller, Hugh (1999): "The Hypertext Home: images and metaphors of home on World Wide Web home pages". Online verfügbar: `http://ess.ntu.ac.uk/miller/cyberpsych/homeweb.htm`. Beitrag für die Konferenz Design History Society Home and Away, Nottingham Trent University, 10.–12. September 1999.

Missikoff, Michele; Velardi, Paola und Fabriani, Paolo (2003): "Text Mining Techniques to Automatically Enrich a Domain Ontology". *Applied Intelligence* 18 (3): S. 323–340.

Mitchell, Tom M. (1997): *Machine Learning*. New York, St. Louis, San Francisco etc.: McGraw-Hill.

Mitkow, Ruslan (Hrsg.) (2003): *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press.

Mitra, Nilo; Gudgin, Martin; Hadley, Marc; Mendelsohn, Noah; Moreau, Jean-Jacques und Nielsen, Henrik Frystyk (2002): "SOAP Version 1.2". Technische Spezifikation, W3C. Besteht aus Part 0 (Primer), Part 1 (Messaging Framework), Part 2 (Adjuncts). Online verfügbar: `http://www.w3.org/2002/ws/`.

Mizoguchi, Riichiro (2004): "Ontology Engineering Environments". In: Staab und Studer (2004), S. 275–295.

Modha, Dharmendra S. und Spangler, W. Scott (2000): "Clustering Hypertext with Applications to Web Searching". In: *Proceedings of the 11th ACM Conference on Hypertext and Hypermedia*. San Antonio, S. 143–152.

Möller, Gerhard (1997): "Relationen in der universellen Dezimalklassifikation". Technischer Bericht, Oldenburger Forschungs- und Entwicklungsinstitut für Informatik-Werkzeuge und -Systeme., Oldenburg. Internes Projektmemo. Online verfügbar: `http://www.gerhard.de/info/dokumente/dokumentation/gerhard/relationen_udkz.pdf`.

Möller, Ralf und Haarslev, Volker (2003): "Description Logic Systems". In: *The Description Logic Handbook – Theory, Implementation and Applications*, herausgegeben von Baader, Franz; Calvanese, Diego; McGuinness, Deborah L.; ; Nardi, Daniele und Patel-Schneider, Peter F., Cambridge: Cambridge University Press, S. 282–305.

Moore, Johanna D. und Pollack, Martha E. (1992): "A Problem for RST: The Need for Multi-Level Discourse Analysis". *Computational Linguistics* 18 (4): S. 537–544.

Morkes, John und Nielsen, Jakob (1998): "Applying Writing Guidelines to Web Pages". In: *Conference on Human Factors in Computing Systems – CHI 98*. ACM, New York, S. 321–322.

Morley, Barry; Renouf, Antoinette und Kehoe, Andrew (2003): "Linguistic Research with XML/RDF-aware WebCorp tool". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Motsch, Wolfgang und Viehweger, Dieter (1981): "Sprachhandlung, Satz und Text". In: *Sprache und Pragmatik*, herausgegeben von Rosengren, Inger, Malmö: Gleerup, S. 125–154.

Murayama, Norifumi; Saito, Suguru und Okumura, Manabu (2004): "Are Web Pages Characterized by Color?" In: *Proceedings of the 13th Conference on World Wide Web (WWW-2004)*. New York, S. 248–249. Poster Track.

Muthusamy, Yeshwant K. und Spitz, Lawrence (1998): "Automatic Language Identification". In: *Survey of the State of the Art in Human Language Technology*, herausgegeben von Cole, Ronald; Mariani, Joseph; Uszkoreit, Hans; Varile, Giovanni Battista; Zaenen, Annie und Zampolli, Antonio, Cambridge: Cambridge University Press, S. 314–317.

Myllymaki, Jussi (2001): "Effective Web Data Extraction with Standard XML Technologies". In: *Proceedings of the 10th International World Wide Web Conference (WWW-10)*. Hong Kong, S. 689–696.

Nagao, Katashi und Hasida, Kôiti (1998): "Automatic Text Summarization Based on the Global Document Annotation". In: *COLING 98 – The 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. Association for Computational Linguistics, Montreal, Quebec, Kanada, Band 2, S. 917–921. 2 Bände.

Nardi, Bonnie A.; Schiano, Diane J.; Gumbrecht, Michelle und Swartz, Luke (2004): "Why We Blog". *Communications of the ACM* 47 (12): S. 41–46.

Naumann, Anja; Waniek, Jacqueline; Brunstein, Angela und Krems, Josef (2003): "Wissenserwerb aus WWW-basierten Informationsbeständen". In: Keitel et al. (2003), S. 71–96.

Nelson, Theodor Holm (1965): "A File Structure for The Complex, The Changing and the Indeterminate". In: *Association for Computing Machinery: Proceedings of the 20th national conference*. Cleveland, S. 84–100.

Nelson, Theodor Holm (1972): "As We Will Think". In: Nyce und Kahn (1991a), S. 245–260. Ursprünglich erschienen in *Online 72 Conference Proceedings*, Band 1. Uxbridge, S. 439–454.

Nelson, Theodor Holm (1987): "Literary Machines". Eigenverlag. Edition 87.1.

Nelson, Theodor Holm (1997): "Embedded Markup Considered Harmful". *The World Wide Web Journal* 2 (4): S. 129–134.

Neuberger, Christoph; Tonnemacher, Jan; Biebl, Matthias und Duck, André (1998): "Online – The Future of Newspapers? Germany's Dailies on the World Wide Web". *Journal of Computer-Mediated Communication* 4 (1). Online verfügbar: `http://www.ascusc.org/jcmc/`.

Neumann, Günter (2001): "Informationsextraktion". In: Klabunde et al. (2001), S. 448–455.

Nielsen, Jakob (1995a): "A Home-Page Overhaul Using Other Web Sites". *IEEE Software* 12 (3): S. 75–78.

Nielsen, Jakob (1995b): *Multimedia and Hypertext – The Internet and Beyond*. Boston, San Diego, New York etc.: AP Professional.

Nielsen, Jakob (1996): "The Importance of Being Beautiful". *IEEE Software* 13 (1): S. 92–94.

Nielsen, Jakob (1999): *Designing Web Usability*. Indianapolis: New Riders.

Nielsen, Jakob und Tahir, Marie (2002): *Homepage Usability – 50 Websites Deconstructed*. Indianapolis: New Riders.

Nordbotten, Joan C. und Nordbotten, Svein (1999): "Search Patterns in Hypertext Exhibits". In: *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32)*.

Noy, Natalya F.; Sintek, Michael; Decker, Stefan; Crubézy, Monica; Fergerson, Ray W. und Musen, Mark A. (2001): "Creating semantic web contents with protégé-2000". *IEEE Intelligent Systems* 16 (2): S. 60–71.

Nunberg, Geoffrey (1997): "Genres in Digital Documents – Introduction". In: *Proceedings of the 30th Hawaii International Conference on Systems Sciences (HICSS-30)*. Band 6, S. 2.

Nürnberg, Peter J. und Ashman, Helen (1999): "What Was the Question? Reconciling Open Hypermedia and World Wide Web Research". In: *Proceedings of the tenth ACM Conference on Hypertext and Hypermedia: Returning to our Diverse Roots*. Darmstadt, S. 83–90.

Nürnberg, Peter J.; Leggett, John J. und Schneider, Erich R. (1997): "As We Should Have Thought". In: *Proceedings of the Eighth ACM Conference on Hypertext (Hypertext 97)*, herausgegeben von Bernstein, Mark; Carr, Leslie und Østerbye, Kasper. Southampton, S. 96–101.

Nusselein, Mark (2003): *Inhaltliche Gestaltung eines Data Warehouse-Systems am Beispiel einer Hochschule*, Band 68 von *Monographien: Neue Folge*. München: Bayerisches Staatsinstitut für Hochschulforschung und Hochschulplanung.

Nyce, James M. und Kahn, Paul (Hrsg.) (1991a): *From Memex to Hypertext – Vannevar Bush and the Mind's Machine*. San Diego: Academic Press.

Nyce, James M. und Kahn, Paul (1991b): "A Machine for the Mind: Vannevar Bush's Memex". In: Nyce und Kahn (1991a), S. 39–66.

O'Neill, Edward T. (1997): "Characteristics of Web Accessible Information". In: *63rd IFLA General Conference*. International Federation of Library Associations and Institutions, Kopenhagen. Online verfügbar: `http://www.ifla.org/IV/ifla63/63onee.htm`.

O'Neill, Edward T.; Lavoie, Brian F. und McClain, Patrick D. (1998): "Web Characterization Project – An Analysis of Metadata Usage on the Web". In: *Annual Review of OCLC Research 1998*, Dublin: OCLC Online Computer Library Center. Online verfügbar: `http://www.oclc.org`.

O'Neill, Edward T.; McClain, Patrick D. und Lavoie, Brian F. (1997): "A Methodology for Sampling the World Wide Web". In: *Annual Review of OCLC Research 1997*, Dublin: OCLC Online Computer Library Center. Online verfügbar: `http://www.oclc.org`.

Orlikowski, Wanda und Yates, JoAnne (1998): "Genre Systems: Structuring Interaction through Communicative Norms". Online verfügbar: `http://ccs.mit.edu/papers/CCSWP205/`. Massachusetts Institute of Technology, Sloan School of Management, Center for Coordination Science: CCS WP #205 SWP #4030.

Orlikowski, Wanda J. und Yates, JoAnne (1994): "Genre Repertoire: The Structuring of Communicative Practices in Organizations". *Administrative Science Quarterly* (39): S. 541–574.

Ovsiannikov, Ilia A.; Arbib, Michael A. und McNeill, Thomas H. (1999): "Annotation Technology". *International Journal of Human-Computer Studies* 50 (4): S. 329–362.

Palmer, David (1994): "SATZ – An Adaptive Sentence Segmentation System". Technischer Bericht CSD-94-846, Computer Science Devision, University of California, Berkeley. Online verfügbar: `http://sunsite.berkeley.edu/TR/UCB:CSD-94-846`.

Palmer, David (2000): "Tokenisation and Sentence Segmentation". In: Dale et al. (2000), S. 11–35.

Palmer, Jonathan W. und Griffith, David A. (1998): "An Emerging Model of Web Site Design for Marketing". *Communications of the ACM* 41 (3): S. 44–51.

Pang, Alex Soojung-Kim (1998): "Hypertext, the Next Generation: A Review and Research Agenda". *First Monday* 3 (11). Online verfügbar: `http://firstmonday.org/issues/issue3_11/pang/index.html`.

Panko, Raymond R. und Panko, David K. (1998): "Where Do You Want to Fly Today? A User Interface Travel Genre Based on Flight Simulators". In: *Proceedings of the 31st Hawaii International Conference on Systems Sciences (HICSS-31)*. Big Island, Hawaii, Band 2, S. 110–118.

Pansegrau, Petra (1997): "Dialogizität und Degrammatikalisierung in E-mails". In: Weingarten (1997a), S. 86–104.

Paradis, François (2000): "Information Extraction and Gathering for Search Engines: The Taylor Approach". RIAO (Recherche d'Informations Assistée par Ordinateur), Paris, France.

Park, Han Woo und Thelwall, Mike (2003): "Hyperlink Analyses of the World Wide Web: A Review". *Journal of Computer-Mediated Communication* 8 (4). Online verfügbar: `http://www.ascusc.org/jcmc/`.

Patel-Schneider, Peter F.; Hayes, Patrick und Horrocks, Ian (2004): "OWL Web Ontology Language – Semantics and Abstract Syntax". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/2004/REC-owl-semantics-20040210/`.

Pemberton, Lyn und Shurville, Simon (Hrsg.) (2000): *Words on the Web – Computer Mediated Communication*. Exeter, Portland: Intellect Books.

Pemberton, Steven (2002): "XHTML 1.0: The Extensible Hypertext Markup Language (Second Edition)". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/xhtml1/`.

Penn, Gerald; Hu, Jianying; Luo, Hengbin und McDonald, Ryan (2001): "Flexible Web Document Analysis for Delivery to Narrow-Bandwidth Devices". In: *International Conference on Document Analysis and Recognition (ICDAR '01)*. Seattle, S. 1074–1078.

Pepper, Steve und Moore, Graham (2001): "XML Topic Maps (XTM) 1.0". Technische Spezifikation, TopicMaps.Org. Online verfügbar: `http://www.topicmaps.org/xtm/1.0/`.

Perfetti, Charles A. (1996): "Text and Hypertext". In: Rouet et al. (1996), S. 157–161.

Perry, Timothy T.; Perry, Leslie Anne und Hosack-Curlin, Karen (1998): "Internet use by university students: an interdisciplinary study on three campuses". *Internet Research* 8 (2): S. 136–141.

Pfaffenberger, Bryan (1995): *The USENET Book – Finding, Using and Surviving Newsgroups on the Internet*. Reading, Menlo Park, New York etc.: Addison-Wesley.

Pfammatter, René (1998): "Hypertext – das Multimediakonzept: Strukturen, Funktionsweisen, Qualitätskriterien". In: *Multi Media Mania: Reflexionen zu Aspekten Neuer Medien*, herausgegeben von Pfammatter, René, Konstanz: UVK Medien, S. 45–75.

Pierre, John M. (2001): "On the Automated Classification of Web Sites". *Linköping Electronic Articles in Computer and Information Science* 6 (1). Online verfügbar: `http://www.ida.liu.se/ext/epa/cis/`.

Pilgrim, Chris J. und Leung, Ying K. (1999): "Designing WWW Site Map Systems". In: *10th International Workshop on Database and Expert Systems Applications*. S. 253–258.

Pitkow, James E. (1998): "Summary of WWW Characterizations". In: *Proceedings of the Seventh International World Wide Web Conference*. Brisbane.

PND (1996): *Normdaten-CD-ROM für Personennamen und Schlagwörter*. Deutsche Bibliothek, Frankfurt/Main.

Pohl, Margit und Purghathofer, Peter (2004): "Hypertext Writing Profiles and Visualistion". *Computers and the Humanities* 38 (1): S. 83–105.

Pollem, Niels (1999): "One size fits all". *iX, Magazin für professionelle Informationstechnik* (7): S. 76–80.

Pollock, Annabel und Hockley, Andrew (1997): "What's Wrong with Internet Searching". *D-Lib Magazine* 3 (3).

Poock, Michael C. und Lefond, Dennis (2001): "How College-Bound Prospects Perceive University Web Sites: Findings, Implications and Turning Browsers into Applicants". *College & University Journal* 77 (1): S. 15–21.

Potok, Thomas E.; Elmore, Mark T.; Reed, Joel W. und Samatova, Nagiza F. (2002): "An Ontology-based HTML to XML Conversion Using Intelligent Agents". In: *Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35)*. Big Island, Hawaii.

Procter, Rob und Goldenberg, Ana (1998): "Genres in Support of Collaborative Information Retrieval in the Virtual Library". *Interacting with Computers* 10 (2): S. 157–175.

Quasthoff, Uta M. (1997): "Kommunikative Normen im Entstehen: Beobachtungen zu Kontextualisierungsprozessen in elektronischer Kommunikation". In: Weingarten (1997a), S. 23–50.

Quek, Choon Yang (1997): "Classification of World Wide Web Documents". Senior Honors Thesis, School of Computer Science, Carnegie Mellon University. Online verfügbar: `http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-11/www/wwkb/`.

Quillian, M. Ross (1967): "Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities". *Behavioral Science* 12 (5): S. 410–430.

Quinlan, J. Ross (1993): *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Rada, Roy (1991): *Hypertext – From Text to Expertext*. London, New York, St. Louis etc.: McGraw-Hill.

Rada, Roy (1995): *Interactive Media*. New York, Berlin, Heidelberg etc.: Springer.

Raggett, Dave (1997): "HTML 3.2 Reference Specification". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/REC-html32/`.

Raggett, Dave; Hors, Arnaud Le und Jacbos, Ian (1997): "HTML 4.0 Specification". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/REC-html40/`.

Raggett, Dave; Hors, Arnaud Le und Jacbos, Ian (1999): "HTML 4.01 Specification". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/html401/`.

Raghavan, Sriram und Garcia-Molina, Hector (2001): "Crawling the Hidden Web". In: *Proceedings of the 27th International Conference on Very Large Databases (VLDB)*. S. 129–138.

Rahman, Fuad; Alam, Hassan und Hartono, Rachmat (2001): "Understanding the Flow of Content in Summarizing HTML Documents". In: *International Workshop on Document Layout Interpretation and its Applications (DLIA '01)*. Seattle.

Raible, Wolfgang (1996): "Wie soll man Texte typisieren?" In: *Texte – Konstitution, Verarbeitung, Typik*, herausgegeben von Michaelis, Susanne und Tophinke, Doris, München, Newcastle: Lincom, Band 13 von *Edition Linguistik*, S. 59–72.

Raisman, Neal A. (2003): "Ah, What Rotten Webs We Weave". *Chronicle of Higher Education* 49 (23). Ausgabe vom 14. Februar 2003.

Raskin, Jef (1987): "The Hype in Hypertext: A Critique". In: *Proceedings of the ACM conference on Hypertext*. Chapel Hill, S. 325–330.

Ratner, Julie; Grose, Eric M. und Forsythe, Chris (1996): "Characterization and Assessment of HTML Style Guides". In: *Conference on Human Factors in Computing Systems – CHI 96*. ACM, New York, Band 2, S. 115–116.

Rauber, Andreas; Aschenbrenner, Andreas und Witvoet, Oliver (2002): "Austrian Online Archive Processing: Analyzing Archives of the World Wide Web". In: *Research and Advanced Technology for Digital Technology – 6th European Conference, ECDL 2002*, herausgegeben von Agosti, Maristella und Thanos, Costantino. Berlin, Heidelberg, New York etc.: Springer, Band 2458 von *Lecture Notes in Computer Science*, S. 16–31.

Rauber, Andreas und Merkl, Dieter (2003): "Text Mining in the SOMLib Digital Library System: The Representation of Topics and Genres". *Applied Intelligence* 18 (3): S. 271–293.

Rauber, Andreas und Müller-Kögler, Alexander (2001): "Integrating Automatic Genre Analysis into Digital Libraries". In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, herausgegeben von Fox, E. A. und Borgman, C. L. Roanoke, S. 1–10.

Ray, Erik T. und McIntosh, Jason (2002): *Perl & XML – XML Processing with Perl*. Beijing, Cambridge, Farnham etc.: O'Reilly & Associates.

Raymond, Eric S. (1996): *The New Hacker's Dictionary*. Cambridge, London: MIT Press, 3. Auflage.

Rehm, Georg (1998): *Vorüberlegungen zur automatischen Zusammenfassung deutschsprachiger Texte mittels einer SGML- und DSSSL-basierten Repräsentation von RST-Relationen*. Magisterarbeit, Studiengang Computerlinguistik und Künstliche Intelligenz, Universität Osnabrück. Online verfügbar: `http://www.uni-giessen.de/~g91063/papers.shtml`.

Rehm, Georg (2001): "`korpus.html` – Zur Sammlung, Datenbank-basierten Erfassung, Annotation und Auswertung von HTML-Dokumenten". In: Lobin (2001b), S. 93–103.

Rehm, Georg (2002a): "Schriftliche Mündlichkeit in der Sprache des World Wide Web". In: Ziegler und Dürscheid (2002), S. 263–308.

Rehm, Georg (2002b): "Towards Automatic Web Genre Identification – A Corpus-Based Approach in the Domain of Academia by Example of the Academic's Personal Homepage". In: *Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35)*. Big Island, Hawaii.

Rehm, Georg (2004a): "Das World Wide Web". In: Carstensen et al. (2004), S. 455–460.

Rehm, Georg (2004b): "Hypertextsorten-Klassifikation als Grundlage generischer Informationsextraktion". In: Mehler und Lobin (2004), S. 219–233.

Rehm, Georg (2004c): "Ontologie-basierte Hypertextsorten-Klassifikation". In: Mehler und Lobin (2004), S. 121–137.

Rehm, Georg (2004d): "Texttechnologie und das World Wide Web – Anwendungen und Perspektiven". In: Lobin und Lemnitzer (2004), S. 433–464.

Rehm, Georg (2004e): "Texttechnologische Grundlagen". In: Carstensen et al. (2004), S. 138–147.

Rehm, Georg und Lobin, Henning (2003): "Multimedia in der Informationsgesellschaft: Von Open Source zu Open Information". In: *Psycholinguistik*, herausgegeben von Gert Rickheit, Werner Deutsch und Herrmann, Theo, Berlin: de Gruyter, Band 24 von *Handbücher zur Sprach- und Kommunikationswissenschaft (HSK)*, S. 889–899.

Reimer, Ulrich (1991): *Einführung in die Wissensrepräsentation*. Leitfäden der angewandten Informatik. Stuttgart: Teubner.

Reiss, Eric L. (2000): *Practical Information Architecture – A Hands-On Approach to Structuring Successful Websites*. Harlow, London, New York etc.: Addison-Wesley.

Renear, Allen (1997): "The Digital Library Research Agenda – What's Missing – And How Humanities Textbase Projects can Help". *D-Lib Magazine* 3 (7).

Renner, Karl-Heinz (2003): "Selbstdarstellung im MUD und auf privaten Homepages – Unterschiede und Gemeinsamkeiten". In: Keitel et al. (2003), S. 263–274.

Resnik, Philip und Smith, Noah A. (2002): "The Web as a Parallel Corpus". Technischer Bericht UMIACS-TR-2002-61, University of Maryland. Online verfügbar: `http://www.umiacs.umd.edu/~resnik/pubs.html`.

RFC 0791 (1981): "Internet Protocol". Network Working Group – Request for Comments (RFC). Jon Postel. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 0793 (1981): "Transmission Control Protocol". Network Working Group – Request for Comments (RFC). Jon Postel. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 0822 (1982): "Standard for the Format of ARPA Internet Text Messages". Network Working Group – Request for Comments (RFC). David H. Crocker. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 0850 (1983): "Standard for Interchange of USENET Messages". Network Working Group – Request for Comments (RFC). Mark R. Horton. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 0959 (1985): "File Transfer Protocol (FTP)". Network Working Group – Request for Comments (RFC). Jon Postel und Joyce Reynolds. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 0977 (1986): "Network News Transfer Protocol – A Proposed Standard for the Stream-Based Transmission of News". Network Working Group – Request for Comments (RFC). Brian Kantor und Phil Lapsley. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 1034 (1987): "Domain Names – Concepts and Facilities". Network Working Group – Request for Comments (RFC). Paul Mockapetris. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 1035 (1987): "Domain Names – Implementation and Specification". Network Working Group – Request for Comments (RFC). Paul Mockapetris. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 1036 (1987): "Standard for Interchange of USENET Messages". Network Working Group – Request for Comments (RFC). Mark R. Horton und Rick Adams. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 1123 (1989): "Requirements for Internet Hosts – Application and Support". Network Working Group – Request for Comments (RFC). Robert Braden. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 1321 (1992): "The MD5 Message-Digest Algorithm". Network Working Group – Request for Comments (RFC). Ronald Rivest. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 1349 (1992): "Type of Service in the Internet Protocol Suite". Network Working Group – Request for Comments (RFC). Philip Almquist. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 1436 (1993): "The Internet Gopher Protocol (a distributed document search and retrieval protocol)". Network Working Group – Request for Comments (RFC). Farhad Anklesaria, Mark McCahill, Paul Lindner, David Johnson, Daniel Torrey und Bob Alberti. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 1625 (1994): "WAIS over Z39.50-1988". Network Working Group – Request for Comments (RFC). Margaret St. Pierre, Jim Fullton, Kevin Gamiel, Jonathan Goldman, Brewster Kahle, John A. Kunze, Harry Morris und Francois Schiettecatte. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 1738 (1994): "Uniform Resource Locators (URL)". Network Working Group – Request for Comments (RFC). Tim Berners-Lee, Larry Masinter und Mark McCahill. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 1808 (1995): "Relative Uniform Resource Locators". Network Working Group – Request for Comments (RFC). Roy Fielding. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 1866 (1995): "Hypertext Markup Language – 2.0". Network Working Group – Request for Comments (RFC). Tim Berners-Lee und Dan Connolly. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 1867 (1995): "Form-based File Upload in HTML". Network Working Group – Request for Comments (RFC). Ernesto Nebel und Larry Masinter. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 1942 (1996): "HTML Tables". Network Working Group – Request for Comments (RFC). Dave Raggett. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 1980 (1996): "A Proposed Extension to HTML: Client-Side Image Maps". Network Working Group – Request for Comments (RFC). James L. Seidman. Online verfügbar: http://www.ietf.org/rfc/.

RFC 1991 (1996): "PGP Message Exchange Formats". Network Working Group – Request for Comments (RFC). Derek Atkins, William Stallings und Philip Zimmermann. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2045 (1996): "Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies". Network Working Group – Request for Comments (RFC). Ned Freed und Nathaniel Borenstein. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2046 (1996): "Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types". Network Working Group – Request for Comments (RFC). Ned Freed und Nathaniel Borenstein. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2047 (1996): "MIME (Multipurpose Internet Mail Extensions) Part Three: Message Header Extensions for Non-ASCII Text". Network Working Group – Request for Comments (RFC). Keith Moore. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2048 (1996): "Multipurpose Internet Mail Extensions (MIME) Part Four: Registration Procedures". Network Working Group – Request for Comments (RFC). Ned Freed, John Klensin und Jon Postel. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2070 (1997): "Internationalization of the Hypertext Markup Language". Network Working Group – Request for Comments (RFC). Frangois Yergeau, Gavin Thomas Nicol, Glenn Adams und Martin J. Duerst. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2083 (1997): "PNG (Portable Network Graphics) Specification". Network Working Group – Request for Comments (RFC). Thomas Boutell. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2109 (1997): "HTTP State Management Mechanism". Network Working Group – Request for Comments (RFC). David M. Kristol und Lou Montulli. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2181 (1997): "Clarifications to the DNS Specification". Network Working Group – Request for Comments (RFC). Robert Elz und Randy Bush. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2279 (1998): "UTF-8, a transformation format of ISO 10646". Network Working Group – Request for Comments (RFC). Francois Yergeau. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2326 (1998): "Real Time Streaming Protocol (RTSP)". Network Working Group – Request for Comments (RFC). Henning Schulzrinne, Anup Rao und Robert Lanphier. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2396 (1998): "Uniform Resource Identifiers (URI): Generic Syntax". Network Working Group – Request for Comments (RFC). Tim Berners-Lee, Roy Fielding und Larry Masinter. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2413 (1999): "Dublin Core Metadata for Resource Discovery". Network Working Group – Request for Comments (RFC). Stuart L. Weibel, John A. Kunze, Carl Lagoze und Misha Wolf. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2616 (1999): "Hypertext Transfer Protocol – HTTP/1.1". Network Working Group – Request for Comments (RFC). Roy T. Fielding, James Gettys, Jeffrey C. Mogul, Henrik Frystyk Nielsen, Larry Masinter, Paul J. Leach und Tim Berners-Lee. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2731 (1999): "Encoding Dublin Core Metadata in HTML". Network Working Group – Request for Comments (RFC). John A. Kunze. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2810 (2000): "Internet Relay Chat: Architecture". Network Working Group – Request for Comments (RFC). Christophe Kalt. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2811 (2000): "Internet Relay Chat: Channel Management". Network Working Group – Request for Comments (RFC). Christophe Kalt. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2812 (2000): "Internet Relay Chat: Client Protocol". Network Working Group – Request for Comments (RFC). Christophe Kalt. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2813 (2000): "Internet Relay Chat: Server Protocol". Network Working Group – Request for Comments (RFC). Christophe Kalt. Online verfügbar: http://www.ietf.org/rfc/.

RFC 2822 (2001): "Internet Message Format". Network Working Group – Request for Comments (RFC). Peter Resnick. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 2854 (2000): "The 'text/html' Media Type". Network Working Group – Request for Comments (RFC). Dan Connolly und Larry Masinter. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 2964 (2000): "Use of HTTP State Management". Network Working Group – Request for Comments (RFC). Keith Moore und Ned Freed. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 2965 (2000): "HTTP State Management Mechanism". Network Working Group – Request for Comments (RFC). David M. Kristol und Lou Montulli. Online verfügbar: `http://www.ietf.org/rfc/`.

RFC 3151 (2001): "A URN Namespace for Public Identifiers". Network Working Group – Request for Comments (RFC). Norman Walsh, John Cowan und Paul Grosso. Online verfügbar: `http://www.ietf.org/rfc/`.

Ricardo, Francisco J. (1998): "Stalking the Paratext: Speculations on Hypertext Links as a Second Order Text". In: *Proceedings of the ninth ACM conference on Hypertext and Hypermedia: Links, Objects, Time and Space – Structure in Hypermedia Systems*. Pittsburgh, S. 142–151.

Richter, Gerd; Riecke, Jörg und Schuster, Britt-Marie (Hrsg.) (2000): *Raum, Zeit, Medium – Sprache und ihre Determinanten. Festschrift für Hans Ramge*. Darmstadt: Hessische Historische Kommission.

Rieffel, Eleanor G. (1999): "The Genre of Mathematics Writing and its Implications for Digital Documents". In: *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32)*.

Riley, Patricia; Keough, Colleen M.; Christiansen, Thora; Meilich, Ofer und Pierson, Jillian (1998): "Community or Colony: The Case of Online Newspapers and the Web". *Journal of Computer-Mediated Communication* 4 (1). Online verfügbar: `http://www.ascusc.org/jcmc/`.

Riloff, Ellen M. (1994): *Information Extraction as a Basis for Portable Text Classification Systems (= Technical Report 95-004, Center for Intelligent Information Retrieval)*. Ph. d. thesis, Department of Computer Science, University of Massachusetts Amherst, Amherst.

Robbins, Stephanie S. und Stylianou, Antonis C. (2003): "Global corporate web sites: an empirical investigation of content and design". *Information & Management* 40 (3): S. 205–212.

Roberts, Gregory F. (1998): "The Home Page as Genre: A Narrative Approach". In: *Proceedings of the 31st Hawaii International Conference on Systems Sciences (HICSS-31)*. Big Island, Hawaii, Band 2, S. 78–86.

Rocha, Cristiano; Schwabe, Daniel und Aragao, Marcus Poggi (2004): "A Hybrid Approach for Searching in the Semantic Web". In: *Proceedings of the 13th Conference on World Wide Web (WWW-2004)*. New York, S. 374–383. Refereed Papers Track.

Romary, Laurent; Bonhomme, Patrice; Bruneseaux, Florence und Pierrel, Jean-Marie (1999): "Silfide: A System for Open Access and Distributed Delivery of TEI Encoded Documents". *Computers and the Humanities* 33 (1–2): S. 31–38.

Rosch, Eleanor (1977): "Human Categorization". In: *Studies in Cross-Cultural Psychology*, herausgegeben von Warren, Neil, New York, London, San Francisco: Academic Press, Band 1, S. 1–49.

Rosch, Eleanor (1978): "Principles of Categorization". In: *Cognition and Categorization*, herausgegeben von Rosch, Eleanor und Lloyd, Barbara B., Hillsdale: Erlbaum, S. 27–48.

Rosenbloom, Andrew (2004): "The Blogosphere". *Communications of the ACM* 47 (12): S. 30–33.

Rosenfeld, Louis und Morville, Peter (1998): *Information Architecture for the World Wide Web*. A Nutshell Handbook. Cambridge, Köln, Paris etc.: O'Reilly & Associates.

Rösner, Dietmar und Stede, Manfred (1993): "Zur Struktur von Texten – Eine Einführung in die Rhetorical Structure Theory". *KI* 2: S. 14–21.

Rosso, Mark A. (2005): *Using Genre to Improve Web Search*. Ph. d. thesis, School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill.

Rouet, Jean-François und Levonen, Jarmo J. (1996): "Studying and Learning With Hypertext: Empirical Studies and Their Implications". In: Rouet et al. (1996), S. 9–23.

Rouet, Jean-François; Levonen, Jarmo J.; Dillon, Andrew und Spiro, Rand J. (Hrsg.) (1996): *Hypertext and Cognition*. Mahwah: Erlbaum.

Roussinov, Dmitri; Crowston, Kevin; Nilan, Mike; Kwasnik, Barbara; Cai, Jin und Liu, Xiaoyong (2001): "Genre Based Navigation on the Web". In: *Proceedings of the 34th Hawaii International Conference on Systems Sciences (HICSS-34)*.

Rui, Vong; Huang, Thomas S. und Chang, Shih-Fu (1999): "Image Retrieval: Past, Present, and Future". *Journal of Visual Communication and Image Representation* 10: S. 1–23.

Rumelhart, David E. (1975): "Notes on a Schema for Stories". In: *Representation and Understanding – Studies in Cognitive Science*, herausgegeben von Bobrow, Daniel G. und Collins, Allan, New York, San Francisco, London: Academic Press, S. 211–236.

Rumelhart, David E. (1980): "On Evaluating Story Grammars". *Cognitive Science* 4: S. 313–316.

Runkehl, Jens; Schlobinski, Peter und Siever, Torsten (1998): *Sprache und Kommunikation im Internet – Überblick und Analysen*. Opladen, Wiesbaden: Westdeutscher Verlag.

Ryan, Terry; Field, Richard H. G. und Olfman, Lorne (2003): "The evolution of US state government home pages from 1997 to 2002". *International Journal of Human-Computer Studies* 59 (4): S. 403–430.

Sæbø, Øystein und Päivärinta, Tero (2005): "Autopoietic Cybergenres for e-Democracy? Genre Analysis of a Web-Based Discussion Board". In: *Proceedings of the 38th Hawaii International Conference on Systems Sciences (HICSS-38)*. Big Island, Hawaii.

Sager, Sven F. (1997): "Intertextualität und die Interaktivität von Hypertexten". In: Klein und Fix (1997), S. 109–123.

Sager, Sven F. (2000): "Hypertext und Hypermedia". In: Brinker et al. (2000), S. 587–604.

Sahuguet, Arnaud und Azavant, Fabien (2001): "Building Intelligent Web Applications Using Lightweight Wrappers". *Data and Knowledge Engineering* 36 (3): S. 283–316.

de Saint-Georges, Ingrid (1998): "Click Here if You Want to Know Who I Am. Deixis in Personal Homepages". In: *Proceedings of the 31st Hawaii International Conference on Systems Sciences (HICSS-31)*. Big Island, Hawaii, Band 2, S. 68–77.

Sandbothe, Mike (1997): "Interaktivität – Hypertextualität – Transversalität: Eine medienphilosophische Analyse des Internet". In: *Mythos Internet*, herausgegeben von Münker, Stefan und Rösler, Alexander, Frankfurt/Main: Suhrkamp, S. 56–82.

Sandig, Barbara (1972): "Zur Differenzierung gebrauchssprachlicher Textsorten im Deutschen". In: Gülich und Raible (1972), S. 113–124.

Sandig, Barbara (1997): "Formulieren und Textmuster – Am Beispiel von Wissenschaftstexten". In: *Schreiben in den Wissenschaften*, herausgegeben von Jakobs, Eva-Maria und Knorr, Dagmar, Frankfurt/Main, Berlin, Bern etc.: Peter Lang, Band 1 von *Textproduktion und Medium*, S. 25–44.

Sandig, Barbara (2000): "Text als prototypisches Konzept". In: Mangasser-Wahl (2000a), S. 93–112.

Sanguanpong, Surasak; Piamsa-nga, Punpiti; Poovarawan, Yuen und Warangrit, Suthiphol (2000): "Measuring Thai Web Using NontriSpider". In: *Proceedings of the International Forum cum Conference on Information Technology and Communication*. Bangkok, S. 123–132.

Santini, Marina (2004a): "Identification of Genres on the Web: a Multi-Faceted Approach". In: *Proceedings of the 26th European Conference on IR Research (ECIR 2004)*, herausgegeben von Oakes, Michael P. University of Sunderland, Band 2. Poster Abstract.

Santini, Marina (2004b): "A Shallow Approach to Syntactic Feature Extraction for Genre Classification". Technischer Bericht ITRI-04-02, Information Technology Research Institute, University of Brighton, Brighton.

Santini, Marina (2004c): "State-of-the-Art on Automatic Genre Identification". Technischer Bericht ITRI-04-03, Information Technology Research Institute, University of Brighton, Brighton.

Santini, Marina (2005a): "Genres in Formation? An Exploratory Study of Web Pages using Cluster Analysis". In: *Proceedings of the 8th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK 05)*. University of Manchester.

Santini, Marina (2005b): "Linguistic Facets for Genre and Text Type Identification: A Description of Linguistically-Motivated Features". Technischer Bericht ITRI-05-02, Information Technology Research Institute, University of Brighton, Brighton.

Sanz, Ismael; Berlanga, Rafael und Aramburu, María José (1998): "Gathering Metadata from Web-Based Repositories of Historical Publications". In: *Ninth International Workshop on Database and Expert Systems Applications*, herausgegeben von Wagner, Roland. Wien, S. 473–478.

Sasaki, Felix (2004): "Secondary Information Structuring – A Methodology for the Vertical Interrelation of Information Resources". In: *Proceedings of Extreme Markup Languages 2004*. Montreal. Online verfügbar: `http://www.mulberrytech.com/Extreme/Proceedings/`.

Sasaki, Felix und Witt, Andreas (2004): "Linguistische Korpora". In: Lobin und Lemnitzer (2004), S. 195–216.

Sassen, Claudia (2000): "Phatische Variabilität bei der Initiierung von Internet-Relay-Chat-Dialogen". In: *Soziales im Netz – Sprache, Beziehungen und Kommunikationskulturen im Internet*, herausgegeben von Thimm, Caja, Opladen, Wiesbaden: Westdeutscher Verlag, S. 89–108.

Saunders, Chad und Chiasson, Mike (2005): "Using Genre Systems to Investigate the Interplay Between Technology-in-Practice and the Knowledge Management Practices of Lawyers". In: *Proceedings of the 38th Hawaii International Conference on Systems Sciences (HICSS-38)*. Big Island, Hawaii.

Schank, Roger C. (Hrsg.) (1975): *Conceptual Information Processing*, Band 3 von *Fundamental Studies in Computer Science*. Amsterdam: North-Holland.

Schank, Roger C. und Abelson, Robert P. (1977): *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Hillsdale: Erlbaum.

Scherner, Maximilian (2000): "Kognitionswissenschaftliche Methoden in der Textanalyse". In: Brinker et al. (2000), S. 186–195.

Schick, Russell (2003): "Personal Home Pages and the Family". Department of Computer Science, University of Toronto. Online verfügbar: `http://www.dgp.toronto.edu/~rschick/`.

Schlobinski, Peter (2000a): "Chatten im Cyberspace". In: *Die deutsche Sprache zur Jahrtausendwende*, herausgegeben von Eichhoff-Cyrus, Karin M. und Hoberg, Rudolf, Mannheim, Leipzig, Wien, Zürich: Dudenverlag, S. 63–79.

Schlobinski, Peter (2000b): "HyperText und Textanalyse". In: Richter et al. (2000), S. 809–826.

Schmid-Isler, Salome (2000): "The Language of Digital Genres – A Semiotic Investigation of Style and Iconology on the World Wide Web". In: *Proceedings of the 33rd Hawaii International Conference on Systems Sciences (HICSS-33)*.

Schmid-Isler, Salome und Oehninger, Thomas (2004): "Products in Genre Discussion. Enhanced Approach with the Media Reference Model (MRM)". In: *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS-37)*. Big Island, Hawaii.

Schmidt, Gurly (2000): "Chat-Kommunikation im Internet – eine kommunikative Gattung". In: *Soziales im Netz – Sprache, Beziehungen und Kommunikationskulturen im Internet*, herausgegeben von Thimm, Caja, Opladen, Wiesbaden: Westdeutscher Verlag, S. 109–130.

Schmidt, Ingrid (2004): "Modellierung von Metadaten". In: Lobin und Lemnitzer (2004), S. 143–164.

Schmitz, Ulrich (1996): "ZAP und Sinn. Fragmentarische Textkonstitution durch überfordernde Medienrezeption". In: Hess-Lüttich et al. (1996), S. 11–29.

Schmitz, Ulrich (1997): "Schriftliche Texte in multimedialen Kontexten". In: Weingarten (1997a), S. 131–158.

Schmitz, Ulrich (2000): "Neue Medien als Arbeitsinstrument der Linguistik". In: Kallmeyer (2000b), S. 253–274.

Schmitz, Ulrich (2001): "Optische Labyrinthe im digitalen Journalismus". In: Bucher und Püschel (2001), S. 207–232.

Schmitz, Ulrich (2003): "Deutsche Schriftsprache in hypermedialer Umgebung". *Zeitschrift für germanistische Linguistik* 31 (2): S. 253–273.

Schoenke, Eva (2000): "Textlinguistik im deutschsprachigen Raum". In: Brinker et al. (2000), S. 123–131.

Schönefeld, Tim (2001): "Bedeutungskonstitution im Hypertext". Networx, Nr. 19. Online verfügbar: `http://www.mediensprache.net/de/networx/`.

Schultze, Ulrike und Boland, Richard J. (1997): "Hard and Soft Information Genres: An Analysis of two Notes Databases". In: *Proceedings of the 30th Hawaii International Conference on Systems Sciences (HICSS-30)*. Band 6, S. 40–49.

Schuster, Hermann Josef (1996): "Leitungsorganisation". In: Flämig et al. (1996), S. 839–858.

Schütte, Daniela (2004a): *Homepages im World Wide Web – Eine interlinguale Untersuchung zur Textualität in einem globalen Medium*, Band 44 von *Germanistische Arbeiten zu Sprache und Kulturgeschichte*. Frankfurt/Main: Peter Lang.

Schütte, Wilfried (2000): "Sprache und Kommunikationsformen in Newsgroups und Mailinglisten". In: Kallmeyer (2000b), S. 142–178.

Schütte, Wilfried (2004b): "Diskursstrukturen in fachlichen Mailinglisten: Zwischen Einwegkommunikation und Interaktion". *Osnabrücker Beiträge zur Sprachtheorie* (68): S. 55–75.

Schütz, Astrid und Machilek, Franz (2003): "Who owns a personal home page? A discussion of sampling problems and a strategy based on a search engine". *Swiss Journal of Psychology* 62 (2): S. 121–129.

Schütz, Astrid; Machilek, Franz und Marcus, Bernd (2003): "Selbstdarstellung auf privaten Homepages – Ausgangspunkt und erste Ergebnisse". In: Keitel et al. (2003), S. 234–262.

Schweiger, Wolfgang (1996): "Gebrauchstexte im Hypertext- und Papierformat". *Publizistik* 41 (3): S. 327–345.

Searle, John R. (1969): *Speech Acts*. Cambridge: Cambridge University Press.

Seibold, Ernst (2001): "Das Onlinemedium als Fortsetzung des Printmediums mit besseren Mitteln". In: Bucher und Püschel (2001), S. 233–255.

Selberg, Erik Warren (1999): *Towards Comprehensive Web Search*. Ph. d. thesis, University of Washington.

Seo, Heekyoung und Choi, Jaeyoung Yang Joongmin (2001): "Knowledge-based Wrapper Generation by Using XML". In: *Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*. Seattle.

Shakes, Jonathan; Langheinrich, Marc und Etzioni, Oren (1997): "Dynamic Reference Sifting: A Case Study in the Homepage Domain". In: *Proceedings of the 6th International World Wide Web Conference*. Santa Clara, S. 189–200.

Shepherd, Michael und Watters, Carolyn (1998): "The Evolution of Cybergenres". In: *Proceedings of the 31st Hawaii International Conference on Systems Sciences (HICSS-31)*. Big Island, Hawaii, Band 2, S. 97–109.

Shepherd, Michael und Watters, Carolyn (1999): "The Functionality Attribute of Cybergenres". In: *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32)*.

Shepherd, Michael; Watters, Carolyn und Kennedy, Alistair (2004): "Cybergenre: Automatic Identification of Home Pages on the Web". *Web Engineering* 3 (3/4): S. 236–251.

Shih, Lawrence K. und Karger, David R. (2004): "Using URLs and Table Layout for Web Classification Tasks". In: *Proceedings of the 13th Conference on World Wide Web (WWW-2004)*. New York, S. 193–202. Refereed Papers Track.

Shin, Christian; Doermann, David und Rosenfeld, Azriel (2001): "Classification of Document Pages Using Structure-Based Features". *International Journal of Document Analysis and Recognition* 3: S. 232–247.

Shneiderman, Bob; Nielsen, Jakob; Butler, Scott; Levi, Michael und Conrad, Frederick (1998): "Is the Web Really Different From Everything Else?" In: *Conference on Human Factors in Computing Systems – CHI 98*. ACM, New York, S. 92–93.

Siegel, David (1999a): *Das Geheimnis erfolgreicher Web Sites – Business, Budget, Manpower, Lizenzen, Design*. Frankfurt/Main: Zweitausendeins. Doppelband. Deutsche Übersetzung von: Siegel, David (1997): *Secrets of Successful Web Sites*. Indianapolis: Hayden Books.

Siegel, David (1999b): *Web Site Design – Killer Web Sites*. Frankfurt/Main: Zweitausendeins. Doppelband. Deutsche Übersetzung von: Siegel, David (1997): *Creating Killer Web Sites*. Indianapolis: Hayden Books, 2. Auflage.

Sigletos, Georgios; Farmakiotou, Dimitra; Stamatakis, Kostas; Paliouras, Georgios und Karkaletsis, Vangelis (2003): "Annotating Web pages for the needs of Web Information Extraction applications". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Simanowski, Roberto (2004): "Der Autor ist tot, es lebe der Autor – Autorschaften im Internet". In: Bieber und Leggewie (2004), S. 190–215.

Singh, Surendra N. und Dalal, Nikunj P. (1999): "Web home pages as advertisements". *Communications of the ACM* 42 (8): S. 91–98.

Slone, Debra J. (2002): "The Influence of Mental Models and Goals on Search Patterns During Web Interaction". *Journal of the American Society for Information Science and Technology* 53 (13): S. 1152–1169.

Smith, Michael K.; Welty, Chris und McGuinness, Deborah L. (2004): "OWL Web Ontology Language – Guide". Technische Spezifikation, W3C. Online verfügbar: `http://www.w3.org/TR/2004/REC-owl-guide-20040210/`.

Smith, P. A.; Newman, I. A. und Parks, L. M. (1997): "Virtual Hierarchies and Virtual Networks: Some Lessons from Hypermedia Usability Research Applied to the World Wide Web". *Human-Computer Studies* 47: S. 67–95.

Smoliar, Stephen W. und Baker, James D. (1997): "Text Types in Hypermedia". In: *Proceedings of the 30th Hawaii International Conference on Systems Sciences (HICSS-30)*. Band 6, S. 68–77.

Soderland, Stephen (1997): "Learning to Extract Text-Based Information from the World Wide Web". In: *Proceedings of the 3rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-97)*, herausgegeben von Heckerman, David; Mannila, Heikki und Pregibon, Daryl. Newport Beach, S. 251–254.

Song, Ruihua; Liu, Haifeng; Wen, Ji-Rong und Ma, Wei-Ying (2004): "Learning Block Importance Models for Web Pages". In: *Proceedings of the 13th Conference on World Wide Web (WWW-2004)*. New York, S. 203–211. Refereed Papers Track.

Sørensen, Carsten (1998): "Where Have You Been Today? Investigating Web Navigation Support". In: *Proceedings of the 21st Information Systems Research Seminar In Scandinavian (IRIS-21)*, herausgegeben von Buch, Niels J.; Damsgaard, Jan; Eriksen, Lars B.; Iversen, Jakob H. und Nielsen, Peter A. Sæby Søbad, Aalborg University, S. 875–890.

Sparks, Colin (2003): "The Contribution of Online Newspapers to the Public Sphere: A United Kingdom Case Study". *Trends in Communication* 11 (2): S. 111–126.

Spence, Robert (1999): "A Framework for Navigation". *International Journal of Human-Computer Studies* 51 (5): S. 919–945.

Sperberg-McQueen, C. M. und Burnard, Lou (Hrsg.) (2002): *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium; Humanities Computing Unit, University of Oxford.

Spink, Amanda; Jansen, Bernard J. und Ozmultu, H. Cenk (2000): "Use of query reformulation and relevance feedback by Excite users". *Internet Research* 10 (4): S. 317–328.

Spink, Amanda; Wolfram, Dietmar; Jansen, Major B. J. und Saracevic, Tefko (2001): "Searching the Web: The Public and Their Queries". *Journal of the American Society for Information Science and Technology* 52 (3): S. 226–234.

Staab, Steffen und Studer, Rudi (Hrsg.) (2004): *Handbook on Ontologies*. International Handbooks on Information Systems. Berlin, Heidelberg, New York: Springer.

Stamatatos, E.; Fakotakis, N. und Kokkinakis, G. (2000): "Text Genre Detection Using Common Word Frequencies". In: *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*. Saarbrücken.

Stamatatos, Efstathios; Fakotakis, Nikos und Kokkinakis, George (2001): "Automatic Text Categorization in Terms of Genre and Author". *Computational Linguistics* 26 (4): S. 471–495.

Storrer, Angelika (1997): "Vom Text zum Hypertext – Die Produktion von Hypertexten auf der Basis traditioneller wissenschaftlicher Texte". In: Jakobs und Knorr (1997), S. 121–139.

Storrer, Angelika (1999a): "Kohärenz in Text und Hypertext". In: Lobin (1999b), S. 33–65.

Storrer, Angelika (1999b): "Was ist eigentlich eine Homepage? Neue Formen der Wissensorganisation im World Wide Web". *Sprachreport* 1: S. 2–8.

Storrer, Angelika (2000a): "Schriftverkehr auf der Datenautobahn: Besonderheiten der schriftlichen Kommunikation im Internet". In: Voß et al. (2000), S. 151–175.

Storrer, Angelika (2000b): "Was ist »hyper« am Hypertext?" In: Kallmeyer (2000b), S. 222–249.

Storrer, Angelika (2001a): "Getippte Gespräche oder dialogische Texte? Zur kommunikationstheoretischen Einordnung der Chat-Kommunikation". In: *Sprache im Alltag. Beiträge zu neuen Perspektiven der Linguistik*, herausgegeben von Lehr, Andrea; Kammerer, Matthias; Konerding, Klaus-Peter; Storrer, Angelika; Thimm, Caja und Wolski, Werner, Berlin etc.: de Gruyter, S. 439–466. Herbert Ernst Wiegand zum 65. Geburtstag gewidmet.

Storrer, Angelika (2001b): "Neue Medien – neue Stilfragen: Das World Wide Web unter stilistischer Perspektive". In: *Perspektiven auf Stil*, herausgegeben von Jakobs, Eva-Maria und Rothkegel, Annely, Tübingen: Niemeyer, S. 101–124.

Storrer, Angelika (2001c): "Schreiben, um besucht zu werden – Textgestaltung fürs World Wide Web". In: Bucher und Püschel (2001), S. 173–205.

Storrer, Angelika (2003): "Kohärenz in Hypertexten". *Zeitschrift für germanistische Linguistik* 31 (2): S. 274–292.

Storrer, Angelika (2004a): "Text-Bild-Bezüge und Nutzermetaphern im World Wide Web". *Mitteilungen des Germanistenverbands* 51: S. 40–57.

Storrer, Angelika (2004b): "Text und Hypertext". In: Lobin und Lemnitzer (2004), S. 13–49.

Strohner, Hans (2000): "Kognitive Voraussetzungen: Wissenssysteme – Wissensstrukturen – Gedächtnis". In: Brinker et al. (2000), S. 261–274.

Strube, Gerhard und Hölscher, Christoph (2000): "Informationssuche und Wissenskommunikation: Wissenschaftlicher Alltag im Zeitalter der Neuen Medien". In: Voß et al. (2000), S. 177–197.

Su, Louise T. (2003a): "A Comprehensive and Systematic Model of User Evaluation of Web Search Engines: I. Theory and Background". *Journal of the American Society for Information Science and Technology* 54 (13): S. 1175–1192.

Su, Louise T. (2003b): "A Comprehensive and Systematic Model of User Evaluation of Web Search Engines: II. An Evaluation of Undergraduates". *Journal of the American Society for Information Science and Technology* 54 (13): S. 1193–1223.

Svatek, Vojtech und Vacura, Miroslav (2003): "Problem-Solving Models of Website Analysis". In: *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. Budapest.

Swales, John M. (1990): *Genre Analysis – English in academic and research settings*. The Cambridge Applied Linguistics Series. Cambridge: Cambridge University Press.

Tatsumi, Yushin und Asahi, Toshiyuki (2005): "Analyzing Web Page Headings Considering Various Presentation". In: *Proceedings of the 14th International World Wide Web Conference (WWW 2005)*. Chiba, S. 956–957. Poster Track.

Taylor, Mark J.; England, David und Gresty, David (2001): "Knowledge for Web site development". *Internet Research* 11 (5): S. 451–461.

Techtmeier, Bärbel (2000): "Merkmale von Textsorten im Alltagswissen der Sprecher". In: Adamzik (2000a), S. 113–127.

Tennant, Roy (1997): "Dublin Core Resource Types". Dublin Core Working Paper #2, besteht aus dem "Minimalist Draft: July 17, 1997" und dem "Structuralist Draft: July 24, 1997".

Tergan, Olaf-Sigmar (1997): "Hypertext und Hypermedia: Konzeptionen, Lernmöglichkeiten, Lernprobleme". In: Issing und Klimsa (1997), S. 123–137.

Tettinger, Peter J. (1996): "Forschungseinrichtungen an der Hochschule". In: Flämig et al. (1996), S. 991–1005.

Thalheim, Bernhard und Düsterhöft, Antje (2000): "The use of metaphorical structures for internet sites". *Data & Knowledge Engineering* 35 (2): S. 161–180.

Thelwall, Mike (2001a): "Extracting Macroscopic Information from Web Links". *Journal of the American Society for Information Science and Technology* 52 (13): S. 1157–1168.

Thelwall, Mike (2001b): "A Web Crawler Design for Data Mining". *Journal of Information Science* 27 (5): S. 319–326.

Thelwall, Mike (2002): "Conceptualizing documentation on the Web: An evaluation of different heuristic-based models for counting links between university web sites". *Journal of the American Society for Information Science and Technology* 53 (12): S. 995–1005.

Thelwall, Mike (2003): "What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation". *Information Research* 8 (3). Paper No. 151. Online verfügbar: `http://informationr.net/ir/8-3/paper151.html`.

Thelwall, Mike (2005): "Text Characteristics of English Language University Web Sites". *Journal of the American Society for Information Science and technology* 56 (6): S. 609–619.

Thelwall, Mike und Harries, Gareth (2003): "The Connection Between the Research of a University and Counts of Links to Its Web Pages: An Investigation Based upon a Classification of the Relationships of Pages to the Research of the Host University". *Journal of the American Society for Information Science* 54 (7): S. 594–602.

Thelwall, Mike und Wilkinson, David (2003): "Three Target Document Range Metrics for University Web Sites". *Journal of the American Society for Information Science* 54 (6): S. 490–497.

Thiele, Wolfgang (2000): "Textlinguistik im englischsprachigen Raum". In: Brinker et al. (2000), S. 132–139.

Thieme, Werner (1996): "Organisationsstrukturen der Hochschulen". In: Flämig et al. (1996), S. 813–838.

Thomas, Bernd (2000): "Token-Templates and Logic Programs for Intelligent Web Search". *Journal of Intelligent Information Systems* 14 (2–3): S. 241–261.

Thompson, Henry S. (2000): "Corpus Creation for Data-Intensive Linguistics". In: Dale et al. (2000), S. 385–401.

Thümmel, Wolf (1978): "Der Herr der schickt den Jockel aus". *Osnabrücker Beiträge zur Sprachtheorie* 7: S. 115–144.

Thümmel, Wolf (1979): *Vorüberlegungen zu einer Grammatik der Satzverknüpfung*, Band 6 von *Europäische Hochschulschriften: Reihe 21, Linguistik*. Frankfurt/Main, Bern, Cirencester: Peter Lang.

Tietz, Heike (1997): "Die Zukunft der Textlinguistik". In: Antos und Tietz (1997a), S. 223–230.

Tiun, Sabrina; Abdullah, Rosni und Kong, Tang Enya (2001): "Automatic Topic Identification Using Ontology Hierarchy". In: *Computational Linguistics and Intelligent Text Processing*, herausgegeben von Gelbkuh, Alexander. Berlin, Heidelberg, New York etc.: Springer, Band 2004 von *Lecture Notes in Computer Science*, S. 444–453.

Todesco, Rolf (1997): "Die Definition als Textstruktur im Hyper-Sachbuch". In: Jakobs und Knorr (1997), S. 109–120.

Toms, Elaine G. (2001): "Recognizing Digital Genre". *Bulletin of The American Society for Information Science and Technology* 27 (2). Online verfügbar: `http://www.asis.org/Bulletin/`.

Toms, Elaine G. und Campbell, D. Grant (1999): "Genre as Interface Metaphor: Exploiting Form and Function in Digital Environments". In: *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32)*.

Trigg, Randall (1983): *A Network-Based Approach to Text Handling for the Online Scientific Community*. Ph. d. thesis, Department of Computer Science, University of Maryland.

Tschichold, Jan (1960): *Erfreuliche Drucksachen durch gute Typographie*. Ravensburg: Ravensburger Buchverlag.

Turau, Volker (1998a): "Eine empirische Analyse von HTML-Dokumenten im WWW". Technischer Bericht 0198, Fachhochschule Wiesbaden, Wiesbaden. Online verfügbar: `http://www.informatik.fh-wiesbaden.de/~turau/`.

Turau, Volker (1998b): "Web Design: Industry vs. University". Technischer Bericht 0498, Fachhochschule Wiesbaden, Wiesbaden. Online verfügbar: `http://www.informatik.fh-wiesbaden.de/~turau/`.

Turau, Volker (1998c): "Web-Roboter". *Informatik Spektrum* 21 (3): S. 159–160.

Tyrväinen, Pasi und Päivärinta, Tero (1999): "On Rethinking Organizational Document Genres for Electronic Document Management". In: *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32)*.

Vargas-Vera, Maria; Motta, Enrico; Domingue, John; Lanzoni, Mattia; Stutt, Arthur und Ciravegna, Fabio (2002): "MnM: Ontology Driven Semi-automatic and Automatic Support for Semantic Markup". In: Gómez-Pérez und Benjamins (2002), S. 379–391.

Vasudevan, Venu und Palmer, Mark (1999): "On Web Annotations: Promises and Pitfalls of Current Web Infrastructure". In: *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32)*.

Vater, Heinz (1994): *Einführung in die Textlinguistik*, Band 1660 von *UTB für Wissenschaft*. München: Fink, 2. Auflage.

Vater, Heinz (2001): *Einführung in die Textlinguistik*, Band 1660 von *UTB für Wissenschaft*. München: Fink, 3. Auflage.

Viégas, Fernanda B.; Wattenberg, Martin und Dave, Kushal (2004): "Studying Cooperation and Conflict between Authors with history flow Visualizations". In: *Proceedings of the 2004 Conference on Human Factors in Computing Systems*. Wien, S. 575–582.

Vila, Tony; Greenstadt, Rachel und Molnar, David (2003): "Why We can't be Bothered to Read Privacy Policies – Models of Privacy Economics as a Lemons Market". In: *Proceedings of the 5th international Conference on Electronic Commerce*. Pittsburgh, S. 403–407.

Volk, Martin (1998): "Markup of a Test Suite with SGML". In: *Linguistic Databases*, herausgegeben von Nerbonne, John, Cambridge: Cambridge University Press, Band 77 von *CSLI Lecture Notes*, S. 59–76.

Volk, Martin (2000): "Scaling up: Using the WWW to resolve PP attachments and ambiguities". In: *KONVENS-2000 / Sprachkommunikation, Vorträge der gemeinsamen Veranstaltung 5. Konferenz zur Verarbeitung natürlicher Sprache*, herausgegeben von Zühlke, Werner und Schukat-Talamazzini, Ernst Günter. Ilmenau: VDE Verlag, S. 151–155.

Volk, Martin (2001): "Exploiting the WWW as a corpus to resolve PP attachment ambiguities". In: *Proceedings of the Corpus Linguistics 2001 Conference*, herausgegeben von Rayson, Paul; Wilson, Andrew; McEnery, Tony; Hardie, Andrew und Khoja, Shereen. Lancaster, S. 601–606.

Volk, Martin (2002): "Using the Web as Corpus for Linguistic Research". In: *Tähendusepüüdja. Catcher of the Meaning. A Festschrift for Professor Haldur Õim*, Publications of the Department of General Linguistics 3, University of Tartu. Online verfügbar: `http://www.ifi.unizh.ch/CL/volk/publications.html`.

Vora, Pawan R. und Helander, Martin G. (1997): "Hypertext and its Implications for the Internet". In: Helander et al. (1997), S. 877–914.

Vossen, Piek (2003): "Ontologies". In: Mitkow (2003), S. 464–482.

Voß, G. Günther; Holly, Werner und Boehnke, Klaus (Hrsg.) (2000): *Neue Medien im Alltag – Begriffsbestimmungen eines interdisziplinären Forschungsfeldes*. Opladen: Leske + Budrich.

VW 96 (1996): "VW96 Schema Description". Teil des Dokuments "A Dictionary of HTML META Tags". Online verfügbar: `http://vancouver-webpages.com/META/VW96-schema.html`.

Wagner, Franc (1998): "Sind Printmedien im Internet Online-Medien?" In: *Multi Media Mania: Reflexionen zu Aspekten Neuer Medien*, herausgegeben von Pfammatter, René, Konstanz: UVK Medien, S. 191–211.

Walker, Derek (1999): "Taking Snapshots of the Web with a TEI Camera". *Computers and the Humanities* 33 (1–2): S. 185–192.

Walker, Katherine (2000): "»It's difficult to hide it«: The Presentation of Self on Internet Home Pages". *Qualitative Sociology* 23 (1): S. 99–120.

Wall, Larry; Christiansen, Tom und Orwant, Jon (2000): *Programming Perl*. Cambridge, Köln, Paris etc.: O'Reilly & Associates, 3. Auflage.

Walsh, Norman (1999): *DocBook – The Definitive Guide*. Peking, Cambridge, Farnham etc.: O'Reilly & Associates.

Walters, Alison (1996): *An Analysis of Purposes and Forms of Personal Homepages on the World Wide Web*. Thesis, Sloan School of Management, Massachusetts Institute of Technology.

Wang, Huei-Long; Wu, Shih-Hung; Wang, I. C.; Sung, Cheng-Lung; Hsu, W. L. und Shih, W. K. (2000): "Semantic Search on Internet Tabular Information Extraction for Answering Queries". In: *Proceedings of the 9th International Conference on Information and Knowledge Management (CIKM 2000)*. McLean, S. 243–249.

Wang, Yalin und Hu, Jianying (2002): "A Machine Learning Based Approach for Table Detection on The Web". In: *Proceedings of the 11th International World Wide Web Conference (WWW 2002)*. Honolulu.

Wardrip-Fruin, Noah (2004): "What Hypertext Is". In: *Proceedings of the fifteenth ACM conference on Hypertext & Hypermedia*. Santa Cruz, S. 126–127.

Waterworth, John A. und Chignell, Mark H. (1997): "Multimedia Interaction". In: Helander et al. (1997), S. 915–946.

Wätjen, Hans-Joachim (1998): "GERHARD – Automatisches Sammeln, Klassifizieren und Indexieren von wissenschaftlich relevanten Informationsressourcen im deutschen World Wide Web". *B. I. T. – Zeitschrift für Bibliothek, Information und Technologie* (4): S. 279–290.

Wätjen, Hans-Joachim; Diekmann, Bernd; Möller, Gerhard und Carstensen, Kai-Uwe (1998): "Bericht zum DFG-Projekt: GERHARD – German Harvest Automated Retrieval and Directory". Technischer Bericht, Bibliotheks- und Informationszentrum (BIS) der Carl von Ossietzky Universität Oldenburg. Online verfügbar: `http://www.gerhard.de`.

Watters, Carolyn und Shepherd, Michael (1997a): "The Role of Genre in the Evolution of Interfaces for the Internet". In: *11th Annual Canadian Internet Conference*. Dalhousie University, Halifax.

Watters, Carolyn R. und Shepherd, Michael A. (1997b): "The Digital Broadsheet: An Evolving Genre". In: *Proceedings of the 30th Hawaii International Conference on Systems Sciences (HICSS-30)*. Band 6, S. 22–29.

Wedeles, Lauren (1965): "Prof. Nelson Talk Analyzes P. R. I. D. E." *Vassar Miscellany News* S. 3–4. Ausgabe vom 3. Februar 1965. Online verfügbar: `http://library.vassar.edu/~mijoyce/Ted_sed.html`.

Weingarten, Rüdiger (Hrsg.) (1997a): *Sprachwandel durch Computer*. Opladen: Westdeutscher Verlag.

Weingarten, Rüdiger (1997b): "Textstrukturen in neuen Medien: Clusterung und Aggregation". In: Weingarten (1997a), S. 215–237.

Weinreich, Harald und Lamersdorf, Winfried (2000): "Concepts for improved visualization of web link attributes". *Computer Networks* 33 (1–6): S. 403–416.

Wenz, Karin (2000): "Vom Leser zum User? Hypertextmuster und ihr Einfluss auf das Leseverhalten". *Sprache und Datenverarbeitung* 24 (1): S. 23–34.

von Westarp, Falk; Ordelheide, Dieter; Stubenrath, Michael; Buxmann, Peter und König, Wolfgang (1999): "Internet-Based Corporate Reporting – Filling the Standardization Gap". In: *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32)*.

Wexelblat, Alan (1999): "History-Based Tools for Navigation". In: *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32)*.

Whitehead, Jim (2000): "As We Do Write: Hyper-terms for Hypertext". *ACM SIGWEB* 9 (2-3): S. 8–18.

Wilde, Erik (1999): *World Wide Web – Technische Grundlagen*. Berlin, Heidelberg, New York etc.: Springer.

Wilks, Yorick und Catizone, Roberta (1999): "Can We Make Information Extraction More Adaptive?" In: *Information Extraction – Towards Scalable, Adaptable Systems*, herausgegeben von Pazienza, Maria Teresa, Berlin, Heidelberg, New York etc.: Springer, Band 1714 von *Lecture Notes in Artificial Intelligence*, S. 1–16.

Williams, Sean D. (2002): "What is Webtextuality?" *Text Technology* 11 (1): S. 131–137.

Witt, Andreas (2004): "Multiple hierarchies: New aspects of an old solution". In: *Proceedings of Extreme Markup Languages 2004*. Montreal. Online verfügbar: `http://www.mulberrytech.com/Extreme/Proceedings/`.

Wolff, Christian (2004): "Systemarchitekturen – Aufbau texttechnologischer Anwendungen". In: Lobin und Lemnitzer (2004), S. 165–192.

Woodruff, Allison; Aoki, Paul M.; Gauthier, Eric Brewer Paul und Rowe, Lawrence A. (1996): "An Investigation of Documents from the World Wide Web". *Computer Networks and ISDN Systems* 28 (7-11): S. 963–980.

Woods, William A. (1975): "What's in a Link: Foundations for Semantic Networks". In: *Representation and Understanding: Studies in Cognitive Science*, herausgegeben von Bobrow, D. G. und Collins, A. M., New York: Academic Press, S. 35–82.

Yang, Guizhen; Mukherjee, Saikat; Tan, Wenfang; Ramakrishnan, I. V. und Davulcu, Hasan (2003): "On the Power of Semantic Partitioning of Web Documents". In: *Proceedings of the IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03)*, herausgegeben von Kambhampati, Subbarao und Knoblock, Craig A. Acapulco, S. 21–26.

Yang, Yiming und Chute, Christopher G. (1994): "An Example-Based Mapping Method for Text Categorization and Retrieval". *ACM Transactions on Information Systems* 12 (3): S. 252–277.

Yang, Yiming und Liu, Xin (1999): "A Re-Examination of Text Categorization Methods". In: *Proceedings of the 22nd Conference on Research and Development in Information Retrieval (SIGIR '99)*. Berkeley, S. 42–49.

Yang, Yiming und Pedersen, Jan (1997): "A Comparative Study on Feature Selection in Text Categorization". In: *Proceedings of the 14th International Conference on Machine Learning*. S. 412–420.

Yang, Yiming; Slattery, Seán und Ghani, Rayid (2002): "A Study of Approaches to Hypertext Categorization". *Journal of Intelligent Information Systems* 18 (2): S. 219–241.

Yates, Joanne und Orlikowski, Wanda J. (1992): "Genres of Organizational Communication: A Structurational Approach to Studying Communication and Media". *Academy of Management Review* 17 (2): S. 299–326.

Yates, JoAnne; Orlikowski, Wanda J. und Rennecker, Julie (1997): "Collaborative Genres for Collaboration: Genre Systems in Digital Media". In: *Proceedings of the 30th Hawaii International Conference on Systems Sciences (HICSS-30)*. Band 6, S. 50–59.

Yates, Simeon J. und Sumner, Tamara R. (1997): "Digital Genres and the New Burden of Fixity". In: *Proceedings of the 30th Hawaii International Conference on Systems Sciences (HICSS-30)*. Band 6, S. 3–12.

Yip, April (2004): "The Effects of Different Types of Site Maps on User's Performances in an Information-Searching Task". In: *Proceedings of the 13th Conference on World Wide Web (WWW-2004)*. New York, S. 368–369. Poster Track.

Yoshioka, Takeshi und Herman, George (2000): "Coordinating Information Using Genres". Online verfügbar: `http://ccs.mit.edu/papers/pdf/wp214.pdf`. Massachusetts Institute of Technology, Sloan School of Management, Center for Coordination Science: CCS WP #214 SWP #4127.

Yoshioka, Takeshi; Herman, George; Yates, JoAnne und Orlikowski, Wanda (2001): "Genre Taxonomy". *ACM Transactions on Information Systems* 19 (4): S. 431–456.

Yu, Hwanjo; Han, Jiawei und Chang, Kevin Chen-Chuan (2002): "PEBL: Positive Example Based Learning for Web Page Classification Using SVM". In: *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge discovery and data mining*. Edmonton, S. 239–248.

Zhang, Jin und Dimitroff, Alexandra (2005a): "The impact of metadata implementation on webpage visibility in search engine results (Part II)". *Information Processing and Management* 41 (3): S. 691–715.

Zhang, Jin und Dimitroff, Alexandra (2005b): "The impact of webpage content characteristics on webpage visibility in search engine results (Part I)". *Information Processing and Management* 41 (3): S. 665–690.

Zhang, Xiaoni; Keeling, Kellie B. und Pavur, Robert J. (2000): "Information Quality of Commericial Web Site Home Pages: An Explorative Analysis". In: *Proceedings of the twenty first International Conference on Information Systems*. Brisbane: Association for Information Systems, S. 164–175.

Ziegler, Arne (2001): "Zur @kronymischen Verwendung der Phraseologismen in Textsorten der Internet-Kommunikation am Beispiel der E-Mail". In: *Wer A sägt, muss auch B sägen. Phraseologie und Parömiologie*, herausgegeben von Hartmann, Dietrich und Wirrer, Jan, Hohengehren: Schneider.

Ziegler, Arne (2002): "E-Mail – Textsorte oder Kommunikationsform? Eine textlinguistische Analyse". In: Ziegler und Dürscheid (2002), S. 9–32.

Ziegler, Arne (2004): "Textstrukturen internetbasierter Kommunikation. Brauchen wir eine Medientextlinguistik?" *Osnabrücker Beiträge zur Sprachtheorie* (68): S. 159–173.

Ziegler, Arne und Dürscheid, Christa (Hrsg.) (2002): *Kommunikationsform E-Mail*, Band 7 von *Textsorten*. Tübingen: Stauffenburg.

Zimmer, Dieter E. (1997): "Text in Tüttelchen – Web-Literatur: Realität? Gerücht? Verheißung? Sackgasse?" *Die Zeit* (46): S. 61. Ausgabe vom 7. November 1997.