Untersuchungen zur Methodik und Effizienz der tastaturbasierten Eingabeverfahren verschiedener Schriftsysteme der Welt

INAUGURAL-DISSERTATION
zur Erlangung des Doktorgrades der Philosophie
des Fachbereichs 05 – Sprache, Literatur & Kultur
der Justus-Liebig-Universität Gießen

vorgelegt von:
WANG Kai

Fontanestr. 3
35606 Solms

März 2019

Dekan:
1. Gutachter: Prof. Dr. Henning Lobin
2. Gutachter: Prof. Dr. Thomas Gloning
Tag der Disputation: 13.02.2019

Erklärung zur Dissertation

Ich habe die vorgelegte Dissertation selbständig und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht.

WANG Kai （王锴）
Solms, 28.03.2019

Inhalt

Vorwort IX
0 Einleitung 1
1 Hintergrund für tastaturbasierte Eingabeverfahren verschiedener Schriftsysteme 7
1.1 Überblick von tastaturbasierten Eingabeverfahren 7
1.2 Schreiben und Textverarbeitung 10
1.2.1 Definition, Funktionalitäten und Konventionen des Schreibens 11
1.2.2 Entwicklung der Schreibtechnologie 13
1.2.3 Funktionalitäten und Arbeitsprozess der Textverarbeitung mit Computer 16
1.3 Schrift, Schriftsystem, Schriftzeichen und Schriftrichtung 19
1.3.1 Definitionen über Schrift, Schriftzeichen und Schriftsystem 19
1.3.2 Klassifikation von unterschiedlichen Schriften und Schriftsystemen 21
1.3.3 Allgemeine Entwicklung der Schriften und der allgemeine Zusammenhang zwischen grammatischem Sprachbau und Schrifttyp 26
1.3.4 Fünf größte Schriftkreise, ihre kulturelle Verankerung und Transkriptionen 28
1.3.5 Analyse der Grundeinheit verschiedener Schriftsysteme 33
1.3.6 Schriftrichtungen – Horizontale und Vertikale 36
1.4 Techniken zur Verarbeitung der Schriftzeichen 41
1.4.1 Definition von Zeichen, Glyphe und Font und der Zusammenhang zwischen den drei Begriffen 41
1.4.2 Vereinheitlichte Austauschcodierungen 46
1.4.3 Zeichen-Tasten-Repräsentation und Inputcodierung 51
1.4.4 Techniken zur typographischen Zeichenausgabe 54
1.4.5 Allgemeine Zusammenhänge zwischen Computerlinguistik, Eingabeverfahren und Textverarbeitung 56
2 Eingabeverfahren der alphabetischen Schriftsysteme 59
2.1 Eingabeverfahren des deutschen Schriftsystems 59
2.1.1 Untersuchung des deutschen Schriftsystems 60
2.1.2 Zeichen-Tasten-Repräsentation des deutschen Tastaturlayouts 64
2.1.3 Verarbeitung von Sonderbuchstaben mit diakritischen Zeichen 71
2.1.4 Untersuchung der Tastenbelegung der 26 lateinischen Grundbuchstaben verschiedener Schriftsysteme 74
2.2 Eingabeverfahren des vietnamesischen Schriftsystems 77
2.2.1 Allgemeines über die vietnamesische Sprache und ihr Schriftsystem 78
2.2.2 Zeicheninventar des vietnamesischen alphabetischen Schriftsystems 79
2.2.3 Zeichencodierung und Eingabeverfahren des vietnamesischen Schriftsystems 82
2.2.4 Eingabemöglichkeiten eines alphabetischen Schriftsystems 85
2.3 Textverarbeitung der Devanagari 87

2.3.1 Allgemeines über die alphasyllabischen Schriften und die drei verschiedenen Modelle der Zeichencodierung ... 87
2.3.2 Einführung zur Devanagari-Textverarbeitung mit Beispielswort ... 89
2.3.3 Zeicheninventar der Devanagari ... 91
2.3.4 Zeichencodierung und Zusammensetzung der Schriftzeichen in einer Silbe ... 97
2.3.5 Hindi-Tastaturlayout und die Reihenfolge der einzugebenden Zeichen ... 99
2.3.6 Fazit zur Textverarbeitung der Devanagari ... 101

2.4 Koreanische Textverarbeitung ... 101
2.4.1 Hangul – Unterschiede zu alphasyllabischen Schriften und allgemeine Übersicht ... 102
2.4.2 Simples sowie komplexes Jamo und Syllabarbildung ... 104
2.4.3 Analyse der koreanischen Textverarbeitung ... 107
2.4.4 Argumente für Silbenzeichencodierungen ... 111

2.5 Arabische Textverarbeitung ... 114
2.5.1 Zeicheninventar des arabischen Schriftsystems ... 114
2.5.2 Schwierigkeiten der arabischen Textverarbeitung und allgemeiner Eingabeprozess ... 122
2.5.3 Decodierungen und Algorithmen der Schriftrichtung ... 126

2.6 Zusammenfassung der alphabetischen Eingabeverfahren und Textverarbeitungen ... 128

3 Linguistische Perspektiven zur chinesischen Schrift ... 133

3.1 Verwendungszweck, Schrifttyp, Ursprung und Entwicklung ... 133
3.1.1 Verwendungszweck ... 133
3.1.2 Schrifttyp ... 135
3.1.3 Historische Entwicklung ... 137

3.2 Orthographie, Zeicheninventar und Codierung ... 139
3.2.1 Orthographie und Standardisierung in verschiedenen Regionen ... 140
3.2.2 Zeichengebrauchs- und Zeichencodierungsstandards in CJK-Regionen ... 142
3.2.3 Vereinheitlichte Austauschcodierung in Unicode ... 145

3.3 Graphischer Aufbau und Konstruktionsprinzipien der Schriftzeichen ... 146
3.3.1 Verschiedene Perspektiven über den Aufbau ... 147
3.3.2 Die sechs Schriften ... 148
3.3.3 Strich, Strichmerkmal, -ordnung und -zahl ... 155
3.3.4 Komponenten der Sinogramme ... 157
3.3.5 Geometrische Konstellationen ... 161

3.4 Einführung in Inputcodierungsentwurf und Phonetik des Chinesischen ... 162
3.4.1 Allgemeine Zusammenhänge zwischen Zeichenform, -sinninhalt und -aussprache ... 162
3.4.2 Grundlegende Eigenschaften der Phonetik des Chinesischen ... 165
3.4.3 Transkriptionen des chinesischen Schriftsystems ... 167

3.5 Einführung und weitere Entwicklung im japanischen und koreanischen Schriftsystem 171
3.5.1 Grundlagen der weiteren Entwicklung ... 171
3.5.2 Kontrastive Analysen von Chinesisch, Japanisch und Koreanisch ... 172

3.5.3 Methoden zur Anwendung der Sinogramme und das Schaffen neuer morphologischen Schriftzeichen 174

4 Methodik, einbezogene linguistische Erkenntnisse und computerlinguistische Anwendungen der chinesischen Eingabemethoden 177

4.1 Verschiedene Eingabemethoden der chinesischen Schrift 177

4.1.1 Überblick über die chinesischen Eingabemethoden 178

4.1.2 Vorstellung der Wubi-Eingabemethode 187

4.1.3 Vorstellung der Pinyin-Eingabemethoden: Allgemeines, Mängel, Codeformen und Entwicklungsphasen 195

4.1.4 Vor- sowie Nachteile und auf künstlicher Intelligenz basierende Funktionen der intelligenten Eingabesoftwares 202

4.2 Methoden und linguistische sowie technische Unterstützungen zur Laut-Zeichen-Konversion der intelligenten Pinyin-Eingabemethoden 208

4.2.1 Überblick über die chinesische Grammatik und Grundlagen der intelligenten satzstufigen Pinyin-Eingabemethoden 208

4.2.2 Vier Hauptmethoden zur satzstufigen Laut-Zeichen-Konversion 213

4.2.3 Elektronisches Wörterbuch und linguistische Wissensdatenbank 217

4.3 Informationsverarbeitung in der Einheit von Zeichen 223

4.3.1 Silben-Zeichen-Zusammenhang 224

4.3.2 Silbenformen und Ambiguitätsfälle von Silbensegmentation 228

4.3.3 Automatische Segmentation von Pinyin-Ketten in Voll-, Halb- und Mischform 232

4.3.4 Allgemeine sowie idiolektale Zeichen- und Worthäufigkeit 237

4.4 Informationsverarbeitung im Wort 242

4.4.1 Chinesische Wörter aus linguistischer Perspektive 242

4.4.2 Automatische Wortsegmentation 251

4.4.3 Part-of-Speech Tagging 261

4.4.4 Erkennung, Festlegung und POS-Tagging unregistrierter Wörter 271

4.4.5 Erkennung und Verarbeitung der Phrasen 280

4.5 Informationsverarbeitung im Satz 284

4.5.1 Arbeitsphasen der satzstufigen Laut-Zeichen-Konversion 285

4.5.2 Allgemeine Eigenschaften der chinesischen Syntax 290

4.5.3 Syntaktische Annotation und Akquisition sprachlicher Regeln für verschiedene Arten von Wissensdatenbanken 295

4.5.4 Anwendung der sprachlichen Regeln zur Laut-Zeichen-Konversion 302

4.6 Zusammenfassung von Methodik und Effizienz der chinesischen Eingabemethoden 305

5 Projektplan für ein multilinguales Eingabesystem 311

5.1 Anforderungsanalyse und Grundkonstruktion 311

5.1.1 Anforderungsdefinition, Anwendungsfälle und Problemanalyse 311

5.1.2 Lösungsvorschläge 314

5.1.3 Gesamte Softwarestruktur 316

5.1.4 Look and Feel der Software 317

5.2 Zielorientierte Modulentwicklung, benötigte Ressourcen und Techniken 319
5.2.1 Design für die Inputcodierung verschiedener Sprachen und das Tastaturlayout 320
5.2.2 Definition von Eingabeniveau und -fall sowie ihre Verarbeitungsprozesse 323
5.2.3 Benötigte sprachliche Ressourcen 330
5.2.4 Techniken zur Kandidatenausgabe und -auswahl 333
5.2.5 Verarbeitung der Sonderzeichen 338
5.3 Weitere Probleme der multilingualen Software 344
5.3.1 Linguistische Theorien über Spracherkennung und Konversion 345
5.3.2 Analysen über die Effizienz der Software 352
5.3.3 Verbesserungsvorschläge 354
6 Fazit, Schlussfolgerungen und Ausblick 361
7 Anhang 369
7.1 Parallelindex der Fachbegriffe in Deutsch, Englisch und Chinesisch 369
7.2 Verzeichnis der Abbildungen 376
7.3 Verzeichnis der Tabellen 379
7.4 Literaturliste 381
I Normen und Werkzeugbücher 381
II. Literatur in Deutsch und Englisch 384
III: Literatur in Chinesisch 391

Vorwort

Als eine am Ende der 1980er Jahre geborene Chinesin habe ich die rasante Entwicklung der chinesischen Eingabemethoden direkt miterlebt: von der seltenen Anwendung in Multimedia-PCs bis zur heutigen Verbreitung in PCs, Tablet-PC sowie Smartphones; von den vielen verschiedenen Inputcodierungen bis zur dominierenden Rolle der phonetischen Variante; von der niedrigen bis zur hohen künstlichen Intelligenz der Eingabemethoden. Mein Interesse für die Erforschung der Eingabemethoden begann, als ich mit acht Jahren zum ersten Mal das Schreiben mit dem Computer kennen gelernt habe. In jener Zeit träumte ich davon, zur Erfindung einer effizienten chinesische Eingabemethode beitragen zu können. Später, als die intelligenten Pinyin-Eingabemethoden das hohe Verarbeitungstempo der Eingabe von westlichen Sprachen erreicht hatten, wünschte ich, dass sie im Sinne des Schutzes der schriftlichen Kultur verbessert werden könnten. Seitdem ich mit dem computerlinguistischen Masterstudium in Deutschland angefangen habe, versuche ich mit meinen erworbenen linguistischen sowie informatischen Kenntnissen verschiedene Eingabevarianten unterschiedlicher Schriften sowie Schriftsysteme kontrastiv zu analysieren. Von der Masterthesis bis zu dieser Dissertation vertiefte ich meine Forschungen in diesem Bereich immer weiter.

2013 entschied ich mich, dieses Thema zur Grundlage und Ausgangsfrage meiner Dissertationsarbeit zu wählen. Die größten Forschungsschwierigkeiten ergaben sich aus der Vielfältigkeit an Eingabemöglichkeiten und der Komplexität verschiedener Schriftsysteme (verschrifteter Sprachen). An dieser Stelle möchte ich mich bei meinem Doktorvater Herrn Prof. Lobin bedanken. Er hat mit seinen professionellen Fachkenntnissen, seinem Forschungsgeist und einer großen Begeisterung für den computerlinguistischen Zweig des Fernen Ostens mein Forschungsvorhaben unterstützt.

Das Praktikum im Gutenberg-Museum in Mainz 2014 hat mich weiter von der Kraft der Schrift überzeugt und dazu beigetragen, diese Dissertation zu verfolgen. Ich bin sehr dankbar, bei dieser Chance kulturelle Einsichten gewonnen und meine Forschungsmethodik verbessert zu haben. Dank gilt auch meiner Familie, vor allem meinen Eltern, meiner Schwester und meinem Freund, die mir in schwierigen Zeiten meiner Promotion immer Vertrauen und Unterstützung geschenkt haben. Nicht zuletzt möchte ich meinem Onkel, meiner Tante und der Daqinger Stadtbibliothek danken, die mir auf unterschiedliche Weise bei der Recherche der von mir benötigten chinesischen Literatur geholfen haben.

0 Einleitung

Wie der Titel andeutet, fokussiert diese Arbeit die Erforschung der Methodik tastaturbasierter Eingabeverfahren mit dem PC und die davon bedingte Schreibeffizienz, die bei verschiedenen Schriftsystemen unterschiedlich ist. Egal, von welcher literalisierten Sprache oder von welchem Land die Rede ist, ob das englische Schriftsystem mit 26 Buchstaben zum Einsatz kommt oder das chinesische mit zehntausenden Zeichen: die schriftliche Eingabe ist eine der alltäglichsten Handlungen der modernen Menschheit. Sowohl in der Büroarbeit als auch in der Freizeit, sowohl in Computern als auch in Smartphones: die Anwendung des Eingabeverfahrens ist buchstäblich in jeder Ecke der Welt zu sehen. Eingabetechniken sind fest mit dem modernen Lebensstil verwurzelt und für zivilisierte Gesellschaften unentbehrlich.

Die bedeutende Rolle der Eingabeverfahren und der Mangel an wissenschaftlichen Arbeiten, die bisher die verschiedenen Schriftsysteme behandelten, haben mich motiviert, eine ausführliche Untersuchung in diesem Bereich durchzuführen. Meine Hauptfragen lassen sich wie folgt formulieren: Welche Unterschiede und Gemeinsamkeiten hat die tastaturbasierte Eingabemethodik verschiedener Schriftsysteme weltweit, wie effizient funktioniert sie bei der computergestützten Textverarbeitung und wie kann sie von Erkenntnissen der Computerlinguistik profitieren?

Wie erwähnt lässt sich der Hauptforschungsgegenstand unter dem Terminus *Eingabeverfahren* zusammenfassen. Da es sich dabei bis dato um keinen festgelegten Begriff handelt, definiere ich ihn innerhalb der Arbeit wie folgt:

> Als Eingabeverfahren wird das Verfahren der Eingabe schriftlicher Informationen in ein elektronisches Medium (wie Computer oder Handy) verstanden, das über ein Eingabegerät wie Tastatur, Maus oder Eingabestift erfolgt. Das Verfahren ist in der Regel der erste Prozess der elektronischen Textverarbeitung.

Der Begriff umfasst nach meiner Definition die Verfahren zur Eingabe der Buchstaben, Ziffern, Satzzeichen usw., die von der Tastaturbelegung abhängig ist (für ein alphabetisches Schriftsystem, wie das englische und deutsche Eingabeverfahren), sowie die Eingabe der Logogramme (wie die chinesischen Schriftzeichen), der Silbenzeichen (wie japanische Kana-Zeichen) und der eigenständig codierten Silbenblöcke (wie das koreanischen Hangul), die mit Unterstützung der Eingabesoftware realisiert werden müssen. Ein durch Software unterstütztes Eingabeverfahren heißt *Eingabemethode*.

Nach dem chinesischen staatlichen Standard ‚Chinese information processing – Vocabulary Part 01' wird die tastaturbasierte Eingabemethode der chinesischen Zeichencodierung (汉字编码键盘输入法) wie folgt definiert: „Verfahren der Eingabe der chinesischen Schriftzeichen in den Computer von PC-Benutzern, die mittels der Zeichencodierung sowie der Computerressourcen und via PC-Tastatur abläuft" (GB 12200.1-90: Kap. 4.1.4.4 [Übersetzung der Verfasserin]). Das eingegebene Objekt ist der Zeichencode, der als Inputcode bezeichnet wird. Die Sammlung von Inputcodes heißt Inputcodierung oder Eingabeschema. Auf Basis der Terminologie der Eingabemethode für die chinesische Schrift lässt sich eine allgemeine Definition der tastaturbasierten Eingabemethode ableiten:

> Eine tastaturbasierte Eingabemethode ist die Methode der Eingabe von segmentalen schriftlichen Symbolen, die mit dem funktionalen Tastaturlayout nicht via entsprechender Tasten direkt zu erzeugen sind und mithilfe von Zeichencodierung und Eingabesoftware vom System abgerufen werden.

Aufgrund dieser zwei Eingabeverfahrenskategorien erfolgt die nachfolgende Erforschung getrennt zwischen alphabetischen und nicht-alphabetischen Schriftsystemen, wobei erstere auf Tasten-Buchstaben-Repräsentationen basieren und zweitgenannte exemplarisch anhand des chinesischen Schriftsystems analysiert werden. Insgesamt strukturiert sich die Arbeit in fünf Kapitel, in denen nacheinander vier Leitfragen behandelt werden.

- Leitfrage 1: Wie können Schriftsysteme der Welt klassifiziert werden und inwiefern unterscheiden sich die Schreibkulturen und -technologien anhand der Verschiedenheiten von Schriften und Schriftsystemen?

Kap. 1 bietet als Einführung der Arbeit einen Überblick über Forschungsaspekte von Eingabeverfahren verschiedener Schriftsysteme der Welt. Die Forschungsgegenstände der Arbeit – Eingabeverfahren, Eingabemethode, Textverarbeitung, Tastaturlayout, Schrift, Schriftsystem usw. – werden eingeführt und in Zusammenhängen erklärt. Der Hauptinhalt des Kapitels umfasst die moderne Textverarbeitung (wovon die tastaturbasierte Eingabe der erste Prozess ist), die Kategorisierung von Schriften und Schriftsystemen (was die Art des Eingabeverfahrens bedingt) und Allgemeines über die Zeichenverarbeitung (dieser Teil gilt als Basis für die Forschungen der nachfolgenden Kapitel).

- Leitfrage 2: Wie funktionieren Eingabeverfahren von verschiedenen alphabetischen Schriftsystemen?

Kap. 2 konzentriert sich auf die Eingabe der alphabetischen Schriftsysteme, die hauptsächlich auf Zeichen-Tasten-Repräsentation des nationalen Tastaturlayouts basieren. Anhand der ver-

schiedenen Unterarten von alphabetischen Schriften – voll-, konsonantenalphabetische, alphasyllabische und alphabetosyllabische – werden fünf exemplarische Schriftsysteme erforscht. Die Forschungsreihenfolge beginnt mit simpleren Schriftsystemen (z.B. das Deutsche) und endet mit komplizierteren (das Hindi-, koreanische und arabische Schriftsystem).

- Leitfrage 3: Worin ist die chinesische Schrift besonders, auf welche Arten kann die chinesische Eingabe realisiert werden und welche Wirkungen und Nachteile bringen Methoden der künstlichen Intelligenz für Eingabetechniken mit sich?

Kap. 3 und Kap. 4 handeln von der chinesischen Schrift im Allgemeinen (Kap. 3) und den Eingabemethoden des Chinesischen (Kap. 4) im Besonderen. Wegen des logographischen Schrifttypus und der begrenzten Größe des Zeicheninventars musste die Eingabe der chinesischen Schrift in drastisch anderer Art und Weise realisiert werden, als die des alphabetischen Schriftsystems. Einerseits können für die chinesische Schrift hunderte verschiedene Inputcodierungen entworfen und für Eingabemethoden eingesetzt werden, ohne dass eine einwandfrei funktionieren könnte. Andererseits sind viele Technologien für computergestützte Inputcode-Zeichen-Konversionen erforderlich, die mithilfe von künstlicher Intelligenz realisiert werden.

- Leitfrage 4: Wie kann eine multilinguale Eingabesoftware funktionieren?

Kap. 1 bis 4 stellen den theoretischen Teil der Arbeit dar, in dem das vorhandene Wissen mithilfe von verschiedenen Forschungsmethoden interpretiert, analysiert und dargelegt wird. Auf ihrer Basis folgt der praktische Teil (Kap. 5), in dem der Entwurf einer multilingualen Eingabesoftware erstellt wird.

Im theoretischen Teil finden vor allem vier Forschungsmethoden Verwendung: Klassifizierung, kontrastive Analysen, Abduktion und Induktion.

Wegen der Kompliziertheit von Schriften, Schriftsystemen und Eingabemöglichkeiten müssen die Forschungsobjekte nach bestimmten Aspekten klassifiziert werden, damit die Forschung generalisierter und zielorientierter wird. In dieser Arbeit ist eine **Kategorisierung** vor allem in vier Fällen notwendig. So lassen sich Schriften zunächst nach ihrer Hauptfunktionalität in drei Schrifttypen (Logographie, Syllabar und Alphabet) gliedern. Alphabete werden weiter in verschiedene Untertypen kategorisiert. Anhand der divergierenden Schriftgruppen, zugehörigen Kulturkreise und Schriftrichtungen werden die Schriftsysteme eingeteilt und nach ihrer Komplexität eingestuft. Der dritte Fall der Klassifikation betrifft die Schriftzeichen innerhalb eines Schriftsystems, welche bei der Zeichenverarbeitung auf verschiedene Art und Weise behandelt werden müssen, wie <ä> und <á> bei der Eingabe mit dem deutschen Tastaturlayout. Zuletzt geht es um die Einteilung der Eingabeverfahren eines einzelnen Schriftsys-

tems. Für manche Schriftsysteme – besonders jene, die nicht auf dem lateinischen Alphabet basieren – werden verschiedene Eingabeverfahren entworfen. In diesem Fall müssen solche Eingabeverfahren nach Faktoren wie Grundarbeitsprinzipien, Inputcodierungsart, eingesetztem Tastaturlayout usw. eingeordnet werden.

Die **kontrastiven Analysen** werden bei der Erforschung von verschiedenen Schriften sowie Sprachen und den für auf dasselbe Schriftsystem orientierten Eingabeverfahren angewendet. Der Vergleich bezieht sich zuerst auf Schriften und Sprachen mit entgegengesetzten Eigenschaften, wie die chinesische Schrift zum lateinischen Alphabet oder Chinesisch zu indogermanischen Sprachen. Durch Gegensätze werden die Grundeigenschaften von Sprachen und Schriften aus der östlichen und westlichen Welt offenkundig. Verglichen werden auch Schriftsysteme, die in derselben oder einer ähnlichen Schrift dargestellt werden, wie Englisch, Deutsch, Französisch und Spanisch (in dem lateinischen Alphabet) sowie Hindi und Hangul (beides Hybride von Syllabar und Alphabet). Anhand der Gemeinsamkeiten können so allgemeingültige Eingabemöglichkeiten festgestellt werden, während die Eingabetechniken einzelner Sprachen auf Basis ihrer Unterschiede national orientiert verbessert werden können. Durch den Vergleich der Eingabeverfahren desselben Schriftsystems können Effizienz, Umsetzbarkeit, Zugänglichkeit usw. manifester betrachtet werden.

Für die Erforschung von Eingabeverfahren eines bestimmten Schriftsystems wird die logische Forschungsmethode der **Abduktion** gebraucht. D.h. die Eingabeprinzipien werden von den linguistischen Theorien über das Schriftsystem und den realistischen Anwendungssituationen des allgemeinen Eingabeverfahrens geschlussfolgert.

Methodik und Effizienz der Eingabeverfahren der meisten Schriftsysteme der Welt, die wegen der Umfangsbegrenzung keine Erwähnung finden, können **induktiv** erschlossen werden: Anhand der Klassifizierung der Schriften und Schriftsysteme kann das Eingabeverfahren eines Schriftsystems durch die Referenz eines vergleichbaren Schriftsystems deriviert werden, bspw. die Eingabe der meisten europäischen Sprachen von dem englischen sowie deutschen Eingabeverfahren, die Eingabe der alphasyllabischen Schriftsystemen von dem Hindi-Eingabeverfahren oder die japanischen von den chinesischen Eingabemethoden.

Thematisch bedingt werden Handbücher und Bibliographien in vielen verschiedenen Sprachen verwendet, vor allem Englisch, Deutsch und Chinesisch. Bei den meisten Fachbegriffen ist die zusätzliche Angabe des englischen und chinesischen Äquivalents obligatorisch. Dabei gilt es zu beachten, dass es für manche Fachbegriffe bisher noch keine festgelegten Entsprechungen im Deutschen gibt, weshalb ich sie selbst aus dem Englischen oder Chinesischen übersetzt habe. Um resultierende Probleme dieser Sprachunterschiede zu vermeiden,

wird im Anhang ein Parallelindex zur Verfügung gestellt, der die betroffenen Fachbegriffe der drei Sprachen kontrastiv gegenüberstellt (vgl. Kapitel 7.1).

Ein chinesisches Sprichwort lautet übersetzt ins Deutsche: „Das Meer beinhaltet hunderte Flüsse und seine Größe ist von der Kapazität bedingt“ (海纳百川，有容乃大). Dasselbe gilt auch für den Bereich der Eingabetechniken, deren durch verschiedene Schriftkulturen ausgelösten tausenden Möglichkeiten in einem großen Umfang vereint betrachtet werden können. In dieser Arbeit versuche ich einen ‚Wassertropfen‘ des ‚zusammengeflossenen Meeres‘ zu analysieren, um einen Beitrag zum Austausch zwischen Ost und West im computerlinguistischen Bereich zu leisten.

1 Hintergrund für tastaturbasierte Eingabeverfahren verschiedener Schriftsysteme

Wie in der Einleitung erwähnt wurde, bietet Kapitel 1 einen Überblick über Eingabeverfahren sowie das Schreiben verschiedener Schriftsysteme per Computer. Es gliedert sich in vier Hauptteile: Kap. 1.1 versteht sich als Einführung, in der die allgemeinen Zusammenhänge zwischen Eingabeverfahren, Textverarbeitung und Schriftsystemen vorgestellt werden. Darauf basierend wird das computergestützte Schreiben in drei Punkten erweitert: Schreibtechnologie (Kap. 1.2), schriftlinguistische Grundtheorien (Kap. 1.3) und Grundlagen zur technischen Realisierung der Textverarbeitung (Kap. 1.4).

1.1 Überblick von tastaturbasierten Eingabeverfahren

Was bedeutet ein tastaturbasiertes Eingabeverfahren und wie funktioniert es je nach den unterschiedlichen Sprachen bzw. Schriftsystemen? Diese Hauptfrage der Dissertation wird in diesem Kapitel im Allgemeinen betrachtet.

Definitionen von Eingabeverfahren sowie -methoden wurden in der Einleitung bereits gegeben. Beim Schreiben mit Computern müssen sie immer zu Textverarbeitung integriert eingeführt werden. Textverarbeitung wird als „das Erstellen und die Manipulation von Textdokumenten mit einem Textprogramm“ definiert (Greulich 2003: 894). Anders formuliert übernimmt das Eingabeverfahren bei der Textverarbeitung die Aufgaben zur Eingabe von den zu schreibenden Zeichen.

Wie in der Einleitung dargestellt, unterscheiden sich Eingabeverfahren in zwei Arten: die von Tastaturbelegung abhängigen Eingabeverfahren und die von Software unterstützten Eingabemethoden. Wenn Zeichen eines Schriftsystems mit dem Eintippen einer Taste – inkl. dem Fall mit Hilfe von Umschalt- oder toten Tasten[1] – erzeugt werden können, zählt das Eingabeverfahren zum ersten Fall. Tastaturbelegung (Tastaturlayout) meint die „Zuordnung der Tasten einer Tastatur zu einzelnen Zeichen oder Befehlen“ (ibid.: 879). Die Unterschiede bei Tastaturlayouts für verschiedene Schriftsysteme liegen bei den Tasten des alphanumerischen Tastenblocks (Schreibmaschinenfeld). Um Texte in jeder literarisierten Sprache effektiv schreiben zu können, ist das Layout der Tastatur entscheidend. Die aktuelle internationale Norm zur Belegung des alphanumerischen Tastensystems ISO 9995-3 wird in Abb. 1-1 angegeben:

[1] Tote Taste kann nicht alleine Zeichen erzeugen, sondern es muss nach ihr eine weitere Taste gedrückt wird, um ein mit Diakritik gebildetem Schriftzeichen einzugeben.

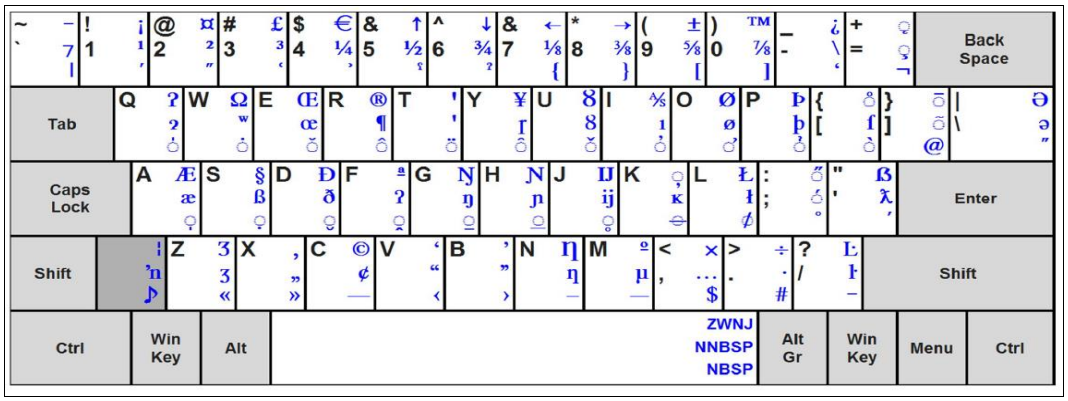


Abb. 1-1: ISO 9995-3: 2010, die internationale Norm für Tastaturbelegung (Pentzlin 2010: 1)[2]

Der Einsatz eines Tastaturlayouts mit 47 oder 48 zeichenrepräsentierten Tasten (exklusive des Leerzeichens, je nachdem ob es die Taste zwischen linkem ‚Shift' und ‚Z' gibt) bietet keine Lösung für nicht im lateinischen Alphabet geschriebenen Schriftsysteme. Für ein aus begrenzten Schriftzeichen bestehendes alphabetisches Schriftsystem wird meistens ein nationales Tastaturlayout entworfen. Für ein aus tausenden Schriftzeichen zusammengesetztes Schriftsystem aber müssen Eingabemethoden entwickelt werden. Diese können nach der Verschiedenheit der Schriften in drei Kategorien zusammengefasst werden:

a) Für alphabetische Schriftsysteme mit einem unangepassten Tastaturlayout. Normalerweise steht für einen PC eine bestimmte mechanische Tastatur zur Verfügung, mit der nur die Symbole einer bestimmten Schrift direkt eingetippt werden können. Bspw. sind im Prinzip nur die lateinalphabetischen Schriftsysteme für das Layout in Abb. 1-1 geeignet. Um ein Schriftsystem trotz der unangepassten mechanischen Tastatur eingeben zu können, wird ein auf Transkription basierendes Eingabeverfahren gebraucht. Dazu zählen bspw. die Eingabemethoden von ‚Russian Cyrillic-Latin conversion' und ‚Arabic-Latin conversion'.[3]

b) Für die koreanische Schrift Hangul (ein aus alphabetischen Komponenten bestehendes Syllabar). Hangul wird von Dürscheid als „alphabetosyllabische Schrift" definiert (Dürscheid 2006: 92). Die Buchstaben, die so genannten Jamo, werden wie der Aufbau der chinesischen Schriftzeichen in quadratischen Silbenblöcken dargestellt. Anders als bei den indischen Schriften werden nicht nur die Buchstaben sondern auch die zehntausenden Silbenblöcke bei der Zeichencodierung als Zeichen codiert. Deswegen sind die Silbenzeichen die kleinsten Einheiten der koreanischen Texte, wobei auf der Tastatur die Buchstaben durch

[2] Die schwarz dargestellten Zeichen entsprechen dem international verbreiteten US-amerikanischen Tastaturlayout, während die blau dargestellten Zeichen zu der üblichen Sekundärgruppe des multilingualen Tastaturlayouts gehören; die Gesamtzahl der belegten Zeichen ist 241 (inkl. Leerzeichen, 48×5+1).

[3] In Webseite von ‚Google Translate' stehen solche Eingabemethoden zur Verfügung.

Tasten verkörpert werden. Aus diesem Grund muss eine Software eingeführt werden, um Silbenblöcke anhand der eingegebenen Buchstaben abzurufen.

c) Für Schriftsysteme mit hunderten oder tausenden syllabischen oder morphologischen Schriftzeichen. Dieser Fall bezieht sich zumeist auf das chinesische und das japanische Schriftsystem. Für die Eingabe müssen Inputcodierungen schematisiert, spezielle Eingabesoftwares mit vielen Funktionen entwickelt und meistens auch manuelle Auswahlen der richtigen Varianten unter verschiedenen Kandidaten durchgeführt werden. Die Inputcodierung für ein chinesisches Schriftzeichen wird entweder nach der Aussprache (wie Pinyin für das Standardchinesische), nach dem Zeichenaufbau oder einem Hybrid der beiden entworfen. Die Eingabeschemata für japanische Eingabemethoden basieren meistens auf 48 Hiragana-Grundzeichen oder Romanji.[4] Die Tastaturlayouts für Chinesisch und Japanisch sind abhängig vom verwendeten Eingabeschema und vielfältig.

Unter der Voraussetzung, dass ein passendes Tastaturlayout eingesetzt und keine Eingabesoftware eingeführt wird, passiert folgendes, wenn eine zeichenrepräsentierte Taste gedrückt wird: Ein Scan-Code wird erzeugt und weitergeleitet. Er wird dann vom PCI-Controller in den entsprechenden Binärcode übersetzt und an die zentrale Recheneinheit (Central Processing Unit, Abk.: CPU) des verbundenen Computers geschickt. Die Recheneinheit verarbeitet nun den Binärcode weiter, ruft die entsprechende Glyphe sowie den entsprechenden Font ab und gibt sie auf Bildschirm aus (vgl. Heilmann 2012: 1f, Gumm/Sommer 2013: 40). Da Binärcodes die einzigen von Computern erkennbaren Signale sind, ist die systematische Festlegung der menschlichen Schriftzeichen in Bitfolge die wesentliche Grundlage für Textverarbeitungen mit Computern (vgl. Lender/Willée 1986: 59). Ein Scan-Code ist ein „Code, der beim Betätigen einer Taste von der Tastatur an den Computer gesendet wird“ (Greulich 2003: 783). Er kann bei verschiedenen Computersystemen und verschiedenen Tastaturaufbauten unterschiedlich sein. Um die Zeichen-Tasten-Repräsentation verschiedener Schriftsysteme eindeutig zu vergleichen, führe ich nach ISO 9995-3 folgende Belegung von Scan-Codes einer Tastatur (Abb. 1-2) als Normung meiner Dissertation durch. Zugleich werden auch die zuständigen Finger nach dem so genannten Zehnfingersystem eingezeichnet.

[4] Hiragana ist eine Silbenschrift des japanischen Schriftsystems, die primär für die Darstellung nativer Wörter verwendet wird. Romanji bezeichnet die Transkription des Japanischen im lateinischen Alphabet nach dem Nationalstandard.

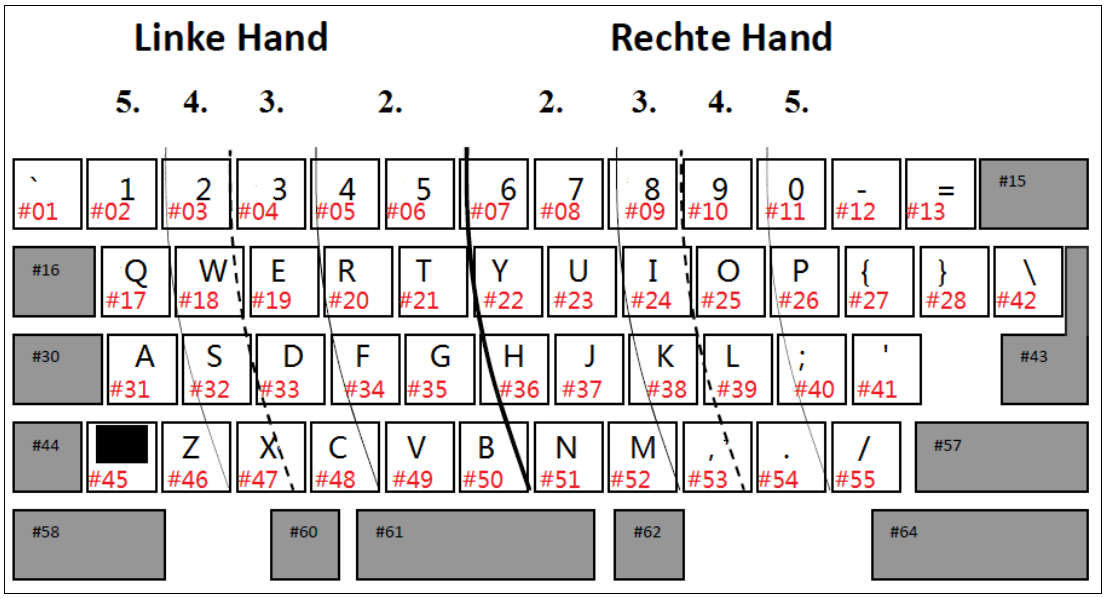


Abb. 1-2: Der vereinheitlichte Scan-Code dieser Dissertation (von dem alphanumerischen Block der PC-Tastatur mit 102 Tasten) und die Tastenkontrolle nach Zehnfingersystem[5]

Zusammenfassend kann der Prozess der Eingabe eines Buchstaben wie folgt erläutert werden: 1) Eintippen der repräsentierten Taste (inkl. Optionen mit Umschalt- oder toter Taste), 2) Codierung in Scan-Code, 3) Codierung in Binärcode, 4) Verarbeitung und Decodierung, 5) Abruf und Ausgabe von Zeichenglyphe sowie -font und Speicherung in Binärcode. Wenn die Schriftzeichen via Inputcodes abgefragt werden müssen, werden drei Prozesse mehr beim Eingabeprozess fortlaufend durchgeführt. Ein Beispiel mit einer Pinyin-Eingabemethode für Chinesisch könnte daher so aussehen: Das Zeichen 王 (IPA: [uáŋ], Pinyin: /wáng/, *König*) muss zuerst von dem PC-Benutzer in Pinyin ‚WANG' decodiert und eingegeben werden. Nachdem das System die vier eingetippten Buchstaben als ein Zeicheninputcode erkannt, alle im Standardchinesisch als /wang/ ausgesprochenen homophonetischen Schriftzeichen abgerufen und in einer Wahlliste auf dem Bildschirm angezeigt hat, muss das richtige Zeichen von dem PC-Benutzer ausgewählt werden. Danach wird das gewünschte Zeichen in Form des internen Codes als Text hinzugefügt und gespeichert, an deren Stelle das entsprechende Font auf dem Bildschirm angezeigt und ausgedruckt werden kann.

1.2 Schreiben und Textverarbeitung

Gemäß der angegebenen Definition kann Textverarbeitung als modernes, hochtechnologisiertes Schreiben mit dem Computer betrachtet werden (vgl. hierzu S. 7). Aus einer Makroper-

[5] Nach Barcodemann; Die Finger zwei bis fünf stehen jeweils für Zeige-, Mittel-, Ring- und kleinen Finger. Scan-Code wird in roter Schrift mit ‚#' angegeben. Das entsprechende Zeichen beim US-amerikanischen Layout wird schwarz in Majuskel (für Buchstaben) sowie ohne Berücksichtigung von Umschaltungen (bei sonstigen Zeichens) dargestellt.

spektive kann das Schreiben als eines der wichtigsten Merkmale der menschlichen Zivilisation angesehen werden, dessen Technologie mit der gesellschaftlichen Entwicklung fest verbunden ist. Aus Mikrosicht ist es eines der häufigsten menschlichen Verhaltensweisen, dessen technische Fortschritte eine entscheidende Rolle für die Erhöhung der Arbeitseffizienz spielen. Kap. 1.2 zielt darauf ab, die Bedeutung des Schreibens und seine technologische Entwicklung zu erforschen. Als das heutzutage am häufigsten gebrauchte Schreibinstrument ist die Computertextverarbeitung der Untersuchungsschwerpunkt.

1.2.1 Definition, Funktionalitäten und Konventionen des Schreibens

Schreiben umfasst im Allgemeinen schriftliche Äußerungen, die durch Zusammensetzung von Schriftzeichen funktionieren (vgl. Günter/Ludwig 1994: XI). Dürscheid definiert Schreiben als „Prozess des schriftlichen Fixierens von Äußerungen“ (Dürscheid 2006: 19). Der Begriff kann nicht isoliert von den Termini *Schrift*, *Schriftzeichen*, *Text* usw. betrachtet werden. Verknappt gesagt ist Schrift das Mittel zur Aufzeichnung der Sprache und die Schriftzeichen sind die kleinsten segmentalen Symbole, aus denen die Schrift zusammengesetzt ist (vgl. Ehlich 1994: 18, Dürscheid 2006: 19). Ein Text ist demzufolge das Produkt des Schreibens. Diese Begriffe werden in Kapitel 1.3.1 erweitert und verfeinert. Nach Plato ist das Schreiben selbst eine Technologie, denn im Vergleich zum Sprechen wird das Schreiben vollkommen künstlich errungen. Anders als das Sprechen erfordert das Schreiben bestimmte Werkzeuge, Ausrüstungen und bestimmte Regeln für Schreibkonventionen (vgl. Ong 1987: 84f). Die Werkzeuge und Ausrüstungen zum Schreiben sind bspw. Stifte und Papier sowie Rechner mit Tastatur, Monitor und Drucker. Die Regeln für das Schreiben sind von einzelnen Sprachen abhängig und umfassen vor allem Wortschatz, Grammatik, Sprachgebrauch und Interpunktionssetzung.

Coulmas hat fünf grundlegende Funktionen des Schreibens aufgelistet: Erinnerung, Distanz, Verdinglichung, soziale Kontrolle und Ästhetik (vgl. Coulmas 1996a: 158ff). Das Schreiben schützt ein Individuum oder die Menschheit insgesamt zunächst vorm Verlust von Informationen und Wissen. Auf individueller Ebene kann es die Einschränkungen des Gedächtnisses überwinden und ein Wissender kann sich besser mit der Produktion neuen Wissens beschäftigen. Auf der Ebene der Gesellschaft wird Wissen niedergelegt, überliefert und die Menschheit kann Generation für Generation Wissen und Kultur kontinuierlich entwickeln. Schreiben kann trotz räumlicher und zeitlicher Distanz funktionieren. Die auf Papier oder in elektronischen Daten übertragenen schriftlichen Informationen können belastbar transportiert und durch Lesen neu entschlüsselt werden. Verdinglichung bedeutet, dass die schriftliche Äußerung als Objekt untersucht und interpretiert werden kann. Mit dieser Funktionalität kann

bspw. die linguistische Forschung einer Sprache auf der Basis von Texten durchgesetzt werden. Soziale Kontrolle heißt, dass Regeln und Gesetze einer Gesellschaft schriftlich fixiert werden, damit sie dauerhafte und stabile Gültigkeit bei den Bürgern erlangen. Ästhetik bezieht sich auf die künstlerischen Aspekte des Schreibens. Schrift als ein sichtbares Objekt der Informationsmitteilung bewirkt beim Lesen auch ästhetische Emotionen, sowohl als handgeschriebene Kalligraphie, als auch als typographische Druckschrift (vgl. ibid, Dürscheind 2006: 39, Ong 1987: 46 & Scheffler 1994: 228).

Textverarbeitung kann im engeren und weiteren Sinne definiert werden. Im Weiteren kann sie alle Schreibtechnologien zur Texterstellung umfassen, wie das manuelle Schreiben auf Papier, Drucktechnik und maschinelles Schreiben mittels Schreibmaschine (vgl. Xu SC 1993: 50). In den meisten Fällen wird der engere Sinn wie folgt verstanden: Textverarbeitung ist sodann die Schreibtechnologie zur Textproduktion mit Computern, die von einem Textverarbeitungsprogramm unterstützt wird (vgl. Coulmas 1996a: 552). Die elektronische Textverarbeitung zielt darauf ab, mithilfe von Computertechnologien manche geistige Schreibarbeit der Menschen zu entlasten, die Qualität und die Geschwindigkeit des Schreibens zu erhöhen und das Schreiben in verschiedenen menschlichen Schriftsystemen zu unterstützen. Da die Grundprinzipien der Schrift und der Schreibkonventionen bei verschiedenen Nationen unterschiedlich sind, mussten die einzigartigen Eigenschaften bei der Entwicklung der Textverarbeitungsprogramme berücksichtigt werden. Die Unterschiede bei Schriftzeichen werden in Kapitel 1.3 erforscht. An dieser Stelle möchte ich zunächst kurz die unterschiedlichen Schreibkonventionen vorstellen.

Zu Schreibkonventionen gehören bspw. die Schriftrichtung und die Markierung von Wort-, Satzgrenze sowie Absätzen. Heutzutage werden die meisten Schriften waagrecht von links nach rechts geschrieben. Ausnahmen davon sind bspw. die arabische und hebräische Schrift, die waagrecht linksläufig dargestellt werden, die sowohl horizontal als auch vertikal darstellbaren chinesische Schrift und die von oben nach unten und spaltenweise von links nach rechts geschriebene mongolische Schrift. Ein mehrere Sprachen unterstützendes Textprogramm muss deswegen verschiedene Schriftrichtungen je nach der Eingabesprache automatisch oder manuell anbieten.

Ein Text wird von Grundeinheiten verschiedener Stufen aufgebaut und modelliert: Buchstaben/Schriftzeichen → Wörter → Phrasen → Sätze → Texte → Skripte → schriftliche Kommunikation (vgl. Ludwig 2005: 16)[6]. Die grenzmarkierenden Symbole sind beim Schreiben zu setzen, wie der Gebrauch der Interpunktionszeichen zur Markierung der syntaktischen

[6] Der Pfeil bedeutet an dieser Stelle, dass das links stehende Objekt Bestandteile von dem rechts ist.

Einheiten. Die Interpunktionszeichen dienen „zur Kennzeichnung von Einheiten, rhetorischen wie grammatischen, die den Wörtern als Grundeinheiten übergeordnet sind“ (ibid.: 114). Früheste Interpunktionszeichen waren Spatien (Leerzeichen), die im 7. Jahrhundert in Europa auftraten, um Worte als graphische Einheiten lesbar zu machen und die Lesegeschwindigkeit effektiv zu erhöhen (vgl. ibid.: 109 & 113). In chinesischen und japanischen Texten ist das Spatium zwischen zwei Wörtern nicht erforderlich. Nach Gallmann sind Interpunktionszeichen „die diskreten graphischen Einheiten, deren Funktion das Segmentieren und/oder Klassifizieren ist“ (Gallmann 1996: 1456). Interpunktionszeichen lassen sich in Wort- und Satzzeichen unterscheiden. Wortzeichen sind die an oder in Wortformen gesetzten Zeichen wie Apostroph <’>, Bindestrich <-> und Trennstrich </>. Satzzeichen sind die bei der Segmentation oder Klassifikation der Sätze eingesetzten Interpunktionssymbole, z.B. Punkt <./。>, Komma <,/，/،>, Fragezeichen <?/¿/؟> usw. (vgl. ibid., Dürscheid 2006: 152). Formen und Regeln zur Setzung der Interpunktionszeichen sind bei verschiedenen Schriftsystemen wegen der Unterschiede bei Schrift, Schriftrichtung, Tradition und Grammatik meistens anders definiert. Bei jedem Eingabeverfahren sind die häufigen Interpunktionszeichen in der Regel mit einem Tastenanschlag direkt zu erzeugen.

Das Schreiben hat hauptsächlich zwei Zwecke: die Produktion und Reproduktion von Texten (vgl. Ludwig 1994: 49). Produktion bezieht sich auf die Textverfassung, während Reproduktion die Wiederherstellung von geschriebenen Schriftwerken meint. Verbindet man diese beiden Zwecke mit Schreibtechnologien in der Geschichte, so kann die Entwicklung des Schreibens in zwei Linien zusammengefasst werden: a) die Produktion der Texte, angefangen beim manuellen Schreiben auf früheren Schriftträgern, übers Schreiben mit Stift und Papier, bis hin zum maschinellen Schreiben mit Schreibmaschine sowie der Textverarbeitung mit Computern; b) die Reproduktion der Texte, angefangen von Abschriften per Hand, über den Buchdruck mit Holztafel, beweglichen Lettern und maschinellem Bleisatz (Typensetzmaschine), bis hin zum heute verwendeten DTP (Desktop-Publishing) (vgl. ibid.: 48-65, Ludwig 2005: 79 & 214f). Die schriftliche Kommunikation kann ebenso in drei Phasen klassifiziert werden: angefangen vom Transport mit menschlichen Kräften, über Telegraphie bis hin zum Mailen und Chatten per Internet.

1.2.2 Entwicklung der Schreibtechnologie

Nach den Entwicklungslinien der Produktion, Reproduktion und Kommunikation der Schriftstücke werde ich in diesem Kapitel vier wichtige technische Erneuerungen als Vorstufen der elektronischen Textverarbeitung vorstellen: die Verwendung des Papiers als Schreibmaterial

und Textträger, die Erfindung der Drucktechnik, die Textübermittlung mittels Telegraphie und maschinelles Schreiben via Schreibmaschine. Die vier Erfindungen beeinflussen heutzutage direkt und indirekt erheblich die moderne Textverarbeitung.

1) Papier

Für das Schreiben ist zuerst ein praktischer und günstiger Schriftträger vorausgesetzt (vgl. Ong 1987: 84). Bevor das Papier erfunden wurde, hatte die Menschheit verschiedene Schreibmaterialien verwendet, wie Steine, Tierknochen, Tontafeln, Papyrus, Bambusstreifen, Seiden, Pergamente usw. Etwa 200 v. Chr. existierte in China schon Hanfpapier. Gegen 105 n. Chr. hat CAI Lun (蔡伦, 50-121 n. Chr.) aus der späten Han-Dynastie die Erfahrung der vorhandenen Papierherstellungstechniken gesammelt und nachgeforscht, durch Experimente die Qualität entscheidend verbessert und neuen Rohstoff zur Papierherstellung entdeckt. Im 4. Jahrhundert hat das Papier in China komplett die anderen Schreibmaterialien ersetzt und wurde allmählich in anderen Ländern verbreitet (vgl. Mazal 1994: 127, Pan 2010: 19).

Als Schreibmaterial hat Papier vor allem folgenden Vorzüge: 1) Die Oberfläche des Papiers ist glatt, weiß und tintenfest, sie hat zudem eine relativ große Fläche und kann mehr schriftliche Informationen übertragen; 2) Das Papier ist leicht, zart und faltbar und passt sowohl zum Pinsel aus dem Osten als auch zum westlichen Federkiel; 3) Es kann für tausende Jahre erhalten bleiben, mit niedrigen Kosten produziert werden und der Rohstoff für Papier ist überall auf der Welt zu finden (vgl. Pan 2010: 3f). Im 12. Jahrhundert haben die Araber die Papiertechnik nach Europa gebracht. Seitdem ist das Papier der Hauptschreibstoff der literalisierten Welt (vgl. Mazal 1994: 127ff).

2) Drucktechnik

Mit der Drucktechnik wurden die ersten Bücher massenmedial verbreitet. Die Holztafeldrucktechnik ist in China spätestens im 7. Jahrhundert entstanden (vgl. Zhang SD/Pang/Zheng 2004: 169). Die Drucktechnik mit beweglichen Lettern geht frühestens auf den chinesischen Schmied BI Sheng (毕昇, 970-1051 aus der Song-Dynastie) im Jahr 1041-1048 zurück. Im Buch ‚Mengxi Bitan' (梦溪笔谈, verfasst im Jahr 1086-1093) von SHEN Kuo wurde diese Drucktechnik aufgezeichnet. Jeder Stempel, der mehrmals verwendbar ist und aus gebranntem Ton hergestellt wurde, wurde mit einem Schriftzeichen eingeprägt, mit Feuer hart gebrannt und nach der Aussprache sortiert. Vor dem Drucken wurden die benötigten Zeichen in einem Eisenrahm auf Platten angeordnet und mit einer erhitzten Mischung aus Wachs und Harz die Stempel miteinander festgebunden. Wegen der Eigenschaften des chinesischen

Schriftsystems muss jedes Schriftzeichen bis zu zwanzig Mal hergestellt werden und für jeden Druck sind mindestens zehntausende Stempel erforderlich (vgl. Shen 1086). [7]

Gegen 1450 hat Johannes Gutenberg den modernen Druck mit beweglichen Lettern erneuert. Das Prinzip war zwar ähnlich dem von Bi Sheng, aber die Techniken, vor allem die typographische Darstellung der Letter, die Herstellung der Typen und das entsprechende maschinegeschützte Druckverfahren, waren fortgeschrittener, so dass die Druckqualität und -quantität deutlich erhöht wurde. Ein anderer wichtiger Grund für den Erfolg war durch das alphabetische Schriftsystem bedingt. Die ca. dreißig Buchstaben konnten im Vergleich zu den tausenden Zeichen viel effektiver als bewegte Letter funktionieren. Diese Technik verbreitete sich rasant in Europa (vgl. Brekle 1994b: 205ff). Aufgrund der von Drucktechnik und Papier gebrachten technischen Voraussetzungen waren Wissenschaft, Wirtschaft, Kultur, Religion usw. so zu neuen Meilensteinen in der Lage.

3) Telegraphie

Mit der Erfindung der Telegraphie wurde Textvermittlung zum ersten Mal technologisiert. Sie wurde auf Basis des elektronischen Impulses und der Codierung von Schriftzeichen entwickelt (vgl. Ludwig 1994: 62f). Der erste Telegraph wurde von Samuel Morse 1837 erfunden. Im Jahr 1858 wurde das erste Seekabel zwischen Europa und Nordamerika verlegt, mit dessen Hilfe die schriftliche Übertragung die Geschwindigkeit von einem Wort pro Minute erreichte (vgl. Heilmann 2014: 19). Später entstand drahtlose Telegraphie via Funk. Das Grundprinzip für die Textübertragung der Telegraphie basiert auf dem Morsecode, also die Zeichencodierung einer Folge von ‚.' (kurzes Signal) und ‚-' (langes Signal), womit ein Text signalisiert werden kann. Auf Seite des Empfängers kann das Signal mithilfe von Aufnahmegerät oder Fernschreiber wieder in schriftlichen Text decodiert werden (vgl. Ludwig 1994: 60). Für die chinesische Schrift wird jedes der insgesamt rund 10.000 Zeichen in vier Ziffern repräsentiert, die weiterhin in Morsecode beschrieben werden (vgl. Sproat 2010: 172f).

4) Schreibmaschine

Die Schreibmaschine hat die Epoche des maschinellen Schreibens eingeleitet. Kunzmann definiert sie als „mechanisches Gerät, mit dem man mittels beweglicher Typen eine sofort lesbare Schrift schnell und sauber zu Papier bringen kann" (Kunzmann 1979: 27). Als Vorläufer des heutigen Computers mit Textprogrammen setzt sich die Schreibmaschine grundsätzlich aus einer Tastatur, einem Übertragungsmechanismus und einem Ausgabegerät zusammen. Beim Schreiben wird die Taste, die das zu schreibende Zeichen überträgt, gedrückt. Dann

[7] Aus dem Kapitel „技艺" *Technologie*: *https://so.gushiwen.org/guwen/bookv_2293.aspx* [Abruf: 2019-03-13].

wird das Signal entschlüsselt und zum Ausgabegerät übertragen. Zuletzt wird das eingegebene Zeichen mit Druckgerät auf das Papier geschlagen und der Papierträgerwaagen oder das Schreibwerk bewegt sich nach links bzw. rechts in einer Zeichenbreite, um sich auf das Schreiben des nächsten Zeichens vorzubereiten (vgl. ibid.: 27-37).

Der Grundaufbau der heutigen PC-Tastatur wurde von der Entwicklung der Schreibmaschinentastatur entlehnt. Das QWERTY-Layout im englischen Sprachraum wurde 1888 in Toronto als erste Normung für internationale Tastaturen festgelegt (vgl. Baier 1996: 1062). Es wurde von dem amerikanischen Buchdrucker Sholes 1868 auf der Basis des Zehnfingersystems (auch Blind- oder Tastschreiben) entworfen. Er hat versucht, die häufigsten Buchstaben des Englischen gleichmäßig zu verteilen und die häufig kombinierten Buchstaben räumlich zu trennen, so dass maschinelle Störungen wegen dem nachfolgenden Tippen der benachbarten Tasten seltener vorkamen (vgl. Kunzmann 1979: 29f & 60ff). Obwohl diese technische Störung im Computerzeitalter nicht mehr existiert, ist das Layout heute eine starke Schreibgewohnheit geworden, so dass alternative Layouts nicht allgemein akzeptabel sein können. Nach diesem Grundaufbau wurden Tastaturlayouts anderer Schriftsysteme entwickelt, wie die deutschen QWERTZ- und die französischen AZERTY- Anordnungen. Ähnlich verfährt das russische Layout im kyrillischen Alphabet mit „ЙЦУКЕН“ (im Latein: JCUKEN).

Die Schreibmaschine wurde so zwar in der westlichen Welt eines der wichtigsten Arbeitsmittel in Büro und Haushalt. Im Kreis der chinesischen Schriftkultur konnte sie aber kaum alltagstauglich sein. Die chinesischen Schreibmaschinen bspw., die tausende zeichentragende Typen enthalten, wurden nur in Druckereien und bei der Presse von berufsgebildeten Arbeitern verwendet (vgl. Sproat 2010: 169f, Chen Y 1955: 1-4).

1.2.3 Funktionalitäten und Arbeitsprozess der Textverarbeitung mit Computer

Die Textverarbeitung unterstützt die geistige Arbeit des Schreibens unter verschiedenen Aspekten. Sie hat nicht nur die Qualität des Schreibens in verschiedenen Schriftsystemen ausschlaggebend verbessert, sondern im weiteren Sinne auch die technischen Voraussetzungen zur Eingabe der nicht-alphabetischen Schriftsysteme gebildet (vgl. Atsuji 1994: 449f). Der Schwerpunkt dieses Kapitels liegt darin, bezüglich der Leistungsmerkmale von Rechnern den allgemeinen Arbeitsprozess der Textverarbeitung zu analysieren.

Die Textverarbeitung stellt eine der Hauptaufgaben des PCs dar. Diese Anwendung hat die Tätigkeit des Schreibens revolutioniert, wie Ong bereits früh in seinem Grundlagenwerk ‚Oralität und Literalität‘ konstatiert: „Die Schrift, der Druck, die Computertechnologie – das sind Meilensteine der Technologisierungsgeschichte des Wortes“ (Ong 1987: 83). Ein Text

kann als elektronische Datei gespeichert, auf Monitoren angezeigt, zu jeder Zeit zugegriffen und verarbeitet und mit Hilfe des Internets blitzschnell weltweit transportiert werden. Unter Berücksichtigung der Eingabeverfahren eines Schriftsystems, die entweder durch den Entwurf von Tastaturlayouts oder die Entwicklung von Eingabesoftwares verwirklicht wurden, ist die Textverarbeitung in jeder literalisierten Ecke der Welt verbreitet.

Die Textverarbeitung unterscheidet sich in folgenden Punkten von dem Schreiben mit Schreibmaschine: Mannigfaltige Zeichen können durch dieselbe Tastatur eingegeben, eine große Menge von Symbolen sowie Informationen in anderen Schriften können in einem Text verwendet und auch unterschiedliche Schriftrichtungen können unterstützt werden (vgl. Coulmas 1996a: 552). Weiterhin verbessern sich Schreibqualität und -effizienz der rechnerbasierten Textverarbeitung mithilfe von verschiedenen neuen Funktionen erheblich. Auf Ebene der Schriftproduktion können sprachliche Symbole in beliebiger Schriftart sowie -größe dargestellt werden. Formatierung und Einrichtung von Schrift, Absätzen, Seiten, Hintergrund usw. lassen sich ebenso standardisieren. Auf Ebene von Orthographie und Morphosyntax können Schreibfehler dank der Korrektursysteme beträchtlich verringert werden. Lexikalisch können Synonyme mit gespeicherten Thesauri effektiv gefunden und ersetzt werden. Viele Textoperationen (wie Ausschneiden, Kopieren, Einfügen, Recherche und Integration) beschleunigen und verbessern das digitale Schreiben entscheidend (vgl. Radvan 2013: 118f). Um die Textverarbeitung mit solchen Funktionen zu untersuchen, wird zuerst die Computerhardware betrachtet.

Die Computerhardware ist von einem Hauptgerät und verschiedenen Ein- und Ausgabegeräten zusammengesetzt. Das Hauptgerät umfasst folgende Komponenten: a) die CPU, auch Hauptprozessor und Zentraleinheit genannt, steht im Zentrum für die Steuerung des Computers und ist für die Verarbeitung der Daten (in Bitfolge) zuständig; b) verschiedene Controller, die „die Subsysteme des Computers oder Peripheriegeräte steuern“ (inklusive dem CPU gehörigen Hauptcontroller, Bus-, Memory-, Grafikcontroller usw.) (Greulich 2003: 191); c) der Arbeitsspeicher, in dem die zu verarbeitenden Daten für kurze Zeit gespeichert werden; d) die Busse, die für den Datentransport zwischen einzelnen Funktionseinheiten zuständig sind und in zwei Gruppen – interne Bussysteme (wie System-, Daten-, Adress- und Steuerbus) und Peripherie-Bussysteme (wie USB-Bus) – eingeteilt werden; e) das BIOS (basic input out system), das die Ein- und Ausgabevorgänge bewirkt; f) der Taktgeber, der den Takt für die Computerarbeiten steuert; g) die Festplatte und andere Speichermedien, die für die dauerhafte Speicherung verantwortlich sind (vgl. ibid.: 184f, Gumm/Sommer 2013: 38-50).

Auf Grundlage dieser Hardwareleistungen kann der Prozess der Textverarbeitung beschrieben werden: Ein Textverarbeitungsprogramm wird voraussichtlich von dem Betriebssystem eingeführt und die Datei für diese zu schreibenden Dokumente begründet und gestaltet. Bei der Textverarbeitung wird zuerst der Scan-Code von der gedruckten Taste in maschinenlesbare Daten (der Binärcode von dem zu schreibenden Zeichen) umgesetzt. Dann werden die zu verarbeitenden Daten im Arbeitsspeicher angelegt und von der CPU mit den entsprechenden Glyphe sowie Font verbunden. Das Zeichen wird nachfolgend nach der Anforderung von CPU und unterstützt vom Controller von der Graphikkarte auf dem Bildschirm angezeigt. So verläuft diese Informationsverarbeitung in einer Spirale, bis Wort, Satz und Text erzeugt werden. Wenn der Text bereitgestellt ist, können die Befehle für Speicherung und Drucken eingegeben werden. Der Prozessor steuert, dass die Datei dauerhaft auf der Festplatte gespeichert wird und der Druckertreiber Druckaufgaben betreibt (vgl. Greulich 2003: 184f, 358f & 735f, Gumm/Sommer 2013: 34-50). Abb. 1-3 stellt diesen Arbeitsprozess der Textverarbeitung graphisch nach:

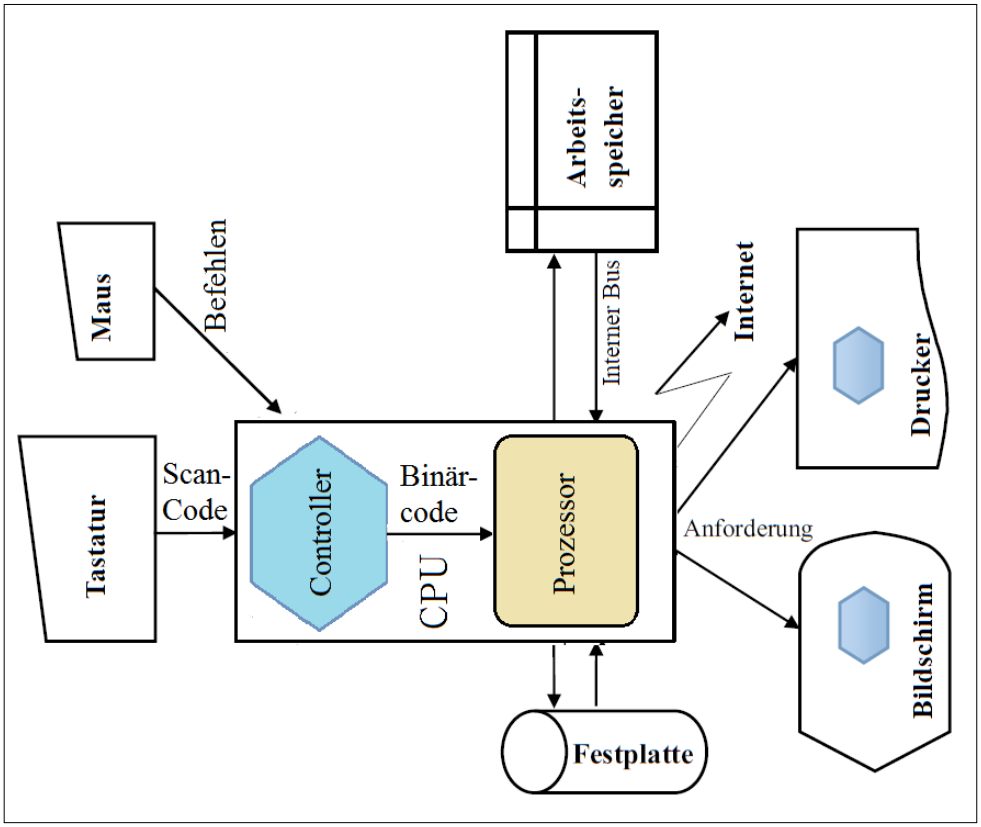


Abb. 1-3: Der Arbeitsprozess des Computers bei Textverarbeitung[8]

[8] In dieser Grafik werden die Diagramme für Hardwares und Datenfluss nach den ‚genormten Elementen eines Datenflussplans' festgelegt. Für den Controller (dazu auch die Graphikkarte und der Druckertreiber) wurde ein sechs- und für den Prozessor ein viereckiges Diagramm gezeichnet, damit sich die verschiedenen Arten der Hardware eindeutig durch die bezeichneten Diagramme unterscheiden lassen.

1.3 Schrift, Schriftsystem, Schriftzeichen und Schriftrichtung

Eingabeverfahren stellen einen der wichtigsten Prozesse der Textverarbeitung dar und werden bei fast jedem PC häufig angewendet. Da auf der Welt zahlreiche verschiedene Sprachen und Schriften existieren, gibt es dementsprechend auch unzählige Varianten von Eingabeverfahren. Für die Erforschung von Schrift, Schriftsystem und Schriftzeichen müssen Lösungsansätze für die folgenden Fragenkataloge formuliert werden: 1) wie sich die Schrift im Allgemein entwickelt hat, 2) wie die Schriftzeichen die sprachlichen Informationen wiedergeben und 3) wie die hunderten heutigen Schriftsysteme kategorisiert werden.

1.3.1 Definitionen über Schrift, Schriftzeichen und Schriftsystem

Ehlich definiert Schrift als „(historisch[es] wie systematisch[es]) Mittel zur Verdauerung des in sich flüchtigen sprachlichen Grundgeschehens, der sprachlichen Handlung“ (Ehlich 1994: 18), während Günter und Ludwig die drei Grundbedeutungen für den mehrdeutigen Begriff wie folgt zusammenfassen:

> (1) die Menge der graphischen Zeichen, mit denen die gesprochene festgehalten wird [...]
> (2) die Gestalt bzw. Form der Schriftzeichen [...]
> (3) das Produkt der Verwendung von Schriftzeichen, d.h. das Schriftstück oder der Text [...]
>
> (Günter/Ludwig 1994: IX f)

In dieser Dissertation bezieht sich Schrift in den meisten Fällen auf den ersten Aspekt von den oben genannten drei Bedeutungen. Um die Beziehungen zwischen Schrift, Schriftzeichen, Schriftsystem und Schrifttyp zu analysieren, möchte ich die terminologische Festlegung samt ihrer relevanten Begriffe nach Dürscheid einführen:

> 1. Schrift = Inventar von Schriftzeichen
> 2. Schriftzeichen = kleinste segmentale Einheit des Schriftsystems
> 3. Schriftsystem = einzelsprachabhängiges Inventar von Schriftzeichen
> 4. Schrifttyp = Gestaltungsprinzip, das einer Schrift zugeordnet liegt [...]
>
> (Dürscheid 2006: 19)

Nach den Terminologien sind Schrift und Schriftsystem systematische Sammlungen von Schriftzeichen und die Unterschiede zwischen den beiden Begriffen liegen vor allem darin, ob das Inventar von einer einzelnen Sprache abhängig ist. Für *Schrift* können die lateinische Schrift, die japanische Hiragana und die chinesische Schrift als Beispiele genannt werden. Die auf derselben Schrift basierten Systeme, die sich auf verschiedene Sprachen beziehen, sind

unterschiedliche Schriftsysteme, wie das deutsche, das englische und das französische Schriftsystem. Im Gegenteil dazu kann ein Schriftsystem auf mehreren Schriften basieren, wie das japanische Schriftsystem, das aus der chinesischen Schrift, Katakana und Hiragana zusammengesetzt ist. *Schrifttyp* lässt sich in untergliedern in den logographischen (Repräsentation der sprachlichen Einheit in Wort bzw. freies Morphem), syllabischen (in Silbe) und alphabetischen (in Phonem) Schrifttyp. Jede Schrift wird nach einem dieser drei Typen oder Hybriden zweier Schrifttypen gestaltet, wobei ein Schriftsystem sowohl überwiegend auf einer Schrift (wie das deutsche Schriftsystem) als auch parallel auf zwei oder mehreren Schriften (wie das japanische) basieren kann. Welcher Kategorie ein Schriftsystem zugeordnet werden soll, ist nach meiner Ansicht vor allem von seinen Hauptschriften und dem Schrifttyp dieser Schrift bedingt (vgl. Dürscheid 2006.: 67ff, Eisenberg 1996: 1371-1374).

Als kleinste segmentale Einheiten der Schrift sind Schriftzeichen in den meisten Fällen auch die einzugebenden Einheiten eines Eingabeverfahrens. So ist der Vergleich verschiedener Typen der Schriftzeichen ein wichtiger Aspekt für die Forschung dieser Dissertation. Im nächsten Kapitel werden die verschiedenen Schriftsysteme kategorisiert, während Kap. 1.3.5 die Grundeigenschaften der Schriftzeichen bei der Wiedergabe gesprochener Informationen vergleicht und analysiert. Nachstehend werden zunächst jedoch die international gebräuchlichen Schriftzeichen vorgestellt.

Da sich die sprachlichen Systeme aller natürlichen Sprachen auf konventionalisierte Regeln von Ausdrucks- und Inhaltsseite beziehen, können Schriftzeichen zuerst aus diesen zwei Aspekten angeführt werden (vgl. Dürscheid 2006: 65):

1) Auf der Ebene der Bedeutung kann jedes Wort oder freies Morphem mit einem eigenständigen Schriftzeichen bezeichnet werden, das Logogramm genannt wird. Die international gebrauchten Logogramme sind bspw. arabische Ziffern (wie <1>, <2>, <3>), mathematische Zeichen (wie <+>, <->, <=>), Währungszeichen (wie <$>, <€>, <£>) und manche Sonderzeichen mit Bedeutung (wie <@>, <&>, <§>). Anders als die auch mit Bedeutung verbundenen Zeichentypen wie Piktogramme (Bilderzeichen) und Ideogramme (Begriffszeichen) haben sie festgelegte sprachliche Bezeichnungen innerhalb einer Sprache. <1>, <2>, <3> etwa wird im Deutschen als <eins>, <zwei>, <drei> produziert.
2) Auf der Ebene der Aussprache kann jedes Phonem oder jede Silbe mit einem bestimmten Phonogramm dargestellt werden. Durch eine Folge von Phonogrammen wird ein Wort oder ein Morphem weiter dargestellt werden. Die 26 lateinischen Grundbuchstaben sind wegen dem kulturellen, wissenschaftlichen und wirtschaftlichen Austausch international

verbreitet worden. Nichtlateinische Schriftsysteme haben zumeist eine lateinische Transkription oder Transliteration, um internationale Kommunikation zu erleichtern.

Nach Dürscheid müssen Schriftzeichen sowohl Laut- als auch Bedeutungsseite wiedergeben, entweder eigenständig (wie internationale Zahlzeichen oder chinesisches Schriftzeichen) oder in einer Phonogrammfolge. So sind die Piktogramme und Ideogramme keine Schriftzeichen.[9] IPA und chinesisches Pinyin können anhand dieser Theorie zu keinem Schriftsystemen gezählt werden (vgl. Dürscheid 2006.: 66).

1.3.2 Klassifikation von unterschiedlichen Schriften und Schriftsystemen

Zusammenfassend lassen sich die aufgefächerten Informationen über Schrift und Schriftsystem wie folgt darlegen: 1) eine Schrift kann mit einem Schrifttyp oder dem Hybrid zweier Schrifttypen gestaltet werden; 2) ein Schriftsystem kann sowohl hauptsächlich auf einer Schrift als auch parallel auf mehreren Schriften basieren; 3) auch das alphabetische Gestaltungsprinzip kann mit unterschiedlicher Art und Weise dargestellt werden. Buchstaben können nicht nur linear nacheinander geschrieben, sondern auch gruppierend in Silbeneinheiten gezeichnet werden. Bei der Wiedergabe von Phonemen mit dem Alphabet gibt es vor allem zwei Möglichkeiten: die schriftliche Repräsentation von Konsonanten und Vokalen (volles Alphabet) und die so genannte Konsonantenschrift (partielles Alphabet). Ausgehend von diesen Thesen werden die Analysen der Schrift und Schriftsysteme in diesem Kapitel weiter vertieft. Es gibt folgende Varianten der Schriften:

A. Logographische Schrift (die auf dem logographischen Schrifttyp basierte Schrift): Eine logographische Schrift besteht in der Regel aus Schriftzeichen, die eine bestimmte Bedeutung übertragen und weder in kleinere bedeutungstragende noch in kleinere distinktive Segmente zerlegt werden können (vgl. Eisenberg 1996b: 1371). Die chinesische Schrift ist die bekannteste Logographie unter den lebenden Schriften der Welt.[10] Ein chinesisches Logogramm kann zwar in vielen Fällen beim Aufbau in Komponenten zerlegt werden, ist aber von der Bedeutung und der Aussprache her unzerlegbar. Das Zeichen <河> (/hé/, *Fluss*) bspw., das von dem bedeutungshinweisenden Radikal <氵> (abgeleitet von <水> /shuǐ/, *Wasser*) und der lauthinweisenden Komponente <可> (/kě/, *können*, *-bar*, *erlauben* etc.) aufgebaut ist, funktioniert als die schriftliche Wiedergabe eines Wortes/Morphems einer bestimmten und von einer einzelnen Sprache abhängigen Phonetik. Im Vergleich zum

[9] Die Piktogramme und Ideogramme unterscheiden sich von zwei der sechs Konstruktionsprinzipien der chinesischen Schriftzeichen (siehe Kap. 3.3.2), die dieselbe Bezeichnung haben.
[10] Präziser kann sie als morphologische Schrift definiert werden.

Phonogramm gilt ein Logogramm ohne phonetische Einschränkung. Dasselbe Zeichen wird in Japanisch mit derselben Bedeutung, aber mit der phonetischen Bezeichnung /kawa/ definiert, obwohl das chinesische und japanische Wort für *Fluss* miteinander nicht verwandt ist.[11]

B. Silbenschrift (die auf dem syllabischen Schrifttyp basierte Schrift, auch Syllabar genannt): Eine Silbenschrift basiert auf Silbenzeichen, in denen ein Zeichen eine bestimmte Silbe wiedergibt und graphisch unzerlegbar ist. Katakana und Hiragana sind typische Silbenschriften (vgl. Eisenberg 1996b: 1371). Zum Beispiel wird *Deutschland* in Japanisch als <ドイツ> /doitsu/ bezeichnet, das aus drei Katakana-Zeichen besteht: <ド> /do/ (gebildet aus dem Grundmora <ト> plus Diakritika); <イ> /i/ und <ツ> /tsu/). In einer Silbenschrift entsprechen sich Syllabar und Schriftzeichen in den meisten Fällen eins-zu-eins.[12]

C. Vollalphabet: Basis ist eine Segmentalschrift, die Buchstaben für Konsonanten und Vokale umfasst. Es handelt sich um ein Alphabet im engeren Sinne (vgl. Coulmas 1996a: 11 & Unicode Glossary: Alphabet). Außer der vollphonemischen Wiedergabe haben solche Schriften zwei weitere Hauptgemeinsamkeiten: Die Buchstaben lassen sich in Majuskel und Minuskel unterscheiden und die Schriftrichtung läuft immer horizontal rechtsläufig. Zu dieser Kategorie gehören fünf heutzutage noch verwendete Alphabete: das lateinische, kyrillische, griechische, armenische und koptische (vgl. Unicode 12.0 Chapters: Kap. 4.2: 164).[13] Die ebenso vollalphabetisch dargestellten Hangul und mongolische Schrift sind Ausnahmen und werden nicht zu dieser Kategorie gezählt (siehe Punkt F).

D. Konsonantenschrift (auch Abjad): Eine Alphabetschrift zählt zu den Konsonantenschriften, wenn die Vokale schriftlich nicht angegeben werden. Die Konsonantenschrift ist im Kontext der semitischen Sprachen entstanden. In solchen Sprachen dienen die Vokale meistens zur Angabe der grammatischen und abgeleiteten Flexionen. So können Informationen in solchen Sprachen im Zusammenhang in der Regel mit wenigen Störungen im Zusammenhang entschlüsselt werden. Buchstaben werden anhand der Position im Wort in verschiedenen Glyphen – Initial-, Medial-, Final- und isolierte Formen – dargestellt (vgl. Coulmas 1996a: 91, Bodmer 1997: 58). Die heute verwendeten Konsonantenschriften sind z.B. die linksläufige arabische und hebräische Schrift.

[11] Vgl. *https://jisho.org/search/%E6%B2%B3* [2018-01-23]; /kawa/ entspricht der Hiragana-Angabe <かわ>.
[12] Vgl. *http://jisho.org/search/%E3%83%89%E3%82%A4%E3%83%84* [2018-01-23].
[13] Unter den anderen genannten Schriften werden das glagolitische, archaisch georgische und Deseret-Alphabet heutzutage kaum benutzt. Die WarangCiti-Schrift in Indien ist eine alphasyllabische Schrift, weshalb sie nicht zu dieser Kategorie zählt.

E. Das zwischen Alphabet und Silbenschrift stehende Alphasyllabar (auch Abugida genannt): Nach Coulmas sind Alphasyllabare Schriften, in der Syllabare als Gestaltungseinheit funktionieren, aber gleichzeitig segmental zerlegt werden können (vgl. Coulmas 1996a: 483). Alle indischen Schriften sind nach diesem Prinzip gebildet, eines der bekanntesten Beispiele ist die Devanagari, die für Sanskrit, Hindi und Nepali in Indien und Nepal verwendet wird. Knapp zusammengefasst gibt es zwei Graphemarten: Grapheme für Vokale, die nur als Anfangsbuchstabe eines Wortes auftreten, und konsonantische Grapheme, die voreingestellt mit dem Vokal /a/ kombiniert sind. Wenn ein anderer Vokal /a/ ersetzen soll, wird ein abhängiges Vokalzeichen bei diesem Konsonantenzeichen beigefügt (vgl. Sen 1996: 1429f). Beispielsweise wird das Syllbar < कि > /ki/ von dem Buchstaben < क > /ka/ und dem diakritischen Zeichen < ि> /i/ zusammengesetzt.

F. Ausnahmenschriften: Hangul für Koreanisch und die mongolische Schrift[14]: Das Hangul ist eine Sonderschrift und kann nicht unumstritten einem Alphabet oder einem Syllabar zugeordnet werden. Es basiert zwar auf einem Vollalphabet, aber die Buchstaben einer Silbe müssen in einem quadratischen Silbenblock, wie die Form der chinesischen Schriftzeichen, dargestellt werden. Hangul ist deswegen mit dem Vollalphabet, der Alphasyllabar und der chinesischen Schrift verwandt. Die klassische mongolische Schrift, die heutzutage noch in der Inneren Mongolei Verwendung findet, kann ebenso wegen ihrer Besonderheiten zu keiner der fünf oben genannten Kategorien klassifiziert werden. Sie funktioniert vollalphabetisch wie das lateinische. Die Buchstabenglyphen lassen sich wie die des arabischen Alphabets anhand der Position im Wort unterscheiden. Sie muss in vertikaler Schriftrichtung geschrieben werden, d.h. sie läuft primär von oben nach unten und sekundär spaltenweise von links nach rechts (vgl. Coulmas 1996a: 343-346).

Unter der Voraussetzung, dass die international verbreiteten Schriftzeichen und die Sonderfälle eines Schriftsystems nicht berücksichtigt werden, lassen sich die Schriftsysteme grundsätzlich zwischen den auf einer Schrift basierten Schriftsystemen und den von zwei oder mehreren verschiedenen Schriften gemischten Schriftsystemen unterscheiden. Bezugnehmend auf die oben genannten fünf Typen können die Schriftsysteme so kategorisiert werden:

A. Logographisches Schriftsystem: Es bezieht sich auf ein auf der logographischen Schrift basierendes Schriftsystem. Unter den heute noch verwendeten Schriftsystemen kann nur das chinesische Schriftsystem zu dieser Kategorie gezählt werden.

[14] In der Äußeren Mongolei wird heute meist das kyrillische Alphabet zur Verschriftlichung des Mongolischen gebraucht, während die klassische mongolische Schrift in der Inneren Mongolei (VR China) beibehalten wird.

B. Syllabisches Schriftsystem: Solch ein Schriftsystem ist grundsätzlich aus einer Silbenschrift zusammengesetzt. Heutzutage sind sie kaum mehr vorhanden. Manche schon auf Linear-B-Schriften basierten aber ausgestorbenen Schriftsysteme (Nachkommen der Keilschrift) zeigen syllabische Eigenschaften (vgl. Robinson 2013: 98). Die Cherokee-Schrift für die Cherokee (eine nationale Minderheitssprache in den USA), die ca. 1820 erfunden wurde, ist eine der wenigen modernen syllabischen Schriften (vgl. Coulmas 1996a: 73).

C. Phonemisches Schriftsystem (auf Vollalphabet basierendes Schriftsystem): Die Kategorie umfasst die meisten Schriftsysteme. Alle in der lateinischen (wie das deutsche Schriftsystem), der griechischen (das griechische Schriftsystem) und der kyrillischen Schrift (wie das russische Schriftsystem) etc. übertragenen Schriftsysteme sind darin untergeordnet.

D. Konsonantenschriftsystem (auf Konsonantenschrift basierendes Schriftsystem, auch Abjad): Ein Konsonantenschriftsystem dient heutzutage meistens einer semitischen Sprache, wie dem arabischen Schriftsystem.

E. Alphasyllabisches Schriftsystem (auch Abugida): Solche Schriftsysteme können auch als indische Schriftsysteme bezeichnet werden und verbreiteten sich im Großteil Südasiens. Beispiele dafür sind das Hindi-, das laotische und das tibetische Schriftsystem (vgl. Coulmas 1996a: 227).

F. Mischschriftsystem: Ein Mischschriftsystem ist ein Schriftsystem, in dem mehrere Schriften parallel funktionieren. Das japanische und koreanische Schriftsystem sind dafür typisch. Im japanischen Schriftsystem wirken sowohl die Silbenschrift Hiragana und Katakana, als auch die sinojapanische Schrift Kanji (vgl. ibid.: 239). Heute setzt sich das südkoreanische Schriftsystem überwiegend aus Hangul und zu einem geringem Teil aus Hanja (sinokoreanischen Schriftzeichen) zusammen (vgl. ibid.: 277).[15] Auf dieser Basis habe ich die in meiner Dissertation erwähnten Schriften und Schriftsysteme kategorisiert (siehe Abb. 1-4).

[15] Kanji und Hanja sind nicht vollständig mit Hanzi (die fürs chinesische Schriftsystem nötigen Schriftzeichen) identisch. Die Japaner und Koreaner haben auch nach den Bildungsprinzipien des Sinogramms morphologische Schriftzeichen für ihre native Sprache erfunden und solche Schriftzeichen gehören ebenso zu Kanji sowie Hanja. An dieser Stelle werden deswegen die Begriffe ‚sinojapanische' und ‚sinokoreanische Schrift' statt der chinesischen Schrift angewendet.

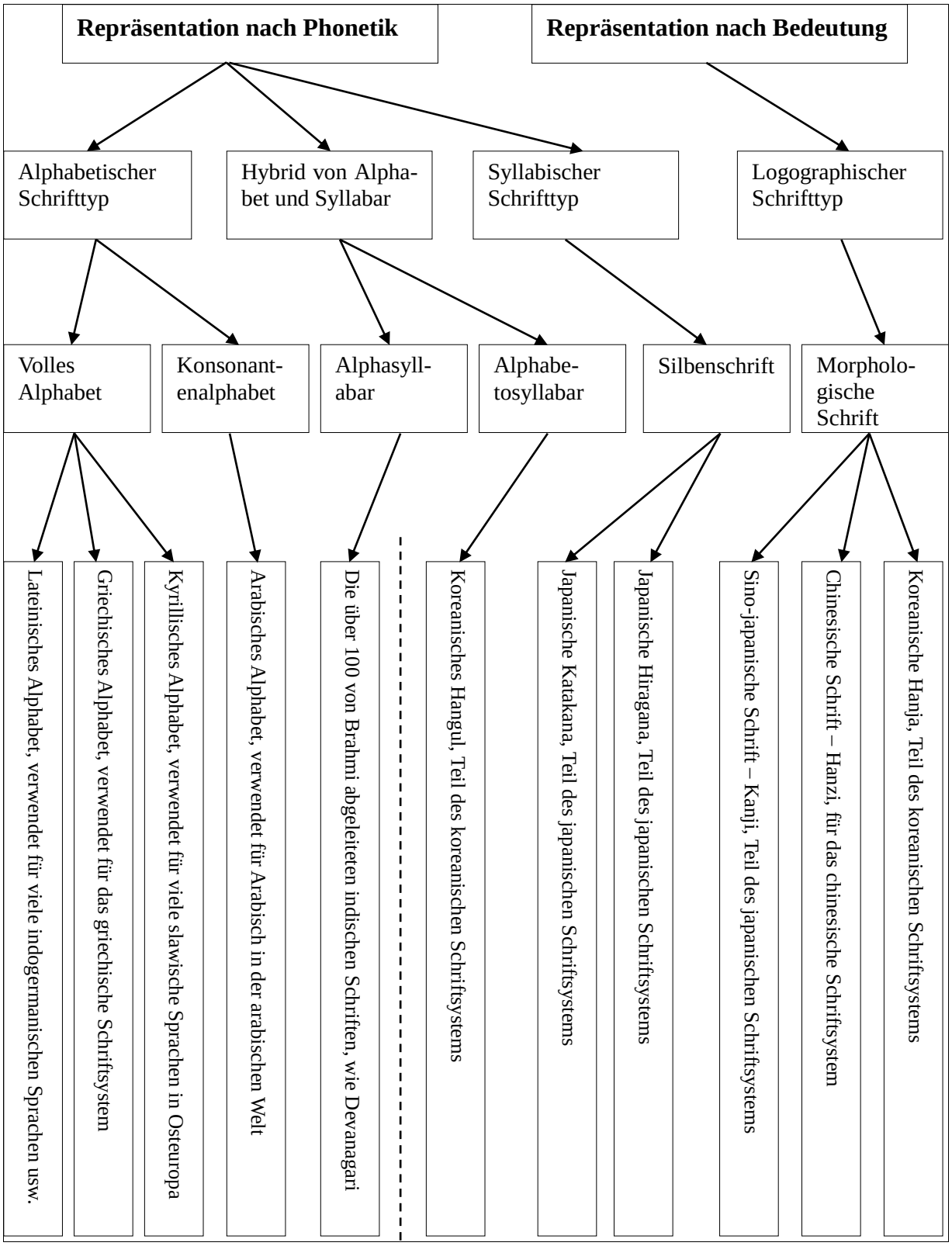


Abb. 1-4: Klassifikation der Schriften und Schriftsysteme[16]

[16] Klassifiziert anhand Coulmas 1996b: 1383 (nach Faber's 1992), Dürscheid 2006: 68f und eigenen Analysen; die gestrichelte Linie grenzt ab, welche Schriften (rechts von der Linie) im Normalfall bei der Textverarbeitung Eingabemethoden erfordern.

1.3.3 Allgemeine Entwicklung der Schriften und der allgemeine Zusammenhang zwischen grammatischem Sprachbau und Schrifttyp

Es gibt zwar hunderte verschiedene Schriften in der heutigen Welt, der Ursprung aller modernen Schriften kann aber auf die drei ältesten Schriften zurückgeführt werden: die ägyptischen Hieroglyphen, die sumerische Keilschrift und die chinesische Schrift. Durch die Analyse der Entwicklung aus den drei ältesten Schriften zieht sich ein roter Faden, der von der Vielfältigkeit der menschlichen Schriften zeugt.

Im Bereich der Sprachwissenschaft ist die evolutionistische Hierarchie der Schrift generell bekannt, wonach sich die menschliche Schrift von Piktogrammen bis hin zu mit sprachlichen Einheiten kombinierten Vollschriften entwickelt habe. Die Schriftevolution verlief demzufolge in drei Phasen: von wortsyllabischen (auch logographischen oder morphosyllabischen), über syllabischen bis zu den alphabetischen (vgl. Gelb 1963: 190f). Betrachtet man diese Theorie jedoch aus Sicht der heute noch verwendeten chinesischen Schrift, ist sie sehr diskussionswürdig. Bei der Entwicklung der Schrift können drei Linien gezeichnet werden.

Linie A: Von den ägyptischen Hieroglyphen, über die archaische Konsonantenalphabeten (wie die phönizische, protosemitische und aramäische Schrift), entwickelte sich zu Vollalphabeten (wie das griechische, lateinische, kyrillische Alphabet), modernen Konsonantenschriften (wie das arabische und hebräische Alphabet) und alphasyllabischen Schriften (wie Devanagari in Indien).

Nach Gelbs Theorie fehlt in dieser Linie ein Syllabar als Übergang zwischen Wortsyllabar und Alphabet. Denn das phönizische Alphabet wird häufiger als Alphabet definiert, im Vergleich zu der Definition als syllabische Schrift. Coulmas definiert das Alphabet als „writing system characterized by a systematic mapping relation between its signs (graphemes) and the minimal units of speech [...]“ (Coulmas 1996a: 9). Im weiteren Sinne betrifft das Alphabet sowohl die Konsonantenschrift, in der nur Konsonanten mit Graphemen für Phoneme dargestellt werden, als auch das so genannte Vollalphabet, in dem alle Phoneme, sowohl Konsonanten als auch Vokale, komplett schriftlich angegeben werden (ibid.: 11). Von dieser Entwicklungslinie ist zu ersehen, dass die Schriftevolution nicht die Phase des Syllabars durchlaufen muss.

Linie B: Von der akkadischen und hethitischen Keilschrift entwickelte sich zu der elamitischen, der hurritischen, der kypro-minoischen syllabischen Schrift etc.

Diese Entwicklungslinie brach schon beim syllabischen Schritt ab und kann diese Schriftevolutionstheorie nicht bestätigen. Linie B hat (trotz des Abbruchs) mit dem phönizischen Alpha-

bet das erste Alphabet hervorgebracht und nachfolgend die meisten heutigen alphabetischen Schriften indirekt beeinflusst.

Linie C: Von der archaischen chinesischen Schrift (ca. 14. bis 2. Jh. v. Chr.) entwickelte sich zu der modernen chinesische Schrift (seit 2 Jh. v. Chr. bis heute), nach deren Prinzipien die logographische Schriften anderer Nationen des Ostasiens (wie Kanji, Hanja und Chữ Nôm) entwickelt wurden.

Bei der Evolution der chinesischen Schrift blieb der morphologische Schrifttyp im Grunde unverändert. In dem chinesischen Schriftsystem funktioniert ein Schriftzeichen im Allgemeinen als ein Morphem mit festgelegtem Silbenwert. In dem japanischen und koreanischen Schriftsystem wurde die chinesische Schrift trotz der Erfindung von nativen Phonogrammen (Hiragana und Katakana für Japanisch und Hangul für Koreanisch) nicht abgeschafft. Die Theorie der Schriftevolution, mit der die chinesische Schrift als rückständig definiert wird, hat im letzten Jahrhundert viele Experten und Politiker motiviert, das chinesische Schriftsystem zu alphabetisieren. Solche Versuche missglückten. Heutzutage wird die These für die Rückschrittlichkeit der chinesischen Schrift von den meisten Sprachwissenschaftlern verneint. Im chinesischen Kulturkreis und Japan herrscht eher die Tendenz, dass sich die chinesische Schrift für lange Zeit fortsetzend weiterentwickeln wird.

Das Ökonomieprinzip – so Gelb – sei als treibende Kraft der Schriftgeschichte dafür verantwortlich, dass Schriften immer weniger Zeichen benötigten (vgl. Gelb 1963: 69). Allgemein angesehen kann ein alphabetisches Schriftsystem mit relativ wenig Engagement erfunden werden. Dies erfordert im Allgemeinen weniger Grundbildungskompetenzen zum Lesen und Schreiben und hat bedeutende Vorzüge bei der Drucktechnik mit Bewegungslettern und dem maschinellen Schreiben. Aber diese Evolutionstheorie kann am Beispiel der heute noch verwendeten chinesischen Schrift keine wissenschaftliche Festlegung sein. Aus neutraler Perspektive haben Alphabet und Logographie beide ihre Vorzüge. Der Typ der gezeichneten Sprache ist eines der wichtigsten Faktoren, warum die drei Linien der Schriftentwicklung verschiedene Richtungen eingegangen sind.

Die altägyptische Hieroglyphenschrift wurde ursprünglich für die Wiedergabe der ägyptischen Sprache erfunden, die grammatisch zum flektierenden Sprachbau zählte und zu der afroasiatischen Sprachfamilie gehörte (vgl. Bußmann 2002: 51). Die Wortbildung besteht aus konsonantischer Wortwurzel und Flexionsmorphem, um auf die grammatische Kategorie hinzuweisen. Für die Schreibung der Flexion waren logographische Schriftzeichen untauglich, weshalb Konsonanten tragende Phonogramme erfunden und angewendet wurden (vgl. Schenkel 1994: 289, Li BJ 1991: 56f). Bei der Verschriftlichung der phönizischen Sprache, die zur

selben Sprachfamilie wie die ägyptische Sprache gehörte und nach ähnlichen Wortbildungs- und Grammatikprinzipien funktionierte, wurden die konsonantischen Phonogramme aufgenommen und weiterentwickelt (vgl. Bußmann 2002: 510 & 596). Das Konsonantenalphabet entstand anhand der afro-asiatischen Sprachen und gilt heute immer noch für Arabisch und Hebräisch, die dieser Sprachfamilie zugehörig sind.

Bei der Entlehnung des phönizischen Alphabets in der griechischen Sprache musste eine Reformierung der vokalischen Buchstaben durchgeführt werden. Die indogermanischen Sprachen funktionieren zwar grammatisch auch mit flektierendem Sprachbau, aber die Wortwurzel besteht sowohl aus Konsonanten als auch Vokalen. Ein Vollalphabet, nämlich das griechische, das lateinische oder kyrillische Alphabet, ist deswegen für die indogermanischen am besten passend (vgl. Li BJ 1991: 60ff).

Im Gegensatz zu der ägyptischen Sprache ist Chinesisch isolierenden Sprachbaus, d.h. dass die einzelnen Wörter unverändert bleiben und die syntaktischen Beziehungen durch ‚grammatische Hilfswörter' oder die Position der Wörter im Satz ausgedrückt werden (vgl. Bußmann 2002: 321). Es ist eine wichtige Ursache dafür, dass es durch die morphologische Schrift effektiv und präzise geschrieben werden kann und sich nicht zu einer phonologischen Schrift entwickelt hat, im Vergleich zu Linie A und B (vgl. Li BJ 1991: 57f).

Bei der Entlehnung der chinesischen Schrift in die agglutierenden Sprachen Japanisch und Koreanisch war die Schreibung der Affixe problematisch. In den beiden Sprachen werden oder wurden viele Wörter mit gemischten Schriften orthographisch aufgezeichnet, d.h. die unflektierbare Wortwurzel wird mit der chinesischen Schrift und das agglutierende Affix mit Hiragana sowie Hangul geschrieben (vgl. ibid.: 60ff, Lu 2009: 19f).

Anhand der Analyse im Zusammenhang zwischen Sprach- und Schrifttyp von Linie A und C ist festzustellen, dass die Bewertung der Schrift von einer bestimmten Sprache abhängig ist. Alphabet, Syllabar und Logographie haben alle ihre Vorzüge und können für Sprachen verschiedenen Sprachbaus geeignet sein und effektiv wirken. Außer dem Entscheidungsfaktor der aufgezeichneten Sprache ist die kulturelle Verankerung ein weiterer wesentlicher Grund für die Beziehung von Schrift zu Sprache. Diese wird im nachfolgenden Kapitel beleuchtet.

1.3.4 Fünf größte Schriftkreise, ihre kulturelle Verankerung und Transkriptionen

Neben den Aspekten der Schrifttypen können die Schriften und Schriftsysteme auch anhand ihrer Schriftkultur kategorisiert werden. Die meisten heutigen Schriften können einem der fünf größten Kreise zugeordnet werden (vgl. Xu SC 1993: 1-38).

1) Der Kreis des lateinischen Alphabets umfasst vor allem Mittel- und Westeuropa, Amerika, Ozeanien, Afrika und einige Länder in Asien.

Das (auch als romanisches Alphabet bezeichnete) lateinische Alphabet ist ca. im 6. Jh. v. Chr. von der Ableitung des griechischen Alphabets entstanden und ist heutzutage die verbreitetste Schrift der Welt. Es wurde zum ersten Mal für romanische und germanische Sprachen eingesetzt. Auch manche Sprachen, die nicht zu den indogermanischen Sprachen zählen, verwenden es, etwa Finnisch (Klassifikation: Uralisch), Vietnamesisch (Austroasiatisch), Türkisch (Turksprachen) etc. (vgl. Coulmas 1996a: 285). Die weite Verbreitung der lateinischen Schrift ist von der westlichen Kirche des Christentums, der Kolonialisierung von Europa in anderen Kontinenten und der dominanten Rolle Europas und den USA bei Wirtschaft, Wissenschaft usw. geprägt. Schriftevolutionär hat sich das lateinische Alphabet beim Einsatz für zahlreiche Sprachen zu verschiedenen Varianten entwickelt. Die archaische lateinische Schrift umfasst 21Grundbuchstaben, die moderne 26. Neben den Grundbuchstaben gibt es viele mit diakritischen Zeichen und Ligaturen variierte Buchstaben, deren Anwendung sich je nach Schriftsystem unterscheiden lässt. Bspw. sind im italienischen Schriftsystem nur 21 Buchstaben eingeschlossen, im deutschen 30 inklusive vier Sonderbuchstaben, im spanischen 27 und im französischen insgesamt 42 Buchstaben (vgl. ibid.: 286). Zur Erneuerung der lateinischen Buchstaben wurden zusammengefasst vier Methoden angewendet (vgl. Bußmann 2002: 70f).

- Die erste Methode heißt Buchstabendifferenzierung, die sich daran richtet, dieselbe Lautung konsonantisch und vokalisch schriftlich zu unterscheiden. Mit dieser Methoden sind bspw. <u> (von <v>) und <j> (von <i>) entstanden.
- Die zweite Methode basiert auf dem Zusammenschreiben eines Diagraphs, z.B. <ß> (von <s> und <z>), <æ> (von <a> und <e>) und <w> (Verdopplung von <v>).
- Mit dem Einfügen diakritischer Zeichen bei einem Grundbuchstaben wurden die meisten lateinischen Schriftzeichen geschaffen. Von einzelner Sprache abhängig haben solche Schriftzeichen unterschiedliche Funktionen. In manchen Schriftsystemen werden sie als wesentliche Buchstaben für die Wiedergabe bestimmter Phoneme verwendet, wie <ä> für das deutsche Schriftsystem. Sie können in vielen Fällen auch Allographe für Grundbuchstaben darstellen, um von homophonetischen Wörter schriftlich zu divergieren, wie <à> und <â> im Französischen. In Tonsprachen sind diakritische Zeichen auch für die Tonkennzeichnung zuständig, wie im Vietnamesischen (vgl. auch Unicode 12.0 Chapters: Kap. 7.1: 290ff).

- In manchen lateinalphabetischen Schriftsystemen wurden auch Buchstaben von einer anderen Schrift übernommen. Bspw. wurde der aus dem Runenalphabet stammende Buchstabe <Þ/þ> im isländischen Schriftsystem aufgenommen, um das Phonem [θ] /th/ zu repräsentieren.

2) Der Kulturkreis des kyrillischen Alphabets findet sich in vielen Ländern Osteuropas, wo die orthodoxe Kirche vorherrscht. Es ist meistens mit slawischen Sprachen verbunden. Wie das lateinische Alphabet wurde das kyrillische Alphabet ebenso ausgehend von dem griechischen Alphabet entwickelt, weshalb die drei Alphabete viele vergleichbare Eigenschaften haben. Es gibt z.B. 33 Buchstaben in dem russischen und 30 in dem bulgarischen Schriftsystem (vgl. Coulmas 1996a: 106-109).

3) Der arabische Schriftkreis, dem über zwanzig arabische Länder und weitere muslimisch geprägte Länder angehörig sind. Das arabische Alphabet ist in der Hauptsache für die Aufzeichnung des Arabischen, Persischen und Urdu zuständig. Die augenscheinlichsten Unterschiede zwischen der arabischen und der lateinischen/kyrillischen Schrift sind, dass es sich um eine Konsonantenschrift handelt und diese von rechts nach links läuft. Jeder Buchstabe wird wortpositionsabhängig in der passenden Form – Independent, Initial, Medial sowie Final – dargestellt (vgl. ibid.: 18ff, Bodmer 1997: 181).

4) Der Kreis der indischen Schriften, die als Alphasyllabar definiert werden. Zu diesen gehören viele Länder und Regionen Südasiens, wo der südliche (Theravada) sowie tibetische Zweig des Buddhismus (Vajrayana) und Hinduismus als Hauptreligionen existieren. Die mit der indischen Schrift geschriebenen Sprachen umfassen die modernen indischen Sprachen (der indogermanischen Sprachfamilie zugehörig), Tibetisch (Sinotibetisch), Tamil (Dravidisch) etc. Es gibt zwar ca. hundert verschiedene Schriften in diesem Schriftkreis, solche Schriften haben aber denselben Stamm, nämlich die Brahmi-Schrift, die nach dem Grundprinzip des Alphasyllabars funktionieren (vgl. Coulmas 1996a: 227ff).

5) Der chinesische Schriftkreis mit der chinesischen Schrift und davon entwickelten Schriften. Zu diesem Kreis zählen die Länder und Regionen, wo die chinesischen Sprachen (Festlandchina, Taiwan, Hongkong, Macao, Malaysia und Singapur), Japanisch (Japan) und Koreanisch (heute nur Südkorea) gesprochen werden.[17] Die chinesischen Sprachen gehören zu der sinotibetischen Sprachfamilie und vom grammatischen Standpunkt zu den isolierenden Sprachen. Die japanischen und koreanischen Sprachen werden zu keiner bestimmten

[17] In Vietnam und Nordkorea wurde im 20. Jahrhundert die chinesische Schrift abgeschafft. Die beiden Länder gehören deswegen heutzutage nicht mehr zu dem chinesischen Schriftkreis.

Sprachfamilie addiert und besitzen einen agglutinierenden Sprachbau (vgl. Bodmer 1996: 180-191). Die chinesische Schrift ist nach allgemeiner Ansicht unter den heutigen Schriften am komplexesten, gleichzeitig aber nach der Bevölkerung (neben dem lateinischen Alphabet) die zweithäufigste Schrift (ca. 1,6 Milliarden)[18] auf der Welt. Der Konfuzianismus, der chinesische Zweig des Buddhismus (Mahayana) sowie Taoismus fungieren als Symbolträger dieses Schriftkreises.

Von den fünf Schriftfamilien ist herauszuziehen, dass die Eigenschaften der Sprache, die Kultur und die Religion drei wichtige Entscheidungsfaktoren für die schriftliche Repräsentation einer Sprache sind. Der politische Antrieb ist daneben auch ein wichtiger Faktor, dessen Wirk- und Durchsetzungskraft in der Realität jedoch bezweifelt werden muss. Viele Schriftreformen mit politischem Hintergrund sind gescheitert. Bspw. wurden in der Zeit der Sowjetunion in vielen Teilrepubliken Schriftreformen hin zum kyrillischen Alphabet durchgeführt. Nach ihrem Untergang haben viele Länder jedoch ihre ursprüngliche Schrift wieder eingeführt (vgl. Bergerhausen 2011: 039f). In den 1950er Jahren wurde in China der Versuch der Alphabetisierung unter Führung von Mao Zedong unternommen. Ergebnisse von dieser Reform waren die Vereinfachung der chinesischen Schriftzeichen und die Verwendung des Pinyins als alphabetische Transkription. Obwohl die chinesische Schrift vor den 70er Jahren aus technischen Gründen nicht mit Computern zu verarbeiten war, erwies sich die Alphabetisierung der ältesten noch gebräuchlichen Schrift als undurchführbar. Die Hauptgründe für den Misserfolg der Schriftrevolution war einerseits die feste Verwurzelung der chinesischen Schrift in der chinesischen Kultur, deren Ablösung einen unüberschaubaren Verlust der Kultur hätte verursachen können. Zudem kann ein Alphabet die chinesische Sprache nicht präzise und eindeutig wiedergeben, da zahlreiche Homophone und ein isolierender Sprachbau vorliegen (vgl. Atsuji 1994: 445ff).

Aufbauend auf diesen Ausführungen lassen sich die folgenden Fazits ziehen: 1) Schrift ist nicht nur die schriftliche Äußerung einer Sprache, sondern auch eines der wichtigsten Symbole der Kultur einer Nation. 2) Schrift gehört dem Volk und Schriftreformen müssen die Eigenschaften der Sprache und den kulturellen Hintergrund respektieren. 3) Schreib- und Textverarbeitungstechniken dienen zum Schreiben und müssen zuerst je nach verschiedenen Schriften entwickelt werden. Es wäre eine falsche Richtung, für die vorhandenen Schreibtechniken die Schrift zu reformieren.

[18] Nach den Statistiken der Länderdatenbank der „Deutschen Stiftung Weltbevölkerung“ (DSW): Daten Mitte 2017.

Innerhalb eines Schriftkreises können Wortentlehnungen quasi ‚barrierefrei' durchgeführt werden. Zum Beispiel weisen die Wortformen sowie -aussprachen des Nomens *Computer* und des Eigennamens *Berlin* in allen lateinalphabetischen Sprachen kaum Unterschiede auf. Die Schriftformen der chinesischen sowie japanischen Ortsnamen sind ebenso im ganzen CJK-Schriftkreis annähernd identisch. Dem hingegen treten mehr Probleme beim sprachlichen sowie kulturellen Austausch zwischen verschiedenen Schriftkreisen auf, wie bei Termini, Eigennamen, Fremdsprachlernen, Tastaturlayout usw. Um diese Schwierigkeiten zu überwinden und internationale Kommunikation zu verstärken, sind phonetische Transkriptionen vonnöten. Transkription ist „Vorgang und Ergebnis der Wiedergabe eines Textes beliebiger Verschriftung in Form eines alphabetischen Textes" (Bußmann 2002: 712). Ein nicht-lateinalphabetisches Schriftsystem wird phonetisch bspw. in lateinische Buchstaben umgeschrieben, so dass auf die Wortaussprache trotz weniger entsprechenden sprachlichen sowie schriftlichen Kenntnisse hingedeutet werden kann. Das zu transkribierende Alphabet kann das international vereinheitlichte IPA, eine verbreitete vorhandene Schrift (wie die Schriften der oben genannten fünf Kreise) oder eine speziell für phonetische Übertragung erfundene Schrift (chinesisches Zhuyin) sein. Wegen der Verbreitung des lateinischen Alphabets weltweit werden Transkriptionen sowie Transliterationen im lateinischen Alphabet besonders häufig gebraucht.

Transkription und Transliteration sind Begriffe für phonetische Umschriften, die für die weitere Forschung dieser Arbeit unterschieden werden müssen. Transliteration wird wie folgt definiert: „A one-to-one conversion of the graphems of one writing system into those of another writing system [...]" (Coulmas 1996a: 509). Transliteration ist deswegen ein Unterbegriff von Transkription und hat Begrenzungen für manche Schriftsysteme, in denen die Eins-zu-eins-Übertragung in alphabetische Symbole unmöglich ist, wie das chinesische und japanische Schriftsystem mit vielen homophonetischen morphologischen Zeichen. Nach Bußmann muss Transliteration „die Übertragung eines in alphabetischer oder syllabischer Schrift geschriebenen Textes [...]" sein (Bußmann 2002: 713). Um Eindeutigkeit zu gewährleisten, sind diakritische Sonderbuchstaben häufig in einer im lateinischen Alphabet dargestellten Transliteration vorhanden. So wird bspw. der russische Buchstabe ‚**Ж**/ж' als ‚ž' dargelegt. Hingegen kann er als Graphem ‚sch' transkribiert werden, was bei der Wortrepräsentation Ambiguität verursachen kann.

Trotz der möglichen Mehrdeutigkeit wird die Transkription für Textverarbeitung bevorzugt. Denn im Gegenteil zu Transliteration kann eine Transkription innerhalb von ASCII-Zeichen (hauptsächlich 26 Grundbuchstaben) entworfen werden, weshalb sie relativ leicht zu erwerben und mit dem internationalen US-amerikanischen Tastaturlayout einwandfrei zu be-

arbeiten ist. Deswegen stellt eine Transkription häufig die Inputcodierung dar. Transliteration wird wegen der präziseren Repräsentation bei linguistischen Erforschungen bevorzugt.

1.3.5 Analyse der Grundeinheit verschiedener Schriftsysteme

In den Kapiteln 1.3.1 bis 1.3.4 wurden die Schriften und Schriftsysteme dargelegt und anhand Schrifttyp und Kulturkreis gegliedert. In diesem Kapitel werden Analysen und Vergleiche verschiedener Schriften vor dem Aspekt der schriftlichen Grundebene weiter durchgeführt. Die Grundebene eines Schriftsystems kann aus verschiedenen Perspektiven mit verschiedenen Begriffen betrachtet werden. Nach Dürscheid ist ein Schriftzeichen die „kleinste segmentale Einheit“ und ein Graphem „die kleinste bedeutungsunterscheidende Einheit“ eines Schriftsystems (Dürscheid 2006: 19 & 129).

Eisenberg hat statt Schriftzeichen ‚graphematische Grundformen‘ als die kleinste segmentale Einheit angegeben (vgl. Eisenberg 1996b: 1371). Im Normalfall wird ein Graphem als die kleinste distinktive Einheit eines Schriftsystems betrachtet, obwohl die Schriftzeichen in den meisten Fällen die grundlegenden Einheiten beim Schreiben, bei der Textverarbeitung und Zeichencodierung sind.

Ein Graphem ist die kleinste funktionale Einheit der Schrift, die das Schriftsystem in ein strukturelles Niveau der Sprache einwirkt. Der grundlegende Unterschied der Grapheme bei den drei Schrifttypen bezieht sich auf die zu repräsentierenden sprachlichen Einheiten: ein Phonem, eine Silbe oder ein Morphem/Wort (vgl. Coulmas 1996a: 174f, Bußmann 2002: 264). Ein alphabetisches Graphem kann sowohl ein Buchstabe als auch eine Buchstabenfolge sein, der/die ein Phonem wiedergibt. In einer logographischen oder syllabischen Schrift sind Graphem und Schriftzeichen meistens miteinander übereinstimmend. Die Unterschiede zwischen den Graphemen sowie Schriftzeichen verschiedener Schriftsysteme werden in der unten stehenden Graphik (Abb. 1-5) bezeichnet. Sie zeigt drei Koordinatenachsen: Schriftform, Aussprache und Bedeutung. Je komplizierter ein Graphem durchschnittlich beim Aufbau eines Schriftsystems ist, desto weiter ist es bei der Schriftformachse vom Startpunkt entfernt. Die Ausspracheachse bewegt sich von kleineren Einheiten (Phonemen), über Silben, bis zu einer betonten Silben oder einer Folge der Silben. Punkt (1) in der Schriftform-Aussprache-Fläche repräsentiert das Graphem der Alphabetschrift. In derselben Fläche steht Punkt (2) für Silbenzeichen. Punkt (3) (für Logogramme) ist der einzige Punkt, der mit allen drei Achsen koordiniert. Da ein Schriftzeichen im chinesischen Schriftsystem als eine betonte Silbe und im japanischen als eine oder mehrere unbetonte Silben ausgesprochen wird, wird die Aussprache für Logogramme mit höherer Stufe als eine unbetonte Silbe markiert.

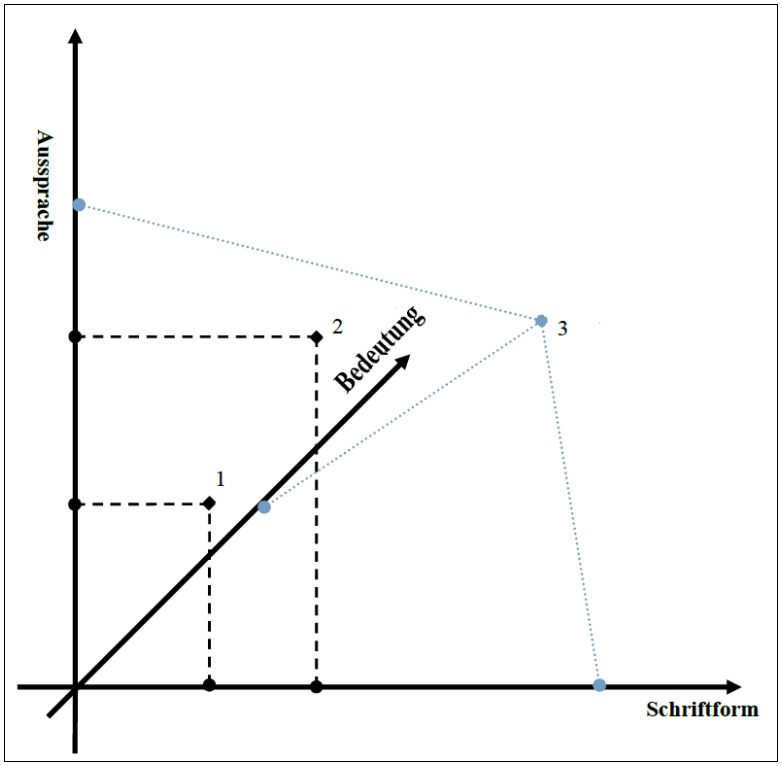


Abb. 1-5: Vergleich der Grapheme in verschiedenen Schriftsystemen

Wenn die Graphem-Laut-Korrespondenz der Alphabetschrift betrachtet wird, können alphabetische Grapheme weiter in vier Stufen klassifiziert werden: reinphonographische, phonemische, morpho- und partiellphonemische (vgl. Coulmas 1996b: 1382: Fig 118.1). Zu der reinen Phonologie gehört kein Schriftsystem, das einer natürlichen Sprache zugehörig ist. Die reine Phonographie IPA, in der eine Eins-zu-eins-Entsprechung zwischen Laut und Graphem eingesetzt wird, ist für keine natürliche Sprache tauglich, da mit reiner Phonologie bspw. die Dialekte einer Sprache schriftlich nicht vereint werden können (vgl. Eisenberg 1996b: 1373). Die am meisten an Lauten orientierenden Grapheme stammen aus der so genannten phonemischen Schrift (wie dem Finnischen), in der Grapheme strikt die bedeutungsunterscheidenden Phoneme in der Sprache repräsentieren. Anders als im IPA werden unterschiedliche lautliche Varianten eines Phonems mit einem identischen Graphem dargestellt, die als Allophone bezeichnet werden (vgl. Dürscheid 2006: 313). Die Grenze zwischen den Graphemen in phonemischer und morphophonemischer Schrift wird durch Allomorphe markiert, die Varianten eines Morphems sind (vgl. Bußmann 2002: 69). Ein Morphophonem bezeichnet phonologische Einheiten, die Allomorphe eines bestimmten Phonems vertreten (vgl. ibid.: 452). Modernes Deutsch und modernes Englisch werden mit morphophonemischen Schriften geschrieben. In den beiden Schriftsystemen ist die Wortwurzel eines Wortes trotz der lautlichen Änderung durch Flexion in den meisten Fällen graphematisch unverändert, wie *Kind* [ˊkint] und *Kinder* [ˊkindɐ] im Deutschen. Ein Schriftsystem mit Konsonantenschrift, in der nur Konsonanten

und manche langen Vokale mit Graphemen repräsentiert werden, wird auch als partielles phonemisches System bezeichnet, d.h. dass die Phoneme teilweise mit Graphemen dargestellt werden. Die vokalische Aussprache eines Wortes in Konsonantenschrift muss deswegen im grammatischen Kontext entschlüsselt werden (vgl. Coulmas 1996b: 1382).

Syllabische Zeichen sind unter den heute noch existierenden Schriftsystemen meistens nur für die im japanischen Schriftsystem verwendeten Kana-Zeichen charakteristisch. In der japanischen Sprache herrscht eine relativ einfache Silbenstruktur, so dass bei Hiragana und Katakana jeweils nur 48 Grundzeichen und insgesamt ca. einhundert Silbenzeichen in Gebrauch sind.

Nach der Definition von Dürscheid und Coulmas ist es problematisch die Grapheme von Alphasyllabar und Hangul festzulegen. Ein Buchstabe in Hangul überträgt ein bestimmtes Phonem und ein Buchstabe aus Alphasyllabar eine zu modifizierende Silbe. Gleichzeitig aber tritt in solcher Schrift eine Silbe als funktionale Einheit auf. Für die Verarbeitung von Hangul werden die Buchstaben auf der koreanischen Tastatur belegt, während in Unicode tausende Silbenblöcke eigenständig codiert werden. Für die Eingabe der Devanagari werden die Grundbuchstaben und abhängigen vokalischen Zeichen sowohl auf der Tastatur als auch in Unicode codiert. Nach meinem Verständnis haben sowohl indischen Schriften als auch Hangul Graphemen in zwei Niveaus – alphabetischem und syllabischem. Die Markierung für ein indisches Graphem in Abb. 1-5 schwankt nach dem Zeichentyp zwischen Punkt-1 und -2. Für Hangul kann Punkt-1 auf einen koreanischen Buchstaben und Punkt-2 auf einen von Buchstaben zusammengesetzten Silbenblock hindeuten.

Punkt-3 entspricht hauptsächlich Sinogrammen in dem chinesischen, japanischen und koreanischen Schriftsystem. Sie sind sowohl Grapheme als auch Schriftzeichen. Die oben stehende Graphik zeigt, dass ein Logogramm im Vergleich zu Phonogrammen bei der Wiedergabe der Sprache eine Dimension mehr besitzt, weshalb solch ein Schriftzeichen auf drei Ebenen – Zeichenform, Aussprache und Bedeutung – zu berücksichtigen ist. Bezogen auf die chinesische Schrift sind Zeichenform, Aussprache und Sinninhalt die Grundattribute eines Schriftzeichens. So überträgt ein Logogramm im Allgemeinen die meisten Informationen, ist aber vom Aufbau am kompliziertesten. Die Anzahl der Schriftzeichen einer logographischen Schrift kann deswegen unbegrenzt hoch sein. Im heutigen chinesischen Schriftsystem der VR China gibt es mehr als 70.000 eigenständige Sinogramme, inklusive seltenen, ausgestorbenen und Variantenzeichen (vgl. Li J 1996: 1409).

1.3.6 Schriftrichtungen – Horizontale und Vertikale

Schriftrichtung (auch Schreibrichtung; eng.: direction of script) lässt sich definieren als „the conventional course in which a script is written and (usually) read“ (Coulmas 1996a: 130). Sie ist erheblich von dem Schrifttyp abhängig. Bei einer alphabetischen Schrift, in der einzelne Buchstaben linear nacheinander (manchmal verbindend) folgen, muss eine bestimmte Schriftrichtung festgelegt werden. Eine logographische oder syllabische Schrift, wie die chinesische Schrift und das japanische Kana, kann sowohl horizontal als auch vertikal geschrieben und gelesen werden. Bei der Schriftniederlegung wirken eine primäre und eine sekundäre Schriftrichtung. So ist z.B. die primäre Richtung der lateinischen Schrift von links nach rechts, sekundär zeilenweise von oben nach unten. Wenn sich die Symbole einer Schrift primär horizontal anordnen lassen, ist die Schrift horizontal. Im Gegenteil ist sie dementsprechend vertikal (vgl. ibid.: 131). Heutzutage finden vier Arten von Schriftrichtungen Verwendung:

1) Die nur horizontal von links nach rechts dargestellten Schriften (Abk.: LR, eng.: Left to Right, LTR). Das Schreiben beginnt entsprechend beim linken Punkt der obersten Zeile eines Textlayouts. Zu dieser Gruppe zählen die meisten heutigen Schriften, inklusive des lateinischen, kyrillischen und griechischen Alphabets und die alphasyllabischen Schriften.
2) Die horizontal von rechts nach links geschriebenen Schriften (Abk.: RL, eng.: Right to Left, RTL), wie die arabische und hebräische Schrift. Der Startpunkt des Schreibens befindet sich im Textlayout oben rechts. Da die arabischen Ziffern rechtsläufig gezeichnet werden, ist ein arabischer oder hebräischer Text in den meisten Fällen bidirektional, mit RL als Haupt- und LR als Nebenrichtung. Wie sich die Bidirektionalität technisch unterstützen lässt, ist eine der größten Schwierigkeiten für die Textverarbeitung solcher Schriften.
3) Die in unterschiedlichen Richtungen (sowohl horizontal als auch vertikal beschreibbar) darstellbaren Schriften, die vor allem das chinesische, japanische und koreanische Schriftsystem einbeziehen (vgl. Bergerhausen 2011: 616f & 620). Beeinflusst von frühen Schriftträgern wie Holz- und Bambusplättchen wurden die CJK-Schriften traditionell vertikal von oben nach unten und die Spalte (auch die vertikale Zeile oder Kolumne genannt) von rechts nach links aufgezeichnet. Diese traditionelle Richtung des Schreibens kann abgekürzt als TB-RL (*top to bottom* und *right to left*) bezeichnet werden. Wenn der Text nur in einer Zeile darstellbar wäre und waagrechte Richtung bevorzugt würde, liefe die Schrift von rechts nach links, wie auf Plaketten historischer Gebäude. Aus Einflüssen aus dem Westen ab dem 19. Jh. ist die horizontale LR-Richtung immer beliebter geworden und ersetzte allmählich die traditionelle Schriftrichtung in offiziellen und inoffiziellen Schreibsi-

tuationen (vgl. Li J 1996: 1410, Stalph 1996: 1423, Chen DH 1984: 41). Im heutigen CJK-Schriftkreis herrschen deswegen zwei Schriftrichtungen vor: die traditionelle TB-RL und die moderne LR-Richtung (vgl. Xu SC 1993: 117).

4) Die von oben nach unten und sekundär von links nach rechts geschriebenen Schriften (TB-LR), wie die mongolische und die mandschurische Schrift. Die mongolische Schrift wird in der Inneren Mongolei als die offizielle Schriftträgerin der mongolischen Sprache beibehalten. Die mandschurische Schrift ist heute fast ausgestorben. Diese Schriftrichtung wird daher heute selten gebraucht.

Wie die chinesische Schrift waren die ausgestorbenen altägyptischen Hieroglyphen sowohl vertikal als auch horizontal darstellbar. Das phönizische und protosemitische Alphabet – die von den altägyptischen Hieroglyphen abgeleiteten, ältesten Alphabete – wurde festgelegt horizontal von rechts nach links geschrieben. Als das griechische Alphabet durch Anlehnung des phönizischen Alphabets erfunden wurde, war die Schriftrichtung hauptsächlich zeilenweise abwechselnd links-rechts sowie rechts-links, was als ‚Boustrophedon' bezeichnet wird (vgl. Coulmas 1996a: 49). Über die Jahrhunderte wurde die griechische Schriftrichtung horizontal von links nach rechts festgelegt. Diese Reform bedingte, dass die meisten vollalphabetischen Schriften horizontal rechtsläufig sind. Im indischen Schriftkreis herrscht ebenso die LR-Schriftrichtung vor und die aus Indien stammenden digitalen Ziffern werden trotz der getrennten Entwicklung in verschiedenen Kulturkreisen immer rechtsläufig geschrieben. Bei den Konsonantenalphabeten wird dagegen immer noch die Schreibkonvention des protosemitischen Alphabets in RL-Schriftrichtung beibehalten (vgl. ibid.: 131).

Da die meisten international gebrauchten Schriften (die arabische Zahlschrift, musikalische Notenschrift und das lateinische Alphabet) nur in Richtung LR geschrieben werden können, verursacht dies bei in anderen Richtungen dargestellten Schriften in vielen Fällen Ambiguitäten, Lesestörungen und technische Schwierigkeiten bei der Textverarbeitung. Wegen solcher internationalen Schriften werden in vielen arabischen, hebräischen, chinesischen, japanischen und mongolischen Texten etc. zwei oder mehr Schriftrichtungen verwendet. Bei solch einem Text müssen die Haupt- und Nebenschriftrichtung je nach Häufigkeit der beiden Richtungen bestimmt werden (vgl. Xu SC 1993: 118). Wenn die Hauptschrift vertikal geschrieben würde, gäbe es die Möglichkeit, den horizontalen Text um 90° zu rotieren. Die linksläufigen arabischen sowie hebräischen Schriften, deren Richtung 180° entgegengesetzt zur LR-Richtung ist, erfordern mehr technische Voraussetzungen (vgl. Davis / Lanin / Glass 2019:

Kap. 1). Die gemischten Schriftrichtungen von LR und RL innerhalb eines Text heißen ‚bidirektional' (Abk.: BIDI) (vgl. Unicode Glossary: Bidi).

Von derselben dezimalen Zahlschrift, die original aus Indien stammt und über Arabien in Europa verbreitet wurde, entstanden bei der geographischen Verbreitung und historischen Entwicklung verschiedene graphisch-variierte Systeme. Trotz der Formunterschiede wird die Schriftrichtung dieser Zahlschrift ohne Ausnahme immer horizontal von links nach rechts geschrieben. Der Gebrauch der Ziffer im Kontext einer linksläufigen Schrift ist deswegen problemhaft. In Arabisch z.B. wird das Datum <٢٠١٥/٠٤/١٩> (in ostarabischen Ländern) oder <2015/04/19> (in westarabischen Ländern) angegeben. Die logische Ordnung behält die Reihenfolge Tag, Monat und Jahr bei, die im Text von rechts nach links angegeben wird, während die Anzahl links-rechts-sequenziert wird. Wie das Datumtrennzeichen ist die Position der mathematischen Zeichen auch gegenüberstehend zu der Konvention der anderen Schriften. So kommt das Vorzeichen (Plus- und Minuszeichen) rechts, Prozent- und Quadratzeichen etc. aber links von dem Zahlzeichen vor (vgl. Li GB 1993: 15).

Ebenso herrscht bei der rechtsläufigen musikalischen Notation Konkurrenz, wenn ein arabischer Liedtext dazu komponiert wird. Die Reversrichtung bestimmt, dass es unmöglich ist, unter dem Parameter der Melodie passende Wörter anzuordnen. Dazu gibt es drei Lösungsmöglichkeiten: 1) den arabischen Text ins lateinische Alphabet transkribieren und unterhalb von Noten anordnen; 2) den Text in einzelne Wörter segmentieren und wörtlich von links nach rechts der Notation entsprechend angeben; 3) die Notenschrift und den arabischen Liedtext getrennt schreiben und die Sänger nach eigenem Gefühl den Text zur Melodie verbinden lassen (vgl. ibid.: 15f).

Die Übertragung von Liedtexten ist auch in der mongolischen Schrift ein komplexes Problem, für die sich ausgehend von den Lösungen des arabischen Raums vier Möglichkeiten postulieren lassen: 1) Übertragung in die kyrillische, chinesische oder lateinische Schrift; 2) die vertikal kombinierten Silben oder Wörter eigenständig unterhalb der Notation anordnen; 3) die Schrift einer ganzen Zeile gegen den Uhrzeigersinn um 90° rotieren; 4) den Liedtext und die Notenschrift getrennt schreiben. Da die Schriften des CJK-Kreises in unterschiedlichen Schriftrichtungen niedergeschrieben werden, gibt es weniger Konkurrenz bei den gemischten Texten mit einer internationalen Schrift. Probleme der CJK-Schriften liegen hauptsächlich in den Ambiguitäten, die wegen der unklaren Schriftrichtung auftreten können.

In gleicher Schriftgröße und -art (in Song- oder Regelschriftart) sind Höhe und Breite aller CJK-Schriftzeichen wegen der quadratischen Gestalt nahezu gleich. So könnten die Zeichen verschiedener Zeilen sowohl spalten- als auch zeilenweise arrangiert werden (sofern

keine Interpunktionszeichen gesetzt würden). Ästhetisch wäre deswegen verwirrend, ob ein Text horizontal oder vertikal läuft. Es ist deswegen erstens entscheidend, Abstand zwischen Nachbarzeichen und Zeilen eindeutig beim Textlayout zu unterscheiden. Zweitens muss per Kontext die verwendete Schriftrichtung entschlüsselt werden (vgl. Lunde 2009: 480f).

Wegen des logographischen Schrifttyps der chinesischen Schrift und des isolierenden grammatischen Aufbaus des Chinesischen kann Ambiguität bei der Schriftrichtung eines kurzen Texts auftreten. Ein Beispiel wird in Abb. 1-6 ersichtlich, in der die vier Schriftzeichen in alle vier Richtungen gewendet einen sinntragenden Text ergeben. Würde er in LR-Schriftrichtung (Richtung 1) gelesen, ergibt sich der Name eines der ältesten klassischen Schriftwerke über Sinographie: ‚Shuowen-Jiezi' von Xu Shen. Würde hingegen die traditionelle Schriftrichtung TB-RL (2) durchgeführt, wäre /wénzì jiěshuō/ mit der Bedeutung *Schrifterklärung* zu lesen. Erfolgte die Entzifferung via Schriftrichtung (3), könnte man auf Deutsch die Wörter *Textbeschreibung* und *Zeichenerklärung* (nur in klassischer Schriftsprache, /wénshuō zìjiě/) erkennen. In Richtung 4 (links-rechts) ergäbe sich die Handlung *Schrift erklären* /shuōjiě wénzì/.

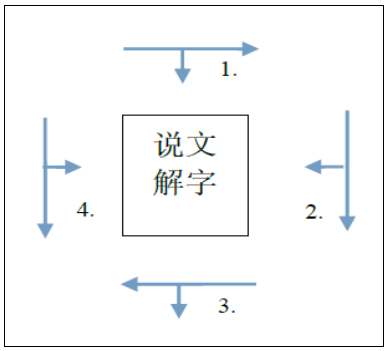


Abb. 1-6: Die Ambiguitäten des Chinesischen wegen verschiedenen Schriftrichtungen[19]

Der Abstand zwischen Nachbarzeichen und Zeilen ist in diesem Fall das wesentliche Kriterium, um die Schriftrichtung zu entschlüsseln. In Abb. 1-6 ist so zu erkennen, dass der Text horizontal verfasst wurde. Da Richtung 3 meistens nur bei einzeiligen chinesischen Texten verwendet wird, ist Richtung 1 am wahrscheinlichsten.

Theoretisch kann eine Schrift in der horizontalen Schriftrichtung effektiver beim Lesen funktionieren als in der vertikalen. Da die Augen waagrecht stehen, haben Leser auf horizontalen Zeilen prinzipiell einen weiteren Blick, als bei einer vertikalen Spalte. Aus verschiedenen Gründen (wie Lesenseffizienz, Verbreitung der westlichen Kultur, Einführung der arabischen Ziffern etc.) wurde die horizontale Schriftrichtung ab dem 19. Jh. in CJK-Regionen

[19] Die vier Zeichen werden in der Kurzform der Song-Schriftart angegeben. Die Aussprache in Pinyin nach der Reihenfolge links-rechts und oben-unten lautet: /shuō/, /wén/, /jiě/ und /zì/.

immer mehr gefördert und verwendet. In der VR China wurde die chinesische Schrift ab 1955 offiziell waagrecht geschrieben und als Standard festgelegt (vgl. Zhou YG 1954: 43). In Taiwan, Hongkong, Singapur, Japan und Korea wird sie heutzutage auch in den meisten Fällen bevorzugt. In manchen Fällen – etwa Publikationen historischer Schriftstücke, Zeitungen, Magazine, Werbungen, Ladenlogos, Banner usw. – findet die vertikale Schriftrichtung bis heute Verwendung (vgl. Chen DH 1984: 29). Die möglichen Schriftrichtungen verwirren beim Lesen auf Chinesisch, besonders unter Grundschülern und Fremdsprachlernern. Im CJK-Kreis sind Textverarbeitungsprogramme nötig, in denen die vertikale oder horizontale Schriftrichtung nach eigenen Bedürfnissen umgeschaltet werden kann. Manche Interpunktionszeichen wie Gedankenstrich, Buchtitelzeichen oder Klammer werden bei vertikaler Richtung in der variierten Form dargestellt und 90° in Uhrzeigerrichtung von ihrer horizontalen Form gedreht wird (vgl. Lunde 2009: 484-492).

Beim Einfügen einer horizontalen Schrift in einen vertikal dargestellten Text (z.B. Gebrauch arabischer Ziffern in einem vertikal geschriebenen chinesischen Text) gibt es vor allem zwei Methoden. Meistens wird der eingefügte Text um 90° rotiert (siehe Abb. 1-7). Wenn der Text aus einem kurzen Wort besteht, das kaum breiter als ein Sinogramm ist, kann das Wort wie ein selbstständiges Schriftzeichen behandelt werden. Wie in der zweiten Variante der chinesischen vertikalen Datumsangabe von Abb. 1-7 können die Ziffern <2015>, <04> und <24> sowohl insgesamt um 90° rotiert, als auch richtungsunverändert in der Worteinheit im Text hinzugefügt werden.

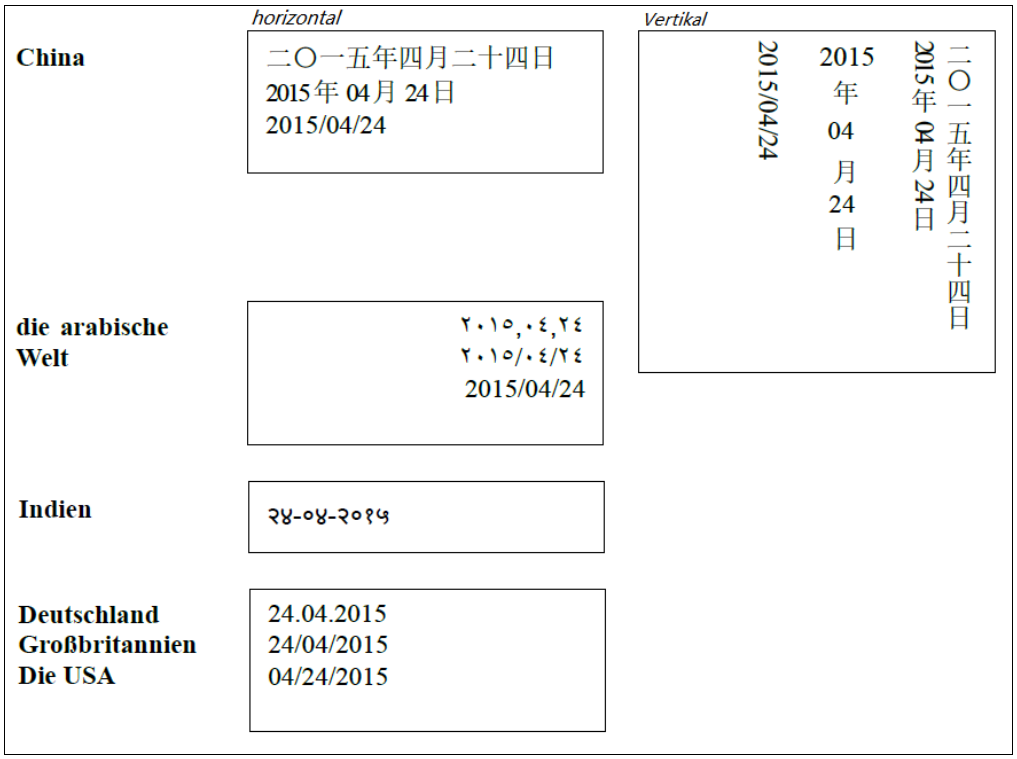


Abb. 1-7: Varianten für die Datumsangabe in verschiedenen Regionen[20]

1.4 Techniken zur Verarbeitung der Schriftzeichen

Um die Textverarbeitung in jeder literalisierten Sprache durchführen zu können, sind Techniken zur Informationsverarbeitung einzelner schriftlicher Einheiten die grundlegende Voraussetzung. Bei der Informationsverarbeitung sind die Prozesse der Eingabe, Decodierung, Encodierung bis zur Ausgabe eingeschlossen. In Kap. 1.3.5 wurde die linguistische Bedeutung von Schriftzeichen als kleinste segmentale Einheit der Schrift vergleichend vorgestellt. Kap. 1.4 fokussierte hingegen erweitert die technischen Aspekte von Zeichen aus verschiedenen Schriften. Dies beinhaltet die verschiedenen Codierungsarten sowie Darstellungsweisen bei Tastenbelegung, Prozessor, Bildschirm, Drucker und Speicher.

1.4.1 Definition von Zeichen, Glyphe und Font und der Zusammenhang zwischen den drei Begriffen

Zur Schriftverarbeitung mit Computer sind zuerst drei Fachbegriffe einzuführen: Zeichen, Glyphe und Font.

[20] Das Symbol <〇> (/líng/, *null*) ist die Kleinschreibung von <零>, das generell nur bei Jahr- und Seitenangabe verwendet und linguistisch meistens nicht als ein chinesisches Schriftzeichen angesehen wird. Die drei arabischen Schreibweisen müssen in der Reihenfolge *Jahr*, *Monat* und *Tag* im Computer eingegeben werden, wenn sie unter der arabischen Textverarbeitung abläuft, die umgekehrt von der sprachlichen Reihenfolge ist.

Der Begriff *Zeichen* (eng.: character) weist unterschiedliche Bedeutungen im sprachwissenschaftlichen und computertechnologischen Kontext auf. Linguistisch betrachtet ist ein Zeichen „one of the elementary signs of a written language [...]“ (Coulmas 1996a: 72), das entweder ein Schriftzeichen (Buchstabe, Syllabar oder Logogramm) oder ein Hilfszeichen (Satz- und Interpunktionszeichen) darstellt. Aus Perspektive der Textverarbeitung hingegen sind ‚abstract character‘ gemeint, die wie folgt definiert werden können: „[A] unit of information used for the organization, control or representation of textual data“ (Unicode Glossary: Abstract Character). Des Weiteren ist es die Grundeinheit der Zeichencodierung im Unicode oder in anderen Codierungsstandards (vgl. ibid.: Character). Nach dieser Auslegung wird zu jedem Codepunkt des Unicodes ein unterschiedliches Zeichen definiert. Zeichen sind unabhängig von Sprache, Schriftart, Schreibtechnik, Betriebssystem und Programm. Die Großschreibung des lateinischen Buchstabens <A> ist deswegen immer dasselbe Zeichen, egal in welcher Schriftgröße und -art es dargestellt und ob es im Deutschen, Englischen oder anderen Schriftsystemen geschrieben wird.

Die zur elektronischen Textverarbeitung verwendeten Zeichencodierungen können in vier Sorten klassifiziert werden: 1) interner Code (eng.: internal code, chi.: 内码) im Binär, der innerhalb der CPU zur Datenbearbeitung und -speicherung eingesetzt wird; 2) Austauschcode (interchange code, 交换码) im Binär, welcher als Mittel zur Textvermittlung funktioniert; 3) Ausgabecode (display code, 输出码) im Binär, der mit auszugebenden Informationen (wie Glyphe und Font) verknüpft ist; 4) Inputcode (input code, 输入码), der als Eingabemittel eines nicht auf der Tastatur belegten Zeichens (wie die für die chinesischen Schriftzeichen) funktioniert und im Normalfall nicht binär ist (vgl. Feng 1999: 158f). Innerhalb von einem Computer können der interne, der Austausch-, der Ausgabe- und der Inputcode eines Schriftzeichens zwar übereinstimmen, aber in den meisten Fällen unterscheiden sie sich. Die Austauschcodierung muss wegen den Anforderungen des nationalen sowie internationalen Datenaustauschs bei allen Computern vereinheitlicht sein (vgl. ibid.). Unicode ist heute der international verbreitetste Standard für Austauschcodierung, der die Zeichen aller modernen verwendeten Schriften umfasst. Zur Ausgabe der Zeichen auf Bildschirm und Druckpapier müssen Glyphen und Fonts im Computer gespeichert werden.

Die Glyphe (eng.: glyph) ist die graphische Darstellung vom Zeichen. Sie bleibt selbstständig bei verschiedenen Schriftstilen und -größen. Eine Glyphe kann ein Zeichen, einen Zeichenteil oder eine Zeichenfolge repräsentieren. Ein Zeichen kann ebenso mehreren Glyphen oder einem Teil der Glyphe entsprechen (vgl. Zhang ZC 1999: 73, Whistler et al. 2008: Kap. 2.2). Darauf bezogen sind die Begriffe Glyphencode (glyph code), -identifizierer (glyph

identifier) und -bild (glyph image) zu nennen. Mithilfe von Glyphencode sowie -identifizierer kann die konkrete, auf dem Bildschirm ausgegebene Glyphe innerhalb eines Fonts (Glyphenbilds) mit entsprechenden Zeichen kombiniert werden. Je nach den verschiedenen Schriften können Zeichen-Glyphen-Beziehungen in die folgenden sieben Fälle klassifiziert werden, zu denen in Tab. 1-1 Beispielschriftzeichen angegeben werden.

1) Die Eins-zu-eins-Entsprechung in einer bestimmten Schriftart.

In verschiedenen Schriftarten sind die Glyphen eines Zeichens manchmal variiert. Unter der Voraussetzung einer selben Schriftart entspricht bei manchen Schriftgruppen ein Zeichen eins-zu-eins seiner Glyphe. Die lateinischen Grundbuchstaben sind Beispiele für diesen Fall.

2) Entsprechung einer Glyphe zu zwei oder mehreren Zeichen aus verschiedenen Schriften.

Die Glyphe <A> kann bspw. die graphische Darstellung von drei unterschiedlichen Zeichen sein: dem ersten lateinischen Buchstaben (U+0041), dem griechischen Buchstaben Alpha (U+0391) und dem kyrillischen Buchstaben A (U+0410) in Majuskel. Die drei Buchstaben haben zwar denselben Stamm und sind in derselben Schriftart graphisch identisch, werden aber wegen der Zugehörigkeit zu verschiedenen Schriften als unterschiedliche Schriftzeichen definiert (vgl. Bergerhausen 2011: 036 & 50f).

3) Die Entsprechung eines Zeichens zu mehreren vom Kontext abhängigen Glyphen.

Dieser Fall bezieht sich vor allem auf das arabische, hebräische und mongolische Alphabet, in denen ein Buchstabe situationsabhängig in verschiedenen Formen dargestellt wird. Bspw. hat ein arabischer Buchstabe vier verschiedene Glyphen – jeweils für isolierende, Initial-, Medial- und Finalform – obwohl er mit einem einzigen Codepunkt definiert wird. Computergestützt wird die korrekten Form/Glyphe eines Buchstabens anhand seiner Wortposition ausgegeben (vgl. Whistler et al. 2008: Kap. 2.2f).

4) Die Entsprechung eines Zeichens zu mehreren, von verschiedenen Schriftsystemen sowie Regionen bedingten Glyphen.

Dies bezieht sich vor allem auf die chinesische Schrift. Wegen der getrennten Schriftentwicklung und verschiedenen nationalen Orthographien etc. haben dieselben Schriftzeichen in der VR China, Taiwan, Hongkong, Japan, Korea und Vietnam unterschiedliche typographische Standards. Im Allgemeinen liegen die typographischen Unterschiede desselben Zeichens in der Darstellung einzelner Striche sowie der Verhältnisse der Komponenten im Zeichen. In den verschiedenen Regionen werden verschiedene Glyphensysteme mit den CJK-vereinten Ideogrammen des Unicodes verknüpft (vgl. Zhang ZC 1999: 26f).

5) Die Glyphe für eine Ligatur.

Ligaturen sind „häufig aus ästhetischen Gründen vorgenommene Verbindung[en] zweier oder mehrerer Buchstaben, so dass ein einziges in sich geschlossenes Gebilde entsteht“ (Bußmann 2002: 409). Bei einer alphabetischen Schrift sind die mit Glyphen angezeigten Ligaturen manchmal einzusetzen. Die Buchstabenfolge <*fi*> bspw. kann optional mit zwei Glyphen von jedem Buchstaben eigenständig oder mit einer Glyphe verbunden angezeigt werden (vgl. Bergerhausen 2011: 036 & 50f).

6) Die zusammengesetzten Glyphen für ein mit Diakritika gebildetes Schriftzeichen.

Diakritische Zeichen sind die zusätzlich an oder in einem Grundzeichen gesetzten Symbole, „mit denen bestimmte Unterscheidungen getroffen werden sollen“ (Bußmann 2002: 161f). Diakritische Zeichen werden häufig in einer vollalphabetischen Schrift (für Bildung neuer Buchstaben), einer Konsonantenschrift (für Angabe der vokalischen Phoneme), einer Alphasyllabar (als abhängiger Vokal) oder in japanischen Kana-Schriften verwendet.

Die Codierung solcher Zeichen kann mit zwei unterschiedlichen Methoden einhergehen: als ein Gesamtzeichen und als Zeichenfolge von einem Grundzeichen und einem oder mehreren diakritischen Zeichen. Das Schriftzeichen <À> beispielsweise kann sowohl als ein Zeichen (U+00C0), als auch als die Folge von zwei Zeichen (U+0041 für <A> und U+0300 für Grave) repräsentiert werden. Unabhängig davon, mit welcher Methode <À> codiert wird, wird es bei der Ausgabe mit der Sequenz von zwei Glyphen (<A> und die kombinierende Glyphe < ` >) repräsentiert (vgl. Whistler et al. 2008: Kap. 2.2). Bei der ersten Methode herrscht die Entsprechung zwischen einem Zeichen zu zwei oder mehreren Glyphen. Bei der zweiten Methode gibt es zwar Eins-zu-eins-Entsprechungen zwischen Zeichen und Glyphe, aber die Zeichen repräsentierenden Glyphen werden nicht linear angeordnet, sondern mit bestimmten Kombinationsregeln ausgegeben. In diesem Fall können sich Zeichen in Basiszeichen (base character, wie <A> von dem Kombinationszeichen <À>) und kombinierende Zeichen (combining mark, wie Gravis von <À>) unterscheiden (vgl. Unicode 12.0 Chapters: Kap. 3.6: 106). Kombinierende Zeichen können weiter in drei Klassen eingeteilt werden: 1) nonspacing mark, wie das lateinische diakritische Gravis, das bei der Repräsentation über oder unter einem Grundzeichen steht und keinen Raum bei der visuellen Grundlinie erfordert; 2) enclosing mark, wie das kyrillische Hunderttausendzeichen < ҈ > (U+0488), das keinen Raum braucht und ein Grundzeichen umschließt; 3) spacing mark, das Platz bei der visuellen Grundlinie benötigt, wie das abhängige Vokalzeichen der Devanagari < ि > (/i/, U+039F) (vgl. ibid: 106ff). Bei verschiedenen Schriften werden die kombinierenden Zeichen nach den

Schrifteigenschaften und technischen Gründen mit verschiedenen Methoden behandelt (siehe Kap. 2.1.3 & 2.2.4). In den lateinischen Schriften bspw. wird die Codierung des gesamten Kombinationszeichens bevorzugt, während hingegen in den indischen Schriften und der arabischen Schrift die Codierung der kombinierenden Zeichen eine bedeutende Rolle spielen.

7) Mehrere Glyphenvarianten eines Zeichens oder einer orthographischen Silbe in einer alphasyllabischen Schrift.

Im Kap. 1.3.2 [S. 22f] wurden die Grundprinzipien der alphasyllabischen Schrift vorgestellt. Die Schrift wird zwar in Silben gruppiert, die syllabischen Einheiten sind aber in Grundbuchstaben und abhängigen Vokalzeichen zu segmentieren. Letztgenannte sind spacing marks (siehe letzten Absatz, die dritte Variante der kombinierenden Zeichen). Zur Darstellung eines Grundbuchstaben in verschiedenen Kontexten, wie z.B. als eigenständige Silbe, Kombination mit einem Vokalzeichen, in einem Konsonantencluster sowie in graphischen Ausnahmen bei der Kombination mit bestimmten Zeichen, muss mit bestimmten Glyphen repräsentiert werden (vgl. ibid.: Kap. 12.1: 457). Mehr Informationen darüber werden in Kap. 2.3 eingeführt.

Die Glyphe kann auf dem Bildschirm und beim Drucken in verschiedenen Schriftarten, -größen etc. ausgegeben werden und die sichtbare Graphik heißt Font. Im engeren Sinne bezieht sich ein Font auf das konkrete Aussehen eines Schriftzeichens, während im weiteren Sinne die Sammlung von Schriftzeichengraphiken in gleicher Schriftart gemeint ist (vgl. Zhang ZC 1999: 73). Font unterscheidet sich in Schriftart und manchmal auch in Größe. Z.B. wird für den Hauptinhalt der Arbeit das Font ‚Times New Roman' in 12 pt angewendet.

Fall	Beispielszeichen[21]	Unicode	Glyphe & Erklärung	Font
1)	Lateinischer BS. <A> in Min.	U+0061	a (in 'Arial')	a, **a**, *a* [22]
			a (in 'Courier New')	a, **a**, *a*
2)	Lateinischer BS. <A> in Maj.	U+0041	A (in 'Times New Roman')	A, **A**, *A*
	Griechischer BS. <A> in Maj.	U+0391		
	Kyrillischer BS. <A> in Maj.	U+0410		
3)	Aarabischer BS. /'Ayan/	U+0639	ع (isolierende Form)	<ع> /'/
			ع (Initialform)	in <عين> /'jn/
			ع (Medialform)	in < لعب > /l'n/
			ع (Finalform)	in < سلع > /sl'/

[21] Abkürzungen dieser Spalte: BS.: Buchstaben, Min.: Minuskel, Maj.: Majuskel, chi_m: Standardchinesisch, chi_yue: Kantonesisch, jap: Japanisch, kor: Koreanisch, vie: Vietnamesisch und Lig.: Ligatur

[22] Das Zeichen wird in verschiedenen Schriftstilen von derselben Schriftart dargestellt: normal, fett und kursiv. Ohne weitere Anmerkung gilt es gilt auch für andere Zeilen in dieser Spalte.

Fall	Beispielszeichen[21]	Unicode	Glyphe & Erklärung	Font
4)	das CJKV-Zeichen für Moral; (Aussprache in chi_m: /dé/; chi_yue: /dak1/; jap: /oshie/ (Kun), /toku/ (On); kor: <덕> /deok/; vie: <đức>).	U+5FB7	德 (aus VR China, G0-3542)[23]	德德德德德
			德 (aus Hongkong, HB1-BC77)	Unterscheidung bei Schriftarten wie oben
			德 (aus Taiwan, T1-6C61)	Wie oben
			德 (aus Japan, J3-7445)	Wie oben
			德 (aus Südkorea, K0-536C)	Wie oben
			德 (aus Vietnam, V1-5471)	Wie oben
5)	Lig. von den lateinischen BS. <F> und <I> in Min.	U+0066 + U+0069	fi (in 'Times New Roman')	fi, **fi**, *fi*
6)	<a> mit Umlaut	U+00EA	a +¨	ä, **ä**, *ä*
		U+0061 + U+0308		
7)	BS der Devanagari für /ka/	U+0915	क /ka/ (isolierende Form)	क [24]
			Cl-Form wie in <कु> /ku/ (kombiniert mit einem abhängigen Zeichen)	कु
			Cd-Form in <क्> /k/ (Konsonant ohne Vokal)	क्
			Ch-Form, wie in <क्त> /kta/ (in einem Konsonantencluster)	क्त
			Die Glyphe für Silben <क्ष> /kṣa/	क्ष

Tab. 1-1: Die Beziehungen zwischen Zeichen, Glyphen und Font mit Beispielzeichen aus verschiedenen Schriften[25]

Zeichen, Glyphe und Font eines schriftlichen Symbols sind bei der Textverarbeitung mit einem bestimmten Code und Identifizierer miteinander fest verknüpft. *Zeichen* bezieht sich auf die Schrift: Es ist die Grundeinheit für Eingabe, Dateispeicherung und Datenaustausch. *Glyphe* bezieht sich auf die Zeichenstruktur und definiert die typographische Darstellung eines Zeichens. *Font* ist die sichtbare Graphik auf Bildschirm oder Druckpapier. Unter den drei Begriffen sind Zeichen am abstraktesten, Fonts am konkretesten.

1.4.2 Vereinheitlichte Austauschcodierungen

Datenverarbeitung mit Computer basiert auf Binärcode, also der Folge von Nullen und Einsen. Wie im letzten Kapitel erwähnt, müssen nationale sowie internationale Normungen für Austauschcodierungen festgelegt werden. Ihr Inventar enthält neben schriftlichen Symbolen auch Steuer- (eng.: control character) und Formatzeichen (format character). Zeichen, die mit

[23] Reihenfolge in Spalte ‚Glyphenerklärung' und ‚Font': Herkunftsort, abgekürzter Name des national-gestammten Codierungsstandards und der Codewert in diesen Standards. Der Font der chinesischen Schrift wird 0,5 cm hoch (entspricht ca. 12 Pt.) dargestellt. Die dargestellten Schriftarten in Font-Spalte sind nachfolgend: Song- (宋体), Regel- (楷体), Rennen- (行书), Grass- (草书), Kanzleischrift (隶书).

[24] In Schriftart Mangal.

[25] Vgl. Unicode 12.0 Chapters: Kap. 2.2: 15ff, Davis/Whistler 2019: Kap. 1, Unicode-Datei: „ucd.unihan.flat"

Hand geschrieben und von Computer ausgegeben werden können, heißen graphische Zeichen (graphic character) (vgl. Unicode 12.0 Chapters: Kap. 3.6: 106). Wie die Codierung für unbegrenzte graphische Zeichen aus verschiedenen Schriften in einem internationalen Standard vereinheitlicht wird, ist der Schwerpunkt dieses Abschnitts.

Graphische Zeichen lassen sich in sechs Typen klassifizieren: schriftliche Basiszeichen verschiedener Schriftsysteme (inklusive Buchstaben, Silbenzeichen und logographischen Schriftzeichen), kombinierende Zeichen, Ziffern (inklusive den internationalen und verschiedenen lokalen sowie nativen Zahlzeichen), Interpunktionszeichen, sonstige Symbole (Währungs-, mathematische Zeichen usw.) und Trennzeichen (ibid.). Die arabischen Zahlzeichen, die international verwendeten sonstigen Symbole und die meisten Interpunktionszeichen sind unabhängig von einer bestimmten Schrift und werden generell in allen menschlichen Schriftsystemen verwendet. Wie in Kap. 1.4.1 im Zusammenhang von Zeichen und Glyphe eingeführt wurde, gibt es zwei Methoden der Behandlung von einem mit kombinierenden Zeichen gebildeten Symbol: 1) die Codierung des Gesamtsymbols als Zeichen; 2) die getrennte Codierung von Basiszeichen und kombinierenden Zeichen, gesetzt dem Fall, dass die Glyphen von zwei oder mehreren Zeichen kombiniert dargestellt werden müssen. Mit welcher Methode die Zeichencodierung durchgeführt wird, ist ein wichtiger Faktor für die Größe des Zeichenvorrats eines Schriftsystems. Kap. 2 geht tiefer auf kombinierende Zeichen am Beispiel verschiedener Schriftsysteme ein.

Die Entwicklung der Standardisierung der internationalen Austauschcodierung kann hauptsächlich in drei Phasen unterteilt werden: das ASCII mit grundsätzlichen Zeichen innerhalb des englischsprachigen Raums, die erweiterten Zeichencodierungen im ISO für verschiedene Schriften und der international vereinheitliche Standard Unicode.

ASCII (American Standard Code for Information Interchange) war einer der frühesten vereinheitlichten Standards zur Zeichencodierung für den Informationsaustausch im Internet. Es handelt sich um eine 7-Bit-Codierung, die somit 2^7 Zeichen enthält. Unter den somit 128 Gesamtzeichen befinden sich 95 graphische (94 sichtbare und das Leerzeichen) sowie 33 Steuerzeichen (C0-Controls). Die codierten graphischen Zeichen umfassen die 26 lateinischen Grundbuchstaben in Majuskel und Minuskel, die arabischen Ziffern, generell verwendete Interpunktions-, Wort- und Mathematikzeichen und das US-amerikanische Währungszeichen ($). ASCII hat zwar regionale Einschränkungen, wurde aber als das Grundblock der später entstandenen internationalen Zeichencodierungsstandards – wie ISO-8859 und Unicode – übernommen (vgl. Unicode Glossary: ASCII).

Ein Binärcode kann in hexadezimaler und dezimaler Nummerierung beschrieben werden. Im Hexadezimalsystem wird eine Folge von vier Bits (16 Positionen) mit einem Wert von 0 bis F festgelegt. Im Dezimalsystem wird der Wert dezimal nummeriert. Der Buchstabe <A> z.B., der als ‚100 0001' in ASCII codiert wird, hat den hexadezimalen Wert 41 (2^2+0+0=4, 0+0+0+2^0=1) und den dezimalen Wert 65 (2^6+0+0+0+0+0+1=65).

Mit der Verbreitung des Computers und des Internets war es notwendig, die Zeichencodierung für verschiedene Schriften sowie Schriftsysteme zu erweitern. Der erste Schritt war, ein Bit mehr für die Zeichencodierung einzusetzen und die 8-Bit-Codierung durchzuführen. Auf der Basis von ASCII wurden die Normungen für verschiedene alphabetische Schriften in 8-Bits-Codierung entworfen.

In Sprachräumen, wo logographische und syllabische Schriften gebraucht werden, erforderte die Zeichencodierung wegen des großen Zeichenvorrats das Vielfache von 7-Bits. So wurde im CJK-Schriftkreis Zeichencodierung mit einer zwei- oder dreifachen ASCII-Codierung durchgeführt, ergo 14 oder 21 Bits, wobei ein Zeichen durch zwei oder drei sichtbare ASCII-Zeichen repräsentiert wird (vgl. Lunde 2009: 8ff). In einem 14-Bits-Codierungsstandard können insgesamt 8.836 (94×94) Zeichencodes definiert werden. So wurden Zeichencodierungen mit 14 Bits in der VR China, Japan und Südkorea und 21 Bits (830.584) in Taiwan entworfen und angewendet. Bspw. wird das Schriftzeichen <德> (/dé/, *Moral*) in GB2312-80 (der nationale Standard der VR China) als 34. Code (1-94 Zeichen in jeder Zone) von Zone 21 (insgesamt Zone 1-94 Zone) angelegt und sein Zeichencode kann nach dem Dezimalsystem als 2134 dargestellt werden. Um die Kompatibilität zwischen Zeichencodierungen aus dem CJK-Schriftkreis und ASCII zu ermöglichen, wird in den chinesischen, japanischen sowie koreanischen Zeichencodierungsstandards häufig die Codierungsmethode EUC (Extended UNIX Coding) eingesetzt. Grundprinzip der Methode ist, jede 7-Bit-Stelle hexadezimal auf 8-Bit zu erhöhen, damit diese Codierung Kapazität für ASCII hat. In diesem Fall kann die CPU einen aus doppelten 7-Bits zusammengesetzten Zeichencode von einem ASCII-Code unterscheiden und fehlerfrei verarbeiten. Je nach Anwendung in verschiedenen nativen Standards gibt es vier Hauptvarianten: EUC-CN (in der VR China), EUC-TW (in Taiwan), EUC-JP (in Japan) und EUC-KR (in Südkorea) (vgl. Lunde 2009: 242-252). Im GB2312 wird die Zone von 1 bis 94 sowie die nummerierten Zeichen 1 bis 94 mit dem hexadezimalen Wert von 0×A1 bis 0×FE (die Zonen-Nummerierung plus A0) definiert. Der hexadezimale Wert von <德> lautet z.B. in GB2312-80: 0×B5-0×C2 (21 +A0 = B5; 34 +A0 = C2).

Die verschiedenen Normen der Zeichencodierung haben den internationalen Datenaustausch erheblich behindert. So ist die Idee eines universalen Codierungssystems entstanden. 1991 wurde die erste Version von Unicode (Unicode 1.0.0) eingeführt. Die neueste Version (Unicode 12.0) erschien im März 2019 und umfasst über hundert Kategorien (150 Schriften) und über einhunderttausend Zeichen (137.929) Zeichen (vgl. Unicode 12.0 Chapters: Preface: xxi). Das Grundprinzip ist die vereinheitliche 16-Bit-Codierung (USC-2 genannt, wobei UCS für Universal Character Set steht) auf der fundamentalen multilingualen Ebene (Abk.: BMP, für Englisch: basic multilingual plane). Auf den weiteren 16 Ergänzungsebenen (von 1. bis 16. Plane, U+10000 - U+10FFFF) werden Codes aus bis zu 33 Bits (UCS-4 genannt) eingeschrieben (vgl. Unicode 12.0 Chapters: Kap. 2.8: 44). Theoretisch kann der Unicode insgesamt 1.114.112 Zeichen ($2^{16} \times 17$) tragen. Nach aktuellen Stand werden 12,38% der Codepunkte (137.929/1.114.112) besetzt.

Die Vereinigung aller Schriftzeichen auf der Welt in einem Codierungsstandard war ein langer Prozess, für den zahlreiche Experten aus verschiedenen Sprachräumen zusammenarbeiten mussten, Entwicklungsarbeit geleistet haben – und leisten werden. Er basierte zum einen auf vorhandenen Normen wie ASCII und Teilnormen des ISO-8859. Manche Normen entsprechen auch den Kategorien des Unicodes. So ist der erste Unicode-Block mit ASCII mit 128 Codes übereinstimmend. Der zweite Block wurde von ISO-8859-1 als Norm für Westeuropa übernommen. Die Codierung der arabischen Buchstaben basiert ebenso auf ISO-8859-6 (vgl. Unicode 12.0 Characters: U+0080-U+00FF & U+0600-U+06FF).

Es gibt drei verschiedene vom Unicode Standard angebotene Encodierungsformen, die als UTF (UCS Transformation Format) bezeichnet werden: UTF-8, UTF-16 und UTF-32, je nachdem in wie vielen Bits ein Zeichen primär verarbeitet wird. UTF-8 ist unter den drei Formen die am häufigsten verwendete und das vorherrschende Format für Webseiten, bei denen die Unicodezeichen je nach erforderlichen Bits in einem bis zu vier Bytes repräsentiert werden (vgl. Unicode 12.0 Chapters: Kap. 2.6: 40ff). Die Zeichen von der BMP-Ebene werden mit UTF-8 in drei Stufen klassifiziert. Von U+00 bis U+7F (ASCII-Zeichen) werden die Codepunkte innerhalb von einem Byte gespeichert und transportiert. Weitere lateinischen Buchstaben, Buchstaben von vielen anderen Alphabeten, weitere Sonderzeichen und sonstigen Zeichen (U+0080-U+07FF) werden in zwei Bytes behandelt. Zu den in drei Bytes angelegten Symbolen gehören die Codepunkte U+0800-U+FFFF (wie alphasyllabische Zeichen, japanische Kana, koreanische Silben, CJK-Ideogramme etc.) (vgl. Umamaheswaran 2002: Kap. 3). Im Vergleich zu UTF-8 werden alle Zeichen der BMP-Ebene in UTF-16 mit 16 Bits (2 Bytes) und die Zeichen auf anderen Ebenen mit vier Bytes definiert. Als Beispiel taugen

drei Zeichen aus ASCII, dem kyrillischen Alphabet und den CJK-Ideogrammen in beiden Encodierungsformaten: <A> (UTF-8: 41; UTF-16: 0041), <И> (D0 98; 0418) und <德> (E5 BE B7; 5FB7). Welches Format effektiver ist, ist deshalb von der vorwiegend im Text verwendeten Schrift abhängig. In UTF-32 werden alle Zeichen in 32 Bits anerkannt und verarbeitet. Für die meisten heute verwendeten Schriften (die Schriften aus BMP) ist UTF-32 deswegen uneffektiv und unökonomisch.

Nach Unicode Standard 12.0 wird die Zeichencodierung von denen in der Dissertation erforschten Schriften in der folgenden Tabelle aufgelistet. Die Hauptkategorien solcher Schriften werden zudem beschrieben.

Schrift	**Attribut:** 1) Richtung 2) Sprache 3) Gebiet[26]	**Hauptkategorie mit Blockname, Umfang & Zeicheninventar**			**Erweiterte Kategorien**
die lateinische Schrift	1) Rechts; 2) EN, DE, FR, ES, IT, TR, VI usw.; 3) Europa, Amerika, weltweit	C0 Controls & Basic Latin	0000 - 007F	94 graphische Zeichen (exklusive Leerzeichen)	C1 Controls & Latin-1 Supplement; Latin Extended-A, -B, -C und -D[27]
die arabische Schrift	1) Links & bidirektional; 2) AR, FA, UR 3) die arabische Welt, Westasien	Arabic	0600 - 06FF	254, darunter 64 für Buchstaben (28 Grundbuchstaben, manche modifizierte Schriftzeichen, diakritische, Zahl- & Hilfszeichen	Arabic Supplement; Arabic Extended-A; Arabic Extended-B
die Devanagari	1) Rechts; 2) HI, MR, NE; 3) Indien, Nepal	Devanagari	0900 - 097F	127, darunter 36 Konsonantenzeichen, 17 für unabhängige und 16 für abhängige Vokalzeichen	Devanagari Extended
die chinesische Schrift	1) untersch.; 2) ZH, JA, KO, VI (his.); 3) VR China, Taiwan, Singapur, Japan, Südkorea, Nordkorea (his.), Vietnam (his.)	CJK Unified Ideographs	4E00 - 9FCC	20.941 (inklusive der gebräuchlichen traditionellen sowie vereinfachten chinesischen, sinojapanischen und sinokoreanischen Schriftzeichen)	CJK Extension-A, -B, -C, -D, -E und -F; CJK Compatibility Ideographs Supplement; CJK Radicals

[26] Die drei Attributen einer Schrift werden wie folgt angegeben: a) Schriftrichtung (rechtsläufig, linksläufig, unterschiedlich oder senkrecht); b) die Sprachen, die in dieser Schrift niedergeschrieben werden (Abkürzungen der Sprachen werden nach ISO 639-3 angegeben); c) die Verwendungsgebiete dieser Schrift.

[27] Das ‚Latin-1' umfasst die variierten Buchstaben aus Mittel- und Westeuropa; ‚Latin Extended-A' sammelt die anderen Sonderbuchstaben aus Europa. Die Blocks Latin Extended-B, -C und -D dienen vor allem für historische und nichteuropäische Verwendungen des lateinischen Alphabets.

Schrift	**Attribut:** 1) Richtung 2) Sprache 3) Gebiet[26]	**Hauptkategorie mit Blockname, Umfang & Zeicheninventar**			**Erweiterte Kategorien**
die Kana	1) untersch.; 2) JA; 3) Japan	Hiragana; Katakana	3040 - 309F; 30A0 - 30FF	84 Hiragana-Zeichen von 93 Zeichen; 90 Katakana-Zeichen von 96 Zeichen	Kana Supplement
das Hangul	1) untersch.; 2) KO; 3) Süd- und Nordkorea	Hangul syllables	AC00 - D7AF	11.252 Syllabare	Hangul Compatibility Jamo; Hangul Jamo Extended-A; Hangul Jamo Extended-B

Tab. 1-2: Zeichencodierung verschiedener Schriften im Unicode[28]

1.4.3 Zeichen-Tasten-Repräsentation und Inputcodierung

In Kapitel 1.1 habe ich knapp die Funktionsweise der Tastatur und in Kap. 1.2.1 die Geschichte der Tastatur im Zeitalter der Schreibmaschine vorgestellt. Im letzten Kapitel wurde die allgemeine Situation der Zeichencodierung im Unicode eingeführt. Wenn die gebrauchten graphischen Zeichen eines Schriftsystems weniger sind als die, die die alphanumerischen Tasten einer PC-Standardtastatur mithilfe von Umschalttasten (Shift, AltGr, Shift + AltGr und Caps Lock) insgesamt repräsentieren können[29], so ist das Eingabeverfahren vor allem von der Zeichen-Tasten-Repräsentation abhängig. Wenn ein Schriftsystem diese Bedingung nicht erfüllt, müssen spezielle Softwares und in vielen Fällen auch Eingabeschemata entwickelt werden, um die Konversion von Inputcodes zu Schriftzeichen zu verwirklichen.

Bei den meisten alphabetischen Schriftsystemen wurde das Tastaturlayout von der Schreibmaschine auf den Computer mit wenigen Änderungen übernommen. Typisch dafür ist das amerikanische Standardtastaturlayout, bei dem die 94 sichtbaren Zeichen (exklusive des Spatiums) mithilfe von doppelter Umschaltung bei 47 zeichentragenden Tasten belegt werden (wie Tab. 1-3). Die Erforschung der Tastaturen für alphabetische Schriftsysteme bezieht sich vor allem darauf, die Zeichen-Tasten-Repräsentation zu nationalisieren und eine höhere Schreibeffizienz zu gewährleisten. Die chinesische und japanische Schreibmaschine mit tausenden Tasten jedoch könnte kaum mit einem alltagstauglichen PC verbunden werden. Diese Einschränkung beim Eingabegerät bedeutet, dass das tastenbelegungsbasierte Eingabeverfahren nicht zum Schreiben der beiden Sprachen eingesetzt werden kann. Die computergestützten Verarbeitungstechniken ermöglichen aber mithilfe mehrmaliger Encodierung und Decodierung die Zeicheneingabe durch eine Art der Sondercodierung, die mit dem US-amerikanischen Tastaturlayout repräsentierbar ist und als Inputcodierung bezeichnet wird.

[28] Nach Unicode 12.0 Character: U+0000-U+D7AF.
[29] Im höchsten Fall mit vollständiger fünf-maliger Belegung insgesamt 240 (48×5, exkl. Leerzeichen)

Scan-Code	Ohne Umschaltung		Umschaltung mit Shift		Scan-Code	Ohne Umschaltung		Umschaltung mit Shift	
	Form	ASCII	Form	ASCII		Form	ASCII	Form	ASCII
#1	`	60	~	7E	**#28**	]	5D	}	7D
#2	1	31	!	21	**#31**	a	61	A	41
#3	2	32	@	40	**#32**	s	73	S	53
#4	3	33	#	23	**#33**	d	64	D	44
#5	4	34	$	24	**#34**	f	66	F	46
#6	5	35	%	25	**#35**	g	67	G	47
#7	6	36	^	5E	**#36**	h	68	H	48
#8	7	37	&	26	**#37**	j	6A	J	4A
#9	8	38	*	2A	**#38**	k	6B	K	4B
#10	9	39	(	28	**#39**	l	6C	L	4C
#11	0	30	)	29	**#40**	;	3B	:	3A
#12	-	2D	_	5F	**#41**	'	27	"	22
#13	=	3D	+	2B	**#42**	\	5C	|	7C
#17	q	71	Q	51	**#45**				
#18	w	77	W	57	**#46**	z	7A	Z	5A
#19	e	65	E	45	**#47**	x	78	X	58
#20	r	73	R	52	**#48**	c	63	C	43
#21	t	74	T	54	**#49**	v	76	V	56
#22	y	79	Y	59	**#50**	b	62	B	42
#23	u	75	U	55	**#51**	n	6E	N	4E
#24	i	69	I	49	**#52**	m	6D	M	4D
#25	o	6F	O	4F	**#53**	,	2C	<	3C
#26	p	70	P	50	**#54**	.	2E	>	3E
#27	[	5B	{	7B	**#55**	/	2F	?	3F

Tab. 1-3: Die Belegung des US-amerikanischen Standardtastaturlayouts[30]

In Festlandchina ist die Standardtastatur mit QWERTY-Belegung zur Textverarbeitung am verbreitetsten. Allgemein betrachtet liegt der Schwerpunkt der Eingabe der chinesischen Schrift in der Entwicklung der Eingabesoftware und der Inputcodierungsentwürfe, was im Kontrast zur Eingabe der typischen Alphabetschriften steht.

Der Aufbau der Computertastaturen ist trotz der Vielfältigkeit der Schriften im Prinzip ähnlich. Die Abweichung liegt hauptsächlich in der unterschiedlichen Repräsentation einzelner Tasten (vgl. Greulich 2003: 877). Bei der Verarbeitung der diakritischen Zeichen und anderen kombinierenden Zeichen gibt es ebenso bei verschiedenen nationalen Tastaturlayouts drei verschiedene Varianten.

1) Die erste Möglichkeit ist die Belegung eines mit diakritischen Zeichen gebildeten Buchstaben, wie <ä>, <ö> und <ü> beim deutschen Tastaturlayout.
2) Die zweite Möglichkeit ist die Repräsentation der diakritischen Zeichen mit toter Taste. Sie ist die „Taste, die einen Buchstaben mit einem Akzent versieht. […] Wenn erst nach oder

[30] Vgl. Abb. 1-2, S. 10.

mit ihr eine weitere Taste gedrückt wird, erzeugt sie das gewünschte Zeichen“ (ibid.: 903). Die diakritischen Zeichen ‚Akut‘, ‚Gravis‘ und ‚Zirkumflex‘ werden im deutschen Tastaturlayout als tote Taste belegt. In diesem Fall wird ein Zeichen zwar bei der Eingabe zerlegt codiert, bei dem Datenaustausch und der Textverarbeitung aber als ein gesamter Zeichencode behandelt (wie die Eingabe von ‚á‘, siehe Kap. 2.1.3).

3) Bei der dritten Möglichkeit werden die kombinierenden Zeichen auf der Tastatur belegt und verarbeitet. Sie haben einen eigenständigen Codewert und Glyphen und werden bei der Eingabe immer nachfolgend von einem Basiszeichen eingetippt.

Die ersten zwei Möglichkeiten werden meistens bei den diakritischen Zeichen eines alphabetischen Schriftsystems mit begrenztem Zeichenvorrat angewendet. Die letzte Variante ist hauptsächlich für die Tonzeichen eines alphabetischen Schriftsystems (wie das vietnamesische Eingabeverfahren, siehe Kap. 2.2) und die von Basiszeichen abhängigen Vokalzeichen der Konsonanten- und der alphasyllabischen Schriften zuständig (wie das Hindi-Eingabeverfahren, siehe Kap. 2.3).

Ein Computer kann im Prinzip für die Verarbeitung aller Schriftsysteme universal sein, aber eine mechanische Tastatur ist auf ein bestimmtes Schriftsystem eingeschränkt. In diesem Fall lassen sich die Begriffe *mechanisches*, *visuelles* und *funktionales Tastaturlayout* unterscheiden. Ein mechanisches Tastaturlayout wird auf der sichtbaren und berührbaren Tastatur verzeichnet und bleibt unverändert. Das visuelle Tastaturlayout bezieht sich auf die Ausstattung der zusätzlichen Zeichenangabe. Das funktionale Layout ist von der Software bestimmt und kann unabhängig vom mechanischen im System beliebig umgestellt und eingesetzt werden (vgl. Daley 2016: keyboard layout). Ohne die sichtbare Tastaturbelegung aber direkt nachschlagen zu können, ist es für PC-Benutzer eine Herausforderung, die Position jedes Zeichens auswendig zu lernen. Für die Lösung dieses Problems wurden Eingabesoftwares für die nicht-lateinalphabetischen Schriftsysteme entwickelt, die mittels lateinischer Transkription ablaufen.

Die daraus zu Tage tretenden Unterschiede sollen abschließend anhand eines exemplarischen Vergleichs der Eingabefunktionsweisen des Englischen (als Stellvertreter der alphabetischen Schriftsysteme) und des Chinesischen (als Stellvertreter der nicht-alphabetischen Schriftsysteme) in drei Punkten zusammengefasst werden: 1) mehr geistige Arbeit für die Inputcodierung von einem Schriftzeichen sowie für die Auswahl unter den Kandidaten von einem PC-Benutzer, der beim Schreiben des Chinesischen benötigt wird; 2) höhere Anforderungen an die künstliche Intelligenz chinesischer Eingabeprogramme, um den Umgang mit dem Computer für Textverarbeitungen zu vereinfachen und die Effizienz des Schreibens zu

erhöhen; 3) Einschränkungen bei Schriftzeichen, die nicht eingegeben werden können. Denn die meisten chinesischen Eingabesoftwares verfügen über eine eingeschränkte chinesische Zeichendatenbank, die viel weniger Zeichen als das gesamte Inventar beinhaltet (vgl. Xu SC 1993: 158f). Manche seltenen Schriftzeichen können deswegen nicht abgerufen werden. Im Gegensatz dazu sind alle englischen Wörter durch 26 Buchstaben eingebbar.

1.4.4 Techniken zur typographischen Zeichenausgabe

Wie in Kap. 1.4.1 erwähnt, ist die Begründung der Glyphe und des Fonts die technische Grundlage für die Ausgabe schriftlicher Informationen per Computer. Ob das Ausgabegerät ein Zeichen in gewünschter Schriftart sowie -größe anzeigen kann, ist es von den gespeicherten Fonts im Computer abhängig. Die technischen Bedingungen für Fonts, vor allem der Speicherplatz und die Qualität von Monitor und Drucker, sind fest von der Anzahl und vom Aufbau der Schriftzeichen abhängig. Aus diesem Grund werden nachfolgend das englische und das chinesische Schriftsystem bei der Zeichenausgabe als Beispiele herangezogen und verglichen. Font unterscheidet sich zwischen Bitmap- und Vektorfont. Bei Bitmapfonts (auch Raster- oder Pixelfont genannt) wird eine Glyphe von bestimmter Schriftart und -größe in Bildpunkten (Pixeln) gespeichert. Demgegenüber wird bei Vektorfonts eine Graphik in bestimmtem Schriftstil mit Vektoren beschrieben und definiert (vgl. Greulich 2003: 358f).

Seit Entstehung der Textverarbeitung mit Computern geriet die chinesische Textverarbeitung lange Zeit wegen der Qualität des Monitors, des Druckers und des Speichermediums in Schwierigkeiten. Während ein lateinisches Zeichen im Minimum 7×9 px benötigt, um angezeigt zu werden, liegt dieser Wert bei chinesischen Schriftzeichen mindestens bei 16×16 px. Die Menge der sichtbaren ASCII-Zeichen beträgt insgesamt 94 (exklusive des Leerzeichens). In China aber gibt es (trotz der sog. Schriftreformvereinfachung) laut der 1988 veröffentlichten Liste der gemeingebräuchlichen Schriftzeichen im modernen Chinesisch 7.000 Schriftzeichen, die im Mindesten von den Medien benötigt werden. Das heißt, dass allein im Basisniveau der chinesischen Textverarbeitung in nur einer Schriftart und -größe mindestens 303-mal mehr Speicherplatz wie bei englischen Bitmaps-Fonts erforderlich ist. Legt man die angegebenen Zahlen zugrunde, beträgt ein chinesisches Font von Song-Ti (Standarddruckschriftart) mindestens 224 KB (16×16×7000÷8; ein Pixel beträgt in Schwarzweiß ein Bit, wobei ein Byte 8 Bits entspricht). Nur die Arbeitsspeicher weniger Computer vor 1970 konnten dies leisten. Wenn chinesische Schriftzeichen außerhalb der gemeingebräuchlichen Zeichenliste in hoher Qualität oder für mehr Schriftarten der chinesischer Schrift gewünscht sind, könnte ein Hundertfaches des Speicherplatzes nötig sein. Das Verhältnis zwischen englischen und chine-

sischen Fonts könnte deswegen einen Faktor von über zweitausend erreichen (vgl. Xu SC 1993: 146ff). Des Weiteren müssen Bildschirme und Drucker mit relativ hoher Auflösung (1024×768 px bei Bildschirmen und 600 dpi bei Druckern) verwendet werden, um Schriftzeichen komplizierten Aufbaus leserlich auszugeben (vgl. Cai 2005: 211, Greulich 2003: 78). Diese hohe technische Voraussetzung war eines der wichtigsten Argumente vieler Experten, warum die chinesische Textverarbeitung mit Computern nicht möglich sei.

Um das Problem zu lösen, wurden Entwürfe hauptsächlich in drei Richtungen eingeführt: Verbesserung der Hardware (vor allem Speichermedium, Monitor und Drucker), Erforschung der Datenkompression sowie Entwicklung des Vektorfonts. In allen drei Richtungen hat man große Fortschritte erreicht. Die höchste Auflösung der Ausgabegeräte hat sich seit den 1970er Jahren vervielfacht. Bis 2013 hat ein im Haushalt verbreiteter Computermonitor mit Flüssigkristallanzeigen eine Auflösung von 1600×2560 px erreicht. Ein Drucker konnte 1200 dpi unterstützen (vgl. Gumm/Sommer 2013: 56f, Greulich 2003: 711f). Die Effizienz des Speichermediums hat sich rasant erhöht, während seine Kosten vielfach sanken. Die Schwierigkeiten der Eingabe wurden zugleich durch Entwicklung spezieller Software und Entwürfen der Inputcodierung gelöst. Der erste Computer, der die Textverarbeitung in der chinesischen Schrift unterstützte, entstand 1978 in Japan (vgl. Xu SC 1993: 155).[31] In den 80er und 90er Jahren ist die Textverarbeitung endgültig in Druckwesen, der Büroarbeit und im Alltagsleben in China angekommen. Einen bedeutenden Beitrag hat ein Projektteam der Universität Peking geleistet, das von dem Informatiker WANG Xuan (王选, 1937-2006) geleitet wurde. Mithilfe des von dem Team entwickelten Computerlasersatzsystems gelang es, den Speicherplatz der chinesischen Fonts ums 50fache zu komprimieren und blitzschnell zu dekomprimieren (vgl. ibid.: 148). Verschiedene Schriftarten, -größen, -stärken und -lagen konnten nun in einem Computer gespeichert und für Textverarbeitung angewendet werden.

Wegen der notwendigen großen Speicherkapazitäten behinderten Bitmapfonts ebenso die Textverarbeitung mehrerer Schriftsysteme und die Darstellung vielfältiger Schriftstile. Seit den 1990er Jahren wurden sie sukzessive von Vektorfonts ersetzt (vgl. Greulich 2003: 125). Das Grundprinzip der Vektorschrift ist, mit mathematischen Formeln vor allem die Beschreibung von Koordinaten, Winkeln und Radien sowie Umrissen und Flächen eines Fonts zu definieren. Mit diesen Techniken kann ein Schriftbild in beliebiger typographischer Größe von hoher Qualität ausgegeben und die Schriftschnitte künstlich kursiv oder fett erzeugt werden (vgl. ibid.: 359). Der größte Unterschied zwischen lateinischen und chinesischen Vektorfonts

[31] Da im modernen japanischen Schriftsystem nur ca. 2000 Kanji allgemeingebräuchlich sind, musste diese Technik für die chinesische Textverarbeitung signifikant weiterentwickelt werden.

ist, dass ein chinesisches Schriftzeichen meistens in einzelnen Strichen beschrieben und mit einer Folge von vektoriellen Linienzügen dargestellt wird. Diese Methode wird als Strichfont bezeichnet, das sich von dem Outlinefont bei einem Buchstaben unterscheidet. Mit dieser Technik können die chinesischen Fonts viel Speicherplatz sparen. In der westlichen Welt ist Outlinefont in den meisten Fällen synonym für Vektorfont (vgl. ibid. & Gao 1995: 42-47).

1.4.5 Allgemeine Zusammenhänge zwischen Computerlinguistik, Eingabeverfahren und Textverarbeitung

Es liegt auf der Hand, dass computerlinguistische Anwendungen für Arbeitsprozesse und Leistungsmerkmale der Textverarbeitung von großer Relevanz sind. Die theoretischen und praktischen computerlinguistischen Anwendungen für die Textverarbeitung unterscheiden sich erheblich bei verschiedenen Sprachen/Schriftsystemen, insbesondere was deren spezifischen Eingabeverfahren anbelangt. In diesem Abschnitt sollen die allgemeinen Zusammenhänge zwischen Computerlinguistik, Eingabeverfahren und Textverarbeitung knapp erläutert werden. Diese Erforschung wird in den darauffolgenden Kapiteln anhand bestimmter Schriftsysteme vertieft. Die Beziehungen zwischen den drei Objekten sind vor allem aus Perspektive der allgemeinen Textverarbeitung und den Eingabemethoden mancher Schriftsysteme zu betrachten.

1) Einsatz der computerlinguistischen Anwendungen zur Ergänzung des menschlichen Geistes beim Schreiben mit dem Computer.

Text ist eine Art der schriftsprachlich fixierten Erscheinungsformen der Sprache und zugleich eines der wichtigsten Forschungsobjekte der Computerlinguistik, die sich mit der Verarbeitung der natürlichen Sprachen mit Computern beschäftigt (vgl. Lobin 2010: 10). Bei einem Textverarbeitungsprogramm sind deswegen viele computerlinguistische und sprachtechnologische Methoden erforderlich, um Effektivität und Qualität des Schreibens zu verbessern. Eine wichtige computerlinguistische Anwendung ist bspw. die Funktion der Korrektursysteme, die die Rechtschreibung- und Grammatikfehler untersuchen.

Die in dem Textprogramm ‚Microsoft Word' eingesetzten Korrektursysteme bieten zusammengefasst Leistungen in zwei Punkten: a) Fehlermarkierung einer Zeichenkette, die im maschinenlesbaren Wörterbuch einer bestimmten Sprache nicht eingetragen ist, und Suche nach dem ähnlichsten String; b) Fehleranzeige bei Flexion und Derivationsaffixen, die mithilfe morphologischer Beschreibungsweisen, bspw. der Zwei-Ebenen-Morphologie, ermöglicht wird (vgl. Fliedner 2010: 556ff).

Korrektursysteme des Chinesischen stehen bei ‚Microsoft Word' derzeit nicht zur Verfügung. Die große Herausforderung ist durch die Besonderheiten des Sprach- und Schriftaufbaus bedingt. Im Chinesischen entstehen die häufigsten Schreibfehler bei einzelnen Schriftzeichen im Sprachgebrauch, anders als in einer indogermanischen Sprache, in der orthographische und Flexionsfehler überwiegen. So müssen chinesische Korrektursysteme dazu fähig sein, Wortsegmentation und POS-Tagging mit relativ hoher Korrektheit durchzuführen und die Regeln für Wortbildung und Pragmatik jedes gemeingebräuchlichen Schriftzeichens zu ‚beherrschen'. Das ist zurzeit sowohl auf linguistischer als auch technischer Ebene problematisch zu verwirklichen (vgl. Luo / Luo / Gong 2004: 244f).

2) Computerlinguistische Anwendungen als technische Grundlage der Eingabemethoden des chinesischen und japanischen Schriftsystems.

Bei den heute am meisten verwendeten Eingabemethoden des Chinesischen und Japanischen werden die auf der Phonetik basierten Inputcodierungen am meisten gebraucht. Einer der größten Nachteile von solchen Methoden ist die große Menge von Homophonen. Um die Kandidaten zu reduzieren und die Schreibgeschwindigkeit zu beschleunigen, wird der Inputcode in den meisten Fällen im Kontext eingegeben, als Wort, Phrase, Satzteil oder Satz. Zeichenketten mit höherer Häufigkeit werden auch in der Wahlliste vorne angezeigt. Die nicht im Lexikon existenten, aber oft geschriebenen Wörter oder Zeichenketten werden automatisch eingetragen. So werden bei einer solchen Eingabesoftware viele computerlinguistische Methoden angewendet, wie z.B. Silben- sowie Wortsegmentation, Häufigkeitserrechnung von Zeichen/Wörtern und die Wahrscheinlichkeit einer festgelegten Zeichen- sowie Wortkette, Satzgeneration anhand Inputcodes usw. (vgl. Feng 2001: 5, Liu ZY/Wu/Li W 2008: 156).

Wegen der gehäuften Schwierigkeiten bei der Eingabe sind Erforschung und Entwicklung von Eingabemethoden in Ostasien wichtige Ansätze der Computerlinguistik. Diese hat in China zwei spezielle Teilbereiche: die Informationsverarbeitung des Chinesischen (eng.: Chinese information processing, chi.: 中文信息处理) und die Informationsverarbeitung der chinesischen Schrift (Chinese character information processing, 汉字信息处理). Informationsverarbeitung des Chinesischen meint „die Verarbeitung der Phonetik, Schriftform und Bedeutung der chinesischen Sprache" (GB 12200.1-90: Kap. 4.1.1.2 [Übersetzung der Verfasserin]). Informationsverarbeitung der chinesischen Schrift indes lässt sich als „die Manipulation und Bearbeitung der chinesischen Schrift mittels Computer, bspw. durch Eingabe, Ausgabe und Erkennung der chinesischen Schrift" definieren (ibid.: Kap. 4.1.1.3). Zusammengefasst liegt der Schwerpunkt der Informationsverarbeitung der chinesischen Schrift auf der Verarbeitung der Schriftzeichen, inklusive der Codierung, Eingabe, Ausgabe, Speicherung, Redaktion und

Distribution etc. (vgl. Sheng 2006: 78). Sie gilt für den gesamten CJK-Schriftkreis und hängt nur wenig mit sprachlichen Kenntnissen zusammen. Hingegen ist Informationsverarbeitung des Chinesischen bei der chinesischen Sprache eingeschränkt und umfasst die Textverarbeitung auf der Ebene des Schriftzeichens, des Wortes und des Satzes. Selbiges gilt für Informationsverarbeitung des Japanischen und des Koreanischen. Die modernen intelligenten Eingabemethoden des Chinesischen und des Japanischen, in denen Verarbeitungen in der Einheit von Sätzen möglich sind, hängen mit beiden Teilbereichen zusammen.

Die Computerlinguistik fand ihren Ursprung zuerst in Europa und den USA, weshalb zunächst indogermanische Sprachen erforscht wurden. Wegen der Verschiedenheit der menschlichen Sprachen und Schriften sind die computerlinguistischen Theorien und Anwendungen fest von der einzelnen Sprache und deren jeweiligen Schriftsystemen abhängig. Es ist für Computerlinguisten somit eine wichtige Aufgabe, die Computerlinguistik anhand der Besonderheiten einzelner natürlicher Sprache zu erforschen und anzuwenden. Ein internationaler wissenschaftlicher Austausch ist hierfür unerlässlich. Ein Ziel dieser Dissertation ist es daher, anhand der Eingabemethoden die Aspekte der Computerlinguistik aus Ostasien vorzustellen, zu analysieren und mit denen der indogermanischen Sprachen zu vergleichen.

2 Eingabeverfahren der alphabetischen Schriftsysteme

In der Einleitung wurden die tastaturbasierten Eingabeverfahren in zwei Gruppen unterteilt: die alphabetischen bzw. logographischen/syllabischen. Ziel des folgenden Abschnittes ist, die Eingabemöglichkeiten der verschiedenen alphabetischen Schriftsysteme zu erforschen. Dazu gehören zunächst die prinzipiellen alphabetischen, sprich die voll- und konsonantenalphabetischen Schriftsysteme. Zugleich sind die Eingabemöglichkeiten der zwischen Alphabet und Syllabar liegenden Schriften (namentlich Alphasyllabar und Hangul) Forschungsgegenstände des Kapitels.

Um diese Forschungsziele zu erreichen, werden fünf verschiedene Schriftsysteme anhand exemplarischer Beispiele erforscht: 1) das deutsche Eingabeverfahren als Vertreter für die auf dem lateinischen Alphabet basierten europäischen Schriftsysteme (Kap. 2.1); 2) das vietnamesische Eingabeverfahren als Beispiel für die alphabetischen Schriftsysteme mit vielen diakritischen Zeichen (Kap. 2.2); 3) das Hindi-Eingabeverfahren als Vorbild von Eingabemöglichkeiten der alphabsyllabischen Schriften in Südasien (Kap. 2.3); 4) das koreanische Eingabeverfahren als Ausnahmefall, da die Konversion von Buchstabenfolgen zu Silbenzeichen obligatorisch ist (Kap. 2.4); und 5) das arabische Eingabeverfahren als Beispiel der Textverarbeitung in linksläufiger Schriftrichtung (Kap. 2.5). In den Analysen jedes Schriftsystems gibt es vor allem sechs Leitfragen: a) die Grundschrifteigenschaften, b) das Zeicheninventar, c) die Zeichen-Tasten-Repräsentation, d) die Zeichencodierungsmöglichkeiten, e) die Zeichenausgabe und f) sonstige Herausforderungen bei der Textverarbeitung. Die Forschungsergebnisse werden in Kap. 2.6 zusammengefasst.

2.1 Eingabeverfahren des deutschen Schriftsystems

In Kap. 1.4.3 wurde das US-amerikanische QWERTY-Layout vorgestellt, das im Allgemeinen als Grundlage für die Zeichen-Tasten-Repräsentation der lateinalphabetischen Schriftsysteme gilt. Wegen der zusätzlichen Schriftzeichen außerhalb von ASCII und der Abweichung der Buchstaben- sowie Bigrammhäufigkeit wurde der internationale Standard für ein bestimmtes Schriftsystem in Details reformiert. Ziele von Kap. 2.1 sind, das QWERTY-Layout sowie seine Varianten anhand seines Einsatzes für bestimmte Schriftsysteme zu evaluieren. Das deutsche Schriftsystem und sein Standardlayout werden dafür zunächst exemplarisch erforscht. Danach wird in Kap. 2.1.4 das QWERTY-Layout anhand kontrastiver Analysen von Zeichen- und Bigrammhäufigkeit in vier unterschiedlichen lateinischen Schriftsystemen bewertet.

2.1.1 Untersuchung des deutschen Schriftsystems

Das deutsche Schriftsystem ist ein morphophonemisches Schriftsystem (siehe auch Kap. 1.3.5 [S. 34]). Anders ausgedrückt ist das deutsche Schriftsystem ein relativ „tiefes System“, das sowohl auf Graphem-Phonem-Korrespondenzen als auch auf „morphologische[n] Regularitäten“ basiert (Dürscheid 2006: 139). Zur Erforschung des deutschen Tastaturlayouts wird das deutsche Schriftsystem aus folgenden Blickwinkeln betrachtet: a) dem Zeicheninventar, inklusive der Buchstaben der Fremdwörter, Interpunktions- und Sonderzeichen; b) häufigen Buchstabenverbindungen, die durch festgelegte Grapheme, Diphthonge, Konsonantencluster, Morpheme usw. bedingt sind.

Neben den 26 Grundbuchstaben des lateinischen Alphabets gehören zusätzlich vier variierte Buchstaben zum deutschen Schriftsystem: die drei mit Umlaut gebildeten Buchstaben <Ä/ä>, <Ö/ö> und <Ü/ü> und das von Ligaturen stammende Schriftzeichen ‚Eszett‘ in Minuskel <ß> (ausgenommen ist hierbei das Schweizerdeutsche).[32] Aufgrund von Fremdwörtern treten in deutschen Texten in manchen Fällen auch Sonderbuchstaben auf, die mit einem Grundbuchstaben und einem diakritischen Zeichen (Gravis, Akut oder Zirkumflex) gebildet werden. Sie sind theoretisch nicht dem deutschen Schriftsystem zugehörig, z.B. <é> in *Café* und <à> in der Redewendung *eine Kiste à zehn Stück*. Die Eingabe solcher Zeichen anhand einer deutschen Tastatur ist daher gefordert.

Nach der DIN-Norm ‚32743-8‘ für den Schriftzeichenvorrat umfassen die mit der deutschen Standardtastaturlayout (DIN 2137) eingebbaren Zeichen das nachstehende Inventar. Dabei gehören die sonstigen Schriftzeichen unter 1.4 nicht zum generellen Vorrat. (Sie werden im deutschen Eingabeverfahren unter Verwendung einer toten Taste erzeugt.)[33]

[32] Die Esszett-Majuskel ‚ß‘ gehört nicht zu dem deutschen Schriftsystem und wird nur in seltenen Fällen gebraucht, z.B. in ‚GIEßENER ZEITUNG‘; es wird deswegen nicht im deutschen Tastaturlayout belegt.

[33] Aufgrund ihres relativ geringeren Gebrauchs und der alternativen Möglichkeit des Umschreibens in Grundbuchstaben in vielen Fällen werden die Sonderbuchstaben mit Akut, Gravis und Zirkumflex an dieser Stelle nicht diskutiert.

2 Schriftzeichenvorrat

2.1 Schriftzeichen

2.1.1 Kleinbuchstaben

a bis z, ä, ö, ü, ß

2.1.2 Großbuchstaben

A bis Z, Ä, Ö, Ü

2.1.3 Ziffern

0 bis 9

2.1.4 Sonstige Schriftzeichen

2.1.4.1 mit Akut (´)

á, ć, é, ǵ, í, Í, ń, ó, ŕ, ś, ú, ý, ź

Á, Ć, É, Í, Ĺ, Ń, Ó, Ŕ, Ś, Ú, Ý, Ź

2.1.4.2 mit Gravis (`)

à, è, ì, ò, ù

À, È, Ì, Ò, Ù

2.1.4.3 Mit Zirkumflex (^)

â, ĉ, ê, ĝ, ĥ, î, ĵ, ô, ŝ, û, ŵ, ŷ

Â, Ĉ, Ê, Ĝ, Ĥ, Î, Ĵ, Ô, Ŝ, Û, Ŵ, Ŷ

2.2 Schriftsonderzeichen

Leerzeichen
, Komma
. Punkt
: Doppelpunkt
; Semikolon
! Ausrufzeichen
? Fragezeichen
" Anführungszeichen
- Mittestrich (Trennstrich, Bindestrich)
' Apostroph
% Prozentzeichen
² Exponent 2
³ Exponent 3
µ My-Zeichen
° Gradzeichen
& Zeichen für kommerzielles „und“
§ Paragraphzeichen
$ Dollarzeichen
+ Pluszeichen
* Sternzeichen
# Nummerzeichen
= Gleichzeichen
(Klammer auf
) Klammer zu
/ Schrägstrich
´ Akut
` Gravis
^ Zirkumflex
€ Eurozeichen

Abb. 2-1: DIN 32743, Teil 8, Endgeräte für die Textkommunikation; Nationaler Teletex-Schriftzeichenvorrat (DIN-Taschenbuch 210: 91)[34]

Exklusive der sonstigen Schriftzeichen (Kategorie 2.1.4) gibt es insgesamt 109 sichtbare Zeichen (inklusive des Leerzeichens) bei der deutschen Textverarbeitung; 14 mehr als die graphischen Zeichen von ASCII. Neben den sieben Zeichen der vier Sonderbuchstaben treten sieben Schriftsonderzeichen zusätzlich hinzu: <°> (Gradzeichen), <€> (Währungszeichen für Euro), <§> (Paragrafenzeichen), <µ> (Mikro-Zeichen), <²> (Symbol zum Quadrieren), <³> (Symbol der Kubik) und <´> (Akut). Außer solchen auf Tasten belegten Zeichen werden 32 (nur die Vokale) bis 59 (alle sonstigen Schriftzeichen unter 1.4 nach DIN 32743, Teil 8) mit den nachfolgenden Tastenkombinationen von einer toter Taste und dem entsprechenden Grundbuchstaben erzeugt. Wegen der 14 zusätzlichen Zeichen gibt es im deutschen Tastaturlayout im Ver-

[34] Neunter Teil des DIN 32 743 (‚Schriftzeichendarstellung bei Geräten mit eingeschränktem Zeichenvorrat‘) sind die sonstigen Schriftzeichen (Kategorie 2.1.4) nicht eingeschlossen. Im Normalfall können nicht alle Schriftzeichen dieser Art erzeugt werden. Die eintippbaren Zeichen umfassen normalerweise nur die von den Grundbuchstaben <a>, <e>, <i>, <o>, <u> mit Akut, Gravis sowie Zirkumflex zusammengesetzten Sonderbuchstaben sowie <Ý/ý>. 2002 wurde das Währungszeichen ‚€‘ dieser DIN-Norm hinzugefügt.

gleich zu dem US-amerikanischen vor allem drei Reformen beim Aufbau: 1) das Beifügen einer Taste zu der Tastatur (Scancode #45, zwischen dem linken Shift und ‚Y/Z'); 2) die Verwendung der Drittbelegung der Tasten, die bei der Umschaltung via ‚AltGr' erfolgt; 3) die Repräsentation der drei diakritischen Zeichen mit toten Tasten.

Zu Forschungsdesideraten der Graphemik werden Fremdwörter im Deutschen im Normalfall ausgeklammert (vgl. Dürscheid 2006: 139). Daher werden sie bei den nachfolgenden Untersuchungen des Kapitels nicht mehr betrachtet.

Die meisten Buchstaben entsprechen eigenständig einem bestimmten Graphem. Ausnahmen davon sind <C>, <Q>, <V>, <X> und <Y>. <C> und <Q> müssen zusammen mit anderen Buchstaben Grapheme bilden, wie <ch> <ck>, <sch> und <qu>. <V> kann sowohl als [f] (wie in <von>) wie auch als [v] (wie in <Vase>) ausgesprochen werden. Aus diesem Grund hat Eisenberg <V> nicht als ein Graphem definiert, sondern als Allograph für <F> oder <W>. <X> steht in der deutschen Sprache für die Lautfolge [ks], wie in <Hexe> oder <Nixe>. Der Buchstabe <Y> tritt in der deutschen Schrift nur bei Fremdwörtern oder deutschen Eigennamen (meistens in Form von <ay> oder <ey> für das Diphthong [ai]) auf (vgl. Eisenberg 1996a: 1452, Dürscheid 2006: 132ff). Außer solchen Einzelzeichengraphemen gibt es nach Eisenberg im deutschen Grapheninventar die vier Diagraphen <ie>, <ch>, <qu> und <pf> sowie den Trigraph: <sch> (vgl. Eisenberg 1989: 60).

Es gibt ca. 40 Phoneme im Hochdeutschen, aber 29 Grapheme (neun für Vokale und zwanzig für Konsonanten), weshalb nicht ein Graphem eins-zu-eins mit nur einem Phonem korrespondiert. So kann ein Graphem in verschiedenen Kontexten verschiedenen Phonemen entsprechen. Bspw. wird das Graphem <s> als [z] vor einem Vokal, als [s] im Silbenauslaut und als [ʃ] im Zusammenhang von <st> oder <sp> im Silbenanlaut ausgesprochen (vgl. Eisenberg 1996a: 1452f). Die Differenzierung der Homophone in der Schrift ist ein wichtiges Prinzip der deutschen Graphemik. Gleich ausgesprochene deutsche Wörter werden in den meisten Fällen mit Allographen unterschiedlich dargestellt, wie *das – dass*, *ist – isst*, *Saite – Seite*, *Ferse – Verse* etc. (vgl. Altmann 2010: 128).

Ein Diphthong ist eine „Vokalfolge mit Gleitbewegung der Artikulationsorgane, so dass sich auditiv zwei Lautphasen unterscheiden lassen [...]" (Bußmann 2002: 167). Ein Konsonantencluster ist eine Sequenz zweier oder mehrerer konsonantischer Laute, das sowohl von einem Buchstaben als auch von einer Zeichenfolge dargestellt werden kann (vgl. Unicode Glossary: Consonant Cluster). Im deutschen Schriftsystem wird die Repräsentation mit Buchstabenfolgen bevorzugt. Der Konsonantencluster [ks] bspw. kann als <x> (wie in *Hexe*), <ks> (wie in *Keks*) oder <chs> (wie in *Sechs*) geschrieben werden. Viele festgelegte Buchstaben-

folgen sind von Diphthongen und Konsonantencluster bedingt, wie Analysen des Silbenbaugesetztes verdeutlichen.

Das Grundschema der deutschen Sprechsilbe lautet ‚KKVKK' (vgl. Eisenberg 1989: 62; K steht für Konsonant und V für Einzelvokal oder Diphtong). In dem Schema können die ersten beiden und das letzte K weggelassen und die Folge VK (K für den vorletzten Konsonant) durch einen Diphthong oder Langvokal ersetzt werden, wie *da*, *Ei*, *nah*, *es* usw. Eine Silbe setzt sich aus zwei Teilen zusammen: Silbenansatz (auch Anfangsrand genannt, sie besteht maximal aus drei Phonemen und entspricht KK am Anfang des Schemas) und Silbenreim (Silbenkern + Endrand, in dem Kern befindet sich immer der Hauptvokal). Die Theorie der Sonorität ist ein wichtiger Punkt für die Analyse der Silbenbildung. Die Sonoritätshierarchie für das Deutsche fängt mit Plosiven wie [b], [p], [d], [t] oder [k] an, gefolgt von Frikativen (wie [s], [z], [ʃ]), Nasalen ([m], [n]) und Liquiden ([l], [r]). Sie endet als Höhepunkt mit dem Vokal (Silbenkern), der in dieser Hierarchie weiter nach hohen und niedrigen Vokalen unterteilt werden kann. Der Anfangs- und Endrand einer Silbe wird bevorzugt gebaut, da sie „zum Kern hin monoton zunimmt […]" (ibid.: 61).

Ein deutscher Diphthong befolgt fast immer das Silbenbaugesetz, das von einem höheren Vokal vorne und einem niedrigeren Vokal hinten zusammengesetzt ist. Manche häufigen Bigramme wie <au>, <ei>, <eu> und <äu> können so erklärt werden (vgl. ibid.: 67f, Altmann 2010: 143).

Das Silbenbaugesetz gilt ebenso meisten für Konsonantencluster. Die nach dem Silbenbaugesetz gebildeten Silbenränder werden als Standardränder bezeichnet, wie *Schrank*, *Knall*, *Traum*, *Tschüss* und *Wurst*. Das Gegenteil von Standardrändern (bzw. ein nicht nach der Sonoritätshierarchie gebauten Rand) wird als Nebenrand bezeichnet, z.B. im Silbenansatz <sp> [ʃp] und <st> [ʃt]. Nach Altmann sind die folgende Konsonantenverbindungen im Deutschen am häufigsten: <pf> – [pf], <z, ts> – [ts], <dsch, g> – [dʒ], <x, chs, cks, gs> – [ks], <qu> – [kv], <st> – [ʃt] sowie <sp> – [ʃp] (vgl. Eisenberg 1989: 61f, Altmann 2010: 150).

Morpheme werden im Allgemeinen als „kleinste bedeutungstragende Elemente einer Sprache [definiert], die als phonologisch-semantische Basiselemente nicht mehr in kleinere Elemente zerlegt werden können […]" (Bußmann 2002: 448). Die deutsche Sprache gehört zu den flektierenden Sprachen und ihr Schriftsystem basiert auf einem morphophonemischem Alphabet. Analog zu Sprach- sowie Schrifttypologie können viele häufige Buchstabenfolgen als festgelegte Morpheme analysiert werden.

Morpheme können in lexikalische (Morpheme mit lexikalischer Bedeutung, die die Wortbasis darstellen) und grammatische Morpheme (Morpheme, die grammatische Bedeu-

tungen tragen) klassifiziert werden (vgl. Kessel/Reimann 2010: 94). Nachfolgend werde ich wegen ihrer überwiegenden Häufigkeit die grammatischen Morpheme vorstellen, die sich nach Wortbildungs- und Flexionsmorphemen weiter untergliedern lassen.

Der Terminus *Wortbildungsmorphem* (auch Derivationsmorphem) bezieht sich in den meisten Fällen auf Wortbildungsaffixe, bei denen Präfix (wie *un-, ent-, ab-, zu-, wider-* ect.), Suffix (wie *-heit, -keit, -ung* und *-in* für Nomen und *-lich, -isch, -ig* und *-bar* für Adjektive) und Zirkumfix (wie *be-...-t, ver...-ern*) unterschieden werden (vgl. Plath 2014: 9). Außer den Affixen gibt es ebenso Fugenmorpheme, die als hinzugefügtes Element bei der Kompositabildung zwischen den Wortteilen gebraucht werden. In der deutschen Sprache sind dies vor allem *-e, -s, -es, -n, -en, -er* und *-ens* (vgl. ibid.: 11f).

Gegenüber den Wortbildungsmorphemen, die für die Bildung neuer Wörter (Derivation) zuständig sind, stehen die Flexionsmorpheme für grammatische Änderungen zur Verfügung. Sie umfassen Flexionssuffixe (wie *mach<u>en</u>, fröhlich<u>en</u>, Übung<u>en</u>*) und -zirkumfixe (wie *<u>ge-mach</u>t*). Nach den flektierbaren Wortarten und Deklinationsarten können bestimmte Flexionsaffixe aufgelistet werden. Die Flexion der Substantive ist von Numerus, Kasus und Genus abhängig. Die morphologischen Kategorien des Verbs sind Tempus, Genus (Aktiv oder Passiv), Modus, Numerus und Person. Adjektive sowie Teile der Adverbien sind nach Grad (Positiv, Komparativ und Superlativ) flektierbar. Daneben müssen grammatische Änderungen eines Adjektivs je nach Art von Kasus und Numerus der Nominalphrase sowie Genus der modifizierten Substantive durchgeführt werden. Folgende Flexionsmorpheme können nach Deklinationsarten einer Wortart angegeben werden: *-e, -es, -en, -er, -s, te, ge-...-t, -et, -end, -t, -st* etc. (vgl. Clément 2000: 40ff & 132ff).

2.1.2 Zeichen-Tasten-Repräsentation des deutschen Tastaturlayouts

Die deutschen Tastaturlayouts unterscheiden sich zwischen jenen Ländern, in denen Deutsch als einzige nationale Muttersprache (Deutschland und Österreich) gesprochen wird, und multilingualen Ländern, in denen das Deutsche eine Amtssprache unter mehreren ist (die Schweiz und Belgien). Wegen der Europäisierung und der Internationalisierung werden in solchen Ländern in manchen Fällen mehrsprachige Tastaturlayouts, die für mehrere Schriftsysteme zuständig sein können, angewendet, wie die Europa- und Expertentastatur. Zur Untersuchung des deutschen Eingabeverfahrens in dieser Dissertation nehme ich hauptsächlich das deutsche nationale Tastaturlayout als Forschungsobjekt (Normen T1 ‚DIN 2137-2', vgl. Abb. 2-2).

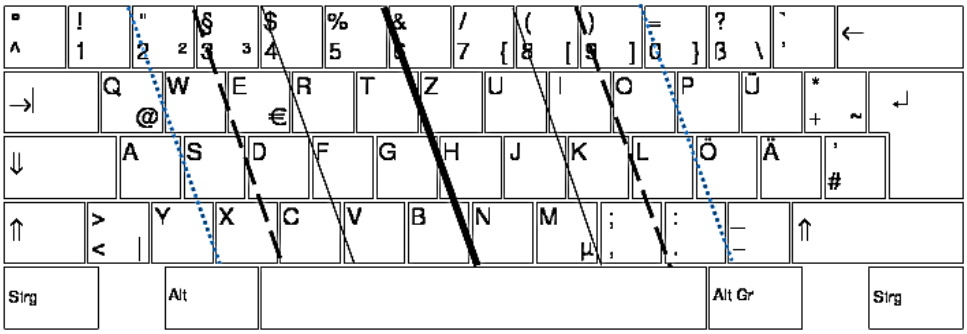


Abb. 2-2: Die deutsche Tastaturbelegung nach DIN 2137-2 (DIN-Taschenbuch 210: 21)[35]

Das deutsche QWERTZ-Tastaturlayout lässt sich auf die amerikanische QWERTY-Belegung nach Sholes im Jahr 1868 (vgl. S. 16 dieser Arbeit) und die Tastatur der deutschen Schreibmaschine zurückführen. Obwohl die QWERTY-Tastaturbelegung und ihre Varianten wiederholt kritisiert wurden, sind Versuche der Tastaturoptionalisierung wegen der ‚verwurzelten Eingewöhnung' der Schreibmaschinen- sowie PC-Benutzer meist misslungen. Nach Light und Anderson können die Kriterien für funktional hochwertige Tastaturlayouts in drei Komponenten zusammengefasst werden.

- The relative frequency of every letter pair for English […].
- The 'travel times' between positions on the keyboard.
- The mapping of the letters of the keys.

(Sproat 2010: 164)

Bei der Reform des QWERTY-Layouts für ein anderes lateinalphabetisches Schriftsystem wie das deutsche müssen weitere Faktoren berücksichtigt werden, wie z.B. die Einflüsse aus dem internationalen Standard für PC-Tastaturen, die Tastaturlayouts anderer Sprachen und die Schreibtraditionen seit Beginn der Schreibmaschinenära. Von einem einzelnen Schriftsystem abhängig sind Zeichenvorrat und Buchstabenhäufigkeit variiert und die Repräsentation der mit Diakritika gebildeten Sonderschriftzeichen unterscheidet sich ebenso erheblich, wie der Unterschied der Eingabe von <ü> und <á> mit einer deutschen Tastatur belegt. Aus Effektivitätsgründen ist es vorteilhaft, die häufigen Zeichen mit Zeige- sowie Mittelfinger erreichen zu können oder in der Mittelreihe der alphabetischen Tasten zu positionieren. Kriterien für ein rationales Tastaturlayout können anhand der Psychomotorik wie folgt genannt werden:

Möglichst häufiger (am besten rhythmischer) Handwechsel.
Belegung der 8 Grundtasten (Ausgangstasten) mit sehr häufigen Buchstaben.
Ungefähr gleichstarke Inanspruchnahme der Hände mit Bevorzugung der Rechten.

[35] Im 1993 erschienenen DIN-Taschenbuch ist das Normen inaktuell, deswegen wird diese Abbildung auf der Basis vom Normen im Taschenbuch nach dem aktuellen deutschen Tastaturlayout ergänzt, wie ‚€'

> Starke Belegung der Mittelreihe, dann Oberreihe, danach der Unterreihe.
> Entlastung des 4. und 5. Fingers durch Belegung der zugehörigen Tasten m. seltenen Buchstaben.
>
> Einschränkung der einander unmittelbar folgenden Anschläge des 3./4. u. 4./5. Fingers.
> Vermeidung von weiten Spanngriffen, sowohl auswärts wie einwärts.
> Vermeidung von Sprunggriffen zwischen Ober= und Untertastenreihe und umgekehrt.
> Vermeidung des wandernden Fingers, der unmittelbar nacheinander zwei verschiedene (Tasten anschlägt.
> (Meier 1978: 339)

Einzelne Tastenanschläge werden nach Position und Arbeitsstärke der Hand sowie Finger bewertet: Grundreihe > Oberreihe > Unterreihe; Zeigefinger > Mittelfinger > Ringfinger > kleiner Finger; rechte Hand > linke Hand (für Linkshänder umgekehrt). Für die nachfolgenden Anschläge zweier Tasten gibt es hauptsächlich drei Klassen: Handwechsel (Kriterium 1), Fingerwechsel (inklusive Spann- und Sprunggriffen; Kriterium 7 und 8) und Fingerwandel (Kriterium 9). Anhand experimentaler Statistiken der Psychomotorik dauert der Handwechsel zweier Anschläge durchschnittlich 0,02 Sekunden, der Fingerwechsel einer Hand 0,03 S und der Fingerwandel 0,09 S (vgl. Wang YM 2005: 20).

Nach Meiers Kriterien und der Häufigkeitsforschung des deutschen Schriftsystems werde ich das Standardtastaturlayout evaluieren (vgl. unten stehend). Analog zu diesen Kriterien und Praktika hat Meier den Schwierigkeitsgrad des Fingerwechsels sowie -wanderns in acht Stufen unterteilt:

> Schwierigkeitsstufen:
>
> | | |
> |---|---|
> | 1.) Grundstellung | 2.) Normal-Hochgriffe |
> | 3.) Normal-Tiefgriffe | 4.) Sprunggriffe |
> | 5.) Innenspreizgriffe | 6.) Außenspreizgriffe |
> | 7.) Umschaltung | 8.) Griffe in der Zifferreihe |
>
> (Meier 1978: 340)[36]

Zu dieser Schwierigkeitshierarchie, die auf Basis der Schreibmaschinentastatur erstellt wurde, muss bezüglich Punkt 7.) ergänzt werden, dass die Schwierigkeit der Umschaltung mit ‚Shift' und ‚AltGr' nicht einer bestimmten Schwierigkeitsstufe zugeordnet werden kann, da die Umschaltung primär mit der ‚Shift'-Taste vollzogen wird. Sie wird am häufigsten mit dem klei-

[36] Zu Stufe 2) und 3) gehören sowohl Griffe zwischen Grund- und Ober-/Unterreihe, als auch die Fingerwechsel der nachfolgenden Anschläge in einer Reihe (Ober-/Unterreihe); Spreizgriff meint die Handspreizung wegen zwei der weit entfernten Tasten, darunter mindestens eine, die sich außerhalb der Grundtastenzone jeder Reihe befindet. Unterschieden werden Innenspreizgriff (wie ‚ge'), Außenspreizgriff (wie ‚mü') und beide betroffene Varianten (wie ‚hä').

nen Finger gedrückt und ihre Umschaltung kann durch Handwechsel oder Außenspreizgriffe erledigt werden. Die Taste ‚AltGr' hingegen wird nur rechts von der Leertaste belegt und normalerweise mit dem rechten Daumen erreicht. Dieser Griff ist ungleich schwerer, gehört der rechten Zone an (wie <~> im deutschen Tastaturlayout) und sollte daher tendenziell eher hinter Punkt 8.) folgen (Taste ohne Umschaltung der Ziffernreihe).

Um die Zusammenhänge zwischen Tastenbelegung und Buchstaben- sowie Bigrammhäufigkeit zu verdeutlichen, dient die unten stehende Tabelle (Tab. 2-1). Zu der statistischen Größe einzelner Buchstaben werden drei Statistiken für Häufigkeit und die Verteilungswahrscheinlichkeit eingeführt, so dass die Abweichung möglichst gering ausfällt. Zu jedem Buchstaben werden zudem drei seiner häufigsten Bigramme angegeben. Anhand dieses Vorgehens wird ersichtlich, welche anderen Buchstaben bei der Belegung eines Buchstabens am meisten berücksichtigt werden sollten. Alle zitierten und ausgerechneten Statistiken für Zeichenhäufigkeit in dieser Arbeit werden bis auf die zweite Nachkommastelle gerundet.

Rg.	**Buchstabe**	**Häufigkeit(1)**[37]	**Häufigkeit (2)**[38]	**Häufigkeit (3)**[39]	**Verteilung (Nx)**[40]	**Fingergriff**[41]	**drei häufigsten Bigramme**[42]		
1	E/e	16,65%	17,48%	17,50%	86	L_3_O	ER: 4,09%	EN: 4,01%	DE: 2,27%
2	N/n	10,36%	9,84%	9,80%	49	R_2_U	EN: 4,01%	ND: 1,87%	IN: 1,68%
3	I/i	8,14%	7,73%	7,70%	45	R_3_O	EI: 1,93%	IN: 1,68%	IE: 1,63%
4	R/r	7,94%	7,54%	7,50%	33	L_2_O	ER: 4,09%	RE: 1,12%	RA: 0,80%
5	S/s	5,57%	6,83%	6,80%	34	L_4_G	ES: 1,40%	ST: 1,16%	SE: 0,99%
6	T/t	5,43%	6,13%	6,10%	32	L_2_O	TE: 1,85%	ST: 1,16%	NT/TI: 0,59%
7	A/a	5,15%	6,47%	6,50%	25	L_5_G	AN: 1,02%	RA: 0,80%	AU: 0,75%
8	H/h	4,76%	4,23%	4,20%	26	R_2_G	CH: 2,42%	HE: 1,02%	HA: 0,70%
9	D/d	4,21%	4,83%	4,80%	29	L_3_G	DE: 2,27%	ND: 1,87%	DI: 0,93%
10	U/u	4,01%	4,17%	4,20%	19	R_2_O	UN: 1,19%	AU: 0,75%	RU/US: 0,48%
11	G/g	3,70%	3,06%	3,10%	20	L_2_G	GE: 1,47%	NG: 0,94%	EG: 0,50%
12	L/l	3,68%	3,49%	3,50%	21	R_4_G	LE: 0,65%	EL: 0,63%	LI: 0,62%
13	C/c	2,95%	2,68%	2,70%	15	L_3_U	CH: 2,42%	SC: 0,89%	IC: 0,76%
14	B/b	2,38%	1,93%	1,90%	6	L_2_U	BE: 1,01%	EB: 0,45%	AB: 0,31%
15	M/m	2,36%	2,58%	2,60%	9	R_2_U	ME/EM: 0,50%	MI: 0,44%	MA: 0,40%
16	F/f	2,25%	1,65%	1,70%	4	L_2_G	UF: 0,27%	EF: 0,26%	FE: 0,25%

[37] Quelle: Meier 1978: 334; mit Sonderbuchstabenberücksichtigung.
[38] Quelle: Bauer, FL 1997: 286; ohne Sonderbuchstabenberücksichtigung (jeweils als ae, oe, ue und ss umgeschrieben).
[39] Quelle: Pommerening 2007, zitiert von [2] nach F. W. Bauer; ohne Sonderbuchstaben-Berücksichtigung.
[40] Quelle: Best 2006: 74.
[41] Der Anfangsbuchstabe L/R steht jeweils für linke/rechte Hand, die mittige Ziffer für den jeweiligen Finger und der letzte Buchstabe für die Tastenreihe: Grund-/Mittel- (G), Ober- (O), Unter- (U) und Zifferreihe (Z).
[42] Basiert auf Statistiken von Pommerening 1999, zitiert von F. W. Bauer. Die Statistiken für <ä>, <ö>, <ü> und <ß> habe ich anhand einiger Zeitungsartikel mit 2.070 Wörtern und 12.345 Buchstaben erhoben. Sehr häufige Bigramme (über 1%) werden mit Unterstrich markiert.

Rg.	Buch-stabe	Häufig-keit(1)[37]	Häufig-keit (2)[38]	Häufig-keit (3)[39]	Verteil-ung (Nx)[40]	Finger-griff[41]	drei häufigsten Bigramme[42]		
17	O/o	2,25%	2,98%	3,00%	14	R_4_O	ON: 0,64%	OR: 0,50%	VO: 0,43%
18	W/w	1,58%	1,48%	1,50%	12	L_4_O	WE: 0,48%	WI: 0,36%	WA: 0,34%
19	K/k	1,41%	1,46%	1,50%	18	R_3_G	KA/KE: 0,26%	NK: 0,25%	KO: 0,24%
20	Z/z	1,20%	1,14%	1,10%	8	R_2_O	ZU: 0,43%	ZE: 0,28%	TZ: 0,26%
21	V/v	1,08%	0,94%	0,90%	5	L_2_U	VO: 0,43%	VE: 0,37%	NV: 0,18%
22	Ü/ü	0,79%			3	R_5_O	ÜR: 0,22%	FÜ: 0,19%	ÜN: 0,11%
23	P/p	0,73%	0,96%	1,00%	5	R_5_O	PR: 0,23%	SP: 0,22%	PA: 0,16%
24	Ä/ä	0,55%			3	R_5_G	ÄN: 0,15%	LÄ: 0,11%	RÄ: 0,08%
25	ß	0,32%			5	R_5_Z	Iß: 0,06%	Oß: 0,05%	Eß: 0,03%
26	Ö/ö	0,27%			1	R_5_G	ÖR: 0,09%	HÖ: 0,08%	ÖH: 0,06%
27	J/j	0,18%	0,27%	0,30%		R_2_G	JA/JE: 0,09%	JU/NJ: 0,05%	RJ: 0,04%
28	Q/q	0,04%	0,02%	0,00%		L_5_O	QU: 0,02%	EQ: 0,01%	0
29	X/x	0,03%	0,04%	0,00%		L_4_U	EX: 0,02%	XI/XO/XT: 0,01%	0
30	Y/y	0,03%	0,08%	0,01%		L_5_U	SY: 0,02%	YE/YA/YB: 0,01%	0

Tab. 2-1: Der Buchstabenhäufigkeits-Tastenbelegungs-Zusammenhang im Deutschen

Es ist unmöglich, die häufigsten Verbindungen von jedem Buchstaben gleichzeitig zu berücksichtigen. Zum Beispiel ist <be> das am meisten vorkommende Bigramm für <b>, aber für <e> gibt es zehn Verbindungsmöglichkeiten mit höherer Wahrscheinlichkeit. So ist es erforderlich, eine Rangliste von den häufigsten deutschen Bigrammen heranzuziehen (siehe Tab. 2-2). Da die Reihenfolge zweier Buchstaben bei nachfolgenden Anschlägen kaum ins Gewicht fällt, wird auch in dieser Tabelle die gesamte Häufigkeit zusammen mit der umgekehrten Buchstabenfolge berechnet.

Rang	Bigramm	Häufig-keit[43]	Typ der Folge[44]	Umg. Form & ihre Häufigkeit	Gesamt-Häufigkeit & Rg.	Schwierig-keitsstufe[45]
1	ER	4,09%	M.	RE: 1,12%	5,21%; 2	2
2	EN	4,01%	M.	NE: 1,22%	5,23%; 1	0
3	CH	2,42%	G., Teil G.	HC: 0,01%	2,43%; 5	0
4	DE	2,27%	M. & A.	ED: 0,51%	2,78%; 4	2w
5	EI	1,93%	D.	IE: 1,63%	3,56%; 3	0
6	ND	1,87%	KK.	DN: 0,06%	1,99%; 9	0
7	TE	1,85%	M. & A.	ET: 0,55%	2,40%; 6	5
8	IN	1,68%	M.	NI: 0,65%	2,33%; 8	4&5
9	IE	1,63%	G.	EI: 1,93%	3,56%; 3	0
10	GE	1,47%	M.	EG: 0,50%	1,97%; 10	2&5
11	ES	1,40%	M.	SE: 0,99%	2,39%; 7	2

[43] Quelle: Pommerening, 2007: Bigramme, zitiert von [2] nach F. W. Bauer.
[44] Vgl.: Kap. 2.1.1; M. - Morphem, G. - Graphem, D. - Diphthong und KK. - Verbindung zweier Konsonanten. A. steht für sonstige Verbindungen, die zu keiner erwähnten Typologie gehören.
[45] Handwechsel wird mit 0, die sonstige Stufen nach der von Meier genannten Schwierigkeitsstufe (S. 65f) markiert. Wenn Fingerwandern nötig ist, wird hinter der Nummer ein ‚w' zugefügt.

Rang	Bigramm	Häufigkeit[43]	Typ der Folge[44]	Umg. Form & ihre Häufigkeit	Gesamt-Häufigkeit & Rg.	Schwierigkeitsstufe[45]
12	NE	1,22%	A.	EN: 4,01%	5,23%; 1	0
13	UN	1,19%	M., Teil M. & A.	NU: 0,33%	1,52%; 14	4w&5w
14	ST	1,16%	KK.	TS: 0,50%	1,66%; 12	3&5
15	RE	1,12%	A.	ER: 4,09%	5,21%; 2	2
16	HE	1,02%	A.	EH: 0,57%	1,59%; 13	0
17	AN	1,02%	M. & A.	NA: 0,68%	1,70%; 11	0
18	BE	1,01%	M.	EB: 0,45%	1,46%; 15	4&5
19	SE	0,99%	A.	ES: 1,40%	2,39%; 7	2
20	NG	0,94%	G., Teil M.	GN: 0,05%	0,99%; 19	0
21	DI	0,93%	A.	ID: 0,20%	1,13%; 18	0
22	SC	0,89%	Teil G.	CS: 0,01%	0,90%; 20	3
23	IS	0,79%	A.	SI: 0,65%	1,44%; 16	0
24	IT	0,78%	A.	TI: 0,59%	1,37%; 17	0
25	IC	0,76%	A.	CI: 0,01%	0,77%; 21	0

Tab. 2-2: Die 25 häufigsten Bigramme im Deutschen[46]

Nach den angegebenen Kriterien und Statistiken funktioniert das deutsche Standardtastaturlayout nicht hochwertig. Bspw. können folgende irrationale Belegungen genannt werden.

1) Aspekt der Buchstabenhäufigkeit: Die beiden häufigsten Buchstaben <E> und <N> werden in der Oberreihe mit Mittelfinger und in der Unterreihe mit Zeigefinger erreicht, aber nicht als eine Grundtaste belegt. Im Gegenteil: der viertseltenste Buchstabe <J> (0,18%) wird mit dem Zeigefinger in der Grundstellung des Zehnfingersystems gegriffen.

2) Aspekt der Bigrammhäufigkeit: Manche relativ häufigen Bigramme wie ‚TE', ‚IN', ‚UN' gehören zu den Kriterien, die nach Meier vermieden werden sollten (vgl. S. 65f). Nur die Hälfte der häufigsten Bigramme wird mit Handwechsel angeschlagen. Zwei davon erfordern wandernde Finger, deren durchschnittliche Dauer das 4,5-fache des Handwechsels beträgt.

3) Aspekt der Arbeitsverteilung beider Hände: Wenn die Beanspruchung beider Hände für die Buchstabenschläge berechnet wird, ist die linke Hand für 58,96% der Buchstabenschläge zuständig, hingegen nur 41,04% von der rechten Hand. Bei der Analyse der acht Grundtasten einer deutschen Standardtastatur ist die Verteilung noch eindeutiger. So werden mit der linken Hand 17,18% der Buchstaben erreicht (A, S, D und F), entgegen 5,27% mit der rechten (J, K, L und Ö) (vgl. Meier 1978: 342). Die Arbeitsverteilung der beiden Hände ist damit entgegengesetzt, da die Mehrheit zu den Rechtshändern zählt.

[46] Die hellgrau markierten Bigramme sind die umgekehrte Form eines oben angegeben Bigramms; Wenn die Bigramme aus denselben Buchstaben in beiden Formen als ein Bigramm betrachtet werden, werden insgesamt 21 Bigramme aufgelistet.

4) Aspekt der Beanspruchung der Finger: Nach der obigen Rechnung sind die beiden Zeigefinger für zwölf Buchstaben mit einer Häufigkeit von 45,65% zuständig, die Mittelfinger für fünf Buchstaben (33,36%), die Ringfinger für fünf Buchstaben (13,11%) und die kleinen Finger für acht Buchstaben (7,88%). Im Normalfall sind Zeige- sowie Mittelfinger am stärksten und die kleinen Finger am schwächsten beim Eintippen. Genau diese abfallende Hierarchie tritt auch bei den Statistiken zu Tage. Ein auffälliger Schwachpunkt liegt jedoch bei der Belastung der kleinen Finger mit exponentiell mehr Tasten, als die Stärke/Schwäche des Fingers vermuten ließe. Zum rechten kleinen Finger allein werden fünf Buchstaben (ß, P, Ü, Ö und Ä), vier weitere Zeichen und häufig auch die rechten Funktionstasten definiert, d.h. dass er trotz geringer Häufigkeit relativ viel wandern muss. Im Vergleich dazu werden dem stärkeren Mittelfinger nur drei zeichentragende Tasten (E, D und C bei dem linken und I, K und Komma beim rechten) zugeordnet.
5) Aspekt der Beanspruchung der drei Reihen: Anhand statistischer Analysen sind der Grundreihe 31,73% der Buchstaben zugeordnet, wobei auf die Oberreihe 48,76% und auf die Unterreihe 19,19% entfallen. Die Statistiken stehen im Kontrast zum Grundprinzip der Tastaturbelegung, nach dem im Idealfall die Grundreihe am stärksten zu belegen ist.[47]

Nach dem Vergleich mit dem internationalen US-amerikanischen Layout ist die Tastenposition der 24 Grundbuchstaben bei der deutsche Variante (‚QWERTZ-Belegung') unverändert. Der einzige Unterschied bei den Grundbuchstaben, der Umtausch zwischen <Y> und <Z>, hat vor dem Hintergrund der Zeichenhäufigkeit Vorzüge. Zunächst kommt <Z> (1,20%) in der deutschen Schrift öfter vor als <Y> (0,03%). Wegen der Vertauschung wird der Griff zur Z-Taste beschleunigt, der mit dem Zeigefinger in die Oberreihe führt. Zweitens ist <Z> mit <U> und <T> benachbart, mit denen er am häufigsten zusammen auftritt.

Neben den Buchstaben muss auch die Belegung der Sonderzeichen auf Ebene der Häufigkeit erforscht werden. Außer dem Leerzeichen kann die Repräsentation der Sonderzeichen im Allgemeinen in drei Stufen klassifiziert werden: Eingabe ohne Umschaltung (erste Belegung, für häufige Zeichen), mit ‚Shift' (zweite Belegung, für weniger häufige Zeichen) und mit ‚AltGr' (Abkürzung für alternative Graphik, dritte Belegung, für seltene Zeichen). Mithilfe der ‚Java Letter Frequency' habe ich anhand eines selbst erstellten Textkorpus die Häufigkeit von Interpunktions-, Wort- sowie Sonderschriftzeichen errechnet (vgl. Tab. 2-3).

Zeichen	Häufigkeit	Auftreten	Belegung	Zeichen	Häufigkeit	Auftreten	Belegung
.	0,42%	1264	#54	]	0,01%	24	#10+AG
,	0,32%	942	#53	!	0,01%	17	#02+SH

[47] Die Statistiken der vier Aspekte basieren auf den Statistiken von Meier, angegeben in Spalte 3 von Tab. 2-1.

Zeichen	Häufigkeit	Auftreten	Belegung	Zeichen	Häufigkeit	Auftreten	Belegung
-	0,14%	410	#55	|	0,00%	11	#45+AG
"	0,08%	250	#02+SH	+	0,00%	9	#28
)	0,07%	219	#10+SH	=	0,00%	9	#11+SH
:	0,06%	183	#54+SH	&	0,00%	6	#07+SH
(	0,05%	155	#09+SH	'	0,00%	4	#42+SH
<	0,02%	48	#45	$	0,00%	3	#05+SH
>	0,02%	48	#45+SH	€	0,00%	3	#19+AG
/	0,01%	39	#08+SH	%	0,00%	2	#06+SH
?	0,01%	37	#12+SH	*	0,00%	2	#28+SH
;	0,01%	35	#53+SH	{	0,00%	1	#08+AG
[	0,01%	24	#09+AG				

Tab. 2-3: Die Häufigkeit der Sonderzeichen in der deutschen Schriftsprache[48]

Die Häufigkeit der Sonderzeichen stimmt mit ihrer Tastenbelegung im Allgemeinen überein. Punkt, Komma und Bindestrich mit überwiegender Häufigkeit werden mit den Tasten einer alphabetischen Reihe ohne Umschaltung eingegeben. Bei der dritten Belegung werden nur die Sonderzeichen mit geringer Häufigkeit (nach der Berechnung des Korpus weniger als 0,01%) definiert.

Nach den fünf genannten Aspekten für irrationale Belegung (siehe S. 69f) ist das deutsche Standardtastaturlayout unvollkommen, eine Rationalisierung des Layouts ist aber kritisch und wurde trotz mehrmaliger Verbesserungsvorschläge schwer durchgeführt. Die folgenden drei Gründe könnten die Ursache dafür sein. Zunächst wird diese Tastaturbelegung für die deutsche Textverarbeitung seit hundert Jahren eingesetzt und eine Reform der Schreibgewohnheit könnte große Störungen mit sich bringen. Zweitens sind die Tastenanschläge meistens vom Schreibgedankengang abhängig und eingeschränkt, weshalb eine erhöhte Effizienz kaum auffiele. Drittens müssen zahlreiche Deutschmuttersprachler Texte in einer Fremdsprache schreiben, in deren Schriftsystem die Zeichenhäufigkeit variiert. So kann die Nationalisierung des Tastaturlayouts (der Entwurf ohne Berücksichtigung internationaler Normen) das Schreiben in einer anderen Sprache stark behindern. Aus solchen Gründen bin ich der Meinung, dass die standardisierte Belegung trotz vieler Kritikpunkte zurzeit nicht ersetzbar ist.

2.1.3 Verarbeitung von Sonderbuchstaben mit diakritischen Zeichen

Zur Eingabe der Sonderbuchstaben wie <é> und <ü> mit dem deutschen Eingabeverfahren lässt sich konstatieren, dass es zwei unterschiedliche Methoden gibt: 1) die Belegung als gesamtes Zeichen auf einer Taste; 2) die Zerlegung in Grundbuchstaben und diakritisches Zei-

[48] Das Tool ‚Java Letter Frequency' ist verfügbar unter: *http://www.imn.htwk-leipzig.de/~dborkman/offtopic/letter_frequency/letter.html* [Abruf: 2014-05-31]. Texte des Korpus mit 298.016 N-Grammen sind 20 Zeitungsartikel von ‚Spiegel.de' und ‚Zeit.de' und der 1. Kapitel dieser Dissertation. Wegen der Einschränkung des Korpus sind nicht alle Sonderzeichen im Korpus eingeschlossen.

chen, das mit einer toten Taste repräsentiert wird. Analog zu dem deutschen Alphabet werden die Sonderbuchstaben <Ä>, <Ö> und <Ü>, bei denen das diakritische Zeichen als Umlaut auf den Grundvokal <A>, <O> und <U> gesetzt wird, als eigenständige Zeichen auf der Tastatur codiert. Die nicht zu dem deutschen Schriftsystem zählenden Sonderbuchstaben indes, die nur in manchen Fremdwörtern auftreten, sind nur mit Beteiligung einer toten Taste zu erreichen.

Wie in Kap. 1.4.1 erwähnt, werden ebenso bei der lateinischen Schrift die beiden Verfahren bei der Zeichencodierung für Datenaustausch angewendet. In Unicode werden solche lateinischen Sonderbuchstaben sowohl als Gesamtzeichen als auch als Zeichensequenz von Basis- und kombinierenden diakritischen Zeichen codiert. Ein Beispiel: Der Buchstabe <Ü> kann einerseits mit einem Codepunkt (U+00DC) verknüpft und dargestellt werden, andererseits kann er auch mit zwei Zeichencodes (U+0055 für <U> und U+0308 für Diakritikum *Umlaut*) repräsentiert werden. Aus diesen beiden Methoden resultieren zwar kaum Unterschiede für das menschliche Lesen auf Bildschirm und Druckpapier, sie sind aber für den Computer grundsätzlich unterschiedliche Objekte. Von linguistischer und technischer Warte aus ist vorzuschlagen, für eine bestimmte Schrift bei Sonderbuchstaben einheitlich mit derselben Methode zu verfahren. Unter dieser Voraussetzung können Softwares und die computergestützten linguistischen Forschungen mit weniger technischen Störungen ablaufen. Die lateinischen Sonderbuchstaben werden in den meisten Fällen mit eigenständigen Zeichen codiert. Um die Konkurrenz zwischen den zwei verschiedenen Codierungsformen im Unicode desselben Zeichens zu lösen, wird die Äquivalenz zwischen der Zeichensequenz und dem eigenständigen Zeichencode begründet. Der Prozess der Äquivalenz im Unicode wird als ‚Normalisierung' bezeichnet (vgl. Whistler 2019: Kap. 1.1). Die Verarbeitung von <Ü> z.B. ist die Zeichenfolge vom Codewert U+0055 <U> und U+0308, das Umlautzeichen äquivalent mit dem Codewert U+00DC für <Ü>. Die Eingabe mancher Sonderbuchstaben erfordert ebenso ‚Normalisierung', <é> z.B. wird im deutschen Eingabeverfahren im Umwandlungsprozess von <´> + <e> zu <é> unterstützt.

Zur Untersuchung der Sonderbuchstabeneingabe vergleiche ich drei verschiedene Tastaturlayouts, die für die deutsche Textverarbeitung zuständig sein könnten: die Standardtastatur des Deutschen (siehe Abb. 2-2), des Belgischen (Abb. 2-3) und des Schweizerischen (Abb. 2-4). Die Ergebnisse werden in Tab. 2-4 wiedergegeben.

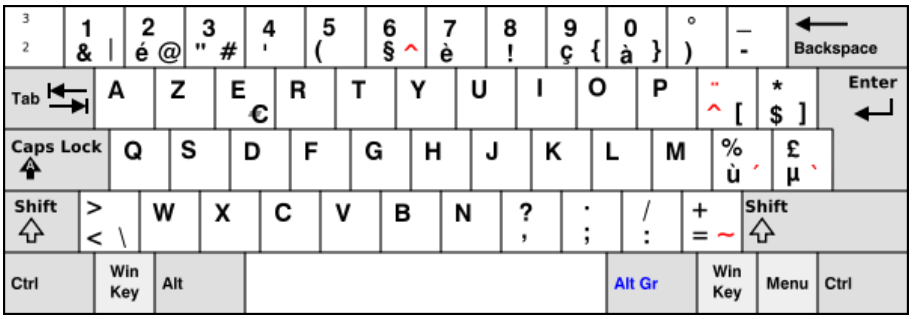


Abb. 2-3: Das belgische Tastaturlayout[49]

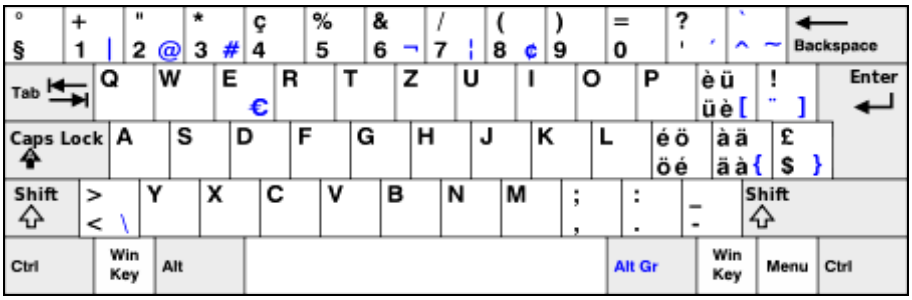


Abb. 2-4: Das schweizerische Tastaturlayout[50]

Tastatur	**Anzahl der Zeichenbelegung**	**Umschaltung**	**belegte Sonderbuchstaben**	**Mit toter Taste bezogene Zeichen**	
				Diakritika	Buchstaben
Deutsche	109	3 Stufen	Ä/ä, Ö/ö, Ü/ü, ß	` ´ ^	Á/á, Û/û, Ì/ì, Ý/ý etc.
Belgische	110	3 Stufen	é, è, à, ç, ù	¨ ` ´ ^ ~	Ä/ä, Á/á, Ì/ì, Ĩ/ĩ etc.
Schweizerische	117	5 Stufen	Ä/ä, Ö/ö, Ü/ü, Ç/ç, À/à, É/é, È/è	` ´ ^	Í/í, Ù/ù, Ê/ê, Á/á, Ý/ý etc.

Tab. 2-4: Der Vergleich zwischen der Verarbeitung der Sonderbuchstaben in deutschen Texten

Durch den Vergleich kristallisiert sich heraus, dass die Eingabe der Sonderbuchstaben hauptsächlich von der Sprachsituation einer Nation abhängig ist. In Belgien sind Niederländisch[51], Französisch[52] und Deutsch Amtssprachen, während Französisch im offiziellen Schriftverkehr am meisten gebraucht wird. Bei der belgischen Tastatur werden deswegen die häufigen Sonderbuchstaben des Französischen stärker berücksichtigt. In der Schweiz dagegen ist Deutsch die Muttersprache von mehr als der Hälfte der Bevölkerung, die parallel mit Französisch, Italienisch und Rätoromanisch als Amtssprache verwendet wird. Bei der schweizerischen Tastatur werden die Sonderbuchstaben aus den vier Amtssprachen betrachtet. Der Einsatz von toten

[49] Die als tote Tasten belegten diakritischen Zeichen werden im Layout rot markiert; die belegten Zeichen betragen 110 (35×2+13×3+1 [13 Tasten mit 3-mal Belegung]; inkl. Leerzeichen).

[50] Die als tote Tasten belegten diakritischen Zeichen finden sich auf #12 (3. Belegung), #13 (alle Belegungen) und #28 (2. Belegung); die belegten Zeichen betragen 117 (33×2+12×3+1×4+2×5+1 [eine Taste mit 4-mal sowie zwei Tasten mit 5-mal Belegung]; inkl. Leerzeichen); die 3./4. Belegung des schweizerischen Tastaturlayouts erfolgt mit ‚Umschaltsperre' (Caps Lock) ohne oder mit ‚Shift'-Kombination.

[51] Sonderbuchstaben im Niederländischen: <Ä/ä>, <Ö/ö>, <Ü/ü>, <Ë/ë>, <Ï/ï> und <Ÿ/ÿ>.

[52] Sonderbuchstaben im Französischen: <à>, <â>, <æ>, <ç>, <è>, <é>, <ê>, <ë>, <î>, <ï>, <ô>, <œ>, <ù>, <û>, <ü> und <ÿ>. Sie treten in der Schrift kaum in Großschreibung auf.

Tasten für die drei Diakritika ‚Akut', ‚Gravis' und ‚Zirkumflex' ist deswegen ökonomisch und praktisch. In Tab. 2-5 verdeutliche ich die Verarbeitung der auf den fünf Vokalen basierten Kombinationszeichen unter dem Aspekt der Zeichencodierung:

diakritisches Zeichen		**Umlaut**	**Akut**	**Gravis**	**Zirkumflex**
Unicode		U+0308	U+0301	U+0300	U+0302
1. Gruppe	Grundbs.	A/a			
	Form	Ä/ä	Á/á	À/à	Â/â
	Unicode	U+00C4/E4	U+00C1/E1	U+00C0/E0	U+00C2/E2
	Sprache[53]	DE, NL	HU, CS	FR, IT, PT	PT, FR
2. Gruppe	Grundbs.	E/e			
	Form	Ë/ë	É/é	È/è	Ê/ê
	Unicode	U+00CB/EB	U+00C9/E9	U+00C8/E8	U+00CA/EA
	Sprache	FR, NL	FR, ES, RM	FR, IT, RM	PT, FR, RM
3. Gruppe	Grundbs.	I/i			
	Form	Ï/ï	Í/í	Ì/ì	Î/î
	Unicode	U+00CF/EF	U+00CD/ED	U+00CC/EC	U+00CE/EE
	Sprache	FR, NL	FO, ES	IT	FR
4. Gruppe	Grundbs.	O/o			
	Form	Ö/ö	Ó/ó	Ò/ò	Ô/ô
	Unicode	U+00D6/F6	U+00D3/F3	U+00D2/F2	U+00D4/F4
	Sprache	DE, NL, RM	IS, PL	IT	FR, PT
5. Gruppe	Grundbs.	U/u			
	Form	Ü/ü	Ú/ú	Ù/ù	Û/û
	Unicode	U+00DC/FC	U+00DA/EA	U+00D9/F9	U+00DB/FB
	Sprache	DE, NL, RM	FO, ES	FR	FR

Tab. 2-5: Die Zeichencodierung mancher mit diakritischen Zeichen abgeleiteten Sonderbuchstaben, die in der lateinischen Schrift vorhanden sind

2.1.4 Untersuchung der Tastenbelegung der 26 lateinischen Grundbuchstaben verschiedener Schriftsysteme

In Kapitel 2.1.2 wurde erschlossen, dass der Entwurf der deutschen Standardtastatur von der internationalen Norm und unterschiedlichen Schriftsystemen stark beeinflusst ist. Dasselbe gilt für fast alle lateinalphabetischen Schriftsysteme, deren Tastaturlayouts von der QWERTY-Belegung abgeleitet wurden. Das lateinische Alphabet ist die häufigste Schrift in der heutigen Welt. Die Standardtastenbelegung der 26 Grundbuchstaben dient deswegen in fast allen Ländern als Referenz.

Bei der Tastenbelegung der 26 Grundbuchstaben gibt es bei den meisten lateinischen nationalen Standardtastaturen keine großen Unterschiede, obwohl die Buchstabenhäufigkeit und die Häufigkeit der Bigramme stets variiert. In Tab. 2-6 vergleiche ich nachfolgend die Buchstaben-Tasten-Repräsentation auf US-amerikanischen, deutschen, französischen und spani-

[53] Nur europäischen Sprachen werden in dieser Tab. Als Beispiel genannt; DE: Deutsch, NL: Niederländisch, FR: Französisch, HU: Ungarisch, CS: Tschechisch, ES: Spanisch, FO: Färöisch, IS: Isländisch, IT: Italienisch, PT: Portugiesisch und RM: Rätoromanisch (nach ISO 639-1).

schen Standardtastaturen. Die durchschnittliche Häufigkeit eines Buchstaben in den vier Sprachen wird berechnet.

Buch-staben	**EN**[54]	**DE**	**FR**[55]	**ES**	**Im DS.**	**Rang**	**TB. ge-nerell**	**TB. var-iiert**[56]
A/a	8,09%	5,15%	7,84%	11,26%	8,09%	2.	#31	#17
B/b	1,52%	2,38%	1,06%	1,30%	1,57%	18.	#50	
C/c	3,26%	2,95%	2,97%	4,58%	3,44%	12.	#48	
D/d	3,49%	4,21%	4,18%	5,61%	4,37%	10.	#33	
E/e	12,41%	16,65%	14,66%	13,77%	14,37%	1.	#19	
F/f	2,48%	2,25%	1,12%	0,81%	1,67%	17.	#34	
G/g	1,78%	3,70%	1,27%	1,12%	1,97%	16.	#35	
H/h	4,73%	4,76%	0,92%	0,72%	2,78%	13.	#36	
I/i	6,46%	8,14%	7,26%	5,56%	6,86%	6.	#24	
J/j	0,11%	0,18%	0,31%	0,32%	0,23%	25.	#37	
K/k	0,56%	1,41%	0,05%	0,04%	0,52%	22.	#38	
L/l	4,10%	3,68%	6,01%	4,83%	4,66%	9.	#39	
M/m	2,57%	2,36%	2,96%	3,12%	2,75%	14.	#52	#40
N/n	6,41%	10,36%	7,13%	6,94%	7,71%	3.	#51	
O/o	8,13%	2,25%	5,23%	8,41%	6,01%	8.	#25	
P/p	2,14%	0,73%	3,01%	2,60%	2,12%	15.	#26	
Q/q	0,18%	0,04%	0,99%	0,86%	0,52%	22.	#17	#31
R/r	7,13%	7,94%	6,55%	6,84%	7,12%	4.	#20	
S/s	6,41%	5,57%	8,08%	8,40%	7,12%	5.	#32	
T/t	8,90%	5,43%	7,07%	4,42%	6,46%	7.	#21	
U/u	3,60%	4,02%	5,70%	3,93%	4,31%	11.	#23	
V/v	1,10%	1,08%	1,33%	1,02%	1,13%	19.	#49	
W/w	2,00%	1,58%	0,04%	0,01%	0,91%	20.	#18	#46
X/x	0,20%	0,03%	0,46%	0,18%	0,22%	26.	#47	
Y/y	2,14%	0,03%	0,30%	0,92%	0,85%	21.	#22	#46
Z/z	0,10%	1,20%	0,12%	0,23%	0,41%	24.	#46	#22/#18
Gesamt	100,00%	98,08%	96,62%	97,80%	98,13%			

Tab. 2-6: Die Häufigkeit und die Tastenbelegung der 26 Grundbuchstaben des Englischen, Deutschen, Französischen und Spanischen

Auf Basis der errechneten durchschnittlichen Häufigkeit der Grundbuchstaben evaluiere ich im Anschluss die internationale Standardtastatur bei den vier Schriftsystemen. Die Ergebnisse der Bewertung werden in Tab. 2-7 angegeben. Die variierten Tastenbelegungen des deutschen und französischen Tastaturlayouts werden ignoriert.

	Zeigfinger		**Mittelfinger**		**Ringfinger**		**kleiner Finger**		**Gesamt; horizontal**
Linke H.	RTFGVB	19,90%	EDC	22,19%	WSX	8,24%	QAZ	9,02%	59,34%
Rechte H.	YUHJNM	18,64%	IK	7,37%	OL	10,66%	P	2,12%	38,79%
Oberr.	RTYU	18,73%	EI	21,23%	WO	6,91%	QP	2,64%	49,51%

[54] Meier 1978: 334; vgl. auch die statistische Quelle für Deutsch und Spanisch in Spalte ‚DE' und ‚ES'.
[55] Müller 2003 [Abruf: 2017-11-15].
[56] Abkürzung für *Tastenbelegung variiert*, die mit der US-amerikanischen (TB. generell) nicht übereinstimmen (bei der Angabe der Buchstabenhäufigkeit hellgrau markiert). Wenn es für einen Buchstaben mehr als eine Variante gibt, werden die Varianten nach der Reihfolge der Sprachen in der Tabelle aufgelistet.

	Zeigfinger		Mittelfinger		Ringfinger		kleiner Finger		Gesamt; horizontal
Grundr.	FGHJ	6,65%	DK	4,89%	SL	11,77%	A	8,08%	31,39%
Unterr.	VBNM	13,16%	C	3,44%	X	0,22%	Z	0,41%	17,23%
Gesamt; vertikal	...	38,54%	...	29,56%	...	18,90%	...	11,14%	98,13%

Tab. 2-7: Die Evaluation der QWERTY-Belegung im Buchstaben-Tasten-Zusammenhang für das englische, deutsche, französische und spanische Schriftsystem

Die Evaluation der standardisierten Tastenbelegungen wird nachfolgend durch eine Betrachtung der Bigramme in den vier Schriftsystemen erweitert. In Tab. 2-8 werden die 35 häufigsten Bigramme der vier Sprachen berechnet und absteigend aufgelistet. Bei der Analyse werden die zwei Formen der Zeichenverbindung, die aus denselben beiden Buchstaben zusammengesetzt sind, zu einer Variante summiert.

Rang[57]	Form1	Form2	EN[58]	DE	FR	ES[59]	Im DS.	G_S[60]
1	ER	RE	3,02%	5,21%	3,28%	3,06%	3,64%	2
2	EN	NE	1,84%	5,23%	3,82%	3,47%	3,59%	0
3	ES	SE	2,29%	2,39%	4,46%	3,05%	3,05%	2
4	DE	ED	1,46%	2,78%	3,53%	2,74%	2,63%	2w
5	LE	EL	1,18%	1,28%	4,32%	2,50%	2,32%	0
6	TE	ET	1,74%	2,40%	2,34%	1,56%	2,01%	2&5
7	AN	NA	2,26%	1,70%	1,66%	1,91%	1,88%	0
8	AL	LA	1,11%	1,04%	1,78%	3,56%	1,87%	0
9	AR	RA	1,58%	1,31%	1,64%	2,75%	1,82%	2
10	ON	NO	2,10%	0,82%	2,01%	2,24%	1,79%	4&5
11	IN	NI	2,06%	2,33%	1,11%	1,44%	1,74%	4&5
12	TI	IT	2,16%	1,37%	2,18%	1,09%	1,70%	0
13	IE	EI	0,77%	3,56%	1,15%	0,84%	1,58%	0
14	AT	TA	1,80%	1,03%	1,54%	1,46%	1,46%	2&5
15	IS	SI	1,48%	1,44%	1,38%	1,38%	1,42%	0
16	ST	TS	1,53%	1,66%	1,14%	0,99%	1,33%	2&5
17	OR	RO	1,67%	0,80%	1,18%	1,62%	1,32%	0
18	AS	SA	1,42%	0,89%	1,01%	1,87%	1,30%	1
19	NT		1,10%	0,59%	1,72%	1,65%	1,26%	0
20	ME	EM	0,91%	1,05%	2,04%	1,02%	1,26%	0
21	HE	EH	2,66%	1,59%	0,45%	0,25%	1,24%	0
22	TH	HT	3,37%	0,61%	0,91%	0,00%	1,22%	0
23	DA	AD	0,60%	0,65%	0,89%	2,41%	1,14%	1
24	ND		1,18%	1,99%	0,65%	0,65%	1,12%	0
25	OS	SO	1,08%	0,47%	0,73%	1,94%	1,05%	0
26	EC	CE	1,19%	0,26%	1,66%	1,05%	1,04%	4w
27	RI	IR	0,98%	0,69%	1,15%	1,20%	1,00%	0
28	LI	IL	0,96%	0,86%	0,99%	1,06%	0,97%	2
29	TR	RT	0,93%	0,80%	1,13%	0,78%	0,91%	2w
30	IC	CI	0,70%	0,77%	0,28%	1,87%	0,90%	0

[57] Vgl. Meier 1978: 336, Müller 2012 [Abruf: 2014-06-12].
[58] Statistiken für Englisch, Deutsch und Französisch nach: Pommerening 2014 [Abruf: 2014-06-11].
[59] Müller 2002a [Abruf: 2014-06-12].
[60] ‚G_S' steht für Schwierigkeitsstufe und wird gleich wie in Tab. 2-2 angegeben: 0 – Handwechsel; 1-6 für Fingerwechsel einer Hand in verschiedener Stufe; 1w-6w für Fingerwandern in verschiedener Stufe (siehe S. 65f).

Rang[57]	Form1	Form2	EN[58]	DE	FR	ES[59]	Im DS.	G_S[60]
31	CA	AC	0,83%	0,29%	0,87%	1,62%	0,90%	3
32	UN	NU	0,45%	1,52%	0,61%	0,94%	0,88%	4w
33	GE	EG	0,52%	1,97%	0,52%	0,47%	0,87%	2&5
34	CH		0,46%	2,43%	0,31%	0,17%	0,84%	0
35	CO	OC	0,77%	0,17%	0,83%	1,58%	0,84%	0
Summe			50,16%	53,95%	55,27%	56,18%	53,89%	

Tab. 2-8: Die 35 häufigsten Bigramme der vier Sprachen

Nach dieser Berechnung werden insgesamt achtzehn der 35 häufigsten Bigramme mit Handwechsel (Grad 0) realisiert. Dies entspricht 26,41% und im Verhältnis zu der Gesamthäufigkeit der 35 Bigramme 49,01%. Nur 2,44% (4,98% relativ) beziehen mit nachfolgenden Anschlägen zweier Grundtasten ein (Grad 1). Grad 2 (Normal-Hochgriffe) sind die Bigramme mit absoluter Häufigkeit (17,72%) sowie relativer Häufigkeit (32,88%) zugeordnet. Die Buchstabenfolgen, deren Eingabe Sprunggriffe (Grad 4) erfordern, betragen insgesamt 5,45% (10,11% relativ). Vier Bigramme erfordern wandernde Finger (5,46%, 10,13% relativ). Der Innenspreizgriff wird in der Tabelle immer gemischt mit einem anderen Schwierigkeitsgrad gebraucht. Aus diesem Grund wird Grad 5 an dieser Stelle nicht errechnet.

Anhand der aus Tab. 2-7 sowie Tab. 2-8 gewonnenen Schlussfolgerungen zeigt sich, dass die internationale Norm des Tastaturlayouts auch für die vier am meisten verwendeten lateinalphabetischen Schriftsysteme im Durchschnitt nicht rational ist. Vor der Perspektive der Beanspruchung beider Hände muss die relativ schwache linke Hand mehr Arbeit leisten (60,14% vs. 38,82%). Bei der Analyse der drei alphabetischen Reihen der Tastatur wird die Oberreihe am häufigsten belegt (50,20%), obwohl die Grundstellung des Tastaturschreibens nach dem Zehnfingersystem auf der Grundreihe (31,53%) stattfindet. Bei den zehn häufigsten Buchstaben der vier Sprachen können mindestens zwei davon rationalisiert werden: der dritthäufigste <N> (Unterreihe mit Zeigfinger) und der achthäufigste <O> (Oberreihe mit Ringfinger). Viele der 35 häufigsten Bigramme der vier Schriftsysteme haben einen relativ hohen Schwierigkeitsgrad. Wie in Kap. 2.1.2 erwähnt, ist die Rationalisierung des QWERTY-Layouts (inklusive seiner Varianten) trotz der Ineffektivität schwer durchzuführen.

2.2 Eingabeverfahren des vietnamesischen Schriftsystems

Auf Basis von Kap. 2.1.1 bis 2.1.4 lassen sich Faktoren für den Entwurf des Tastaturlayouts eines alphabetischen Schriftsystems formulieren: 1) Bestimmung des benötigten Zeichenvorrats des Schriftsystems; 2) Errechnung der Buchstaben- sowie Bigrammhäufigkeit; 3) Verarbeitung der Sonderbuchstaben mit Diakritika (welche Methode wäre effektiver: die Zerlegung in Basis- sowie diakritische Zeichen oder die Belegung kompletter Zeichen?); 4) Berücksich-

tigung des internationalen und anderer nationaler Tastaturlayouts des lateinischen Schriftkreises. Bei der Verbreitung des allgemeinen Tastaturlayouts für die Eingabe der sonstigen alphabetischen (vor allem nichteuropäischen) Schriftsysteme gibt es außer den vier Punkten noch viele weitere Probleme zu lösen. An dieser Stelle wird das Eingabeverfahren des vietnamesischen Schriftsystems dafür exemplarisch herangezogen.

2.2.1 Allgemeines über die vietnamesische Sprache und ihr Schriftsystem

Das moderne, auf dem lateinischen Alphabet basierte vietnamesische Schriftsystem weicht wegen der sprachlichen Unterschiede zu den indogermanischen Sprachen stark von den europäischen Schriftsystemen ab. Im Kontrast zu den alphabetischen Buchstaben liegt der größte Unterschied bei den häufig verwendeten diakritischen Zeichen, mit denen phonemische Vokale und der Silbenton wiedergegeben werden.

Vietnamesisch (vie.: Tiếng Việt, archaische Schreibung in der chinesischen Schrift: 啃越) zählt nach seinem Sprachbau zu den monosyllabisch-isolierenden Sprachen. Phonetisch betrachtet handelt es sich um eine Tonsprache mit sechs verschiedenen Tönen. Sie gehört zur austroasiatischen Sprachfamilie. Wegen kultureller Einflüsse sind 70% der Wörter chinesischer Herkunft (vgl. Nguyễn 1990: 49-54, Müller-Yokota 1994: 399-402, Coulmas 1996a: 85f & 543). Drei verschiedene Schriften wurden zur vietnamesischen Aufzeichnung eingeführt:

1) die chinesische Schrift mit denen aus China stammenden chinesischen Schriftzeichen, die auf Vietnamesisch als ‚Chữ Hán' (Schreibung in der chinesischen Schrift: 字漢, chi. Lesart: /zì hàn/, ca. 111 v. Chr. bis 1910) bezeichnet werden;
2) die einheimische morphologische Schrift ‚Chữ Nôm' (字喃 /zì nán/, 13. Jh. bis 1910) mit 6.285 nativen Schriftzeichen, die nach den Konstruktionsprinzipien der chinesischen Schrift erfunden wurden;
3) die auf dem lateinischen Alphabet basierte ‚Quốc Ngữ' (國語 /guó yǔ/), die im 17. Jh. entstanden ist und seit 1910 offiziell benutzt wird. Chữ Hán und Chữ Nôm wurden dementsprechend abgeschafft (vgl. ibid.).

Chữ Hán und Chữ Nôm gehören zu dem chinesischen Schriftkreis, deren Konstruktionsprinzipien bei der schriftlinguistischen Erforschung der sinographischen Grammatologie (Kap. 3) vorgestellt werden. In diesem Kapitel wird nur das moderne vietnamesische Schriftsystem berücksichtigt, das den phonologischen Aspekt fokussiert.

Die vietnamesischen Silben folgen der Struktur „$C_1(C_2)VC_3$“[61] (Coulmas 1996a: 543). C_2 ist häufig der Semivokal /w/. Einer der insgesamt sechs Töne entscheidet über die Silbenaussprache. Als V dieser Formel können sowohl Einzelvokale (insgesamt zwölf) als auch Diphthonge sowie Triphtonge eingesetzt werden. In den meisten Fällen ist eine betonte vietnamesische Silbe ein eigenständiges Wort oder ein sinninhalttragendes Morphem, das zur polysyllabischen Wortbildung verwendet werden kann. Aus diesem Grund wird Vietnamesisch schriftlich hauptsächlich monosyllabisch geschrieben, weshalb zwischen zwei Silben in den meisten Fällen ein Leerzeichen gesetzt wird. Anhand des isolierenden Sprachaufbaus sind die vietnamesischen Wörter unflektierbar (vgl. Boscher 1989: 11f).

2.2.2 Zeicheninventar des vietnamesischen alphabetischen Schriftsystems

Anhand der Phoneme wurden die 22 lateinischen Grundbuchstaben für Quốc Ngữ ausgewählt, darunter 16 für Konsonanten und sechs für Vokale.

> A/a, B/b, C/c, D/d, E/e, G/g, H/h, I/i, K/k, L/l, M/m, N/n, O/o, P/p, Q/q, R/r, S/s, T/t, U/u, V/v, X/x, Y/y
>
> (Boscher 1989: 9)[62]

Zur Wiedergabe der weiteren vokalischen Phoneme stehen die vokalischen Grundbuchstaben mit diakritischen Zeichen zur Verfügung. Der Buchstaben-Phonem-Zusammenhang solcher Kombinationsvokale wird in der folgenden Tabelle angegeben.

Grundvokal		**Variante 1**		**Variante 2**	
Buchstabe	IPA	Buchstabe	IPA	Buchstabe	IPA
A/a	[a]	Ă/ă	[ă]	Â/â	[ə]
E/e	[ɛ]	Ê/ê	[e]		
I/i	[i]				
O/o	[ɔ]	Ô/ô	[o]	Ơ/ơ	[ɤ]
U/u	[u]	Ư/ư	[ɨ]		
Y/y	[i]				

Tab. 2-9: Der vokalische Buchstaben-Phonem-Zusammenhang des vietnamesischen Alphabets[63]

Die Grapheme für weitere Konsonanten werden hauptsächlich durch Dia- oder Trigraphen kreiert, wie <th>, <kh>, <ph>, <nh>, <gh>, <tr>, <ch> <gi>, <qu>, <ng> und <ngh>.[64] Eine Ausnahme stellen die variierten konsonantischen Buchstaben <Đ/đ> für den Laut [ɗ] dar, deren Form sich von <D/d> durch einen Querstrich unterscheidet (vgl. Boscher 1989: 10).

[61] C steht für Konsonant und V für Vokal; Tonzeichen werden immer auf dem Vokal gesetzt.
[62] Anmerkung: <F>, <J>, <W> und <Z> gehören nicht zu dem vietnamesischen Alphabet.
[63] Vgl. Boscher 1989: 10; drei diakritische Zeichen werden verwendet.
[64] Im Vietnamesischen gibt es 10 Ligaturen aus zwei und eine aus drei konsonantischen Buchstaben.

Die Tonmarkierung auf dem Hauptvokal bestimmt die Tonhöhe einer Silbe. Wie bei sinotibetischen Sprachen ist der Ton ein wesentlicher Faktor zur Bedeutungsunterscheidung. Unter den sechs Tönen bleibt der erste Ton (der Normalton) unmarkiert. Für die anderen fünf werden dementsprechend fünf Tonzeichen angewendet. Die Tonzeichen, der Höheverlauf, die Bezeichnungen und Beispielwörter der sechs Töne werden in Tab. 2-10 angegeben. Zur Vorstellung von Textverarbeitung und Eingabeverfahren des Vietnamesischen werden die technischen Daten, die Zeichencodes in Unicode sowie die Tastenposition/-repräsentation von drei Eingabeverfahren dementsprechend in der Tabelle aufgeführt.

Attribut vom Ton	**1. Ton**	**2. Ton**	**3. Ton**	**4. Ton**	**5. Ton**	**6. Ton**
Zeichenform		◌̀	◌́	◌̣	◌̉	◌̃
Beschreibung	Normal	fallend	steigend	tief	fallen-steigend	unterbrochen-steigend
Tonhöhe[65]	33	21	35	32 oder 31	313	35
Bezeichnung VI	không	huyền	sắc	nặng	hỏi	ngã
Zeichenname		Gravis	Akut	Unterpunkt	Haken	Tilde
Beispielwort	ma (*Geist*)	mà (*aber*)	má (*Mutter*)	mạ (*Reissämling*)	mả (*Grab*)	mã (*Code*)
Unicode		U+0340	U+0341	U+0323	U+0309	U+0303
TB_TCVN[66]		#06 (‚5‘)	#09 (‚8‘)	#10 (‚9‘)	#07 (‚6‘)	#08 (‚7‘)
RP_Telex[67]		F	S	J	R	X
RP_VNI[68]		2	1	5	3	4

Tab. 2-10: Die sechs Töne der vietnamesischen Sprache[69]

Mit Berücksichtigung des Tons gibt es insgesamt 72 Varianten für vokalische Schriftzeichen (inkl. Buchstaben mit dem unmarkierten ersten Ton, siehe Tab. 2-11), die auf Basis von zwölf vokalischen Buchstaben (siehe Tab. 2-9) gebildet werden. Deswegen gehören 60 weitere mit Tonzeichen gebildete Buchstaben und je nach Majuskel und Minuskel 120 Schriftzeichen zusätzlich zum vietnamesischen Schriftsystem.

Buchstabe	**Zeichen**	**1. Ton**	**2. Ton**	**3. Ton**	**4. Ton**	**5. Ton**	**6. Ton**
A/a	Form	A/a	À/à	Á/á	Ạ/ạ	Ả/ả	Ã/ã
	Unicode	U+0041/ U+0061	U+00C0/ U+00E0	U+00C1/ U+00E1	U+1EA0/ U+1EA1	U+1EA2/ U+1EA3	U+00C3/ U+00E3
Ă/ă	Form	Ă/ă	Ằ/ằ	Ắ/ắ	Ặ/ặ	Ẳ/ẳ	Ẵ/ẵ
	Unicode	U+0102/ U+0103	U+1EB0/ U+1EB1	U+1EAE/ U+1EAF	U+1EB6/ U+1EB7	U+1EB2/ U+1EB3	U+1EB4/ U+1EB5
Â/â	Form	Â/â	Ầ/ầ	Ấ/ấ	Ậ/ậ	Ẩ/ẩ	Ẫ/ẫ

[65] Nach Ziffertonsystem von Zhao Yuanren.
[66] Nach dem nationalstandardisierten Tastaturlayout; bei der Darstellung der belegten Tasten wird zuerst der Scan-Code mit #, dann das entsprechend belegte Zeichen im internationalen Tastaturlayout in Anführungszeichen angegeben.
[67] Im Fall der auf dem US-amerikanischen Layout basierten Eingabemethode ‚Vietnamesisch Telex‘.
[68] Im Fall der auf dem US-amerikanischen Layout basierten Eingabemethode ‚Vietnamesisch VNI‘.
[69] Vgl. Boscher 1989: 13, Nguyễn 1990: 55, Unicode 12.0 Character: U+0300-U+036F.

Buchstabe	Zeichen	1. Ton	2. Ton	3. Ton	4. Ton	5. Ton	6. Ton
	Unicode	U+00C2/ U+00E2	U+1EA6/ U+1EA7	U+1EA4/ U+1EA5	U+1EAC/ U+1EAD	U+1EA8/ U+1EA9	U+1EAA/ U+1EAB
E/e	Form	E/e	È/è	É/é	Ẹ/ẹ	Ẻ/ẻ	Ẽ/ẽ
	Unicode	U+0045/ U+0065	U+00C8/ U+00E8	U+00C9/ U+00E9	U+1EB8/ U+1EB9	U+1EBA/ U+1EBB	U+1EBC/ U+1EBD
Ê/ê	Form	Ê/ê	Ề/ề	Ế/ế	Ệ/ệ	Ể/ể	Ễ/ễ
	Unicode	U+00CA/ U+00EA	U+1EC0/ U+1EC1	U+1EBE/ U+1EBF	U+1EC6/ U+1EC7	U+1EC2/ U+1EC3	U+1EC4/ U+1EC5
I/i	Form	I/i	Ì/ì	Í/í	Ị/ị	Ỉ/ỉ	Ĩ/ĩ
	Unicode	U+0049/ U+0069	U+00CC/ U+00EC	U+00CD/ U+00ED	U+1ECA/ U+1ECB	U+1EC8/ U+1EC9	U+0128/ U+0129
O/o	Form	O/o	Ò/ò	Ó/ó	Ọ/ọ	Ỏ/ỏ	Õ/õ
	Unicode	U+004F/ U+006F	U+00D2/ U+00F2	U+00D3/ U+00F3	U+1ECC/ U+1ECD	U+1ECE/ U+1ECF	U+00D5/ U+00F5
Ô/ô	Form	Ô/ô	Ồ/ồ	Ố/ố	Ộ/ộ	Ổ/ổ	Ỗ/ỗ
	Unicode	U+00D4/ U+00F4	U+1ED2/ U+1ED3	U+1ED0/ U+1ED1	U+1ED8/ U+1ED9	U+1ED4/ U+1ED5	U+1ED6/ U+1ED7
Ơ/ơ	Form	Ơ/ơ	Ờ/ờ	Ớ/ớ	Ợ/ợ	Ở/ở	Ỡ/ỡ
	Unicode	U+01A0/ U+01A1	U+1EDC/ U+1EDD	U+1EDA/ U+1EDB	U+1EE2/ U+1EE3	U+1EDE/ U+1EDF	U+1EE0/ U+1EE1
U/u	Form	U/u	Ù/ù	Ú/ú	Ụ/ụ	Ủ/ủ	Ũ/ũ
	Unicode	U+0055/ U+0075	U+00D9/ U+00F9	U+00DA/ U+00FA	U+1EE4/ U+1EE5	U+1EE6/ U+1EE7	U+0168/ U+0169
Ư/ư	Form	Ư/ư	Ừ/ừ	Ứ/ứ	Ự/ự	Ử/ử	Ữ/ữ
	Unicode	U+01AF/ U+01B0	U+1EEA/ U+1EEB	U+1EE8/ U+1EE9	U+1EF0/ U+1EF1	U+1EEC/ U+1EED	U+1EEE/ U+1EEF
Y/y	Form	Y/y	Ỳ/ỳ	Ý/ý	Ỵ/ỵ	Ỷ/ỷ	Ỹ/ỹ
	Unicode	U+0059/ U+0079	U+1EF2/ U+1EF3	U+00DD/ U+00FD	U+1EF4/ U+1EF5	U+1EF6/ U+1EF7	U+1EF8/ U+1EF9

Tab. 2-11: Die Vokal-Varianten mit Tonzeichen[70]

Außer Buchstaben und betonten Vokalen gehören zu dem vietnamesischen Schriftsystem 50 sonstige ASCII-Zeichen (94-22×2=50). Die vier ungebrauchten lateinischen Grundbuchstaben <F>, <J>, <W> und <Z> werden zur Verschriftlichung von Fremdwörtern verwendet, weshalb sie ebenso zum Zeichenvorrat zählen. Die Wort-, Satz- und Schriftsonderzeichen sind fast identisch mit denen des ASCII-Block im Unicode. Weiterhin ist das Zeichen <₫> für die vietnamesische Währung Đồng ein zusätzliches Sonderzeichen. Demnach besitzt das vietnamesische Schriftsystem insgesamt 229 Zeichen (2×(22+6+1) + 2×60 + 50 + 1 = 229), wenn die mit Tonzeichen markierten Buchstaben als separate Zeichen definiert werden. Werden die Tonzeichen nur eigenständig als kombinierende Zeichen betrachtet, gibt es insgesamt 114 (2×(22+6+1)+5+50+1=114). Wenn die drei diakritischen Zeichen <˘>, <ˆ> und <ʾ> auf Grundvokalen ebenso einzeln verarbeitet und codiert werden, beträgt das Gesamtinventar 105 Zeichen (2×(22+1)+5+3+50+1=105).

[70] vgl. Boscher 1989: 14, Unicode 12.0 Character: U+0000-U+007F, U+0080-U+00FF, U+0100-U+017F & U+1E00-U+1EFF.

2.2.3 Zeichencodierung und Eingabeverfahren des vietnamesischen Schriftsystems

Wie in den Kapiteln 1.4.1, 1.4.3 und 2.1.3 skizziert wurde, können die mit diakritischen Zeichen und Tonzeichen markierten Schriftzeichen sowohl bei der Zeichencodierung als auch bei der Tastenbelegung mit zwei verschiedenen Methoden bearbeitet werden. Solche Sonderbuchstaben können entweder durch die Zusammensetzung von einem Grundbuchstaben und einem oder mehreren kombinierenden Zeichen repräsentiert oder einheitlich codiert werden. Da in der vietnamesischen Schrift ein vokalischer Grundbuchstabe in vielen Fällen zweimal gesetzt (Diakritika für phonetische Änderungen und Tonwerte) werden kann, gibt es drei Möglichkeiten zur vietnamesischen Zeichencodierung:

1) Nur die lateinischen Grundbuchstaben werden berücksichtigt und die Kombinationszeichen durch Verbindung von Basiszeichen und kombinierenden Zeichen (inklusive der zum Zweck der Wiedergabe der Phoneme und Töne zuständigen Diakritika) ausgedrückt. Diese Zeichencodierung basiert auf dem Standard ASCII. Die Varianten- und Kombinationsbuchstaben werden durch die Sequenz von Grundbuchstaben und Sonderzeichen vertreten, bspw. wird <ă> als <a(> und <ằ> als <a(`> ausgedrückt. Durch Programmierung eines Textverarbeitungsprogramms werden die den Codes entsprechenden Glyphen und Font abgerufen und ausgegeben. Mit dieser Art der Zeichencodierung kann das internationale Tastaturlayout für das vietnamesische Eingabeverfahren übernommen werden. Diese Eingabemethode wird ‚VIQR' (Vietnamese Quoted-Readable) genannt (vgl. Đỗ 2005: 11).
2) Alle phonemischen Buchstaben werden codiert, wohingegen die Tonzeichen isolierend als kombinierende Zeichen behandelt werden. In diesem Fall beträgt das minimale Zeicheninventar 114. Zusammen betrachtet mit dem Unicode erfordert diese Methode fünf Blöcke: a) ‚Basic Latin', b) ‚Latin-1 Supplement' (wie Â, Ê und Ô), c) ‚Latin Extended-A' (wie Ă und Đ), d) ‚Latin Extended-B' (wie Ư und Ơ) und e) ‚Combining Diacritical Marks' (alle fünf Tonzeichen). Die Zeichen-Tasten-Repräsentation von dem vietnamesischen, nationalstandardisierten Tastaturlayout TCVN 6064 (Vollzeichnung: Tiêu chuẩn Việt Nam 6064, siehe Abb. 2-5) bezieht sich ebenso auf dieses Zeicheninventar.
3) Die dritte Möglichkeit lautet, alle Zeichen (minimal 229) eigenständig zu codieren. Da die 5-mal Belegung bei allen graphischen Tasten in der Praxis unökonomisch ist, die insgesamt 240 Zeichen beträgt (48×5), werden die mit Tonzeichen angelegten Schriftzeichen in diesem Fall nicht mit Taste repräsentiert, sondern zerlegt eingegeben. Solche Sonderschriftzeichen befinden sich im Unicode im Block ‚Latin Extended Additional' (U+1E00-1EFF). Die 8-Bits-Codierung des vietnamesischen Nationalstandards VISCII (Vietnamese

Standard Code for Information Interchange) enthält alle 229 Zeichen (vgl. Lunde 2009: 974).

Die Eingabe mit TCVN 6064 ist eine der am häufigsten verwendeten Eingabeverfahren für Vietnamesisch. Technisch kann es sowohl mit einem Zeicheninventar aus 114 als auch mit 234 (229 Zeichen plus fünf Tonzeichen) verbunden werden. Wenn das Eingabeverfahren auf dem größeren Inventar basiert, erlauben es die spezifischen Programmkomponenten, aus einer Buchstaben-Ton-Sequenz die entsprechenden Sonderschriftzeichen zu erzeugen. Die tonrepräsentierenden Tasten (repräsentiert mit den Tasten ‚5' - ‚9') sind keine toten Tasten. Durch das eigenständige Eintippen können die abhängige Glyphe und das entsprechende Font angezeigt werden. Mit dem zuvor getippten Buchstaben (egal ob konsonantisch oder vokalisch) wird ein Kombinationszeichen erzeugt. Solche Tasten stehen im Gegensatz zu den toten Tasten des deutschen, belgischen sowie schweizerischen Tastaturlayouts (siehe Kap. 2.1.3, Tab. 2-4). Bei TCVN 6064 werden vor allem die Tasten der numerischen Reihe viermal belegt, wie Abb. 2-5 darstellt.[71]

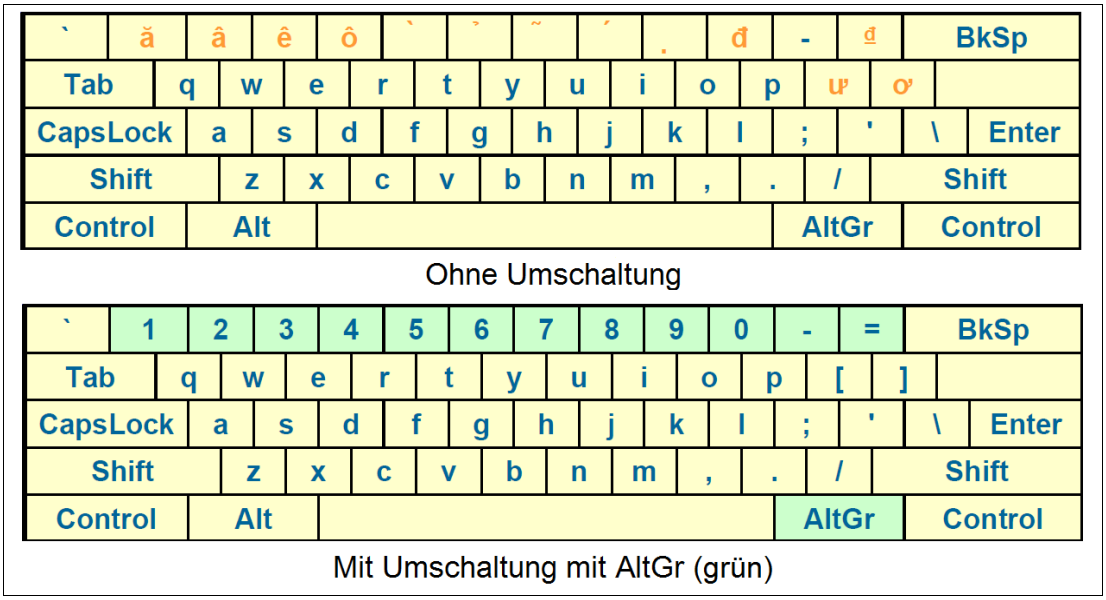


Abb. 2-5: Das Tastaturlayout des vietnamesischen Nationalstandards – TCVN 6064:1995 (Đỗ 2005: 14)[72]

Beim Entwurf der Inputcodierung der mit Ton gebildeten Buchstaben gibt es neben der graphischen Zusammensetzung (Basis- + kombinierendes Zeichen) noch weitere Varianten. Eine Möglichkeit kann darin bestehen, die selten gebrauchten Grundbuchstaben einzusetzen. Die nach diesem Prinzip entworfene Eingabemethode heißt ‚Tiếng Việt Telex', welche auf dem

[71] Nach Eingabetest mit Eingabemethode Vietnamesisch TCVN 6064 von Google Übersetzer.

[72] Hier werden nur die erste und dritte Belegung geschildert. Zu der zweiten Belegung (mit Shift) gehören die Buchstaben in Majuskel. Die vierte Belegung (mit AltGr + Shift) betrifft wie das dritte nur die Nummernreihe und ist identisch mit der zweiten Belegung des US-amerikanischen Tastaturlayouts.

internationalen Tastaturlayout basiert. Die Funktionsweise dieser Methode kann zusammengefasst wie folgt dargestellt werden:

1) Die ‚Variantenbuchstaben' werden durch Wiederholung eines Grundbuchstabens oder durch Beifügung von dem ungebrauchten Buchstaben <W> repräsentiert.

   aa → â; ee → ê; oo → ô; dd → đ;
   aw → ă; ow → ơ; uw/w → ư. [73]

2) Jedes Tonzeichen wird mit einem bestimmten Buchstaben repräsentiert und wird immer nach dem Hauptvokal getippt.

   f – fallender Ton < ` >, wie <mà>;
   s – steigender Ton < ´ >, wie <má>;
   j – tiefer Ton < ̣ >, wie <mạ>;
   r – fallend-steigender Ton < ̉ >, wie <mả>;
   x – unterbrochen-steigender Ton < ̃ >, wie <mã>. [74]

Wenn ein Sonderbuchstabe in Majuskel gebraucht wird, muss der Grundbuchstabe in Großschreibung eingetippt werden. Der für Diakritika oder Ton stehende Buchstabe kann beliebig in Klein- oder Großschreibung eingetippt werden. So werden die mit Tonzeichen angelegten Schriftzeichen bspw. wie folgt codiert:

aaf → ầ; aas → ấ; aaj → ậ; aar → ẩ; aax → ẫ;
Aaf → Ầ; Aas → Ấ; Aaj → Ậ; Aar → Ẩ; Aax → Ẫ.[75]

Zwei ohne Änderung des internationalen Tastaturlayouts funktionierende Eingabemethoden liefert der ‚Google Translate'. Die erste Möglichkeit lautet, sowohl Tonzeichen als auch diakritische Zeichen (inklusive des Querstrichs für <Đ/đ>, das in VNI mit der Taste ‚9' repräsentiert wird), mit einer Nummer zu repräsentieren. Das entsprechende Verfahren heißt ‚Tiếng Việt VNI' (Bảng dấu VNI) und lässt sich graphisch wie Abb. 2-6 skizzieren.

Die diakritischen Zeichen zur Bildung von phonetischen Buchstaben (Taste ‚5' bis ‚9') müssen dem Grundbuchstaben nachfolgen. Da ein Ton für eine ganze Silbe gilt, kann das Zeichen sowohl nach dem Hauptvokal als auch am Ende einer Silbe eingetippt werden. Demgemäß erfolgt die Eingabe der beiden Beispielwörter nach dem unten stehendem Prinzip:

T-i-e-6-1-n-g / T-i-e-6-n-g-1 → Tiếng (*die Sprache*);
V-i-e-6-5-t / V-i-e-6-t-5 → Việt (*Vietnam*).

[73] Nach Eingabetest mit der Eingabemethode Vietnamesisch Telex von ‚Google Translate' & Đỗ 2005: 9.
[74] Ibid.
[75] Ibid.

Wie bei den chinesischen und japanischen Eingabemethoden ist es weiterhin möglich, variierte Buchstaben in ihrer Grundform umzuschreiben. Durch die Anwendung eines Konversionslexikons werden alle möglichen Kandidaten von der Eingabesoftware abgerufen und je nach der Worthäufigkeit absteigend aufgelistet. Der PC-Benutzer wählt eine Zahl durch Eintippen (siehe Abb. 2-7).

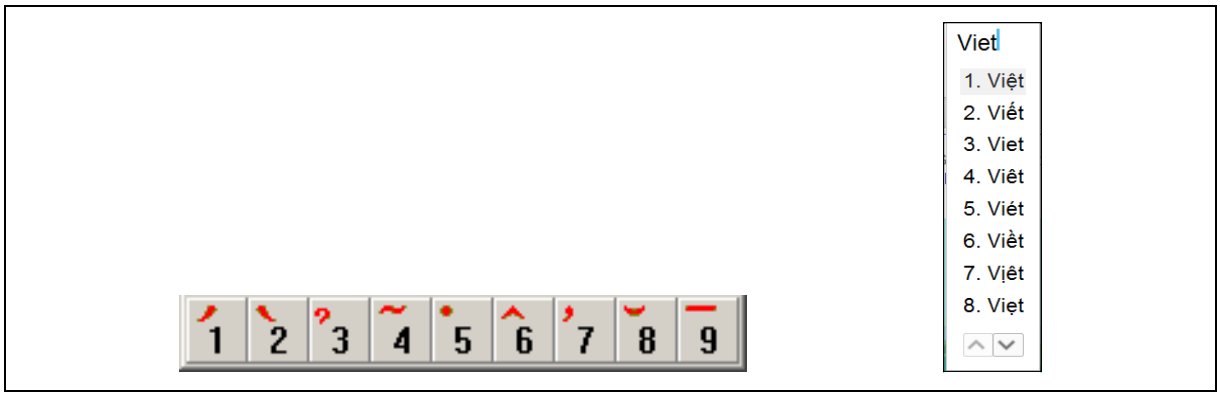


Abb. 2-6 [links]: Die Inputcodierung der Eingabemethode Tiếng Việt VNI[76]

Abb. 2-7 [rechts]: Beispiel für die ‚Eingabemethode per automatischer Worterkennung'[77]

2.2.4 Eingabemöglichkeiten eines alphabetischen Schriftsystems

Zusammenfassend lassen sich die erwähnten fünf Eingabeverfahren des Vietnamesischen in zwei Hauptkategorien gliedern:

1) Reformierung und Orientierung des Tastaturlayouts nach dem nativen Schriftsystem. In dieser Situation werden die dem Schriftsystem zugehörigen Zeichen möglichst auf der Tastatur belegt. Das auf TCVN 6064 basierende Eingabeverfahren ist dafür beispielgebend. Es ist im Prinzip am effektivsten, hat aber meist regionale Einschränkungen.
2) Anwendung des internationalen US-amerikanischen Tastaturlayouts und Eingabe von Inputcodierung. ‚VIQR', ‚Tiếng Việt Telex', ‚Tiếng Việt VNI' und ‚Eingabe per automatische Worterkennung' gehören zu dieser Kategorie. Unterschiedlich sind die vier Methoden vor allem darin, wie die Sonderschriftzeichen codiert werden, ob diakritische sowie Tonzeichen repräsentiert werden und ob der eingegebene Code einen einzigen oder mehrere Kandidaten betrifft.

Wie Vietnamesisch können die meisten alphabetischen oder alphasyllabischen Eingabeverfahren auf ähnliche Weise per PC-Tastatur realisiert und mit Computern verarbeitet werden. So bietet der ‚Google Translate' für das russische, das arabische und das Hindi-Schriftsystem je-

[76] Nach Eingabetest mit der Eingabemethode Tiếng Việt VNI von Google Translate.
[77] Nach Eingabetest mit Eingabemethode Vietnamesisch ‚chào' von Google Translate.

weils zwei unterschiedliche Eingabeverfahren an: 1) Verfahren mit dem nationalisierten Tastaturlayout und 2) Eingabe durch Transkription via international verbreiteter Tastaturen. Bei der zweiten Variante wird ein Buchstabe oder ein Wort zuerst in lateinische Schrift inputcodiert, woraufhin das Computersystem mithilfe eines Konversionslexikons die Umwandlung in native Schrift unterstützt. Voraussetzung zur Eingabemöglichkeit nach der ersten Variante sind dabei schriftlinguistische Forschungsdesiderate.

Die Anwendungssituation eines mit Diakritika gebildeten Sonderbuchstaben in einer Schrift entscheidet, in welcher Art sie behandelt werden. Im vietnamesischen Schriftsystem übertragen die drei diakritischen Zeichen <ˆ>, <ˇ> und <ʾ> keine eigenständige Information, sondern müssen mit einem bestimmten Grundvokal zusammengebildet werden, um bestimmte Phoneme zu repräsentieren. Sodann werden derlei Kombinationsbuchstaben bei der Zeichencodierung als gesamter Code bevorzugt. Die fünf Tonzeichen hingegen repräsentieren den Tonhöhenverlauf einer ganzen Silbe und können bei beliebigen Vokalen angewandt werden. Deswegen können sie auf linguistischer Ebene sowohl als eigenständige Zeichen, aber auch zusammen mit einem Vokal betrachtet werden. Manche vietnamesischen, mit Ton markierten Vokalzeichen jedoch sind graphisch identisch mit den als Buchstaben gebrauchten Zeichen aus einem anderen, auf dem lateinischen Alphabet basierten Schriftsystem. Beispielsweise kann <á> in verschiedenen Sprachen ein bestimmtes Phonem, Betonungszeichen, Langvokal oder einen vokalisch benutzten Konsonant symbolisieren (vgl. Kappenberg 2012: 42). Um die Zeichencodierung der lateinischen Schrift zu vereinheitlichen, ist es notwendig, Äquivalente zwischen Zeichensequenzen und Zeichen derselben Symbole zu begründen und für alle mit Ton gebildeten Schriftzeichen des vietnamesischen Schriftsystems bei bestimmten Codepunkten zu definieren (vgl. Coulmas 1996a: 509, Unicode 12.0 Chapters: Kap. 2.12: 62ff, Bußmann 2002: 161f). Bei der Zeichencodierung der kyrillischen Schrift werden die Zeichencodes der Kombinationsbuchstaben bevorzugt, wie die in der lateinischen Schrift. Ausnahmen finden sich nur bei historischen diakritischen und für native Zahlenangaben verwendeten kombinierenden Zeichen (vgl. Unicode 12.0 Character: U+0400-U+04FF). In der arabischen und Devanagari-Schrift werden hingegen die von Grundbuchstaben abhängigen Zeichen hauptsächlich isolierend als kombinierende Zeichen codiert (vgl. Unicode 12.0 Character: U+0600-U+06FF, U+0900-U+097F). Dies hängt einerseits von den schriftlinguistischen Eigenschaften, andererseits auch von technischen Gründen ab. Aus linguistischer Sicht dienen kombinierende Zeichen in den beiden Schriften dazu, bestimmte Vokale zu repräsentieren. Aus technischer Perspektive kann das kleinere Zeicheninventar leichter auf PC-Tastaturen belegt werden. Kap. 2.3 und 2.5 gehen auf diese Punkte dezidiert ein.

2.3 Textverarbeitung der Devanagari

Wie in Kap. 2.1.4 (S. 77) erwähnt, sind die meisten lateinalphabetischen Schriftsysteme wegen der Schreibgewohnheiten und der internationalen Standardisierung auf der Basis von QWERTY-Layout variiert, obwohl diese Variante generell irrational ist. In allen sonstigen Fällen (Schriftsysteme, die auf einer anderen alphabetischen, alphasyllabischen sowie alphabetosyllabischen Schrift basieren) kann die Zeichen-Tasten-Repräsentation im Prinzip unabhängiger von der QWERTY-Form sein und stärker nach der Zeichen- sowie Bigrammhäufigkeit eines Schriftsystems entworfen werden. Betrachtet man die Arbeitsprinzipien, gibt es bei Eingabeverfahren des kyrillischen oder griechischen Schriftsystems kaum neue technische Herausforderungen. Hingegen muss die Eingabe der sonstigen alphabetischen Schriften – alphasyllabische, alphabetosyllabische und linksläufige konsonantenalphabetische – wegen Besonderheiten bezüglich des Zeichen- sowie Glyphenaufbaus und der Schriftrichtung etc. unter die Lupe genommen werden. Die Eingabeverfahren des Hindi-, koreanischen sowie arabischen Schriftsystem werden als Beispiele für verschiedene technische Schwierigkeiten erforscht.

2.3.1 Allgemeines über die alphasyllabischen Schriften und die drei verschiedenen Modelle der Zeichencodierung

In Kapitel 1.3.2 (S. 22f) wurden die Grundprinzipien der alphabsyllabischen Schriften vorgestellt. Effektive Einheit solcher Schriften ist die orthographische Silbe, die in Grundbuchstaben und abhängige Vokale segmentiert werden kann. Schriftzeichen solcher Schriften unterscheiden zwischen Grundbuchstaben/Basiszeichen, die eigenständig auftreten können, und abhängigen Zeichen, die zu Basiszeichen addiert werden müssen. Ein konsonantischer Grundbuchstabe repräsentiert im Allgemeinen eine mit inhärentem Vokal (wie /a/ im Hindi-Schriftsystem) beendete Silbe. Wenn statt des inhärenten ein anderer Vokal, ein Sonderauslaut oder kein Vokal gebraucht wird, wird ein inhärentes Vokalzeichen dem Grundbuchstaben beigefügt. Anhand der Unterschiede verschiedener indischer Schriften in der Silbenbildung finden drei Modelle zur Zeichencodierung Verwendung:

1) Das Devanagari-Modell.

Grundbuchstaben dieses Modells sind entweder Konsonanten oder unabhängige Vokalzeichen, die meistens am Anfang eines Wortes oder als eigenständiges Wort auftreten. In Schriften wird ein spezielles abhängiges Zeichen (Virama) zur Aufzeichnung von Konsonantenclustern bei vorderen Konsonanten beigefügt, welches auf die Auslassung des inhärenten Vokals hindeutet und den abgehängte Konsonant zu einem so genannten ‚Totkonsonanten' (eng.:

dead consonant) wandelt (vgl. Friedrich 2006: 86f). Bei diesem Modell wird Virama im Unicode als kombinierendes Zeichen codiert. Daher wird eine von Konsonantenclustern gebildete Silbe immer als Zeichenfolge mehrerer Zeichen verarbeitet, welche aus konsonantischen Grundbuchstaben, Virama und optional abhängigen Vokalzeichen besteht. Bei der Clusterbildung werden die Bestandzeichen in vielen Fällen verformt und lassen sich in bestimmte proportionale Konstellationen gruppieren (vgl. Unicode 12.0 Chapters: Kap. 12.1: 448, Unicode 12.0 Character: U+0900-U+097F). Beispielsweise wird die Silbe /krē/ schriftlich als < क्रे > realisiert, die nachfolgend aus < क > /ka/, < ् > (Symbol für Nullvokal), < र > /ra/ und dem kombinierenden Vokal < े > /ē/ besteht (vgl. Friedrich 2006: 89). Die Reihenfolge der Zeichen innerhalb einer Silbeneinheit befolgt die logische Ordnung der Phoneme, jedoch unabhängig von der proportionalen Beziehung zwischen Grundbuchstaben und Vokalzeichen. Deswegen wird das für /ē/ stehende Vokalzeichen trotz der obersten Position der Silbe als letztes Zeichen der Zeichenfolge gespeichert und verarbeitet. Das Modell wird unter den indischen Schriften am häufigsten verwendet.

2) Das tibetische Modell.

Im Gegensatz zu Devanagari, in der sowohl Konsonanten als auch Vokale Grundbuchstaben/Basiszeichen sein können, muss ein tibetischer Grundbuchstabe konsonantisch sein. Anders formuliert wird ein vokalisches Phonem immer in abhängiger Form angegeben, während sich die Konsonanten in unabhängiger und abhängiger hinzugefügter Form (subjoined consonants) unterscheiden. Das Virama (tibetische Bezeichnung: Halanta; Form: ྂ; Unicode: U+0F84) ist bei der Konsonantenclusterbildung ungebräuchlich. Die graphische Addierung von abhängigen Symbolen (inkl. Vokalen, abhängigen Konsonanten, Sonderauslaut usw.) auf Basiszeichen (Grundbuchstabe) kann mehrmals und sowohl horizontal (links oder rechts) als auch vertikal (oben oder unten) durchgeführt werden. Aus diesem Grund ist die Silbenstruktur des tibetischen Modells relativ kompliziert (vgl. Peng 1994: 352f). Im komplexesten Fall setzt sich eine tibetische orthographische Silbe aus sieben Zeichen zusammen. Im Durchschnitt besteht sie aus 1,98 Zeichen, was bedeutet, dass Silben aus einem bis zu drei Zeichen die Mehrheit darstellen (vgl. He/Lu 2007: 47). Der bedeutungsunterscheidende Ton wird schriftlich nicht repräsentiert. Die allgemein im Tibetischen gebräuchlichen Konsonantencluster werden zudem mit identifiziertem Codepunkt definiert. Beispielsweise wird das Zeichen < ག྾ > /gha/ einerseits mit dem Codepunkt U+0F43 codiert, andererseits ist es auch als Zusammensetzung mit dem Konsonant < ག > (/ga/, U+0F42) und dem hinzugefügten Konsonant < ྾ > (/-h/, U+0FB7; unabhängige Form: <ཧ>, U+0F67) darstellbar (vgl. Unicode

12.0 Chapters: Kap. 13.4: 522, Unicode 12.0 Character: U+0F00-U+0FFF). Es gibt zur Zeit mehrere Eingabeverfahren für die tibetische Schrift, aber noch kein standardisiertes Verfahren. Die Ordnung der einzugebenden Zeichen (entweder nach der Reihenfolge der Aussprache oder der Glyphen), das Tastaturlayout, die basierte Zeichencodierung und weitere variieren bei verschiedenen Verfahren ebenfalls erheblich (vgl. He/Lu 2007: 46ff, Peng 1994: 354-358).

3) Das Thai-Modell.

Anders als im Devanagari- und tibetischem Modell, in denen Grundbuchstaben als Basiszeichen einer orthograohischen Silbe stehen müssen, kann in der Thai-Schrift ein Grundbuchstabe auch in der Position eines addierenden Präfix- und Auslautzeichens stehen. Das Basiszeichen ist immer ein konsonantischer Grundbuchstabe. Virama wird nicht gebraucht. Konsonantische Buchstaben lassen sich ohne Modifikation aus Virama zu Konsonantenclustern bilden. In der Schrift werden zwar die Grundbuchstaben mit abhängigen Vokalzeichen (falls nötig) beigefügt, lassen sich aber weiter eigenständig linear nachfolgend (horizontal von links nach rechts) anordnen. Ton wird mithilfe von vier Tonzeichen angegeben, deren Gebrauchssituationen relativ kompliziert sind. Bspw. besteht das Wort < ปลา > (/plā/, *Fisch*) aus der Monosilbe < ป > /p/ (Präfixzeichen), < ล > /l/ (Basiszeichen) und < า > (rechts addierendes Vokalzeichen) (vgl.: Coulmas 1996a: 496ff, Unicode 12.0 Chapters: Kap. 16.1: 631ff, Unicode 12.0 Character: U+0E00-U+0E7F).

2.3.2 Einführung zur Devanagari-Textverarbeitung mit Beispielswort

Devanagari (Hindi: देवनागरी /devanāgarī/) ist die verbreitetste alphasyllabische Schrift und wird in dieser Arbeit als Beispiel für die Textverarbeitung derselben diskutiert. Sie ist und war für die Aufzeichnung mehrerer Sprachen zuständig, wie das archaische Sanskrit und heutige Sprachen (Hindi, Marathi und Nepali). Die zu analysierenden Beispielwörter sind aus Hindi und die Standardtastatur stammt aus Indien. Die Silbenstruktur behält die Formel ‚(((K)K)K) V' (K für Konsonant und V für Vokal) bei, weshalb hat eine orthographische Silbe drei Möglichkeiten: ein konsonantischer oder vokalischer Grundbuchstabe, eine Ligatur von einem Konsonant und einem abhängigen Vokalzeichen oder ein Konsonantencluster aus zwei oder drei Konsonanten (vgl. Unicode 12.0 Chapters: Kap. 12.1: 448, Friedrich 2006: 7ff). Eine orthographische Silbe ist sowohl linguistisch als auch bei der Textverarbeitung zu segmentieren und als Folge von einem bis zu mehreren Zeichen zu analysieren, verarbeiten und mit der Tastatur einzutippen. Eine aus mehreren Zeichen bestehende Silbe kann normalerweise in einem

Textverarbeitungsprogramm (wie Microsoft Word) nur als eine Einheit markiert und in identischem Format definiert werden. Zur Einführung in die Textverarbeitung nehme ich zunächst folgendes Beispielwort:

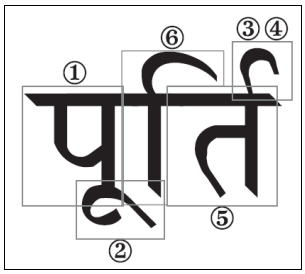


Abb. 2-8: Schriftliche Darstellung des Hindi-Wort /pūrti/ (Unicode 12.0 Chapters: Kap. 2.2: 17)

Das Wort < पूर्ति > /pūrti/ (*Erfüllung*) setzt sich aus sechs Zeichencodes bzw. zwei Silben zusammen. Zeichen ① und ② bilden die erste Silbe (/pū/) und die restlichen Zeichen die zweite. Zeichen < प > (/pa/ ①) und < त > (/ta/ ⑤) sind Grundbuchstaben (texttechnisch: Basiszeichen) und die Hauptbestandteile ihrer Silben. Zeichen ② und ⑥ sind abhängige Vokale, die in einer bestimmten Position zu einem Grundzeichen addiert werden müssen. Das Phonem < र् > /r/ bildet eine Ausnahme. Zusammen mit dem Konsonantencluster < त > /ta/ wird das Phonem zu einer hakenähnlichen kombinierenden Form (siehe ③, ④). Schreibreihenfolge sowie die Eingabe des Wortes bleiben identisch (siehe Nummerierung in Abb. 2-8). Der Eingabeprozess per PC-Tastatur und die schriftliche Ausgabe werden in Tab. 2-12 angegeben.

Prozess [78]	**Zeichen**	**phon. Wert**	**Unicode**	**Tastenbelegung**[79]	**Silbenbildung**	**Zeichenfolge**	**Glyphen**
1	प	/pa/	092A	#36 (H)		**प**	प /pa/
2	ू	/ū/	0942	#21 (T)	पू /pū/	**प ू**	पू /pū/
3	र	/ra/	0930	#37 (J)		**प ू र**	पूर /pūra/
4	्	Null	094D	#33 (D)	र् /r/	**प ू र ्**	पूर् /pūr/
5	त	/ta/	0924	#39 (L)		**प ू र ् त**	पूर्त /pūrta/
6	ि	/i/	093F	#34 (F)	ति /ti/	**प ू र ् त ि**	पूर्ति /pūrti/

Tab. 2-12: Eingabeprozess des Hindi-Worts /pūrti/ *Erfüllung*

Anhand des Beispielwortes lassen sich vier Feststellungen konstatieren: 1) ein Wort wird typographisch zusammengeschrieben und ist graphisch in Silben zu segmentieren; 2) in einer Silbe steht ein Grundbuchstabe (Konsonantenbuchstabe sowie -cluster oder Vokalbuchstabe in Initialform) im Zentrum und ein abhängiges Zeichen kann oben, unten, links oder rechts vom

[78] Die Nummer der Spalte entspricht einerseits der Eingabereihenfolge, andererseits auch der Nummer von Abb. 2-10.
[79] Der Scan-Code und die entsprechende Taste im internationalen Tastaturlayout stehen in Klammern.

Grundzeichen auftreten; 3) die spezielle Schriftform einer orthographischen Silbe kann stark von der Zeichenform der Bestandteilzeichen abweichen; 4) die logische Reihenfolge der Aussprache entscheidet über die Zeichenordnung in elektronischen Texten und bei der Eingabe.

2.3.3 Zeicheninventar der Devanagari

Die Eigenschaften der Devanagari bedingen, dass es zwei Möglichkeiten zur Verarbeitung der orthographischen Silben gibt: a) eine Silbe als eine Zeichenfolge, in der Zeichen und Zeichenglyphen nach bestimmten Regeln verbunden werden; b) eine Silbe als ein eigenständiger Zeichencode. In der Textverarbeitung haben beide Verfahren Vorzüge und Einschränkungen. Verfahren a) ist linguistisch betrachtet primär auf die alphabetischen Eigenschaften der Devanagari bezogen. Das zu codierende Zeicheninventar ist begrenzt, relativ klein und kann auf der PC-Standardtastatur belegt werden. Alle möglichen Silben der Devanagari sind per Zeichenkombination repräsentierbar. Zur Repräsentation der Silben sind jedoch weitere Techniken der Zeichenausgabe erforderlich. Verfahren b) basiert auf der syllabischen Ebene der Devanagari. Ein großer und unbegrenzter Zeichenvorrat wird zur Codierung gebraucht, so dass die Zeichen-Tasten-Repräsentation nicht verfügbar ist und spezielle Eingabemethoden entwickelt werden müssen. Eine höhere typographische Qualität der Ausgabe ist trotz reduzierter technischer Schwierigkeiten verwirklichbar. Zur Verarbeitung der Devanagari wird Verfahren a) eingesetzt, wie im Unicode. Ein Beispiel der Textverarbeitung auf diese Art der Zeichencodierung wird in Abb. 2-9 und Tab. 2-12 geschildert. Nachstehend werden die beiden Verfahren aus linguistischer und technischer Perspektive kontrastiv analysiert.

Es gibt insgesamt 48 Grundbuchstaben in der Devanagari, darunter 13 für Vokale und 35 für Konsonanten (vgl. Coulmas 1996a: 125). Davon werden insgesamt 11 Vokale und 33 Konsonanten im Hindi-Schriftsystem gebraucht (vgl. Klemm 1997: 11). Jeder vokalische Buchstabe (außer /a/) hat zwei verschiedene orthographische Symbole: die Initialform, die immer am Anfang eines Wortes oder als eigenständiges Wort auftritt; und die Media- sowie Finalform, die von Konsonanten abhängig sind. Schriftzeichen für Initialvokale werden als Vollform oder unabhängige Vokale (eng.: independent vowel) bezeichnet, hingegen heißt die Media-/Finalform Kurzform oder inhärenter Vokal (eng.: dependent vowel) (vgl. Sen 1996: 1429, Unicode Glossary: Dependent Vowel & Independent Vowel). Vokal /a/ hat nur eine Voll-, aber keine Kurzform. Neben den 13 Vokalen (inkl. /a/) gibt es fünf weitere, variierte vokalische Buchstaben, die hauptsächlich für Fremdwörter zuständig sind (Unicode 12.0 Character: U+0900-U+097F). So betragen die abhängigen Vokale insgesamt 17 und die unabhängigen 18. Darunter werden im Hindi-Schriftsystem 15 Kurz- sowie 16 Vollformen der vokali-

schen Buchstaben benutzt. Eine Übersicht ihrer Formen, Tastenbelegungen, des Unicodes und ein Beispiel listet Tab. 2-13. Die sechs variierten Buchstaben werden hellgrau markiert.

Transliteration (Transkription)	**Kurzform**					**Vollform**		
	Zeichen	Unicode	Scan-Code[80]	Ps[81]	Kom mit क[82]	Zeichen	Unicode	Scan-Code
a					क	अ	0905	#33+SH (D)
ā (aa)	ा	093E	#19 (e)	R	का	आ	0906	#19+SH (E)
i	ि	093F	#34 (f)	L	कि	इ	0907	#34+SH (F)
ī (ii)	ी	0940	#20 (r)	R	की	ई	0908	#20+SH (R)
u	ु	0941	#35 (g)	U	कु	उ	0909	#35+SH (G)
ū (uu)	ू	0942	#21 (t)	U	कू	ऊ	090A	#21+SH (T)
ṛ (r)	ृ	0943	#13 (=)	U	कृ	ऋ	090B	#13+SH (+)
ṝ (r) (nicht in hi.)	ॄ	0944	#13+AG	U	कॄ	ॠ	0960	#13+SH+AG
ḷ (l) (nicht in hi.)	ॢ	0962	#34+AG	U	कॢ	ऌ	090C	#34+SH+AG
ḹ (l)	ॣ	0963	#20+AG	U	कॣ	ॡ	0961	#20+SH+AG
ê (ai)	ॅ	0945	#03+SH (@)	O	कॅ	ऍ	090D	#02+SH (!)
ĕ (ai)	े	0946	#46 (z)	O	के	ऎ	090E	#46+SH (Z)
e	े	0947	#32 (s)	O	के	ए	090F	#32+SH (S)
ai	ै	0948	#18 (w)	O	कै	ऐ	0910	#18+SH (W)
ô (o)	ॉ	0949	#42 (\)	R	कॉ	ऑ	0911	#42+SH (|)
ŏ (o)	ॊ	094A	#01 (`)	R	कॊ	ऒ	0912	#01+SH (~)
o	ो	094B	#31 (a)	R	को	ओ	0913	#31+SH (A)
au	ौ	094C	#17 (q)	R	कौ	औ	0914	#17+SH (Q)

Tab. 2-13: Vokalische Buchstaben der Devanagari[83]

Die Konsonantenbuchstaben der Devanagari werden in den meisten Fällen nach der Reihenfolge der Artikulationsart angeordnet und aufgelistet. Auf Basis mancher Grundbuchstaben können mit Nuqta-Zeichen einige weitere Konsonanten abgeleitet werden. Solche Zeichen können, sowohl bei der Eingabe als auch bei der Zeichencodierung, nicht nur zerlegt sondern auch einheitlich codiert werden. Bspw. wird < क़ > /qa/ als gesamtes Zeichen mit dem Wert ‚U+0958' oder als Zeichenfolge von < क > /ka/ (U+0915) und < ़ > (Nuqta, U+093C) repräsentiert. In Tab 2-14 werden nur die 33 konsonantischen Grundbuchstaben des Hindi-Schriftsystems angegeben.

Schriftzeichen	**Transl.** (Transk.)[84]	**IPA**	**Unicode**	**Scan-Code**	**Beispielsilben mit /i/**
क	ka	[k]	0915	#38 (k)	कि
ख	kha	[kʰ]	0916	#38+SH (K)	खि
ग	ga	[g]	0917	#24 (o)	गि

[80] In der Klammer wird das entsprechende Zeichen des Scan-Codes im US-amerikanischen QWERTY-Layout angegeben.
[81] Ps steht für Position: Oben (Abk.: O), Unten (U), Links (L) und Rechts (R) eines konsonantischen Buchstaben.
[82] Die Glyphe von der Ligatur, die aus dem Buchstaben /ka/ und dem Vokal zusammengesetzt ist.
[83] Vgl.: Unicode 12.0 Character: U+0900-U+097F, Friedrich 2006: 18 & 41f.
[84] Falls die Transliteration mit einem Sonderbuchstaben außerhalb der ASCII-Codierung dargestellt wird, wird die entsprechende Transkription in der Klammer angegeben, auf der die Hindi-‚namaste'-Eingabemethode basiert.

Schriftzeichen	Transl. (Transk.)[84]	IPA	Unicode	Scan-Code	Beispielsilben mit /i/
घ	gha	[gʱ]	0918	#24+SH (O)	घि
ङ	ṅa (na)	[ŋ]	0919	#23+SH (U)	ङि
च	ca	[c, t͡ʃ]	091A	#40 (;)	चि
छ	cha	[cʰ, t͡ʃʰ]	091B	#40+SH (:)	छि
ज	ja	[ɟ, d͡ʒ]	091C	#26 (p)	जि
झ	jha	[ɟʱ, d͡ʒʱ]	091D	#26+SH (P)	झि
ञ	ña (na)	[ɲ]	091E	#28+SH (})	ञि
ट	ṭa (ta)	[ʈ]	091F	#41 (')	टि
ठ	ṭha (tha)	[ʈʰ]	0920	#41+SH (")	ठि
ड	ḍa (da)	[ɖ]	0921	#27 ([ )	डि
ढ	ḍha (dha)	[ɖʱ]	0922	#27+SH ({)	ढि
ण	ṇa (na)	[ɳ]	0923	#48+SH (C)	णि
त	ta	[t̪]	0924	#39 (l)	ति
थ	tha	[t̪ʰ]	0925	#39+SH (L)	थि
द	da	[d̪]	0926	#25 (o)	दि
ध	dha	[d̪ʱ]	0927	#25+SH (O)	धि
न	na	[n]	0928	#49 (v)	नि
प	pa	[p]	092A	#36 (h)	पि
फ	pha	[pʰ, ɸ]	092B	#36+SH (H)	फि
ब	ba	[b]	092C	#22 (y)	बि
भ	bha	[bʱ, β]	092D	#22+SH (Y)	भि
म	ma	[m]	092E	#48 (c)	मि
य	ya	[j]	092F	#55 (/)	यि
र	ra	[r]	0930	#37 (j)	रि
ल	la	[l]	0932	#51 (n)	लि
व	va	[w, ʋ]	0935	#50 (b)	वि
श	śa (sha)	[ɕ, ʃ]	0936	#52+SH (M)	शि
ष	ṣa (sha)	[ʂ]	0937	#53+SH (<)	षि
स	sa	[s]	0938	#52 (m)	सि
ह	ha	[ɦ]	0939	#23 (u)	हि

Tab. 2-14: Konsonantische Buchstaben der Devanagari[85]

Im Hindi sind die aus einem Konsonant und einem Vokal (KV-Modell genannt) zusammengesetzten Silben die am häufigsten verwendeten Strukturen.[86] Schriftlich werden solche Silben entweder als eigenständiger konsonantischer Buchstabe oder als Zusammensetzung von einem Grundbuchstaben und einem abhängigen Vokal dargestellt. Wenn sie als Zeichen codiert würden, wären dafür mindestens 363 (33×11) Codepunkte vonnöten. Wenn die variierten Konsonanten und Vokale mitgezählt werden, wären es 792. Bei aus demselben Konsonanten gebildeten Konsonant-Vokal-Ligaturen gleicher Schriftgröße bleibt der Grundbuchstabe zwar graphisch meist identisch, wird aber abhängig vom Vokal in verschiedenen Größen dargestellt

[85] Vgl.: Unicode 12.0 Character: U+0900-U+097F, Friedrich 2006: 19, Coulmas 1996a: 125.
[86] Vgl. Kap. 2.3.4; anhand der Analysen von Wörtern aus fünf zufällig ausgewählten Seiten, in: Klemm, 1997, Stichwörterliste.

(siehe Spalte ‚Kom. mit /ka/' in Tab. 2-13). Werden die Konsonantencluster zur Zeichencodierung zusätzlich berücksichtigt, könnten tausende Codepunkte mehr erforderlich sein. Allein für die von zwei Grundkonsonanten zusammengesetzten Cluster gibt es mindestens 1.089 Varianten (33×33). Errechnete man alle Cluster von drei Konsonanten, potenzierten sich die Möglichkeiten auf 35.937. Allgemein sind nur ein kleiner Teil davon (ca. 150 Konsonantencluster) im Hindi gebräuchlich (vgl. Friedrich 2006: 89). Inklusive der gebräuchlichen Konsonantencluster in Sanskrit, Marathi und Nepali beläuft sich die Gesamtzahl – Überschneidungen berücksichtigend – auf weniger als 400. Wenn die Konsonantencluster als Zeichencodes verarbeitet würden, wäre die Codierung der gebräuchlichen Cluster obligatorisch und die Codierung aller anderer gültigen Varianten optional zu ergänzen, damit Fremdwörter, Eigennamen usw. in Devanagari umgeschrieben und vom Computer verarbeitet werden können. Wenn die aus einem Konsonantencluster und einem abhängigen Vokal zusammengesetzten Silben noch mit bestimmten Codepunkten definiert würden, wären zahlreiche Zeichencodes (2.163 bis maximal ca. eine Million; der minimale Wert setzt sich aus der Formel 150×12+33×11 zusammen) notwendig, die insgesamt alle 17 Ebenen des Unicodes umfassten. Die Zeichencodierung von Silbeneinheiten ist deswegen technisch und theoretisch untauglich durchzuführen. Durch Analysen des möglichen Zeichenvorrats ist festzustellen, dass es am sinnvollsten ist, die Grundbuchstaben und abhängigen Zeichen als kleinste codierte Elemente bei der Devanagari-Textverarbeitung zu betrachten. Repräsentation und Glyphen-Ausgabe einer Silbe sind deswegen eine große Herausforderung. Die Verbindungsarten zweier konsonantischer Buchstaben können in sechs Fälle untergegliedert werden (vgl. Friedrich 2006: 86):

I) Horizontale Verbindung von zwei Buchstaben mit Weglassung von dem senkrechten Strich oder von dem rechtsstehenden Strich des ersten Buchstaben: < ग्द > /gda/ (von < ग् > /g/ und < द > /da/) und < क्क > /kka/[87] (von < क् > /k/ und < क > /ka/);

II) Vertikale Verbindung von zwei Buchstaben mit Weglassung des Querstrichs des zweiten Buchstabens, dargestellt unter dem ersten: <द्द> /dda/ (von zweimal < द > /da/ und < ् > dazwischen);

III) Umwandlung von /r/ in einen Haken auf dem Querstrich, wenn er als erster Buchstabe einer Konsonantenligatur auftritt: < र्क > /rka/;

[87] /kka/ ist sowohl horizontal als auch vertikal verbindbar.

IV) Umwandlung von /ra/ in einen schrägen Strich, wenn er als zweiter Buchstabe einer Konsonantenverbindung angewendet wird: < क्र > /kra/;

V) Ausnahmen bei der Verformung des Buchstaben < श > /śa/ in manchen Fällen, wie z.B.: < श्च > /śca/ (von < श् > und < च >), < श्न > /śna/ (von < श् > und < न >) und < श्व > /śva/ (von < श् > und < व >);

VI) Ausnahmen, die nicht den oben genannten fünf Kategorien zugeordnet werden können, etwa: < द्ध > /ddha/ (von < द् > /d/ und < ध > /dha/).[88]

Das zu codierende Zeicheninventar beträgt ca. hundert, inkl. Grundbuchstaben, abhängigen Zeichen und weiteren internationalen sowie nativen Sonderzeichen. Mit diesem Verfahren ist einerseits die Zeichen-Tasten-Repräsentation auf der Tastatur möglich, andererseits sind alle möglichen orthographischen Silbeneinheiten, egal ob allgemeingebräuchliche oder seltene Varianten, mit Computern darstellbar. Um die aus Zeichensequenz bestehenden Silben typographisch auszugeben, muss eine Serie von Glyphenregeln zu einem Zeichencode definiert werden. Im Unicode ist ein konsonantischer Grundbuchstabe mindestens mit fünf Glyphen verbunden (vgl. Unicode 12.0 Chapters: Kap. 12.1: 457):

a) die Form als alleinstehender Buchstabe (die nominale Glyphe, Englisch *nominal glyph*) – C_n, wie < क > /ka/;

b) die Form als mit einem Vokal gebildeter Konsonant (lebender Konsonant oder *live consonant*) – C_l, wie < कु > /ku/ (bei der Ausgabe ist diese Schriftform die Kombination von C_l-Form von /ka/ – < क > und die kombinierende Glyphe des abhängigen Vokals /u/ – < ु >);

c) die Form des Konsonanten mit Virama (Totkonsonant oder *dead consonant*) – C_d, wie < क् > /k/ und < त् > /t/;

d) die Form der Halbkonsonanten (*half-consonant*) – C_h, wie <क> in der Kombination <क्क> /kka/ (die Glyphenkombination von < क > in C_h-Form und in C_n-Form), die Glyphenkombination entspricht der Verbindungsart I & II);

e) keine eigene Form, sondern ersetzt durch eine andere, für eine Silbe (in Formel ausgedrückt als $X.Y_n$) stehende Glyphe, die mit Kombination der C_h-Form des vorderen und C_n-Form des letzten Konsonantenzeichens nicht darstellbar ist. Diese Form entspricht der

[88] Es gibt viele Ligaturenausnahmen, die wegen des eingeschränkten Glyphenvorrats des Computers nicht ausgegeben werden können, wie /tta/, /kta/, /dna/ usw.

Verbindungsart VI und kann als L_n bezeichnet werden, wie z.B. < क्ष > /kssa/ von < क् > /k/ und < ष > /ssa/.

Der Konsonant < र > /ra/ ist ein Ausnahmefall, da er meistens in Sonderform als ein kombinierendes Symbol bei der Konsonantenclusterbildung auftritt. Als vorderer Konsonant tritt er wie ein kombinierendes Zeichen in Hakenform auf dem Grundbuchstaben auf, wie bspw. in < र्क > /rka/ (siehe Verbindungsart III). In diesem Beispiel ist die ausgegebene Schriftform von /rka/ die Zusammensetzung der C_l-Form von < क > /ka/ und der kombinierenden Glyphen von < र > /ra/ (RA_{sup} genannt). Als letzter Konsonant wird er als ein nach links unten gezogener Strich links oder unterhalb von dem Grundzeichen dargestellt, wie in < क्र > /kra/ und < ट्र > /ttra/ (der Verbindungsart IV entspricht; RA_{sub} genannt).

Außer den 35 vokalischen und 44 konsonantischen Schriftzeichen gehören des Weiteren sonstige Zeichen zu der Devanagari, die auch encodiert und auf der Tastatur belegt werden müssen. Die Zeichen können drei Arten untergeordnet werden: die abhängigen sekundären Zeichen (siehe Tab. 2-15), die internationalen Sonderzeichen und native Symbole (siehe Tab. 2-16: Devanagari-Ziffer und native Interpunktionszeichen).

Die sekundären Lautzeichen umfassen das Auslassungszeichen ‚Virama', das Lautumwandlungszeichen ‚Nukta' und ein paar Auslautzeichen. Im Hindi häufig als Auslaut einer Silbe auftauchende Konsonanten sind: < ं > /-ṃ/, < ँ > /-m/ oder /-n/ und < ः > /-ḥ/. Sie können wie die Kurzformzeichen auf einen Grundbuchstaben beigefügt werden, wie z.B. < कैं > /kaiṃ/ (vgl. Friedrich 2006: 84f, Unicode 12.0 Character: U+0900-U+097F). Wenn der Auslaut einer Silbe ein weiterer Konsonant ist, steht der entsprechende Konsonantenbuchstabe zur Verfügung.[89] Zur Konsonantenclusterbildung ist Virama in diesen Fällen immer unentbehrlich. Nukta wird zur Änderung des Konsonanten zu einem anderen verwandten Phonem eingesetzt, wie < क़ > /qa/. Manche variierten Konsonanten können sowohl als ein gesamtes Zeichen als auch als eine Folge von Grundzeichen und Nukta eingegeben und verarbeitet werden (siehe Tab. 2-14).

Symbol	Bezeichnung	phon. Wert	Unicode	Scan-Code	Beispiel
ं	Anusvara	-ṃ (Nasallaut im Auslaut)	0902	#47 (x)	कं /kām/
ँ	Candrabindu	-m/-n (Nasallaut im Auslaut)	0901	#47+SH (X)	कँ /kam/

[89] In Fällen, in denen ein konsonantischer Auslaut am Wortende auftritt, wird Virama weggelassen, wie <अंत> /ant/ *Ende.*

Symbol	Bezeichnung	phon. Wert	Uni-code	Scan-Code	Beispiel
ः	Visarga	-ḥ (im Auslaut)	0903	#12+SH (_)	कः /kaḥ/
्	Virama/ Auslas-sungszeichen	(Nullvokal)	094D	#33 (d)	क् /k/
़	Nukta	(Umwandlung in einen neuen Buchstaben)	093C	#28 (})	क़ /qa/ (U+0958)

Tab. 2-15: Die inhärenten sekundären Zeichen der Devanagari[90]

Die aus Indien stammende digitale Zahlschrift hat wegen ihrer tausendjährigen Verbreitung und Entwicklung weltweit viele verschiedene graphische Varianten. Im indischen Schriftkreis sind u.a. Devanagari-, tibetische und Thai-Ziffern zu unterscheiden, die heute parallel zur internationalen arabischen Zahlschrift gebraucht werden.

Es gibt außerdem auch native Interpunktionszeichen (siehe Tab. 2-16), besonders das in der indischen Schrift anstelle eines Punkts am Satzende verwendete Danda < । > und verdoppeltes Danda < ॥ >. Nach Berechnung der aufgelisteten Zeichen (von Tab. 2-13 bis Tab. 2-16) beträgt das minimale Zeicheninventar der Devanagari mindestens 98. Im Unicode werden insgesamt 107 Zeichen im Block ‚Devanagari' codiert.

Zeichen	Wortform/ Bezeichnung	Internationale Form	Unicode	Scan-Code
०	शून्य /śūnyᵃ/	0	0966	#11 (0)
१	एक **/ek/**	1	0967	#02 (1)
२	दो **/do/**	2	0968	#03 (2)
३	तीन **/t**īn/	3	0969	#04 (3)
४	चार **/cā**r/	4	096A	#05 (4)
५	पाँच **/p**ā̃c/	5	096B	#06 (5)
६	छ **/cha/**	6	096C	#07 (6)
७	सात **/s**āt/	7	096D	#08 (7)
८	आठ /āṭh/	8	096E	#09 (8)
९	नौ **/nau/**	9	096F	#10 (9)
।	Danda (für Satzende)		0964	#54+SH (>)
॥	Doppeltes Danda		0965	#54+AG
ऽ	Avagraha		093D	#54+SH+AG
॰	Abkürzungszeichen		0970	#53+AG

Tab. 2-16: Die Zahlzeichen und nativen Interpunktionszeichen der Devanagari

2.3.4 Zeichencodierung und Zusammensetzung der Schriftzeichen in einer Silbe

Sind für die Zeichencodierung nur ASCII- und Devanagari-Zeichen nötig, kann eine 8-Bit-Codierung genügend Codepunkte anbieten (wie im ISCII: Indian Script Code for Information Interchange). In der universalen Codierung Unicode werden die Devanagari-Zeichen mit 16

[90] Vgl. Unicode 12.0 Character: U+0900-U+097F, Friedrich 2006: 84f, Coulmas 1996a: 125f.

Bit definiert; darunter 12 besetzte (U+0900 – U+097F, in UTF-8 drei Bytes sowie in UTF-16 zwei Bytes). Eine Silbe setzt sich mindestens aus einem Zeichen (ein Konsonant mit /a/ oder ein eigenständiger Vokal) und höchstens aus sieben Zeichen zusammen (wie z.B. < स्त्रिं > /striṁ/ aus den Buchstaben /sa/, /ta/, /ra/, zweimal Virama, einmal Vokalzeichen /i/ und ein sekundäres Lautzeichen /-ṁ/). Der Informationsgehalt einer Silbe wird im ISCII mit einer bis 7 Bytes (8-56 Bits) und in UFT-16 mit 2 bis 14 Bytes (16-224 Bits) repräsentiert (vgl. IS 13194:1991, Unicode 12.0 Character: U+0900-U+097F).

Wenn statt den Buchstaben die Silbe als zu codierende Einheit angewandt würde, wäre eine 16-Bits-Codierung einsetzbar, insofern nur die allgemein gebräuchlichen, mehrere hundert umfassenden Silben aufgenommen würden. So könnten die häufigen Silben aufgrund ihres Informationsgehalts identisch groß zwei Bytes betragen. Seltene Silben müssten aber mit vier Bytes oder ohne Codepunkte repräsentiert werden.

Um herauszufinden, welche Art der Zeichencodierung aus Sicht des Informationsgehalts effektiver für Devanagari ist, habe ich eine statistische Analyse mit Wörtern aus vier zufällig ausgewählten Seiten eines Wörterbuchs und einem Hindi-Artikel mit 3.735 Wörtern (ohne Berücksichtigung der Interpunktions- und Zahlzeichen) durchgeführt.[91] Insgesamt 233 Hindi-Wörter im Wörterbuch wurden analysiert, sie enthalten 873 orthographische Silben. Unter solchen Silben gibt es 56 Konsonantencluster, darunter 53, die aus zwei oder drei Clustern aus drei Konsonanten bestehen. Im analysierten Hindi-Artikel kommen insgesamt 16.097 Devanagari-Zeichen vor. Die statistischen Ergebnisse können wie folgt zusammengefasst werden: 1) ein Hindi-Wort hat durchschnittlich 3,75 orthographische Silben; 2) die Konsonantencluster werden mit 6,41% relativ selten verwendet, die aus drei Konsonanten bestehenden Cluster treten nur in 0,03% der Fälle auf; 3) Silben mit der Struktur ‚(K)V' sind mit 93,59% klar in der Mehrheit; 4) ein Wort enthält nach der Wörter- sowie Zeichenzählung des Artikels durchschnittlich 4,43 Zeichen; 5) statistisch setzt sich eine Silbe im Schnitt aus 1,18 Zeichen zusammen. Durch die Ergebnisse der Stichprobenuntersuchung ist zu schlussfolgern, dass die 8-Bits-Codierung wie in ISCII für die Textverarbeitung der Devanagari am effektivsten ist.

Wie vorher erwähnt, ist die Glyphendarstellung einer orthographischen Silbe eine der größten Schwierigkeiten. Wie im letzten Kapitel vorgestellt wurde, sind die Regeln der Silben-Glyphen-Bildung relativ kompliziert und reich an Ausnahmen. Im Prinzip ist es aber möglich, einen allgemeinen proportionalen Silbenaufbau darzustellen, der für die meisten Fälle gültig ist. Anhand des Beispiels in Abb. 2-8 (S. 90), der Form der abhängigen Zeichen in

[91] Die Stichwörter stammen aus Klemm 1997, der Artikel von *http://hi.bharatdiscovery.org/india* [Abruf: 2015-07-20].

Tab. 2-14 (S. 93) und 2-15 (S. 96f) sowie vergleichbarer Erkenntnisse im Tibetischen habe ich ihn wie folgt gezeichnet.

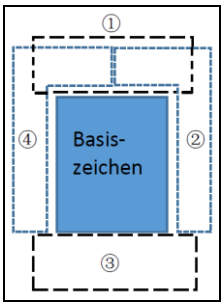


Abb. 2-9: Graphische Silbenstruktur für Glyphen-Zusammensetzungen

Im Zentrum wird der Hauptteil eines Basiszeichens oder eines Konsonantenclusters angepasst. Mit Hauptteil sind die Striche innerhalb der Grundlinie gemeint. So gehört bspw. die kurze Kurve oben von dem Vollvokal < अ > /a/ der ersten Zone an, während der restliche Teil im Zentrum angelegt wird. Die vier Zonen symbolisieren hauptsächlich die möglichen Stellen der Glyphen-Addierung. Anhand der Form eines abhängigen Zeichens wird es graphisch der entsprechenden Position beigefügt, wie etwa im Beispielwort < पूर्ति > (/pūrti/ *Erfüllung*), wo /ū/ der zweiten Zone, die kombinierende Form von /r/ der ersten Zone und /i/ der vierten Zone beigefügt wird (vgl. auch Abb. 2-8). In der ersten und vierten Zone einer Silbe können zwei oder mehrere abhängige Zeichen auftreten.

2.3.5 Hindi-Tastaturlayout und die Reihenfolge der einzugebenden Zeichen

Wie in Abb. 2-9 dargestellt wird, kann ein in Kurzform dargestellter Vokal graphisch sowohl vor dem Basiszeichen (links oder oben) als auch nach dem Basiszeichen (rechts oder unten) vorkommen. Es gibt bei der Zeichenordnung der Eingabe und der Textverarbeitung somit zwei Möglichkeiten: 1) nach der logischen Reihenfolge der Aussprache, d.h. nach der Reihenfolge der Phoneme einer Silbe (Konsonant-Vokal-Auslaut); 2) nach der Ordnung der Ausgabe. Im zweiten Fall werden die kombinierenden Zeichen, die oben sowie links vom Grundbuchstaben eingesetzt werden, vor dem Basiszeichen eingegeben und ihre Bitfolge vor der des Basiszeichens gespeichert. Die erste Variante stimmt mit der logischen Ordnung überein und ist daher umgangsfreundlicher. Die logische Ordnung (eng.: *logical order*) ist „[t]he order in which text is typed on a keyboard. For the most part, logical order corresponds to phonetic order“ (Unicode Glossary: Logical Order). Mit dieser Art der Zeichenordnung als Grundlage sind viele linguistische Analysen und Anwendungen computergestützt leichter durchzuführen, bspw. Text-to-speech, Silbensegmentation usw. Um die logische Ordnung auszuführen, müssen die Textverarbeitungssoftwares in der Lage sein, Zeichenordnungen entsprechend der

auszugebenden Ordnung der Glyphen neu zu gestalten (vgl. Unicode 12.0 Chapters: Kap. 12.1: 461).

Die Eingabe der Devanagari wird in den meisten Fällen mittels des nationalen indischen Tastaturlayouts durchgeführt, welches als Hindi-Inscript bezeichnet wird (siehe Abb. 2-10). Die belegten Zeichen stimmen mit den 98 bis 107 zu codierenden Devanagari-Zeichen im Allgemeinen überein. Da internationale Ziffern und Sondersymbole hinzutreten, wurde beim Hindi-Inscript-Tastaturlayout eine Vierfachbelegung eingeführt. Verknüpft man die Analysen zur Tastenbelegung des nativen Layouts (vgl. Tab. 2-13 & 2-14, S. 92f) ist festzustellen, dass eine bestimmte Regel vorherrscht.

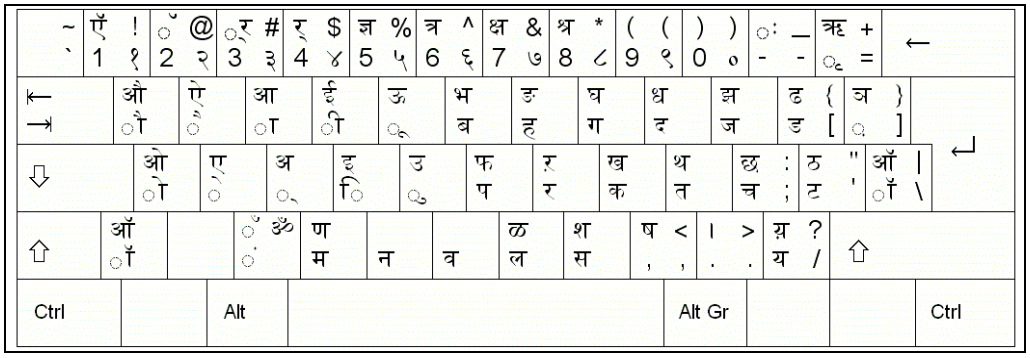


Abb. 2-10: Das Hindi-Inscript-Tastaturlayout[92]

Aus dieser Abbildung sowie den Tabellen 2-13/2-14 wird ersichtlich, dass die linke Hälfte des alphabetischen Blocks hauptsächlich für vokalische und die rechte eher für konsonantische Zeichen zuständig ist. Wie in Kap. 2.1.2 (S. 65) erwähnt wurde, ist der Handwechsel am effektivsten und optimal für längeres Tastaturschreiben. Die Verteilung entspricht einerseits der Regel, dass Konsonanten-Vokal-Wechsel am häufigsten auftreten. Andererseits wird das Auswendiglernen der Tastenbelegung erleichtert, im Vergleich zum QWERTY-Layout.

Die zwei Formen eines Vokals und zwei phonetisch eng verwandte Konsonanten sind häufig auf derselben Taste mit verschiedenen Umschaltungsebenen belegt. Gemäß dieses Prinzips sind < क > /ka/, < ख > /kha/, < क़ > /qa/ und < ख़ > /kha/, welche stets phonetisch oder graphisch denselben Stamm haben, die Zeichen von der ersten bis zu vierten Belegung der Taste mit dem Scancode #38 (Taste: ‚K'). < ि > (Kurzform /i/), < इ > (Vollform /i/), < ऌ > (/ḷ/, ein Sondervokal und wenig verwendet) und < ऌ > (Vollform /ḷ/) werden ebenso alle mit Taste #34 (F-Taste) repräsentiert (siehe Abb. 1-2 [S. 10] & Abb. 2-10).

Der Eingabeprozess von dem Wort < पूर्ति > /pūrti/ in Tab. 2-12 ist ein Beispiel für das Eingabeverfahren der Devanagari. In dem Beispielwort ist die Reihenfolge bei der Zeichen-

[92] Quelle: *http://ildc.in/images/inscript-kb/Devnagari-Inscript-Layout.jpg* [Abruf: 2015-07-21].

ausgabe: C_l-Glyphe von /pa/ – /ū/ in Kurzform – /i/ in Kurzform – Sonderform von /r/ – Cl-Glyphe von /ta/. Die Entwicklung der Schriftform bei Worteingabe wird in Tab. 2-12 Schritt für Schritt aufgezeigt.

2.3.6 Fazit zur Textverarbeitung der Devanagari

Nach den Erforschungen der Devanagari-Textverarbeitung können die folgenden Hauptprinzipien festgehalten werden.

Erstens stellen segmentale Grundelemente die grundlegende Einheit der Textverarbeitung dar, d.h. sowohl die Textverarbeitung als auch die Zeichenbelegung sind auf die alphabetische Ebene der Schrift bezogen, obwohl die Silbe die effektive Einheit darstellt. Die Grundelemente, zu denen die mit /a/ kombinierten Konsonantenzeichen und vokaltragende abhängige sowie eigenständige Zeichen zugehörig sind, haben einen eingeschränkten Umfang mit einem Zeicheninventar unter hundert. Dank des kleinen Zeichenvorrats kann ein Devanagari-Zeichen im nationalen Standard mit einer 8-Bits-Folge repräsentiert und auf einer bestimmten Taste belegt werden. Alle Silben werden (trotz der in manchen Fällen vorkommenden Verformung von der Grundform) als Kombination einer Zeichenfolge verarbeitet.

Zweitens behält die Textverarbeitung und die Eingabe der Devanagari immer die von der Phonetik bedingte Reihenfolge, obwohl kombinierende Zeichen graphisch vor dem Grundbuchstaben (oben oder links davon) angezeigt werden können. Die logische Ordnung erleichtert das Schreiben sowie linguistische Analysen mit dem Computer, aber erfordert mehr Ausgabetechniken.

Das kleine alphabetische Zeicheninventar und die logische Reihenfolge bedingen, dass an die Ausgabe der Devanagari große technische Herausforderungen gestellt werden. Einerseits herrschen komplizierte Zeichen-Glyphen-Beziehungen vor. Ein Konsonantenzeichen hat mindestens vier Glyphen für sich selbst und weitere Kombinationsformen als Konsonantencluster mit einem anderen Konsonant. Zur Ausgabe muss außerdem die Zeichenordnung den Glyphen entsprechend umgestellt werden.

2.4 Koreanische Textverarbeitung

Sowohl in den indischen Schriften als auch in Hangul herrschen effektive syllabische Einheiten vor, die weiter in Komponenten segmentiert werden können. Im Prinzip gibt es für die beiden Schriftarten zwei Möglichkeiten für die zu codierende Einheit der Textverarbeitung, nämlich Buchstaben und Syllabar. In Kap. 2.3 wurde analysiert, warum für die indischen Schriften die Codierung der kleinsten unzerlegbaren phonetischen Symbole (welche Grund-

buchstaben und kombinierende Vokalzeichen umfassen) bevorzugt wird. Im Gegensatz zu Devanagari basiert die koreanische Zeichencodierung primär auf den Silbenblöcken und sekundär auf den alphabetischen Komponenten. Obwohl die koreanischen Buchstaben ebenso codiert werden, stellen die für Silben stehenden Zeichencodes in den koreanischen Texten im Normalfall die Grundeinheit dar (vgl. Unicode 12.0 Chapters: Kap. 18.6: 737). Die Codierung des Silbenblocks bestimmt, dass die Eingabe des Koreanischen von der Buchstaben-Silbenzeichen-Konversion bedingt sein muss, weshalb eine Eingabemethode vonnöten ist (vgl. hierzu Kap. 1.1, S. 8f). Nachfolgend wird der Frage nachgegangen, warum bei der koreanischen Textverarbeitung die Codierung der Silbenzeichen bevorzugt wird und wie die Eingabe abläuft.

2.4.1 Hangul – Unterschiede zu alphasyllabischen Schriften und allgemeine Übersicht

Im Vergleich zu alphasyllabischen Schriften, deren grammatologischen Prinzipien und Eingabeverfahren in Kap. 2.3 erforscht wurden, zeigt Hangul trotz mancher Ähnlichkeiten folgende grundlegende Unterschiede.

1) Die Komponenten der indischen (alphasyllabischen) Schriften unterscheiden sich in Grundbuchstaben (Basiszeichen) und die von Buchstaben abhängigen Symbole (kombinierende Zeichen). Dem hingegen sind die Komponenten des Hanguls gleichwertige Bestandteile der Silbe, die nach einer bestimmten Ordnung und in einer bestimmten proportionalen Lage innerhalb eines Silbenzeichens dargestellt werden. Die alphabetische Komponente des Hanguls, die einen phonetischen Wert, aber keine Bedeutung trägt, heißt Jamo (Hangul: 자모, Hanja: 字母[93]) (vgl. Lunde 2009: 58).

2) Die Buchstaben der indischen Schriften können eigenständig eine Silbe repräsentieren. Im Vergleich dazu muss ein koreanisches Syllabar aus zwei oder drei Jamo bestehen, die jeweils für An-, In- und optional Auslaut stehen. Ein Jamo ist des Weiteren zusammengesetzt aus einem oder zwei Grundbuchstaben (vgl. ibid: 59).

3) In Devanagari gibt es häufig Konkurrenz zwischen der logischen sprachlichen und der graphischen auszugebenden Zeichenordnung. Im Gegensatz dazu stimmen innerhalb eines koreanischen Syllabars die Reihfolge der Jamo auf geometrischer und phonetischer Ebene überein (vgl. Unicode 12.0 Chapters: Kap. 18.6: 737f).

[93] Da es im Koreanischen weder Numerus- noch klare Genus- und Kasus-Unterscheidung der allgemeinen Nomen gibt, wird die Plural- und Genitivform des Wortes *Jamo* in der Arbeit wie seine Sigularform im Nominativ im geschrieben.

Wie in Kap. 1.3.2 bezüglich Hangul skizziert wurde, kann die koreanische Schrift als alphabetosyllabische Schrift bezeichnet werden. Es existiert zwar ein bestimmter Graphem-Phonem-Zusammenhang, aber das aus Jamo zweidimensional zusammengesetzte Silbenquadrat bleibt Hauptoperationseinheit (vgl. Coulmas 1996b: 1358). Dies wirkt sich stark auf die Rezeptionspraxis der Zeichen aus: „Dies bedeutet für die Praxis des Lesens und Schreibens, daß man koreanische Wörter nicht ‚buchstabiert', sondern in Silben gliedert, die man entsprechend schreibt und liest" (Haarmann 1991: 358, nach Dürscheid 2006: 93).

Wie in Kap. 1.3.5 analysiert wurde, ist die grundlegende graphische Einheit der Hangul primär buchstäblich, sekundär syllabisch (siehe S. 35). Da die koreanische Sprache historisch in chinesischer Schrift geschrieben wurde, gibt es zwischen Hangulsyllabar und chinesischen Schriftzeichen (präziser: Hanja) Entsprechungen. In Hangul entspricht ein Silbenzeichen zudem in vielen Fällen mehreren homophonetischen Sinogrammen und kann daher mit mehreren Morphemen verknüpft werden. Besonders bei sinokoreanischen Wörtern (koreanischen Wörtern mit chinesischer Herkunft), die 60% bis 70% des koreanischen Wortschatzes ausmachen, hat ein Silbenzeichen in vielen Fällen Entsprechung zu einem bis mehreren Morphemen (vgl. Sohn 1999: 13, Atsuji 1994: 448).

Die Bezeichnung für die koreanische Schrift Hangul ergibt sich aus der Transkription des südkoreanischen Wortes <한글> (in Hanja: 韓文). Nordkoreanisch heißt sie auch <조선글> (朝鮮文, Choseongul), aus historischem und neutralem Standpunkt aus betrachtet <정음> (正音, Jeongum) sowie <언문> (諺文, Eonmon). Sie wurde in den 1440er Jahren unter der Führung von König Sejong erfunden. Bis dahin wurde die koreanische Sprache seit dem Anfang der literarischen Geschichte (ca. ab 1. Jh.) in der chinesischen Schrift aufgeschrieben, die in Koreanisch als ‚Hanja' Bezeichnung fand (vgl. Coulmas 1996a: 273f). ‚Hanja' ist terminologisch die Gesamtbezeichnung für die im koreanischen Schriftsystem gebrauchten chinesischen Schriftzeichen und die nativ erfundenen morphologischen Schriftzeichen. Seit Ende des Zweiten Weltkriegs ist Hangul in Nord- sowie Südkorea offizielle Hauptschrift. Die tausend Jahre offiziell verwendete Schrift Hanja wurde im Gegensatz dazu im Norden abgeschafft, im Süden in ihrer Verwendung eingeschränkt (ibid.). Politisch bedingt weist das Koreanische sowohl mündlich als auch schriftlich in beiden Teilstaaten erhebliche Unterschiede auf. In dieser Dissertation wird bezüglich der koreanischen Schrift und Textverarbeitung hauptsächlich die Situation in Südkorea fokussiert.

2.4.2 Simples sowie komplexes Jamo und Syllabarbildung

Die graphischen Elemente des Hanguls können in zwei Ebenen klassifiziert werden: die alphabetische Komponente Jamo und das Silbenzeichen. Jamo unterscheidet sich weiterhin in simples sowie komplexes Jamo. Das erstere besteht aus einem einzigen Grundbuchstaben, während das letztere eine Ligatur aus zwei simplen konsonantischen oder vokalischen Jamo sein muss. Wie in einer alphabetischen Schrift ist ein simples Jamo jene graphische Grundform, die segmental nicht zerlegbar und phonemtragend ist. Es gibt im heutigen Hangul insgesamt 24 simple Jamo (Grundbuchstaben), darunter vierzehn für Konsonanten und zehn für Vokale (vgl. Coulmas 1996a: 273f). Die historischen, heute abgeschafften Buchstaben werden nicht berücksichtigt.

Ein Hangulsyllabar setzt sich aus den jeweils für An-, In- und (optional) Auslaut bestehenden Komponenten zusammen. Solche Komponenten können entweder simple oder komplexe Jamo darstellen. Funktionsabhängig können Jamo in drei Komponentenklassen analysiert werden (vgl. Unicode 12.0 Chapters: Kap. 18.6: 735).

Der Anlaut wird ‚Choseong' (Hangul: 초성, Hanja: 初聲) genannt und ist stets der erste, führende Jamo einer Silbe. Als Choseong können alle 14 konsonantischen Grundbuchstaben und fünf Doppelkonsonanten (Verdopplung eines Grundkonsonanten) auftreten, die insgesamt 19 Varianten betragen (siehe Tab. 2-17, vgl. Unicode 12.0 Character: U+1100-U+11FF). Falls der Anlaut einer Silbe leer ist, muss an dieser Stelle der für Nullkonsonant stehende Buchstabe < ㅇ > eingesetzt werden (vgl. Coulmas 1996a: 277).

Der koreanische Terminus für Inlaut heißt ‚Jungseong' (중성, 中聲). Es gibt insgesamt zehn vokalische Grundbuchstaben und elf mögliche komplexe Jamo, die als Silbenkern im modernen Hangul funktionieren. Die 21 Inlaut-Varianten werden in Tab. 2-18 angegeben (vgl. Unicode 12.0 Character: U+1100-U+11FF).

‚Jongseong' (종성, 終聲) bezeichnet den koreanischen Auslaut. Die Bildungsmöglichkeiten umfassen alle konsonantischen Grundbuchstaben, zwei Doppelkonsonanten (angegeben in Tab. 2-17) und elf Konsonantencluster (konsonantisches komplexes Jamo). Die 27 Varianten werden getrennt in Tab. 2-17 und 2-19 angegeben (vgl. ibid.).

Die Anordnung und Gruppierung der zwei oder drei Jamo einer Silbe stimmt mit der phonetischen Reihenfolge überein, d.h. der Anlaut wird immer zuvorderst und der Auslaut im untersten Teil aufgezeichnet. Je nach der Schriftform eines Inlaut-Jamo tritt er entweder horizontal rechts, vertikal unten oder auf beiden Seiten des Anlaut-Jamo auf. Wenn nur die Komponenten aus den drei Jamo-Klassen betrachtet werden, gibt es sechs Varianten der proportio-

nalen Konstellation von Hangulsyllabar, davon jeweils drei für die zwei- sowie dreibestandteiligen Silben. Der Strukturaufbau der Symbole sieht dabei wie folgt aus (Beispielsilben samt Erklärung folgen im Anschluss):

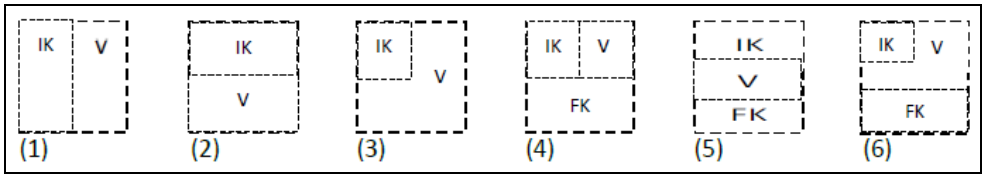


Abb. 2-11: Varianten zur proportionalen Konstellation der Komponenten in Syllabaren[94]

1) Links-rechts-Aufbau, wie <가> /ga/, Strukturbeschreibung: ⿰ (U+2FF0);

2) Oben-unten-Aufbau, wie <고> /go/, Strukturbeschreibung: ⿱ (U+2FF1);

3) Oben/links-unten/rechts-Aufbau, wie <과> /gwa/, Strukturbeschreibung: ⿲ ;

4) Links-rechts-unten-Aufbau, wie <간> /gan/, Strukturbeschreibung: ⿳ ;

5) Oben-mitte-unten-Aufbau, wie <곤> /gon/, Strukturbeschreibung: ⿴ (U+2FF3);

6) Oben/links-mitte/rechts-unten-Aufbau, wie <관> /gwan/, Strukturbeschreibung: ⿵.

Die Varianten der als An-, In- und Auslaut auftretenden Jamo werden jeweils in Tab. 2-17, 2-18 und 2-19 angegeben.

Jamo	**RRK**[95]	**Scan-Code**	**Uni-code**[96]	**Anlaut**	**Auslaut**	**Beispielsilbe & Bedeutung**	**Unicode der Silbe**
ㄱ	G/K–G[97]	#20 (r)	3131	Ja 1100	Ja 11A8	김 /gim/ Nachname *Kim* 약 /jak/ *ungefähr*	AE40 C570
ㄲ	KK	#20+SH (R)	3132	Ja 1101	Ja 11A9	끼 /kki/ *ersetzen* 낚 /nakk/ *Fisch*	AE4C AC02
ㄴ	N	#32 (S)	3134	Ja 1102	Ja 11AB	나 /na/ *ich* 안 /an/ *innen*	B098 C548
ㄷ	D/T–D	#19 (e)	3137	Ja 1103	Ja 11AE	담 /dam/ *Mauer* 받 /bat/ *erhalten*	B2F4 BC1B
ㄸ	TT	#19+SH (E)	3138	Ja 1104	Nein	땅 /ttang/ *Land*	B545
ㄹ	R/L–L	#34 (f)	3139	Ja 1105	Ja 11AF	리 /ri/ Nachname *Lee* 알 /al/ *Ei*	B9AC C54C
ㅁ	M	#31 (a)	3141	Ja 1106	Ja 11B7	마 /ma/ *Pferd* 삼 /sam/ *drei*	B9C8 C0BC

[94] IK für initialen Konsonant (Anlaut), V für Vokal (Inlaut) und FK für finalen Konsonant (Auslaut).
[95] In der Arbeit wird die koreanische Transkription RRK (the Revised Romanization of Korean) verwendet.
[96] Ein konsonantisches Jamo betrifft im Unicode drei Codewerte: jeweils für generelle Fälle (im Block: U+3130-U+318F), als Anlaut und als Auslaut (im Block: U+1100-U+11FF).
[97] Vor einem Vokal wird der Konsonant in /g/ transkribiert; wenn allein ein Auslaut oder ein anderer Konsonant nach ihm folgt, ist /k/ die Transkription. Diese Regel gilt auch für <ㄷ> (D/T), <ㄹ> (R/L) und <ㅂ> (B/P).

Jamo	RRK [95]	Scan-Code	Uni-code[96]	Anlaut	Auslaut	Beispielsilbe & Bedeutung	Unicode der Silbe
ㅂ	B/P–B	#17 (q)	3142	Ja 1107	Ja 11B8	바 /ba/ *Bar* 입 /ip/ *Mund*	BC14 C785
ㅃ	PP	#17+SH (Q)	3143	Ja 1108	Nein	빵 /ppang/ *Brot*	BE75
ㅅ	S	#21 (t)	3145	Ja 1109	Ja 11BA	수 /su/ *Möglichkeit* 맛 /mas/ *Geschmack*	C218 B9DB
ㅆ	SS	#21+SH (T)	3146	Ja 110A	Ja 11BB	in <싸다> /ssa.da/ *packen* in <있다> /iss.da/ *existie-ren*	C2F8 C788
ㅇ	Keins/ NG	#33 (d)	3147	Ja /Ø/ 110B	Ja /ng/ 11BC	오 /o/ *fünf* 왕 /oang/ *König*	C624 C655
ㅈ	J	#18 (w)	3148	Ja 110C	Ja 11BD	자 /ja/ *Zeichen* in <잊다> /ij.da/ *vergessen*	C790 C78A
ㅉ	JJ	#18+SH (W)	3149	Ja 110D	Nein	in <짜다> /jja.da/ *salzig*	C9DC
ㅊ	CH	#48 (c)	314A	Ja 110E	Ja 11BE	창 /chang/ *Fenster* 낯 /nach/ *Gesicht*	CCD0 B0AF
ㅋ	K	#46 (z)	314B	Ja 110F	Ja 11BF	카 /ka/ *Auto* in <키읔> /ki.euk/ *Buch-stabenname*	CE74 C754
ㅌ	T	#47 (x)	314C	Ja 1110	Ja 11C0	탄 /tan/ *Kohle* in <같다> /gat.da/ *gleich*	D0C4 AC19
ㅍ	P	#49 (v)	314D	Ja 1111	Ja 11C1	표 /pyo/ *Karte* 잎 /ip/ *Blatt*	D45C C78E
ㅎ	H	#35 (g)	314E	Ja 1112	Ja 11C2	항 /hang/ *Hafen* in <낳다> /nah.da/ *gebären*	D56D B0B3

Tab. 2-17: Die konsonantischen Buchstaben des Hangul

Historisch gab es vier weitere simple und dutzende weitere komplexe Jamo für Anlaute. Hier werden solche abgeschafften Symbole nicht berücksichtigt. Selbiges gilt für die folgenden Tabellen 2-18 und 2-19.

Jamo	RRK	Scan-Code	Unicode[98]	Rel. z. An.[99]	Beispielsilbe & Bedeutung	Unicode der Silbe
Simple Jamo für Inlaute						
ㅏ	A	#38 (k)	314F, 1161	→	사 /sa/ *vier*	C0AC
ㅑ	YA	#24 (i)	3151, 1163	→	야 /ya/ *Hallo*	C57C
ㅓ	EO	#37 (j)	3153, 1165	→	서 /seo/ *Buch*	C11C
ㅕ	YEO	#23 (u)	3155, 1167	→	혀 /hyeo/ *Zunge*	D600
ㅗ	O	#36 (h)	3157, 1169	↓	소 /so/ *Rind*	C18C
ㅛ	YO	#22 (y)	315B, 116D	↓	교 /gyo/ *Religion*	AD50
ㅜ	U	#51 (n)	315C, 116E	↓	수 /su/ *Weg, Methode*	C218

[98] Vorne der kompatible Codepunkt innerhalb des Umfangs U+3130-U+318F, hinten dem Codepunkt zwischen U+1100-U+11FF. Genauso wird in Tab 2-19 verfahren.

[99] Abkürzung für ‚Relation zum Anlaut-Jamo'; der Anlaut liegt bei der Gegenseite des Pfeils; das Symbol ↓→ heißt, dass sich der Auslaut rechts-unten vom Anlaut befindet.

Jamo	RRK	Scan-Code	Unicode[98]	Rel. z. An.[99]	Beispielsilbe & Bedeutung	Unicode der Silbe
ㄲ	YU	#50 (b)	3160, 1172	↓	균 /gyun/ *Niveau*	ADE0
ㅡ	EU	#52 (m)	3161, 1173	↓	음 /eum/ *Laut*	C74C
ㅣ	I	#39 (l)	3163, 1175	→	시 /si/ *Stadt*	C0D2
komplexe Jamo für Inlaute						
ㅐ	AE	#25 (o)	3150, 1162	→	새 /sae/ *Vogel*	C0C8
ㅒ	YAE	#25+SH (O)	3152, 1164	→	걔 /nyae/ Pronomen für *Kind*	AC54
ㅔ	E	#26 (p)	3154, 1166	→	세 /se/ *Jahre alt*	C138
ㅖ	YE	#26+SH (P)	3156, 1168	→	계 /gye/ *Jahreszeit*	ACC4
ㅘ	WA	#36+#38	3158, 116A	↓→	왕 /wang/ *König*	C665
ㅙ	WAE	#36+#25	3159, 116B	↓→	왜 /wae/ *warum*	C65C
ㅚ	OE	#36+#39	315A, 116C	↓→	쇠 /soe/ *Eisen*	C1E0
ㅝ	WO	#51+#37	315D, 116F	↓→	귀 /gweo/ *Ohr*	ADC0
ㅞ	WE	#51+#26	315E, 1170	↓→	궤 /gwe/ *Schrank*	ADA4
ㅟ	WI	#51+#39	315F, 1171	↓→	뒤 /twi/ *hinter*	B4A4
ㅢ	UI	#52+#39	3162, 1174	↓→	의 /ui/ *…von…*	C758

Tab. 2-18: Vokale und Vokalkombinationen als Inlaute des Hangul

Wie in den Spalten ‚Anlaut' und ‚Auslaut' von Tab. 2-17 angegeben wurde, können 16 der 19 Konsonanten-Jamo als Auslaut auftreten. Neben solchen Möglichkeiten können zusätzlich elf Cluster zu Auslauten angepasst werden.

Jamo für Auslaut[100]		
ㄳ /ks/ U+3133/U+11AA	ㄻ /lm/ U+313B/U+11B1	ㄾ /lp/ U+313F/U+11B5
ㄵ /nj/ U+ 3135/U+11AC	ㄼ /lp/ U+313C/U+11B2	ㅀ /lh/ U+3140/U+11B6
ㄶ /nh/ U+3136/U+11AD	ㄽ /ls/ U+313D/U+11B3	ㅄ /ps/ U+3144/U+11B9
ㄺ /lk/ U+313A/U+11B0	ㄿ /lt/ U+313E/U+11B4	

Tab. 2-19: Die Konsonantencluster als Auslaute des Hangul

Nach den angegebenen Möglichkeiten für An-, In- und Auslaute gibt es theoretisch 399 aus zwei (19×21) und 10.773 aus drei Bestandteilen zusammengesetzte Syllabare (19×21×27). So sind insgesamt 11.172 (399+10.773) Silbenblöcke möglich. In der praktischen Verwendung werden ca. 2.000 davon regelmäßig gebraucht. Im südkoreanisch-nationalen Standard zum Informationsaustausch wurden 2.350 Silbenzeichen aufgenommen (nach KS X 1001:1992). Im Unicode werden alle 11.172 möglichen orthographischen Silben codiert (nach Unicode 12.0 Character: U+AC00-U+D7AF).

2.4.3 Analyse der koreanischen Textverarbeitung

Zur Erforschung der koreanischen Textverarbeitung nehme ich das Wort *Hanja* (Bedeutung: *chinesische Schrift*, in Hangul: 한자, in Hanja: 漢字) als Beispiel, für welches die Zeichen-

[100] In jeder Zelle werden Schriftform, RRK-Transkription und Unicode eines Jamo nachfolgend angegeben.

form, die bestehenden und eingetippten Jamo, der Eingabeprozess und die Zeichencodes analysiert werden.

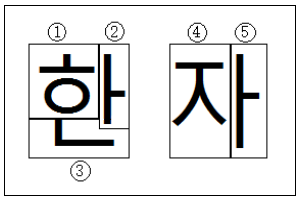


Abb. 2-12: Schriftform des Wortes *Hanja* in Hangul

Pr.	**Nr.**	**Jamo**	**RRK**	**Unicode**	**Relation/ Aufbau**	**Verarbeitung per Computer**
1	①	ㅎ (K)	H	314E	---	Einlesen des K-Jamo, Annahme als IK
2	②	ㅏ (V)	A	314F	①→②	Einlesen des V-Jamo, Kombination mit dem letzten K-Jamo
3	①+②	하	ha	D558	Links-rechts	Silbenbildung – 하
4	③	ㄴ (K)	N	3134	M1: ①→②↓③; M2: ①→②↓③x; M3: ③x.	Einlesen des K-Jamo, Begründung der Hypothesen: M1) allein als FK (siehe Pr. 5); M2) Teil von FK, Wartemodus auf den nächsten Jamo; M3) Anfang einer neuen Silbe, Wartemodus auf den nächsten Jamo.
5	①+②+③	한 (K)	han	D55C	Oben-unten, links-rechts	Durchführung von M1, Silbenbildung – 한
6	④	ㅈ (K)	J	M3: <ㄵ> - 3135; M4: 3148	M3: ①→②↓③+④; M4: ④**x.**	Einlesen des K-Jamo; Negation von M3, Durchführung von M2 & Erweiterung der möglichen Hypothesen: M2) ③+④ als FK der letzten Silbe (siehe Pr. 7); M1) ①+②+③ als eine Silbe festlegen und ④ als IK für neue Silbe.
7	①+②+③+④	핞	hanj	D55D	Oben-unten, links-rechts	Durchführung von M2, Silbenbildung – 핞
8	⑤	ㅏ (V)	A	314F	④→⑤	Einlesen des V-Jamo, Negation von M2 & Rücksetzung; M1 & Pr. 5 festlegen; Kombination von ④⑤**.**
9	④+⑤	자	JA	C790	Links-rechts	Silbenbildung – 자

Tab. 2-20: Eingabeprozess und Zeichencodes von *Hanja*[101]

Bei der koreanischen Eingabe werden alphabetische Komponenten (simple Jamo) für die zu schreibenden Silben eingetippt. Im Beispiel werden die drei Jamo für die Silbe <한> /han/ (Prozess 1, 2 & 4 in Tab. 2-20) und die zwei Jamo für Silbe <자> /ja/ (Prozess 4 & 5) nach der sprachlichen Ordnung eingegeben. Das System kombiniert die Jamo, gruppiert sie und ruft dann den Zeichencode der entsprechenden Syllabare ab.

[101] Verwendete Abkürzungen: Pr.: Prozess, Sy.: Symbol, Rel.: Relation zu dem letzten Zeichen, Aufb.: Aufbau, IK: initialer Konsonant, FK: finaler Konsonant, M: Möglichkeit; nach Eingabetest mit der koreanischen Eingabemethode.

Im Prozess von der Jamo-Eingabe zur Syllabarbildung muss eine automatische Silbensegmentation durchgeführt werden. Dies geschieht dank der geregelten Jamo-Verwendung in den meisten Fällen präzise beim Textschreiben. Nach der Eingabe eines Buchstaben werden zuerst Konsonant oder Vokal anerkannt und nach dem Silben-Modell ‚IK-V-(FK)' analysiert, wobei ihr Wert aus den angegeben Jamolisten (vgl. Tab. 2-17 bis Tab. 2-19) stammen muss.

Wenn es beim Einlesen eines neuen Buchstaben zwei oder mehr verschiedene Möglichkeiten gibt – wie in den Prozessen 4 und 6 in Tab. 2-20 – werden mögliche Hypothesen begründet, die sich mit dem Eintippen neuer Buchstaben überprüfen lassen. Die Silbenbildung mit vorderen Jamo wird zuerst durgeführt, wie z.B. Prozess 5 und 7 in Tab. 2-20. Abhängig von dem nachfolgend eingegebenen Buchstaben kann eine vorher begründete Hypothese als negativ definiert werden, woraufhin die anderen, wahrscheinlicheren Hypothesen eine entsprechende Ergänzung erfahren. Bspw. weist der nach <ㄴ> /n/ eingetippte Buchstabe <ㅈ> /j/ darauf hin, dass <ㄴ> /n/ unmöglich Anfangsbuchstabe einer neuen Silbe sein kann, da <ㄵ> /nj/ nur als Auslaut, aber nicht als Anlaut auftreten kann. Deswegen wird M3 gestrichen, die anderen zwei Hypothesen M1 und M2 weiter berücksichtigt. Mithilfe von dem nach <ㅈ> /j/ auftretenden <ㅏ> /a/ filtert das System die Wahrscheinlichkeiten weiter aus, so dass <ㄴ> /n/ und <ㅈ> /j/ zu verschiedenen Silben gehören müssen, denn ein vokalisches Jamo kann kein Initial sein. M1 wird dann als *wahr* festgelegt. Die Silbentrennung kann auch mithilfe von Leerzeichen oder anderen Hilfszeichen manuell gesteuert werden.

Nach der automatischen oder manuellen Silbensegmentation muss eine koreanische Eingabesoftware die Äquivalenz zwischen den Codes der Jamo und dem Code des Silbenblocks bilden, wie im Beispiel die Konversion von <ㅎ> (/h/, U+314E), <ㅏ> (/a/, U+314F) und <ㄴ> (/n/, U+3134) zum Silbenblock <한> (/han/, U+D55C) (vgl. Unicode 12.0 Chapters: Kap. 18.6: 737).

Alle Grundbuchstaben und ein Teil der komplexen Jamo werden auf der Tastatur belegt. In Abb. 2-13 wird das Standardtastaturlayout dargeboten. Die Scan-Codes und die Entsprechungen im internationalen Tastaturlayout werden in Tab. 2-17 & 2-18 angegeben.

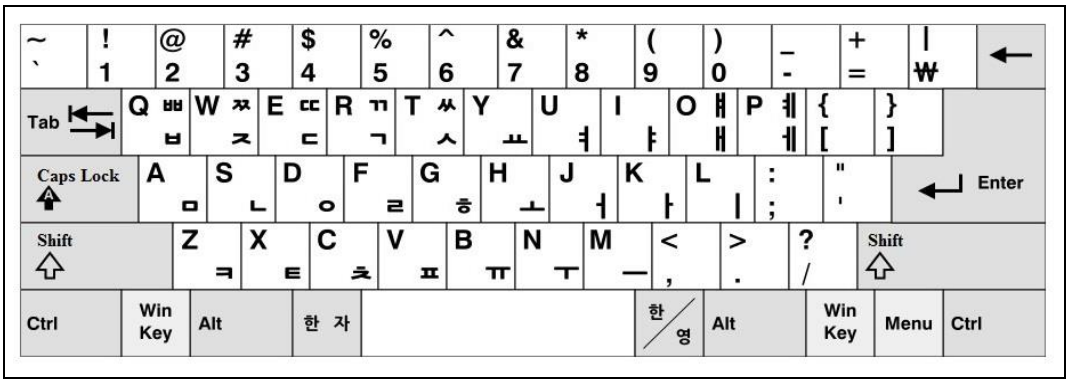


Abb. 2-13: Das koreanische Standardtastaturlayout

Ähnlich wie das Hindi-Tastaturlayout werden die belegten Zeichen nach Konsonanten und Vokalen in beide Handzonen untergliedert. In der Zone der linken Hand werden die konsonantischen Buchstaben verteilt, darunter 14 Grundbuchstaben und fünf Doppelkonsonanten, die als zweite Belegung der für den Grundbuchstaben stehenden Taste auftreten (#17 ‚Q' - #21 ‚T'). Aus Sicht der drei Jamo-Klassen werden alle für Anlaute einsetzbaren Konsonanten auf der Tastatur codiert (siehe Tab. 2-17). Die Zone der rechten Hand steht vor allem für vokalische Buchstaben und Interpunktionszeichen zur Verfügung. Zehn vokalisch simple Jamo und vier komplexe Jamo (< ㅐ > /ae/ und < ㅔ > /e/ als erste Belegung und < ㅒ > /yae/ und < ㅖ > /ye/ als zweite Belegung der Taste #25 ‚O' sowie #26 ‚P') sind die auf der Tastatur codierten Inlaute. Die restlichen komplexen Jamo müssen durch das nachfolgende Eintippen zweier Tasten eingegeben werden. Solche unbelegten komplexen Vokal-Jamo haben eine Gemeinsamkeit: sie sind proportional zerlegbar und schließen den Anlaut zweiseitig (unten und rechts) ein (siehe Tab. 2-18). Die möglichen Konsonantencluster als Auslaute sind ebenso unbelegt und müssen beim Computerschreiben via Zusammensetzung der zwei bestehenden Konsonanten kreiert werden (siehe Tab. 2-19). Wenn ein komplexes Jamo belegt wird, kann es nicht durch Kombination der simplen Jamo eingetippt werden. Es wird vor allem für die Vermeidung des Ambiguitätsfalls bei Silbensegmentation designt. Nach Analysen der belegten Jamo ist zu schlussfolgern, dass bei Eingabe einer Silbe mindestens zwei (bspw. ㄱ + ㅏ → 가 /ga/) und maximal fünf Eintippvorgänge (bspw. ㄱ + ㅗ + ㅏ + ㄱ + ㅅ → 괎 /gwans/) erforderlich sind.

Im Vergleich mit dem internationalen Tastaturlayout werden zwei Funktionstasten mehr eingesetzt, nämlich ‚한자' (*Hanja*, links von der Leerlaste im Tastaturlayout) und ‚한/영' (*HA/EN*, rechts von der Leertaste). Mit *HA/EN* wird schnell zwischen dem koreanischen und englischen Tastaturlayout umgeschaltet. Mit der Funktionstaste *Hanja* lässt sich die Umwandlung von einem Silbenblock zu einem chinesischen Schriftzeichen realisieren.

Aufgrund der dominierenden chinesischen Lehnwörter ist auch das Koreanische entsprechend reich an Homophonen. Im Hangul sind die Homophone schriftlich nicht unterscheidbar. Um diese morphologisch eindeutig darzustellen, ist die Anwendung der Hanja notwendig. Analysiert man das in Abb. 2-12 sowie Tab. 2-20 (S. 108) gezeigte Beispiel näher, wo statt des Silbenzeichens <한> /han/ das entsprechende Hanja erforderlich ist, muss nach der Eingabe von <한> /han/ die Funktionstaste *Hanja* gedrückt werden. Das System bietet dann alle Hanja in einer Zeichenliste an (vgl. Abb. 2-14), die als /han/ ausgesprochen werden. Der PC-Benutzer muss dann mit der kennzeichnenden Nummer das gewünschte Zeichen auswählen.

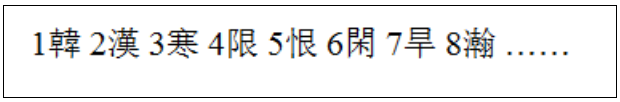


Abb. 2-14: Die dem Silbenzeichen <한> /han/ entsprechende, gebräuchliche Hanja[102]

2.4.4 Argumente für Silbenzeichencodierungen

In Kap. 2.3.3 wurde aus Sicht des Zeicheninventars und der schriftlichen Anwendungen analysiert, warum für die Devanagari-Textverarbeitung die Codierung der segmentalen Buchstaben effektiver ist. Im Gegensatz dazu sind bei der koreanischen Textverarbeitung Silbenzeichen meistens die grundlegende Einheit. Die Gründe für die Bevorzugung der Silbenzeichencodierung können aus verschiedenen Perspektiven argumentativ unterfüttert werden.

1) Der nötige Informationsgehalt zur Textverarbeitung.

Wenn nur die heute verwendeten Jamo codiert und alle Silben als Kombination der Jamo-Zeichen repräsentiert würden, betrüge das zu codierende Zeicheninventar weniger als hundert. Im Fall, dass die konsonantischen Buchstaben in der Form von An- und Auslaut getrennt definiert würden (wie im Unicode-Block ‚Hangul Jamo‘), bräuchte es insgesamt nur 67 (19+21+27) Codepunkte (vgl. Unicode 12.0 Character: U+1100-U+11FF). Berücksichtigte man die Initial- sowie Finalform eines Konsonanten gleich, könnten 51 (19+21+11) Zeichencodes der Jamo für die koreanische Textverarbeitung eingesetzt werden (vgl. ibid.: U+3130-U+318F). Wenn solche koreanischen Jamo in Erweiterungszonen des ASCII belegt würden, könnte eine 8-Bit-Codierung eingesetzt werden, d.h. jeder Jamo würde mit 8 Bits repräsentiert. Eine Silbe, die aus zwei oder drei Jamo besteht, betrüge deswegen 16 oder 24 Bits, wenn die Formatzeichen der Jamo-Kombination nicht berücksichtigt würden.

[102] Grundbedeutung der Zeichen nach der Reihenfolge: 1. *Südkorea*; 2. *chinesische Han-Nation*; 3. *kalt*; 4. *Einschränkung*; 5. *hassen*; 6. *frei*; 7. *Dürre*; 8. *weitgehend*. Weitere Homophone werden in der Abbildung zur Vereinfachung nicht dargestellt; vgl.: Lunde 2009: 854, Row 89 & Hanja keyboard – Yale IME, *https://www.branah.com/hanjayale* [Abruf: 2015-07-20].

Wie in Kap. 2.4.2 vorgestellt, gibt es insgesamt 11.172 mögliche Silben im modernen Hangul, von denen ca. 2.000 allgemeingebräuchlich sind. Zur Codierung solcher Silben im nationalen Standard (inklusive ASCII-Grundzeichen und Jamo) reicht die doppelte Codierung des ASCII (14-Bit-Codierung), die insgesamt 8.836 (94×94) Zeichen aufnehmen kann (siehe Kap. 1.4.2, S. 48). Im Unicode werden alle möglichen Silben mit 16 Bits im Bereich U+AC00-U+D7AF definiert (vgl. Lunde 2009: 144, Unicode 12.0 Character: U+AC00-U+D7AF). Im Vergleich zu der Methode der Jamo-Kombination, bei der mindestens 16 bis 24 Bits für eine Silbe erforderlich sind, ist die Codierung der Silbeneinheit aus Sicht des Informationsgehalts viel effektiver. Weiterhin gehören zu dem koreanischen Schriftsystem heutzutage mindestens 1.800 gebräuchliche Hanja, deren Codierung unmöglich innerhalb von einem Byte zu definieren ist, sondern in 14 Bits (wie in KS X 2001: 1992) oder 16 (wie in Unicode) sein muss.

2) Syllabar als schriftliche, sprachliche und morphologische Grundeinheit.

In Kap. 2.4.1 wurden die Grundeigenschaften des Hanguls analysiert. Auf der Ebene des Jamo handelt es sich um eine alphabetische Schrift und auf Ebene des Silbenzeichens kann es als eine syllabische Schrift definiert werden. Silbenzeichen sind sozusagen die funktionalen Grundeinheiten des Textschreibens und der Wortbildung (vgl. Dürscheid 2006: 91ff). Die Syllabarcodierung stimmt daher stärker mit den Gewohnheiten des Lesens und des Schreibens überein.

Weiterhin wird die koreanische Sprache gemischt in den zwei Schriften Hangul und Hanja geschrieben und ein gebräuchliches Silbenzeichen kann schriftlich einem oder mehreren sinokoreanischen Schriftzeichen entsprechen. Die mit gleichstelligen Bits codierten Silbenzeichen und Hanja bedingen einerseits, dass leichter Kompatibilität bei koreanischen Texten mit gemischten Schriften gewährleistet ist. Andererseits kann effektiver zwischen Silbenzeichen und Hanja konvertiert werden (vgl. Peng 1994: 344f).

3) Keine Unterscheidung bei Basis- und kombinierenden Zeichen, komplizierte Silbenaufbaubeschreibung.

In einer orthographischen Silbe der indischen Schriften sind die Basiszeichen im Zentrum und die kombinierenden Zeichen am Rande zu unterscheiden. Mithilfe der Kombinationsregeln bei Basis- und kombinierenden Zeichen kann eine Silbe typographisch ausgegeben werden (siehe Kap. 2.3.3, S. 95f). Im Gegenteil dazu sind die Jamo für An-, In- und Auslaut einer Silbe gleichwertige Bestandteile.

Um die Jamo einer Silbe graphisch zu kombinieren, gäbe es nach meinen Analysen zwei mögliche Entwürfe: 1) Begründung der quadratischen Modelle für die proportionalen Konstellationen; 2) Mehrere Zeichencodes für ein Jamo als kombinierendes Zeichen in jeder Kombinationsart. Wenn bspw. die Silbe <자> /ja/ durch Kombination der zwei Zeichen sowie deren Glyphen <ㅈ> /j/ und < ㅏ > /a/ repräsentiert würde, müssten sie sich nach der ersten Möglichkeit an das Format ‚□‘ anpassen. Bei der zweiten Möglichkeit müssten die Jamo in allen möglichen Kombinationsformen unterschiedlich codiert werden. Das Jamo <ㅈ> /j/ bspw. hat neun mögliche proportionale Lagen in einer Silbe: drei Möglichkeiten als Anlaut einer zweiteiligen Silbe, drei als Anlaut einer dreiteiligen Silbe und drei als konsonantischer Auslaut. Codierungsentwürfe könnten wie folgt aussehen:

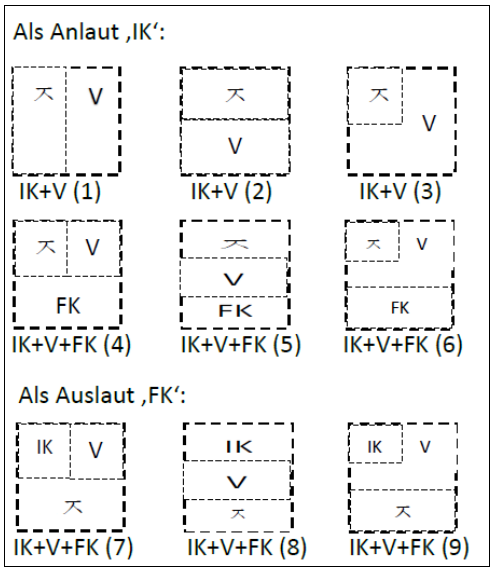


Abb. 2-15: Die neun möglichen Kombinationsarten des Jamo <ㅈ> /j/

Genau wie <ㅈ> /j/ könnte jeder konsonantische, simple Jamo in den neun kombinierenden Formen codiert werden. Jeder vokalische Jamo hat dank seiner bestimmten Relation zum Anlaut zwei mögliche Formen, jeweils in Silben aus zwei oder drei Bestandteilen. In dieser Beispielsilbe muss das System die Relation von < ㅏ > /a/ zum Anlaut (horizontal rechts) erkennen, die beiden Jamo als eine eigenständige Silbe korrekt kombinieren, dann das passende Modell der Silbe (die erste Variante in Abb. 2-15) auswählen und schließlich die beiden Jamo anpassen.

Beide Möglichkeiten der Jamo-Verbindung zu Silben sind im Vergleich zu der Codierung der Silben weniger pragmatisch. Einerseits erforderte der höhere Informationsgehalt mehr Speicherplatz und Vermittlungsaufwand. Wie in der ersten Argumentation aus Sicht der Bitfolge analysiert wurde, hat die auf Silbeneinheiten basierte Codierung im Vergleich zu der auf

Jamo-Kombination repräsentierenden Variante bedeutende Vorzüge. Wenn die speziellen Zeichencodes für Zeichenverbindung, Strukturmodelle oder mehr Zeichencodes für die Kombinationsarten jedes Jamo zum Einsatz gebracht würden, wären die Vorzüge der Silbencodierung deutlicher. Andererseits zieht die Jamo-Kombination mehr technische Schwierigkeiten bei der schriftlichen Ausgabe nach sich, die im Vergleich zu der Ausgabe der indischen orthographischen Silben viel komplizierter ist. Aus typographischer Perspektive könnte auch das aus mehreren Zeichen sowie Glyphen verbundene Schriftbild schwer die gewünschte Ausgabequalität erreichen (vgl. Peng 1994: 344f).

Die Gründe, warum für Hangul die Silbencodierung bevorzugt wird, können demgemäß wie folgt zusammengefasst werden: 1) die Begrenzungen der Silbenvarianten (siehe S. 107); 2) der seltener benötigte Informationsgehalt als Silbencodierung (siehe S. 111); 3) die funktionale Rolle des Silbenzeichens für Schrift und Sprache (siehe S. 102f); 4) das gemischte Schriftsystem von Hangul und der sinokoreanischen Schrift Hanja (siehe S. 110f); 5) die Schwierigkeiten der Jamo-Kombination ohne Unterscheidung von Basis- sowie kombinierenden Zeichen (siehe S. 105).

Kap. 2.3.1 bis 2.4.4 kontrastieren die Textverarbeitung der indischen Schriften und des Hanguls analytisch und zeigen, dass (1) für die erstere Variante die Codierung der segmentalen phonetischen Symbole bevorzugt wird, während Syllabarcodierung im Koreanischen der Kern ist. 2) Beide Tastaturbelegungen basieren jedoch auf Buchstaben. 3) Während bei indischen Schriften die Glyphenbildung wegen der konkurrierenden logischen und Ausgabereihenfolge der Buchstaben technisch herausfordernd ist, basiert die koreanische Textverarbeitung hingegen auf der computergestützten Konversion von einer Jamokette zu Syllabar.

2.5 Arabische Textverarbeitung

Die Erforschung der arabischen Textverarbeitung ist hauptsächlich von ihrer Schriftrichtung bedingt. Wie in Kap. 1.3.6 ausgeführt ist das Alphabet linksläufig, während die Zahlenangabe in Zahlzeichen rechtsläufig geschrieben werden muss. Innerhalb eines arabischen Textes kann es so zwei entgegengesetzte Schriftrichtungen geben. Dies wird als ‚bidirektional/BIDI' bezeichnet (vgl. hierzu Kap. 1.3.6, S. 36ff).

2.5.1 Zeicheninventar des arabischen Schriftsystems

Die arabische Schrift entstand zwischen dem 5. und 7. Jahrhundert aus der aramäischen Schrift. Sie ist das am häufigsten verwendete Konsonantenalphabet und wird in weitläufigen Gebieten zwischen Asien, Afrika und Europa verwendet, wo der Islam vorherrscht. Heute

wird sie neben dem Arabischen auch für die Verschriftung des Persischen im Iran und des Urdu in Pakistan verwendet (vgl. Wilbertz 1994: 312f, Majidi 2006: 1, Habash 2010: 1f). Erforschungen der arabischen Schrift in dieser Dissertation orientieren sich überwiegend an dem Schriftsystem des modernen Standardarabischen (kurz: MSA).

Das MSA ist die offizielle Sprache in der arabischen Welt, wird aber nicht als Umgangssprache verwendet. Es basiert morphologisch und phonologisch auf dem klassischen Arabischen, in dem der Koran geschrieben wurde. Es unterscheidet sich sowohl von den verschiedenen arabischen Dialekten (wie ägyptisches, irakisches, nordafrikanisches Arabisch) als auch von der klassischen Sprache des Korans. Die MSA-basierte Schriftsprache wird in der ganzen arabischen Welt in Medien, Bildung, offiziellen Dokumenten etc. angewendet (vgl. ibid.). Arabisch ist eine semitische Sprache und vom Sprachbau eine flektierende Sprache mit hauptsächlich auf konsonantischen Vorlagen basierten Wortwurzeln, die meistens aus drei Konsonanten bestehen. Anders formuliert unterscheidet sich die Wortbedeutung hauptsächlich bei konsonantischen Phonemen, während die Vokale im Allgemeinen über grammatische Informationen verfügen (vgl. Diab et al. 2007: 160, Wien 1998: 26f). Dies bedingt, dass die Vokale schriftlich nicht angegeben werden müssen.

Das Zeicheninventar des arabischen Schriftsystems kann in fünf Untergruppen klassifiziert werden: die alphabetischen Buchstaben, die mit Hamza gebildeten Schriftzeichen, die diakritischen Zeichen für Vokale, Zahlzeichen und Hilfszeichen.

1) Die alphabetischen Buchstaben und die mit Hamza gebildeten Schriftzeichen.

In MSA gibt es insgesamt 34 Phoneme, darunter 28 konsonantische und drei je lang- sowie kurzvokalische (jeweils von [a], [o] und [u]). Für die schriftliche Niederlegung gibt es 28 Buchstaben und acht weitere variierte Schriftzeichen, die als Grapheme funktionieren. Im Allgemeinen herrscht im arabischen Schriftsystem eine Eins-zu-eins-Entsprechung zwischen Phonem und Graphem, weshalb fast jeder Buchstabe einen bestimmten konsonantenphonetischen Wert hat. Drei Buchstaben (<ا> /ā/, <و> /w/ und <ي> /y/) zählen zur Ausnahme. Der Buchstabe Alif <ا> steht für den Langvokal /ā/ oder den Konsonant [ʔ]; <و> und <ي> sind Semivokale und können sowohl als Langvokale (/ū/ und /ī/) als auch als Konsonanten (/w/ sowie /y/) vertreten sein (vgl. Habash 2010: 31).

Ein Buchstabe hat im Normalfall vier Darstellungsformen: die isolierte, Initial- (nach links verbundene), Media- (beiderseits verbunden) und Finalform (nach rechts verbunden). Bei der Textverarbeitung mit dem Computer werden die vier verschiedenen Formen eines Buchstaben mit unterschiedlichen Glyphen, aber mit demselben Zeichencode repräsentiert.

Vier Buchstaben haben nur eine isolierte und Finalform (Initialform und isolierte, Media- und Finalform sind dabei identisch mit der isolierten Form): <ا> /ā/, <د> /d/, <ر> /r/ und <و> /w/. Ein Buchstabe erscheint nach seiner Position in der entsprechenden Form, weshalb innerhalb eines arabischen Wortes die Buchstaben graphisch kombiniert geschrieben werden können. Ein aus einem Buchstaben bestehendes Wort sowie das Bestimmungsartikel /al-/ (sowohl für Maskulin als auch für Feminin in jedem Kasus) werden immer mit dem folgenden Substantiv zusammengeschrieben (vgl. Majidi 2006: 60).

Im Tastaturlayout wird ein Buchstabe ohne Formunterscheidung belegt, weshalb die passende Buchstabenglyphe nach dem Tastenanschlag im Kontext computergestützt ausgegeben werden muss. Es gibt drei Möglichkeiten der alphabetischen Anordnung (vgl. ibid: 38ff):

A) Nach Zahlwert: jeder arabische Buchstabe wird mit einem bestimmten Zahlwert definiert. Die Bezeichnung für diese Anordnung lautet ‚ʾAbǧadi' (benannt nach den ersten vier Buchstaben). Der Zahlwert wird in der sechsten Spalte von Tab. 2-21 angegeben.

B) Nach Gestalt und Formähnlichkeiten (ʾAbta<u>t</u>i-Reihenfolge): diese Variante wird bspw. bei der Zeichencodierung im Unicode und bei der Tastenbelegung gebraucht (siehe Abb. 2-16 & 2-17). Die 28 Buchstaben sind von 18 Grundgraphen abgewandelt, wie z.B. <ت> /t/ und <ث> /<u>t</u>/. In Tab. 2-21 werden die Buchstaben aufgelistet.

C) Nach Artikulationsstellen der Lautwerte: in diesem Fall werden die Buchstaben von dem Glottal- bis zum Labiallaut platziert, wie die Buchstabenanordnung der indischen Schrift (siehe Tab. 2-14, S. 92f).

Die 28 Buchstaben können sich in Sonnen- und Mondbuchstaben (S./M. in Tab. 21) unterscheiden. Ein mit Sonnenbuchstaben anfangendes Wort wird mit dem Laut [l] assimiliert, wenn es mit dem Bestimmungsartikel <لا> /al/ aufeinander trifft. Im Fall der Mondbuchstaben ist es umgekehrt (vgl. Majidi 2006: 61). Bei der Transliteration des arabischen Alphabets nehme ich DIN 31635 als Grundlage. In der folgenden Tabelle werden die 28 Buchstaben des arabischen Schriftsystems angezeigt. Diese Transliteration gilt für die lateinische Umschrift aller angegeben arabischen Buchstaben sowie Wörter in dieser Arbeit.

Buchstabe	**Name**	**Translit. (Traskri.)[103]**	**IPA**	**S./M..**	**Zahlwert**	**Beispielswort[104]**	**Unicode**
ا	ʾAlif	ʾ / ā (' / a)	[ʔ]/ [ā]	M.	1	<**بدا**>, /bdā/, *auftauchen*	0627

[103] Wenn die Transliteration in lateinischen Sonderbuchstaben erfolgt, wird die Transkription in der Klammer angegeben, auf der das arabische ‚Marhaban'-Eingabeverfahren basiert (vgl. Albrecht 2011: 61f; die angegebene Transliteration gilt für das arabische Schriftsystem).

[104] Bei der Umschrift der Beispielwörter werden nur die verschrifteten Buchstaben transkribiert. Die angegebene Bedeutung verweist auf das Online-Wörterbuch Arabisch-Deutsch – *http://www.lessan.org/de*.

Buch-stabe	Name	Translit. (Traskri.)[103]	IPA	S./M..	Zahl-wert	Beispielswort[104]	Uni-code
ب	Bā’	b	[b]	M.	2	<**باب** >, /bāb/, *Tür*	0628
ت	Tā’	t	[t]	S.	400	<**بات**>, /bāt/, *über-nachten*	062A
ث	θā’	ṯ (th)	[θ]	S.	500	<**رث** >, /rṯ/, *alt*	062B
ج	Ǧim	ǧ (j)	[ǰ]	M.	3	< **جار** >, /ǧār/, *Nachbar*	062C
ح	Ḥā’	ḥ (h)	[ħ]	M.	8	<**حار** >, / ḥār/, *heiß*	062D
خ	Xā’	ḫ (kh)	[x]	M.	600	<**اخ** >, /āḫ/, *Bruder*	062E
د	Dāl	d	[d]	S.	4	<**داكن** >, /dākn/, *dunkel*	062F
ذ	Ḏal	ḏ (z)	[ð]	S.	700	<**ذاكر** >, /ḏākr/, *lernen*	0630
ر	Rā’	r	[r]	S.	200	<**اخبار** >, /aḫbār/, *Anzeige*	0631
ز	Zāy	z	[z]	S.	7	<**خبز** >, /ḫbz/, *backen*	0632
س	Sīn	s	[s]	S.	60	<**درس** >, /drs/, *Lektion*	0633
ش	Šin	š (sh)	[ʃ]	S.	300	<**شرب** >, /šrb/, *trinken*	0634
ص	Ṣad	ṣ (s)	[sˤ]	S.	90	<**صباح**>, /ṣbāḥ/, *Morgen*	0635
ض	Ḍād	ḍ (d)	[dˤ]	S.	800	<**أخضر** >, /aḫḍr/, *grün*	0636
ط	Ṭā’	ṭ (t)	[tˤ]	S.	9	<**طابع** >, /ṭābʿ/, *Siegel*	0637
ظ	Ẓā’	ẓ (z)	[zˤ]	S.	900	<**حظ**>, /ḥẓ/, *herunternehmen*	0638
ع	ʿAyn	ʿ (e)	[ʕ]	M.	70	<**عرب** >, /ʿarb/, *Araber, arabisch*	0639
غ	Ġayn	ġ (gh)	[ɣ]	M.	1000	<**غرب** >, /ġrb/, *Westen*	063A
ف	Fā’	f	[f]	M.	80	<**تفاح** >, /tfāḥ/, *Apfel*	0641
ق	Qāf	q	[q]	M.	100	<**شرق** >, /šrq/, *Osten*	0642
ك	Kāf	k	[k]	M.	20	<**شكرا** >, /škrā/, *danke*	0643
ل	Lām	l	[l]	S.	30	<**رجل** >, /rǧl/, *Bein*	0644
م	Mīm	m	[m]	M.	40	<**سلام**>, /slām/, *Frieden*	0645
ن	Nūn	n	[n]	S.	50	<**نحن** >, /nḥn/, *wir*	0646
ه	Hā’	h	[h]	M.	5	<**هي** >, /hī/, *sie*	0647
و	Wāw	w/ ū	[w]/ [ū]	M.	6	<**نور** >, /nūr/, *Licht*	0648
ي	Yān	y/ ī	[j]/ [i:]	M.	10	<**رأي** >, /ray/, *Meinung*	0649

Tab. 2-21: Die 28 Buchstaben des arabischen Schriftsystems[105]

Außer den 28 Buchstaben gibt es in der arabischen Schrift noch das phonemtragende Schriftzeichen Hamza <ء> (/ʾ/, [ʔ]), das entweder bei den Buchstaben (<ا>, <و> oder <ى>) am Wortanfang oder in der Wortmitte auf oder unter einem Buchstaben gesetzt bzw. isoliert am Wortende verwendet wird. In manchen Fällen wird Hamza auch als arabischer Buchstabe angesehen (vgl. Majidi 2006: 54f). Im arabischen Schriftsystem gibt es sechs mit Hamza gebildete Schriftzeichen: <ء>, <آ>, <أ>, <ؤ>, <إ> und <ئ>. Solche mit Hamza festgelegten Schriftzeichen werden sowohl im Unicode als auch bei der Tastenbelegung als eigenständige Zeichencodes definiert (siehe auch Abb. 2-16 & Abb. 2-17; vgl. Majidi 2006: 54-57).

[105] Vgl. Majidi 2006: XII, 38f & 60f, Unicode 12.0 Caracter: U+0600-U+06FF, DIN-Taschenbuch 343: 326f.

Selbiges gilt auch bei den Schriftzeichen ‚Ta-Marbuta' <ة> und ‚Alif-Maqsura' <ى>, die für bestimmte phonologische und morphologische Funktionen stehen. Ta-Marbuta ist die Kombination von <ه> /h/ und den Punkten von <ت> /t/ und bezeichnet das Femininsuffix /-atun/. Das Alif-Maqsura funktioniert als morphologischer Marker am Wortende. Es entspricht phonologisch dem Buchstaben <ي> und morphologisch <ا>, d.h. es setzt <ا> als letzten Buchstaben eines Wortes ein, um die Derivation des Grundwortes zu markieren (vgl. Majidi 2006: 51-60, Habash 2010: 9ff). Die häufige Ligatur <لا > /lā/ (Verbindung von <ل> /l/ und <ا> /ā/), die sich graphisch deutlich von der Zusammensetzung der Schriftbilder unterscheidet, wird im Gegensatz zu den beiden Schriftzeichen vom Computer als die Folge von zwei Buchstaben repräsentiert und verarbeitet. Beim arabischen Tastaturlayout wird es hingegen mit einer bestimmten Taste belegt (siehe Abb. 2-16 & Abb. 2-17).

2) Diakritische Zeichen für Vokal, Wortstatus und Sonstiges.

Die Eigenschaften der arabischen Morphologie sind einer der wesentlichen Gründe, warum das Konsonantenalphabet für die arabische Sprache praxisorientiert ist. Kurzvokale unterscheiden sich normalerweise nicht in ihren grundsätzlichen Bedeutungen. Das Wort <كتاب > (/ktāb/ [kitāb], *Buch* in Singular) beispielsweise wird in verschiedenen Deklinationsformen trotz derselben Schriftform unterschiedlich ausgesprochen: /kitābu/ (Nominativ, nach dem bestimmten Artikel /al-/) und /kitābun/ (nach unbestimmtem Artikel), im Akkusativ /kitāba/ sowie /kitāban/ und im Genitiv /katābi/ sowie /katābin/. Der Plural *Bücher* und das Verb *schreiben* werden beide als <كتب > /ktb/ geschrieben[106], die sich von dem Wort *Buch* im Singular nur durch den Langvokal /ā/ unterscheiden. Wegen der Charakteristika der Sprache ist es in den meisten Fällen optional, Kurzvokale einzufügen, da die Leser mit fortgeschrittenem Sprachniveau problemlos die Wortart und die Flexionsform eines unvokalisierten Wortes durch den Kontext entschlüsseln können. So sind die alltäglichen Schriftstücke, Zeitungsartikel, offiziellen Dokumente etc. meist unvokalisiert. Im Koran, anderen religiösen Texten, Gedichten und Lehrwerken für Grundschüler ist es aber sinnvoll, Vokale anzugeben, denn die bedeutsamen, veralteten, unbekannten oder seltenen Wörter müssen eindeutig repräsentiert werden (vgl. Bauer 1996: 1433).

Die graphemische Markierung solcher zusätzlichen sprachlichen Informationen ist mithilfe diakritischer Zeichen möglich. Sie umfassen Vokal-, Nunation-Zeichen (Markierung des Status eines Substantivs) und Symbole zur Verdopplung oder Nullvokalisierung eines Konso-

[106] /kutub/ für *Bücher* und /katab/ als Verb (Perfekt, aktiv).

nanten. Die diakritischen Zeichen werden zur Zeichencodierung und Tastenbelegung als eigenständige Zeichensätze behandelt. Sie werden in der folgenden Tabelle verzeichnet.

Diakritische Zeichen	Bezeichnung	Aussprache und Verwendung	Beispielwort[107]	Unicode
◌َ	Vokalzeichen Fatha	/a/, über einem Buchstaben	<الكِتَاب > /al-kitāba/, *das Buch* (/al-/ ist der bestimmte Artikel)	064E
◌ُ	Vokalzeichen Damma	/u/, über einem Buchstaben	<كُتُبْ > /kutub/, *Bücher*	064F
◌ِ	Vokalzeichen Kasra	/i/, unter einem Buchstaben	<الكِتَابِ > /al-kitābi/, *des Buches*	0650
◌ً	Nunation Fathatan	/-an/, über dem letzten Buchstaben eines Nomens, Akkusativ	<كِتَابًا > /kitāban/, *ein Buch* (akk.)	064B
◌ٌ	Nunation Dammatan	/-un/, über dem letzten Buchstaben eines Nomens, Nominativ	<كِتَابٌ > /kitābun/, *ein Buch* (nom.)	064C
◌ٍ	Nunation Kasratan	/-in/, über dem letzten Buchstaben eines Nomens, Genitiv	<كِتَابٍ > /kitābin/, *eines Buches* (gen.)	064D
◌ّ	Taschdid/ Shadd	Verdopplung eines Konsonanten	<كَتَّبَ > /kattaba/, *(jmd. etw.) schreiben (lassen)*	0651
◌ْ	Sukun	kein Vokal bei einem Konsonant	<كِتَبْ > /kitāb/, *Buch* (isolierte Singularform)	0652
◌ٰ	Dagger Alif	das Langvokal [a:], nur in wenigen modernen Wörtern	<لٰا> /lā/, *nicht*	0670

Tab. 2-22: Die diakritischen Hilfszeichen des arabischen Schriftsystems[108]

3) Die Zahl-, Interpunktions- und Sonderzeichen

Die international verbreitete arabische Ziffer wird präziser als ‚westarabische Ziffer' bezeichnet. Im östlichen Orient, von dem aus das moderne Ziffernsystem von Indien aus nach Europa verbreitet wurde, wird meistens die ‚arabisch-indische Zahlschrift' genannt. Im Iran, Pakistan etc. ist heutzutage die ‚ostarabisch-indische Ziffer' verbreitet; in westarabischen Ländern findet die westarabische Ziffer Verwendung (vgl. Habash 2010: 12f, Unicode 12.0 Chapters: Kap. 9.2: 371).

In Kap. 1.3.6 wird die Konkurrenz zwischen der Schreibrichtung arabischer Buchstaben und Zahlschriftzeichen vorgestellt. Sie bedingt, dass die Eingabe der Zahlzeichen innerhalb eines arabischen Texts (wie Datum, mathematische Formeln usw.) komplexer ausfällt. Wenn ein Datum mit Schrägstrich als Trennung (wie <٢٠١٤/٠٨/٢٥> *2014/08/24*) eingetippt wird, muss es mit dem Jahr anfangen und mit dem Tag enden, was mit der Sprechreihenfolge um-

[107] Als Beispielwörter werden die Deklinations- sowie Konjugationsformen von *Buch* und *schreiben* genannt, um die Funktion der Vokalzeichen eindeutig zu schildern.
[108] Vgl. Majidi 2006: 51-60, Habash 2010: 26.

gekehrt ist.[109] Grund dafür ist, dass der Schrägstrich auch als Bruchrechnungszeichen fungieren kann. Daher werden Zahlzeichen von beiden Seiten des Schrägstrichs aus vom Betriebssystem automatisch als eine Zahl in der rechtsläufigen Schriftrichtung anerkannt. Dasselbe gilt auch bei Punkt oder Komma innerhalb einer Ziffer, die als Tausend- sowie Dezimaltrennzeichen verwendet wird.

In der arabischen Schrift können manche mathematischen Schriftzeichen sowohl in der internationalen Form als auch in der nativen Sonderform dargestellt werden, wie das Prozentzeichen sowie Datum-, Prozent-, Dezimal- und Tausendtrennzeichen (vgl. Unicode 12.0 Character: U+0600-U+06FF). Wegen der linksläufigen Schriftrichtung werden manche Interpunktionszeichen umgedreht von der lateinischen Schrift gezeichnet, wie <؟> (Fragezeichen), <؛> (Semikolon) oder <،> (Komma) (vgl. ibid., Li GB 1993: 16). Die im Paar verwendeten Zeichen wie Anführungszeichen werden im Kontrast zur lateinischen Schrift diametral ‚vertauscht', obwohl ihre Darstellungsformen unverändert bleiben. Ein spezielles arabisches Hilfszeichen, das in anderen Schriften kaum zu finden ist, heißt ‚Tatweel' (auch Kaschida genannt). Aus ästhetischen Gründen können Tatweel-Zeichen beliebig oft zwischen zwei Buchstaben eingesetzt werden, damit jede Zeile gleich lang aussieht. Ein Bedeutungswandel geht damit nicht einher, wie <بـا> /bā/ im Vergleich zu <با> zeigt (vgl. Li GB 1993: 12f). Zahl- und Hilfszeichen des arabischen Schriftsystems werden nachfolgend angegeben.

Zahlzeichen	Unicode	Bedeutung	Sonderzeichen	Unicode	Bedeutung
٠	0660	Null	،	060C	Komma
١	0661	Eins	٫	060D	Datumtrennzeichen[110]
٢	0662	Zwei	؛	061B	Semikolon
٣	0663	Drei	؞	061E	Auslassungspunkt
٤	0664	Vier	؟	061F	Fragezeichen
٥	0665	Fünf	٪	066A	Prozentzeichen
٦	0666	Sechs	٫	066B	Dezimaltrennzeichen
٧	0667	Sieben	٬	066C	Tausendtrennzeichen
٨	0668	Acht	٭	066D	Asterisk
٩	0669	Neun	ـ	0640	Tatweel/Kaschida

Tab. 2-23: Zahl-, mathematische und Interpunktionszeichen des arabischen Schriftsystems

Auf Basis der in den Tabellen angegebenen Zeichen aus dem arabischen Schriftsystem kann die Größe des arabischen Zeicheninventars errechnet werden: 28 Buchstaben (vgl.: Tab. 2-21), sechs Hamza-Schriftzeichen, zwei sonstige Sonderschriftzeichen, neun abhängige Zeichen (vgl. Tab. 2-22), zehn Zahlzeichen und native Satz- sowie Wortzeichen (vgl.: Tab. 2-23). Für die Zeichencodierung des arabischen Schriftsystems reicht die 7-Bit-Codierung aus, wie

[109] Nach Eingabetest mit arabischen Eingabeverfahren von Google Translate.
[110] Das Zeichen ist die besagte Alternative zum Schrägstrich und wird seltener auch als ‚/' verwendet.

im Standard ‚ASMO 449' (*Arab Organization for Standardization and Metrology*). Wenn die Zeichencodierung auf ASCII-Basis erweitert würde, stünde das 8-Bits-Codierungssystem zur Verfügung, etwa ‚ASMO 708' (vgl. Wien 1995: 25). Die arabisch-indischen Ziffern können auch als Alternativglyphen von den internationalen repräsentiert werden, wie in ASMO 708 (entspricht ISO-8859-6, Arabic Codeset). Inklusive der Zeichen aus dem persischen und Urdu-Schriftsystem und seltener Zeichen würde ein 16×16-Block im Unicode für die arabische Zeichencodierung notwendig (vgl. Unicode 12.0 Character: U+0600-U+06FF).

Wegen des relativ kleinen Zeicheninventars reichte der Tastaturentwurf für eine zweimalige Belegung. Es gibt mehrere arabische Tastaturlayouts. An dieser Stelle werden zwei davon abgebildet: das Sakhr/MSX-Layout von Sakhr-Computer (spezieller Computer für die arabische Sprache) und das international verbreitete Layout von IBM-PC. Zu Analysezwecken der arabischen Textverarbeitung in dieser Arbeit wird (nachfolgend) das arabische Tastaturlayout von IBM-PC zugrunde gelegt.

Abb. 2-16: Das arabische Tastaturlayout Sakhr/MSX

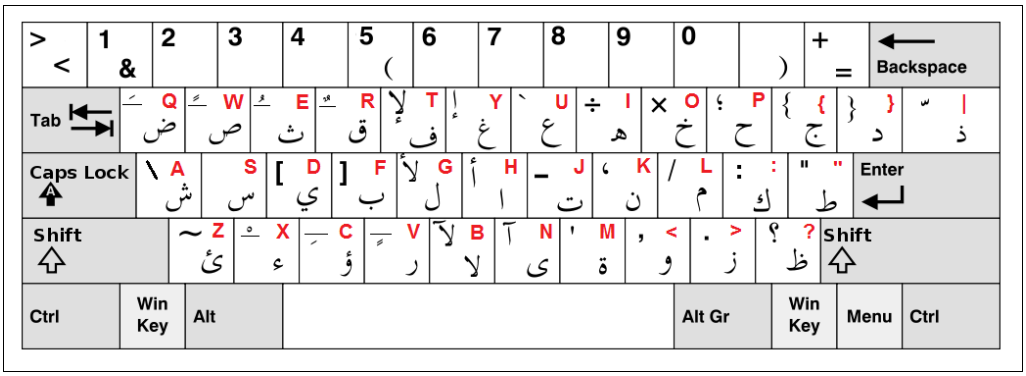


Abb. 2-17: Das arabische Tastaturlayout IBM-PC

Trotz der Unterschiede bei den beiden Layouts gibt es zwei Gemeinsamkeiten: Erstens sind die häufigsten beiden Buchstaben des arabischen Schriftsystems <ا> /ā/ und <ل> /l/ in der Grundreihe mit den Zeigefingern zu erreichen (vgl. Mrayati et al. 2003: 48). Zweitens werden die Buchstaben nach Formähnlichkeit angeordnet, d.h. die von derselben Grundglyphe abge-

leiteten Buchstaben werden auf der Tastatur derart arrangiert, dass sich Nachbarschaften bilden (vgl. z.B. in Abb. 2-17 <ض > /ḍ/ [Taste: ‚Q‘] und <ص > /ṣ/ [Taste ‚W‘]).

2.5.2 Schwierigkeiten der arabischen Textverarbeitung und allgemeiner Eingabeprozess

Aufgrund vieler Unterschiede zwischen der arabischen und lateinischen Schrift gibt es bei der arabischen Textverarbeitung viele technische Herausforderungen, die in den folgenden drei Punkten zusammengefasst werden können (vgl. Wien 1995: 27f)[111]: 1) die linksläufige Schreibrichtung der arabischen Buchstaben, 2) die verschiedenen graphischen Varianten der Buchstaben und 3) die Kompatibilität der Zahlschrift und anderer internationaler Schriften in entgegengesetzter Schriftrichtung als die Hauptrichtung des arabischen Texts.

Bei der linksläufigen computergestützten Verarbeitung gäbe es zwei Möglichkeiten: erstens, Entwurf der linksläufigen Formatierung für das allgemeine Computersystem in rechtsläufiger Schriftrichtung, ergo die Anpassung des Arabischen ans international verbreiteten System; zweitens Entwicklung eines speziellen regionalen Computersystems in RL-Richtung, das in der Literatur schon früh als ‚Arabisierung des Computers‘ (‚arabized systems‘) beschrieben wurde (vgl. Musa 1986, nach Wien 1995: 31). Wegen der rechtsläufigen Digitalschrift, des notwendigen internationalen Datenaustauschs und den Bedürfnissen der Textverarbeitung in anderen Schriften ist die Arabisierung in manchen Fällen unpraktisch und schwer durchzuführen. In vielen Fällen wird die erste Variante eingesetzt und das RL-Textlayout der arabischen sowie hebräischen Schrift durch die Steuer- und Formatzeichen realisiert (vgl. Davis / Lanin / Glass 2019: Kap. 1).

Bei der zweiten erwähnten Schwierigkeit, die Auswahl der Zeichenform je nach dem Kontext, wird hauptsächlich nach der folgenden Methode verfahren: Die isolierte Grundform eines Buchstaben wird voreingestellt dargestellt, während die anderen Varianten erst nach der Buchstabenposition im Wort erzeugt werden (vgl. Wien 1995: 31, Etemad 2005: 24).

Zu Erforschungen des Eingabe- sowie Verarbeitungsprozesses eines arabischen Wortes nehme ich das Wort <كتاب > /ktāb/ (*Buch*) als Beispiel. Die einzelnen Buchstaben werden nach der logischen Ordnung in der folgenden Abbildung nummeriert. Eine Definition für logische Ordnung wurde in Kap. 2.3.5 (S. 99) gegeben. Zur deutlicheren Darstellung werden die vier Buchstaben mit verschiedenen Farben markiert. Ergebnisse der Worteingabeanalyse werden im Anschluss in Tab. 2-24 Schritt für Schritt notiert.

[111] Wien nennt in ihrer Untersuchung insgesamt neun signifikante Unterschiede zwischen Englischen und Arabischen, von denen sich meines Erachtens die genannten drei auf die konkrete Textverarbeitung auswirken.

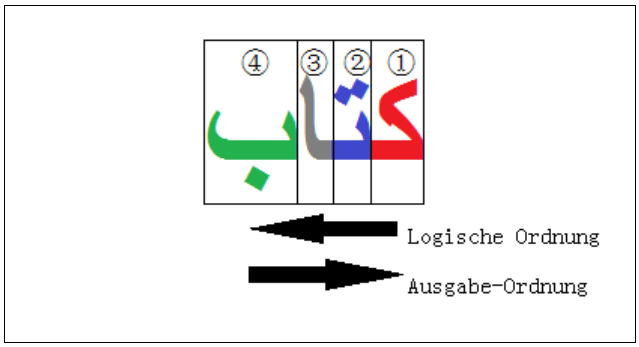


Abb. 2-18: Schriftform vom arabischen Beispielwort /ktāb/

Nr./ log. Ord.	Buch-staben (Bs.)	Uni-code	KT. [112]	AG.-Ord. [113]	Bs.-Pos. & Varianten-formen[114]	Formauswahl & Verarbeitung	Aus-gabe[115]
①	ك /k/	0643	D.	n.	1. Bs.; - Is. (Falls er eigenständig auftritt), - In. (Fall es dahinter weitere Buchstaben im Wort gibt).	Is.	…ك
②	ت /t/	062A	D.	n-1.	2. Bs.; - Fi. (Falls das Wort zu Ende ist) - Me. (Falls es dahinter weitere Buchstaben im Wort gibt).	Umwandlung von ① in In.; Darstellung von ② in Fi.; Verbindung von ①②	…كت
③	ا /ā/	0627	R.	n-2.	3. Bs.; - Fi. (links nicht verbindbar)	Umwandlung von ② in Me.; Darstellung von ③ in Fi.; Verbindung ①②③	…كتا
④	ب /b/	0628	D.	n-3.	4. Bs.; - Is. (Falls das Wort zu Ende ist) - In. (Falls es dahinter weitere Buchstaben im Wort gibt)	Is.; Einfügen von ④ links von der ①②③-Kombination	…كتاب
⑤	Leer-zeichen	0020	-	n-4.	Zeichen für Worttrennung, Markierung für Wortende und Anfang eines neuen Wortes	Festlegung der Is.-Form von ④; Einfügen des Leerzeichens links vom Wort	كتاب

Tab. 2-24: Eingabeprozess von dem arabischen Beispielwort /ktāb/

Bei der Eingabe des Wortes wird die Auswahl der Glyphe und Buchstabenkombination automatisch durchgeführt. Bei der Codierung der arabischen Buchstaben in Unicode wird der Kombinationstyp definiert, um die automatische Kombination benachbarter Buchstaben zu

[112] KT. steht für den Kombinationstyp eines arabischen Buchstabens, dessen Wert entweder eine Rechts- (R), Links- (L) oder Dualkombination (D) sein kann (vgl.: Unicode 12.0 UCD: Datei „ArabicShapping.txt“.)
[113] Abk. für Ausgabeordnung; es wird angenommen, dass der Startpunkt des Texts bei der Ausgabeordnung n liegt.
[114] Die Buchstabenposition im Wort und die möglichen Formen im Kontext; Abkürzungen: Is.: isolierte Form, In.: Initialform, Me.: Medialform und Fi.: Finalform.
[115] Drei Punkte <… > weisen darauf hin, dass an der Stelle weitere Buchstaben des Wortes folgen können.

unterstützen (vgl. Unicode 12.0: UCD: Datei „ArabicShapping.txt“). Wie in der Spalte ‚KT.‘ (Kombinationstyp) in Tab. 2-24 angegeben wurde, sind <ك> /k/ und <ت> /t/ beiderseitig kombinierbar. Je nach ihrer Position im Wort werden sie miteinander ohne Leerraum kombiniert. Hingegen ist <ا> /ā/ nur rechts kombinierbar und bedingt, dass das sich links von ihm befindliche Zeichen <ب> /b/ keine Verbindung aufweist. Das Leerzeichen weist weiter darauf hin, dass <ب> /b/ der letzte Buchstabe des Wortes ist und isoliert geschrieben wird.

Wie in Kap. 1.3.6 und dem letzten Kapitel vorgestellt wurde, ist der arabische Text wegen der von links nach rechts geschriebenen Zahlzeichen in den meisten Fällen bidirektional. Durch die Internationalisierung werden ebenso immer mehr Texte mit gemischten Schriften verfasst, in denen verschiedene Schriftrichtungen vorkommen können. Die bidirektionale Eigenschaft des arabischen Texts bestimmt, dass das Computersystem automatisch die Schriftrichtung anhand der eingegebenen Zeichen umstellen muss. Zur Erforschung des bidirektionalen arabischen Texts nehme ich den Eingabeprozess des Texts <اليوم ٢٠/٠٨/٢٠١٥> (/ālywm 20.08.2015/; *der Tag 20.08.2015*) als Beispiel. Wie in Kap. 1.3.6 skizziert, muss die Eingabe des Datums wegen technischer Einschränkungen in umgekehrter Reihenfolge des Lesens sowie Handschreibens erfolgen. Im Beispiel liegen drei direktionale Typen einzelner Zeichen vor: grundsätzlich linksläufig (wie die Buchstaben, rot markiert in Abb. 2-19), rechtsläufig (wie die Zahlzeichen, schwarz markiert) und in beiden Richtungen darstellbar (wie Leerzeichen und Schrägstrich, blau markiert). Die Ergebnisse und Analysen zu jedem Schritt der Eingabe werden in Tab. 2-25 angegeben.

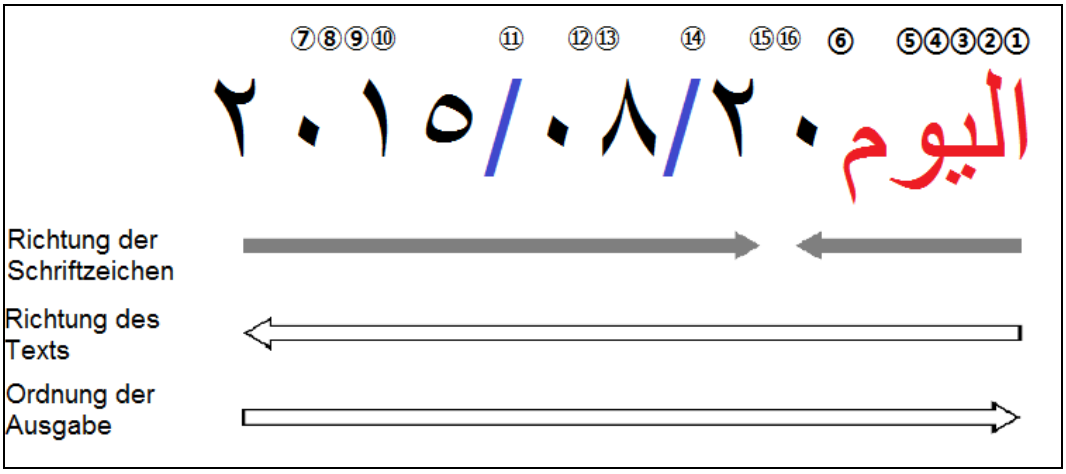


Abb. 2-19: Der arabische Beispieltext von zwei Schriftrichtungen

Nr./ Logische Ordnung	**Zeichen**	**Richtung** des Zeichens [116]	**Pos.** des Curs. [117]	**Relation** zum letzten Zeichen	**Ausgabeordnung**[118]	**Verarbeitung** im Prozess

[116] R für rechtsläufige, L für linksläufige und bidi. für bidirektionale Schriftrichtung.

Nr./ Logische Ordnung	**Zeichen**	**Richtung** des Zeichens [116]	**Pos.** des Curs. [117]	**Relation** zum letzten Zeichen	**Ausgabeordnung**[118]	**Verarbeitung** im Prozess
①–⑤	Bsp.: <ا>/ā/, <ل> /l/, <ي> /y/, <و> /w/ & <م> /m/	R-L	links	←	n bis n-4	Glyphenauswahl, linksläufige Kombination
⑥	Leerzeichen	bidi.	links	←	n-5	Einfügen des Leerzeichens links vom Wort <اليوم >
⑦	Zahlzeichen <2>	L-R	rechts	←	V (n-6)	Einfügen des Zeichens links von ⑥
⑧	Zahlzeichen <0>	L-R	rechts	→	V (n-6); ⑦ zu V (n-7) ändern	Einfügen des Zeichens rechts von ⑦
⑨	Zahlzeichen <1>	L-R	rechts	→	V (n-6) ⑦ und ⑧ auf Basis vom letzten Zustand jeweils -1	Einfügen des Zeichens rechts von ⑧
⑩	Zahlzeichen <5>	L-R	rechts	→	V (n-6) ⑦, ⑧ und ⑨ jeweils -1	Einfügen des Zeichens rechts von ⑨
⑪	Schrägstrich	bidi.	links	←⑦-⑩	V (n-10)	Einfügen des Zeichens links von ⑦
⑫	Zahlzeichen <0>	L-R	rechts	⑦-⑩→⑪→⑫	V (n-6); ⑪ zu V (n-7) ändern; ⑦-⑩ auf Basis vom letzten Zustand jeweils -2	Umstellung von ⑪ in ⑩ rechts, Einfügen des Zeichens rechts von ⑪
⑬	Zahlzeichen <8>	L-R	rechts	→	V (n-6); ⑦ bis ⑫ jeweils -1.	Einfügen des Zeichens rechts von ⑫
⑭	Schrägstrich	bidi. (wie ⑪)	links	←⑦-⑬	V (n-13)	Einfügen des Zeichens links von ⑦
⑮	Zahlzeichen <2>	L-R	rechts	⑦- ⑭ →⑮	V (n-6); ⑭ in V (n-7) ändern; ⑦ bis ⑬ auf Basis vom letzten Zustand -2.	Umstellung von ⑭ rechts von ⑬; Einfügen des Zeichens rechts von ⑭
⑯	Zahlzeichen <0>	L-R	rechts	→	n-6; ⑦ bis ⑮ jeweils -1; ⑦: n-15, ⑧: n-14, ⑨: n-13, ⑩: n-12, ⑪: n-11, ⑫: n-10, ⑬: n-9, ⑭: n-8, ⑮: n-7	Einfügen des Zeichens rechts von ⑮

Tab. 2-25: Der Eingabeprozess für den arabischen bidirektionalen Beispieltext

[117] Position des Cursors nach der Zeicheneingabe; diese ist meistens identisch mit der Schreibrichtung des Zeichens.
[118] Es wird angenommen, dass der Startpunkt des Texts bei der ‚Ausgabeordnung n' liegt; V steht für Variable, d.h. dass sich der Wert im Verlauf der Eingabe ändern wird.

Bei der arabischen Textverarbeitung wird die linksläufige Schriftrichtung als Hauptschrift festgelegt, so dass die bidirektional darstellbaren Zeichen meistens zuerst in dieser Richtung verarbeitet werden, wie etwa Leerzeichen und Schrägstrich. Das bedeutet, dass das System anhand des nach Leerzeichen oder Schrägstrichen eingetippten Zeichens automatisch entscheidet, ob die linksläufige Richtung beizubehalten oder zu korrigieren ist, wie in den Schritten 11 und 12 sowie 14 und 15. Die voreingestellte Richtung des Schrägstrichs wurde im Kontext der Zahlzeichen auf beiden Seiten zurückgenommen und zu einer rechtsläufigen Richtung geändert (siehe Schritt 12 und 15). Bei der Eingabe der Zahlzeichen ist ihre Ausgabeordnung wegen der entgegengesetzten Richtung bis zum Ende immer eine Variable, die sich stets auf Basis von ‚n-6' reduziert (siehe Schritt 7 bis 16).

Im Beispiel gibt es neben den Schriftzeichen noch unsichtbare Format- und Steuerzeichen, die die Verarbeitung und Ausgabe des bidirektionalen Texts unterstützen. Wie solche Sonderzeichen auf die arabische Textverarbeitung einwirken, wird im folgenden Kapitel erforscht.

2.5.3 Decodierungen und Algorithmen der Schriftrichtung

Der moderne Computer kann im Prinzip die Textverarbeitung aller heute verwendeten Schriften bewältigen. Bei den multilingualen sowie -schriftlichen Texten und der Textverarbeitung eines mit verschiedenen Schreibrichtungen dargestellten Schriftsystems wie des arabischen und hebräischen, muss eine bidirektionale Schreibunterstützung durchgeführt werden. ‚Bidirektional' bezieht sich nur auf gemischte Texte mit horizontal links- sowie rechtsläufiger Schriftrichtung (vgl. Unicode Glossary: Bidi). Denn bei der Zusammenstellung einer horizontalen und vertikalen Schrift (wie z.B. bei der chinesischen Schrift im traditionellen Textlayout) orientiert sich das Umdrehen der Textteile in der Regel an einer Hauptschriftrichtung (vgl. Koji/Lunde 2019: Kap. 3.3). Die Konkurrenzen in dieser Situation sind deswegen viel geringer als die Kompatibilität der beiden horizontalen Schreibrichtungen.

Zur Realisierung der bidirektionalen Textverarbeitung muss zuerst für jedes Zeichen ein impliziter ‚bidirektionaler Typ' (eng.: bidirectional type) definiert werden. Die Zeichen, die nach einer bestimmten Richtung – entweder LR (links nach rechts, rechtsläufig) oder RL (rechts nach links, linksläufig) – definiert werden, heißen ‚starke direktionale Zeichen' (eng.: strong directional character) (vgl. Davis / Lanin / Glass 2019: Kap. 3.2). Die starken direktionalen Zeichen mit LR-Richtung umfassen bspw. die Buchstaben der lateinischen, griechischen, kyrillischen und indischen Schriften und die Schriftzeichen der CJK-Schriften im modernen Textlayout. Zeichen des direktionalen Typs ‚stark-RL' sind die Buchstaben der arabi-

schen und hebräischen Schrift (vgl. ibid.). Im Gegensatz zu den starken direktionalen Zeichen gibt es auch manche Zeichen, die in unbestimmter Richtung geschrieben werden können und als ‚schwache direktionale Zeichen' (eng.: weak directional characters) bezeichnet werden. Dazu gehören Ziffern aus verschiedenen digitalen Ziffersystemen, die international benutzten Interpunktionszeichen und Sonderschriftzeichen wie mathematische sowie Währungszeichen etc. (vgl. ibid: Kap 1 & Kap. 3.2). Einzelne Zahlzeichen gehören zwar dem schwachen direktionalen Typ an und sind in jeder Schriftrichtung darstellbar, aber eine Folge von Ziffern muss in den meisten Fällen von links nach rechts geschrieben werden (vgl. ibid.: Kap 2.2). Wenn die Sequenz der Zahlzeichen in einem arabischen oder hebräischen Text auftritt, müssen der Textteil der Zahlschrift und der Textteil der alphabetischen Schrift auf verschiedenen Niveaus verarbeitet werden. Ebenso ist es bei gemischten Texten mit Schriften aus beiden Schriftrichtungen (vgl. ibid.: Kap 3.1.2).

Analysiert man die bidirektionalen Grundprinzipien anhand der Beispiele aus Abb. 2-19 und Tab. 2-25, kann der Verarbeitungsprozess wie folgt erklärt werden: Bei der Eingabe der Buchstaben ① bis ⑤ mit dem starken RL-direktionalen Typ verläuft der Text ohne Konkurrenz von rechts nach links. Nimmt man an, dass Arabisch die Hauptsprache sowie -schrift des Textes wäre, befände sich der Textteil ① bis ⑤ auf dem Niveau-NULL. Das danach eingetippte Leerzeichen ⑥ gehört zu den schwachen direktionalen Zeichen und wird weiterhin in RL-Richtung auf Niveau-NULL eingefügt. Selbiges gilt auch für ⑦ *– 2*, bevor das nächste Zahlzeichen ⑧ *– 0* eingegeben wird. So bilden ⑦, ⑧ und die nach ihnen kommenden Ziffern ⑨ und ⑩ einen übergeordneten Textteil, der in entgegengesetzter Richtung als die Hauptrichtung funktioniert. Zur Unterscheidung wird der linksläufige Textteil als $Text_1$ und der rechtsläufige (mit Zahlzeichen) als $Text_2$ bezeichnet. $Text_2$ muss auf einem höheren Niveau verarbeitet werden, damit das System den Text präzise verarbeiten und ausgeben kann, nämlich auf Niveau-1. Die Richtung des Schrägstrichs ⑪ und ⑭ mit schwachem direktionalem Typ wird dem LR-Kontext angepasst. Die Zeichen von ⑦ bis ⑯ gehören deswegen zu $Text_2$ auf Niveau-EINS, die $Text_1$ (von ① bis ⑥) auf Niveau-0 gegenüberstehen. Die Formatzeichen zur Schriftrichtung in Unicode können in drei Klassen untergliedert werden (vgl. Unicode 12.0 Character: U+2000-U+206C, Davis / Lanin / Glass 2019: Kap. 2):

1) die Allgemeine, wie LRM (Left to Right Mark, U+200E), RLM (U+200F) und ALM (das spezielle Formatierungszeichen der RL-Richtung für arabische Schrift, U+061C);
2) für eingebettete und übergeordnete Texte mit unterschiedlicher Schreibrichtung, wie LRE (Left to Right Embedding, U+202A), RLE (U+202B), LRO (Left to Right Override,

U+202D), RLO (U+202E) und PDF (Pop directional Formatting, Ende des Wirkungsumfangs von LRE, RLE, LRO oder RLO, U+202C);

3) für isolierte Texte mit anderer Richtung, wie LRI (Left to Right Isolate, U+2066) und RLI (U+2067), FSI (First Strong Isolate, U+2068) und PDI (Pop Directional Isolate, Ende des Wirkungsumfangs von LRI, RLI oder FSI, U+2069).

Die Ausgabe des bidirektionalen Beispieltexts visualisiert die folgende Abbildung:

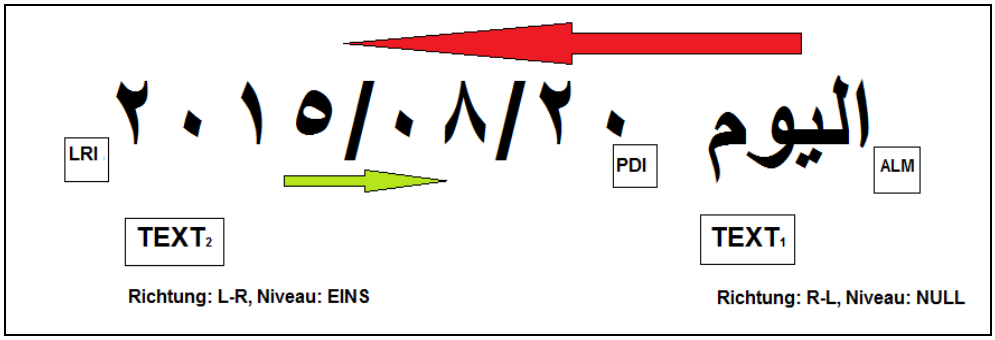


Abb. 2-20: Die Ausgabe des arabischen bidirektionalen Beispieltexts mithilfe von direktionalen Formatzeichen

In diesem Beispieltext sind die Ordnungen der Logik, der Ausgabe und des menschlichen Lesens voneinander differenziert.

Ordnung des menschlichen Lesens: ا ل ي و م SP ٢٠/٠٨/٢٠١٥
Logische Ordnung/Ordnung des Speicherns: ا ل ي و م SP ٢٠١٥/٠٨/٢٠
Ordnung der Ausgabe: ٢٠١٥/٠٨/٢٠ SP م و ي ل ا

Bei Erstellung eines Hypertexts in HTML kann dieser Beispieltext wie folgt formuliert werden. [119]

```
<div dir=“rtl“ lang=“ar“ charset=“utf-8“>
<p> اليوم </p>
<bdi dir=“ltr“> ٢٠١٥/٠٨/٢٠ </dbi>
</p>
</div>
```

2.6 Zusammenfassung der alphabetischen Eingabeverfahren und Textverarbeitungen

In den Kapitelkomplexen von 2.1 bis 2.5 wurden Eingabeverfahren von fünf Schriftsystemen vorgestellt und erforscht, nämlich die des Deutschen, des Vietnamesischen, des Hindi, des

[119] Attribut ‚dir‘ steht für *direction* (Schriftrichtung); das Element ‚bdi‘ für *Bidirection-Isolate* (bidirektional isolierend); der Wert ‚ltr‘ und ‚rtl‘ stehen jeweils für left-to-right und right-to-left.

Koreanischen und des Arabischen. Aus der Perspektive des Schrifttyps sind alle Schriftsysteme hauptsächlich alphabetbasiert. Aus der Perspektive der Textverarbeitung basieren die Tastaturlayouts aller sechs Schriftsysteme auf der Zeichen-Tasten-Repräsentation, aber der Prozess vom Eintippen der Buchstaben/Zeichen, über die Verarbeitung der Zeichencodes im System, bis zur Ausgabe der Glyphen auf Bildschirm und Papier ist bei jedem Schriftsystem und ihren verschiedenen Arten der Schriftzeichen unterschiedlich.

Zusammenfassend gibt es bei der Entwicklung eines Eingabeverfahrens eines (alphabetischen) Schriftsystems vor allem folgende Aufgaben zu lösen: 1) der Entwurf des Tastaturlayouts; 2) die Zeichencodierung der benötigten Zeichen des Schriftsystems; 3) die Begründung von Glyphen und ihre Entsprechungen zu den Zeichencodes, so dass die Zeichen auf dem Bildschirm sowie beim Drucken ausgegeben werden können; 4) die Anordnung der Schriftzeichen, d.h. in welcher Schriftrichtung sie zugeordnet und ob sie kombinierend oder eigenständig dargestellt werden.

Anhand der Unterschiede der verschiedenen Schriftsysteme weichen die Eingabeverfahren deutlich voneinander ab. Die entscheidenden Faktoren sind dabei: A) die Zeicheninventargröße; B) die Codierungsmethode der Kompositionssymbole als Gesamtzeichen oder als Grundzeichen-Diakritika-Folge; C) der Typ des Alphabets: Vollalphabet (wie die lateinische Schrift), Konsonantenschrift (wie die arabische Schrift), alphasyllabische Schrift (wie Devanagari) oder alphabetosyllabische Schrift (wie die koreanische Schrift); D) die Schreibrichtung des Schriftsystems: rechtsläufig, linksläufig oder bidirektional.

Wie viele Zeichen eines alphabetischen Schriftsystems auf der Tastatur belegt werden, ist hauptsächlich von drei Faktoren abhängig: (1) von den benötigten Buchstaben und sonstigen Zeichen, (2) von der Codierungsmethode der Kompositionssymbole und (3) von der Einschränkung der PC-Standardtastatur. Der Schreibmaschinenblock des Standardtastaturlayouts bietet 47 oder 48 zeichentragende Tasten (exklusive Leerzeichen). Mithilfe von ‚SHIFT', ‚AltGr', ‚Shift + AltGr' oder/und ‚Caps Lock', ‚Caps Lock + Shift' kann eine Taste zwei- bis fünfmal belegt werden. Das Maximum der belegbaren Zeichen beträgt deswegen 235 oder 240 (47×5 oder 48×5). Unter den erforschten Schriftsystemen erfordern das amerikanisch-englische, das koreanische und das arabische Tastaturlayout eine zweimalige Belegung. Für die Eingabe des deutschen Schriftsystems ist eine dreimalige und für die des vietnamesischen und der Hindi eine viermalige Belegung nötig. Ob ein Kompositionssymbol eigenständig mit einer Taste oder als Zeichenfolge von einem Grundbuchstaben und einem oder zwei kombinierenden Zeichen repräsentiert wird, ist einerseits von der Einschränkung des Tastaturlayouts bedingt, andererseits aber zugleich von Typ und Häufigkeit der Symbole abhängig. Beim

deutschen Tastaturlayout werden die allgemein gebrauchten Schriftzeichen <Ä>, <Ö> und <Ü> belegt, hingegen müssen die mit <´>, <`> und <^> gebildeten Sonderbuchstaben aus Fremdwörtern separat eingegeben werden (siehe Kap. 2.1.3, 71f). Bei dem Standardtastaturlayout des Vietnamesischen werden die lauttragenden Sonderbuchstaben belegt. Die Tonzeichen, die ebenfalls diakritische Zeichen sind, müssen hingegen zusätzlich nach der Eingabe eines Vokalzeichens eingetippt werden (siehe Kap. 2.2.3, S. 83f).

In welcher Art und Weise die Kombinationssymbole im Zeichencodierungsstandard behandelt werden, ist von dem Typ der kombinierenden Zeichen abhängig. Wenn das kombinierende Zeichen mit einem Buchstaben zusammengebildet werden muss, um einen bestimmten Phonem zu übertragen, so wird bevorzugt, dass die aus ihm bestehenden Kombinationssymbole als eigenständige Zeichencodes definiert werden, wie die diakritischen Zeichen der meisten von der lateinischen Schrift variierten Schriftsysteme. Wenn ein kombinierendes Zeichen allein bestimmte sprachliche Informationen überträgt, wie die abhängigen Vokalzeichen in der arabischen Schrift und den indischen Schriften, wird die getrennte Codierung als Grundzeichen durchgeführt (siehe Kap. 2.3.3, S. 91f & Kap. 2.5.1, S. 118f).

Die belegten Zeichen auf der Tastatur und die codierten Zeichen sind bei einem alphabetischen Eingabeverfahren in den meisten Fällen identisch. Im Detail läuft es bei jedem Schriftsystem jedoch unterschiedlich ab. Beim deutschen und anderen auf dem lateinischen Alphabet basierten europäischen Schriftsystemen sind die Standardtastaturlayouts bei den 26 Grundbuchstaben hauptsächlich identisch mit dem des amerikanisch-englischen Layouts. Die Sonderbuchstaben werden auf der nationalen Standardtastatur belegt, wenn sie dem nativen Schriftsystem angehören und allgemein gebräuchlich sind. Hingegen werden die nur für Fremdwörter oder selten benutzten Sonderbuchstaben in vielen Fällen mithilfe einer toten Taste eingegeben (siehe Kap. 2.1.3, S. 73). Im Gegensatz zu den diakritischen Zeichen der lateinischen Schrift werden die lauttragenden kombinierenden Zeichen der arabischen sowie der indischen Schriften immer getrennt vom Grundbuchstaben codiert und auf der Tastatur belegt. Solche kombinierenden Zeichen werden unabhängig von ihrer Position immer nach dem beigefügten Grundbuchstaben eingetippt und gespeichert (siehe Kap. 2.3.3, 2.3.4 & 2.5.1, S. 92-100 & S. 120f).

Das Eingabeverfahren der alphabetosyllabischen Schrift Hangul gestaltet sich vom Grundprinzip her unterschiedlich zu den Eingabeverfahren der anderen fünf Schriftsysteme. Im Normalfall der alphabetischen Eingabeverfahren sind die eingetippten Einheiten in der Regel identisch mit den zu verarbeitenden Zeichencodes. Aber in dem koreanischen Eingabeverfahren sind die eingetippte Einheiten (simple Jamo) immer Bestandteil der schriftlichen

Einheiten des koreanischen Texts (Silbenzeichen). Aus verschiedenen Gründen ist die Codierung der Silbenblöcke für die koreanische Textverarbeitung effektiver (siehe Kap. 2.4.4, S. 114). Die Umwandlung einer Jamofolge zum Zeichencode der Silbe bestimmt, dass das koreanische Eingabeverfahren eine Eingabemethode ist, wie die Eingabeverfahren für Chinesisch und Japanisch (siehe Definition von Eingabemethode in Einleitung, S. 3). Im Koreanischen stimmt die Reihenfolge von phonetischen und schriftlichen Informationen überein, weshalb die Phoneme immer nach ihrer logischen Reihenfolge im Silbenblock mit einer bestimmten Schriftform dargestellt werden. Aus diesem Grund bleiben dem Koreanischen zwei Schwierigkeiten der chinesischen und japanischen Eingabemethoden erspart. Einerseits ist es nicht notwendig, eine Inputcodierung zu entwickeln. Jamo repräsentiert sowohl sprachlich als auch schriftlich das Segmental von Silben und ist auf Tasten belegbar. Im Vergleich dazu müssen für die Eingabe der chinesischen Schrift spezielle Eingabeschemata nach der Zeichenform oder Aussprache entworfen und von PC-Benutzern erworben werden. Andererseits gibt es keine Überschneidungen koreanischer Silbenzeichen bei identischen Jamofolgen, weshalb das gewünschte Silbenzeichen ohne Ambiguitäten angeboten werden kann. (Bei chinesischen Schriftzeichen gibt es in den meisten Fällen mehrere Kandidaten desselben Inputcodes.) So muss bei den chinesischen und japanischen Eingabemethoden meist eine manuelle Auswahl durch den PC-Benutzer getroffen werden. Für eine höhere Effektivität werden daher auch computerlinguistische Anwendungen eingesetzt, damit der Computer per Kontext des Inputcodes mithilfe von linguistischen Kenntnissen die möglichen Kandidaten verringern kann. Wie die chinesischen sowie japanischen Eingabemethoden je nach den Eigenschaften der Schrift sowie Sprachen funktionieren, wird im vierten Kapitel dieser Arbeit behandelt.

3 Linguistische Perspektiven zur chinesischen Schrift

Wie in der Einleitung angedeutet, werden in Kapitel 3 und 4 die chinesische Schrift und ihre Eingabemöglichkeiten erforscht. Kapitel 3 konzentriert sich dabei auf die Grammatologie, zu der der Schrifttyp (Kap. 3.1), die Verwendung in verschiedenen Schriftsystemen (Kap. 3.2 & 3.5) und die drei Grundattribute der Schriftzeichen (Form, Aussprache und Sinninhalt, siehe Kap. 3.3 & 3.4) zugehörig sind. Auf jeder Ebene des schriftlichen Sprachgebrauchs – wie menschliches Lesen und Schreiben, computergestützte sprachliche Analysen, verschiedene Eingabemethoden usw. – fungieren grammatologische Erkenntnisse als Grundlage.

Methodisch wird dabei anhand zweier Prinzipien verfahren: Kontrastiv wird die chinesische Schrift im Vergleich mit dem lateinischen Alphabet analysiert, bei der Erforschung des Schriftgebrauchs zudem die Situation in verschiedenen Regionen des chinesischen Schriftkreises vergleichend herangezogen. Zweitens werden die komplexen sprachlichen Kategorien der chinesischen Schrift in Tabellen sowie Abbildungen zusammenfassend und musterhaft dargestellt, um notwendige Grundkenntnisse, die in Europa weitgehend unbekannt sind, deutlich zu vermitteln. Dazu zählen z.B. die Schriftentwicklung, der Zeichenaufbau oder die Silbenvarianten.[120]

3.1 Verwendungszweck, Schrifttyp, Ursprung und Entwicklung

In Kap. 1.3 wurden (S. 19-41) überblickshaft allgemeine Informationen über die verschiedenen Schriften sowie Schriftsysteme der Welt vorgestellt. Dabei wurden manche Grundinformationen über die chinesische Schrift – betroffene Schriftsysteme, Schrifttypen, kulturelle Verankerung, Grundeigenschaften der Schriftzeichen etc. – bereits erwähnt. Im Folgenden wird die chinesische Schrift jedoch noch näher fokussiert.

3.1.1 Verwendungszweck

Die chinesische Schrift (auch Sinographie genannt; 漢字/汉字 /hànzì/)[121] wird für das Chinesische, Japanische und Koreanische eingesetzt und von ca. 1,6 Mrd. Menschen als muttersprachliches Schriftsystem gebraucht. Sie ist eine sogenannte Schrift des Selbstoriginals, d.h.

[120] In Kap. 3 werden die chinesischen Begriffe und Eigennamen hauptsächlich in der vereinfachten Schrift angegeben; die aus Taiwan und Hongkong stammenden Begriffe und Titel werden hingegen nur in der traditionellen Schrift angeführt. Im CJK-Schriftkreis generell gültige Begriffe werden zuerst in der traditionellen, dann in der vereinfachten Schrift geschrieben.

[121] In allen drei Schriftsystemen wird *chinesische Schrift* schriftlich als 漢字 (heute als 汉字 verkürzt) dargestellt, allerdings mit phonologischen Unterschieden: *Hànzì* in Putonghua, *Kanji* im Japanischen und *Hanja* im Koreanischen. Die Begriffe *Sinographie* und *Sinogramm* werden von der englischen Übersetzung ‚Sinogram' nach Lu (2008: 4) entlehnt.

sie hat sich in China von vorschriftlichen Bildern auf Basis von Eigenschaften der altchinesischen Sprache zu einer mit Sprache verbundenen Schrift entwickelt (vgl. Lu 2008: 6). Sie hat die chinesische Geschichte seit mindestens 3.500 Jahren schriftlich dokumentiert. Die chinesische Schrift variiert sich heutzutage vor allem in vier Zweigen: die vereinfachte chinesische Schrift in Festland China und Singapur; die traditionelle chinesische Schrift in Taiwan, Hongkong und Macao; die sinojapanische Schrift Kanji in Japan und die sinokoreanische Schrift Hanja in Südkorea (vgl. Müller-Yokota 1994a: 348f).

Die offizielle moderne chinesische Schriftsprache heißt Baihua-Wen (白话文 /báihuàwén/). Sie basiert hauptsächlich auf dem Putonghua (普通话 /pǔtōnghuà/). Putonghua wird auch Mandarin und modernes Standardchinesisch genannt und ist die Standardsprache im kompletten chinesischsprachigen Gebiet. Sie basiert auf der nordchinesischen Sprache mit Beijinger Dialekt als phonetischem Standard und den in Baihua-Wen verfassten Schriftwerken als grammatischem Vorbild. Seit Beginn des 20. Jahrhunderts ersetzte Baihua-Wen die klassische standardisierte chinesische Schriftsprache Wenyan-Wen (文言文 /wényánwén/), die stark von der Umgangssprache abwich. Die moderne Schriftsprache gilt bis heute als der standardisierte Schriftstil in allen chinesischsprachigen Regionen, sowohl in Festlandchina mit Kurz- als auch in Taiwan und Hongkong mit Langzeichen. Die chinesischen Sprachen bilden eine Subgattung der sinotibetischen Sprachfamilie und enthalten neben der nordchinesischen Sprache noch weitere Zweigsprachen, die wegen vieler Unterschiede als andere Sprachen definiert werden. Diese sind bspw. Kantonesisch (gesprochen in Guangdong und Hongkong), Wu- (in Zhejiang, Shanghai und Jiangsu) und Min-Sprache (in Fujian und Taiwan). Die chinesische Schrift hat in diesem Sinne eine übersprachliche/-dialektale und überzeitliche Funktion, was im Vergleich zum Alphabet gewisse Vorteile nach sich zieht: Die Aussprache eines Schriftzeichens veränderte sich erheblich im Laufe der Sprachentwicklung und variiert bis heute regional, wohingegen seine Grundbedeutung sowohl räumlich als auch diachron stabil ist (vgl. Sun 1988: 16f). Dieser Eigenschaft ist zu verdanken, dass in einem großen Land wie China über tausend Jahre (trotz verschiedener Zweigsprachen) schriftliche Kommunikation gelingen konnte. Sie hat auch dazu beigetragen, dass die klassischen Texte von vor zweitausend Jahren zur Alltagslektüre taugen und zur mittelschulischen Ausbildung gehören können. Auch das Japanische, Koreanische und Vietnamesische, die nicht zur selben Sprachfamilie zählen, können mit chinesischer Schrift geschrieben werden.

Die chinesische Schrift hat die koreanische Halbinsel (ab ca. 108 v. Chr.), Japan (ab 1. Jh. n. Chr.) und Vietnam (ab 1. Jh. v. Chr.) literarisiert. Am Anfang wurden die klassischen Werke in chinesischer klassischer Schriftsprache direkt übernommen und für Bildungszwecke einge-

setzt. Es gab auch nach der Verbreitung andauernde Versuche, die jeweils einheimische Sprache mit chinesischer Schrift zu schreiben. In diesem Prozess wurden viele einheimische Schriftzeichen nach den Prinzipien der Sinogramme erfunden, um die Einschränkungen der chinesischen Schrift zu ergänzen. Deswegen gehören zu Kanji und Hanja manche reinjapanischen sowie reinkoreanischen Schriftzeichen, die die chinesische Schrift im engeren Sinne nicht umfasst. Präziser werden sie als sinojapanische und sinokoreanische Schrift bezeichnet. Zur besseren Effektivität schriftlicher Niederlegungen wurden auch einheimische phonographische Schriften erfunden, nämlich Kana (chi.: 假名, jap.: 仮名, Gesamtbezeichnung für Katakana und Hiragana) für das japanische und Hangul für das koreanische Schriftsystem, die gemischt mit der Sinographie in den beiden Schriftsystemen verwendet werden. Heutzutage hat sich der chinesische Schriftkreis verkleinert und umfasst Großchina, Singapur, Japan und Südkorea. In Nordkorea und Vietnam wurde die chinesische Schrift im 20. Jahrhundert komplett abgeschafft und vom Alphabet ersetzt (vgl. Müller-Yokota 1994a: 382-403).

3.1.2 Schrifttyp

Bei der Vorstellung von verschiedenen Schriftsystemen der Welt habe ich die chinesische Schrift dem logographischen Schrifttyp zugeordnet (siehe Kap. 1.3.2, S. 21f). Dies gilt zwar im Allgemeinen als korrekt, kann aber wissenschaftlich nicht als absolute Wahrheit festgelegt werden. Ein wichtiger Grund hierfür ist, dass die chinesische Schrift ein kompliziertes Inventar an Zeichen von verschiedenen Typen besitzt. Es gibt sowohl komplett auf die Bedeutung bezogene (wie Piktogramme und Ideogramme), als auch von der Aussprache abhängige Zeichen (Entlehnungszeichen, siehe Kap. 3.3.2). Die Mehrheit des Inventars bilden so genannte Determinativphonetika, die aus einer bedeutungs- und einer lauthinweisenden Komponente zusammengesetzt sind. Die Konstruktionsprinzipien werden in Kapitel 3.1.3 ausführlich dargestellt. Wie der Schrifttyp der chinesischen Schrift definiert wird, ist in der Schriftlinguistik somit umstritten. Es gibt sowohl Argumente für die Zugehörigkeit zum logographischen, ideophonologischen, als auch morphologischen Typ (vgl. Zou 2004: 24-33).

Die Zuordnung der chinesischen Schrift zur Logographie hat zwar den Hauptunterschied von der chinesischen Schrift zum Alphabet und der Silbenschrift markiert, ist aber aus wissenschaftlichen Ansichten umstritten. Einerseits hat ein Sinogramm im modernen Schriftsystem meistens mehrere Bedeutungen und ist nur im sprachlichen Kontext eindeutig. Zweitens existieren in der chinesischen Schrift auch sehr viele mit phonographischen Symbolen gebildete Schriftzeichen, die nicht ignoriert werden können. Eine gewissenhafte und genaue Be-

zeichnung muss ergänzt werden (ibid.). Müller-Yokota etwa hat die Vollbezeichnung ‚piktographisch-ideographisch-rebusartige Logographie' vorgeschlagen (1994a: 348).

Den Begriff der ideophonologischen Schrift hat der chinesische Linguist Zhou Youguang begründet. Ihm zur Folge gibt es drei Typen von Schrift, die nachfolgend in einer evolutionären Hierarchie stehen: die piktoideographische (形意文字 /xíngyì wénzì/, Bezeichnung der Gestalt und der Bedeutung als Schwerpunkt), ideophonologische (意音文字 /yìyīn wénzì/, Bezeichnung der Bedeutung und der Aussprache als Schwerpunkt) und phonologische Schrift (拼音文字 /pīnyīn wénzì/, Bezeichnung der Aussprache als Schwerpunkt).[122] Die Begründung dieses Arguments ist stark von der Theorie der Schriftevolution (siehe Kapitel 1.3.3, S. 26) beeinflusst und bezieht sich dominant auf Determinativphonetika. So wurde sie im Bereich der Schriftlinguistik kaum anerkannt (vgl. Zou 2004: 24-33).

Im Vergleich zur Logographie bezieht sich der Terminus *morphologische Schrift* (语素文字 /yǔsù wénzì/) mehr auf freie Morpheme, da sie die Grundeinheit der modernen chinesischen Sprache und Schrift bilden. Anhand des modernen Chinesischen/Japanischen wird das Argument innerhalb und außerhalb des sinographischen Kreises unterstützt. Ein Sinogramm ist zwar im Allgemeinen auf ein Morphem bezogen, aber in vielen Lehnwörtern ist dies nicht mehr der Fall. Aus diesen Gründen ist die chinesische Schrift keine hundertprozentig morphologische Schrift (vgl. Zhang YJ 2000: 50f).

Es gibt noch weitere Bezeichnungen für den Schrifttyp der Sinographie. Nach Lu z.B. kann die chinesische Schrift als sachbezogene Schrift (eng.: affair-recorded scripts, chi.: 记事文字 /jìshì wénzì/) charakterisiert werden, die das Gegenteil zu *lautbezogener Schrift* (sound-recorded scripts, 记音文字 /jìyīn wénzì/) darstellt. Die Aufzeichnung von Objekten und Konzepten, also der im Chinesischen gebrauchten konkreten und abstrakten Begriffe, gilt als das Grundprinzip der chinesischen Schrift (vgl. Lu 2008: 5).

Die chinesische Schrift hat ein sehr kompliziertes Inventar von Zeichenbildungsprinzipien, so dass keine Definition für ihren Schrifttyp unumstritten anerkannt und festgelegt werden kann. Außer persönlichen Standpunkten sind auch verschiedene Schriftsysteme und ihre verschiedenen Zeiten bei der Zuordnung entscheidend. Im Chinesischen wird ein Schriftzeichen immer monosyllabisch ausgesprochen und die Zeichen mit zwei oder mehreren Lesarten sind nur ein kleiner Teil des Inventars. Im Japanischen ist ein Sinogramm in vielen Fällen polysyllabisch und hat generell mehrere Lesarten (vgl. Coulmas 1996a: 240f). Im Altertum wur-

[122] Die drei Begriffe stammen aus der englischen Übersetzung, die von Zhou Youguang zu den originalen chinesischen Begriffen angegeben werden: *picto-ideograph*, *ideo-phonograph* und *phonograph* (vgl. 1954: 2-4).

de ein Schriftzeichen meistens für ein Wort erfunden, weshalb Wörter aus Einzelzeichen in der altchinesischen Sprache die Mehrheit darstellten. Mit der Zeit wurden Zeichen immer mehr als freie Morpheme zum Bau neuer Wörter verwendet, so dass unter den heutigen chinesischen Wörtern Bigrammwörter überwiegen (vgl. Lu 2008: 150f). Zusammengefasst gibt es wissenschaftlich-definitorisch keinen völlig zweifelsfreien Begriff, der alle Grundeigenschaften der chinesischen Schrift miteinbezieht und terminologisch fasst. Aber egal auf welchen Ansichten die chinesischen Schriftzeichen basieren: im Normalfall – also ohne Berücksichtigung mancher Fremd- und Sonderwörter – handelt es sich hauptsächlich um Morpheme, die entweder selbst ein eigenständiges Wort repräsentieren oder als Baustein eines solchen fungieren. Aus diesem Grund stimme ich der terminologischen Bezeichnung *morphologische Schrift* am ehesten zu.

3.1.3 Historische Entwicklung

Wann die chinesische Schrift entstanden ist, ist unbekannt. Für den Stamm der chinesischen Schrift existieren verschiedene Erzählungen und Mythen. Als bekannteste davon darf jene gelten, der Cangjie die Erfindung der Schrift zuschreibt. Cangjie (仓颉) war einigen Volkserzählungen zur Folge ein Beamter, der an historischen Aufzeichnungen arbeitete und zu Zeiten Huangdis (chinesischer Urkaiser, der nach allgemeinen historischen Erforschungen im 26. Jh. v. Chr.) lebte. In Wahrheit kann eine Schrift des Selbstoriginals mit großem Zeicheninventar, die zum Niederlegung sprachlicher Ausdrücke geeignet ist, unmöglich von einer Person erfunden worden sein. Cangjie war vermutlich ein Vertreter von zahlreichen Intellektuellen des Altertums, die für die Erfindung, die Entwicklung und die Sammlung der chinesischen Schrift Beiträge leisteten (vgl. Sun 1988: 27ff, Zou 2004: 16f). Archäologische Entdeckungen – wie die Jiahu-Symbole (6510-5619 v. Chr.) und die Yang-shao-Kultur (4800-4200 v. Chr.) – haben zudem darauf hingewiesen, dass es eine Entwicklungszeit von mehreren tausend Jahren vor dem ersten überlieferten chinesischen Dokument geben muss (vgl. Müller-Yokota 1994a: 351, Sun 1988: 36ff). Wegen der mythologisch tradierten Großleistung wurde die erste verbreitete Eingabemethode des Chinesischen nach Cangjie benannt: die ‚Cangjie-Eingabemethode'.

Das älteste heute überlieferte chinesische Schriftsystem lässt sich aufs 14 Jh. vor Chr. aus der Shang-Dynastie (ca. 17-11. Jh. v. Chr.) zurückführen. Aus den Dokumenten dieser Zeit wurden ca. 4.500 Jiagu-Zeichen entdeckt, zu deren Bildung alle sechs Konstruktionsprinzi-

pien verwendet wurden (vgl. Müller-Yokota 1994: 351f, Sun 1988: 78f).[123] Die Entwicklungsphase der chinesischen Schrift kann wie folgt zusammengefasst werden:

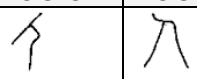
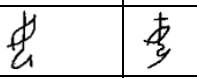
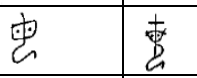
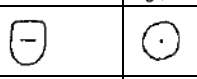
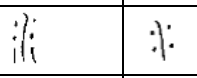
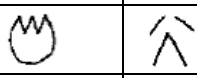
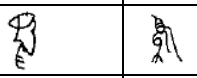
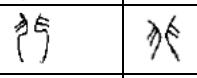
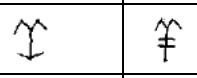
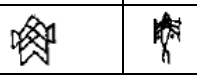
- Jiagu-Zeichen (甲骨文 /jiǎgǔ wén/, auch *Knocheninschrift*, 1300-1028 v. Chr. gebräuchlich): die archaische, meistens auf Tierknochen eingravierte Schrift.
- Jin-Zeichen (金文 /jīnwén/, auch *Bronzeinschrift*, 1066-221 v. Chr.): die auf Bronzegeräten eingravierte archaische Schrift.
- Siegelschrift (篆书 /zhuànshū/), die meistens auf Bambus-/Holzschleifen eingravierte oder geschriebene und heute noch häufig für Siegel gebrauchte archaische Schrift, die sich weiter in der großen (大篆 /dàzhuàn/, ca. 8. Jh.-221 v. Chr.; i.e.S. die Schrift in Qin-Königreich; i.w.S. die Schriften in derzeitigen verschiedenen Königreichen) und der kleinen Siegelschrift (小篆 /xiǎozhuàn/, ca. 221 v. Chr. – 8 n. Chr. im vereinigten China der Qin- und Han-Dynastie) unterscheidet.
- Kanzleischrift (隶书 /lìshū/, offizielle Verwendung 2. Jh. v. Chr. – 3. Jh. n. Chr.): der Beginn der modernen chinesischen Schrift mit Formreformierung bei der Umwandlung von kurvigen zu geraden Strichelementen und der runden zu viereckigen Gestalt.
- Regelschrift (楷书 /kǎishū/, auch *Kai-Ti*, 3. Jh. bis heute): die bis heute gebräuchliche Standardschriftart zum Handschreiben, aus der die Standarddruckschriftart Song-Ti (auch Song- oder Ming-Schrift) abgeleitet wurde. Parallel zur Regelschrift wurden auch das sog. Xingshu (行书 /xíngshū/, Verfahrenstechnik zur Schnellschreibung) und die Stenographie (草书 /cǎoshū/) erfunden und bis heute verwendet (vgl. GB/T 12200.2-94: Kap. 4.1.2, Müller-Yokota 1994a: 351-358).

Um die Schriftentwicklung bei Glyphen und Schreibweise deutlicher aufzuzeigen, zeigt Tab. 3-1 zehn exemplarische piktographische Schriftzeichen in fünf Entwicklungsphasen.

Die Vereinheitlichung der Zeichenformen stellte eines der wichtigsten Merkmale der Vereinigung Chinas dar. Die erste sowie bedeutendste Reform der chinesischen Schrift hat Qin Shihuangdi (秦始皇帝, 259-210 v. Chr. und erster Kaiser des vereinheitlichen Chinas zwischen 221-210 v. Chr.) geleistet. Wegen der fünfhundertjährigen Vielstaaterei (770-221 v. Chr.) variierte die chinesische Schrift in jedem Königreich erheblich. Als China von Qin Shihuangdi vereinigt wurde, behinderten die verschiedenen schriftlichen Varianten stark die Staatsverwaltung und schriftliche Kommunikation. Dies hat ihn motiviert, die chinesische

[123] Von den 4.500 Einzelzeichen wurden bis heute 1.500 Zeichen entschlüsselt.

Schrift mit identischer Zeichenform der kleinen Siegelschrift zu vereinigen (vgl. Zou 2004: 204-262, Müller-Yokota 1994a: 355). Die nachfolgenden Dynastien haben die Schriftvereinheitlichung sowie -standardisierung weiter fortgeführt. Ab etwa 800 bleiben sowohl die Zeichenform als auch die Standardschriftart hauptsächlich unverändert.

Jiagu-Zeichen	**Jin-Zeichen**	**Siegel-schrift (K.)**	**Kanzlei-schrift**	**moderne Form[124]**	**Pin-yin[125]**	**Bedeut-ung**	**Strukturoriginal[126]**
				人	rén	*Mensch*	Bild eines stehenden Menschen in Seitenansicht
				女	nǚ	*Frau*	Bild einer Frau in Seitenansicht
				母	mǔ	*Mutter*	Bild einer stillenden Frau (Mutter)
				日	rì	*Sonne*	Bild der Sonne
				水	shuǐ	*Wasser*	Bild des fließenden Wassers im Fluss
				火	huǒ	*Feuer*	Bild einer Flamme
				鳥/鸟	niǎo	*Vogel*	Bild eines ruhenden Vogels
				飛/飞[127]	fēi	*fliegen*	Bild eines Flügelpaars
				羊	yáng	*Schaf*	Bild eines Schafkopfs mit Horn
				魚/鱼	yú	*Fisch*	Bild eines Fisches

Tab. 3-1: Die Evolution der chinesischen Schrift mit zehn piktographischen Beispielzeichen[128]

3.2 Orthographie, Zeicheninventar und Codierung

Die Prinzipien des Wort-/Zeichenaufbaus, die orthographischen Regelungen, das Zeicheninventar und die Setzungsregeln der Interpunktionszeichen sind die vier Grundelemente einer Schrift (vgl. Li WF 2005: 20). Anhand dieser Prinzipien wird in Kap. 3.2.1 bis 3.3.5 vorgestellt, wie solche Grundelemente bei der chinesischen Schrift funktionieren. Die Zeichenkonstruktionsprinzipien sind der Schwerpunkt von Kapitel 3.3 und die Theorien der Wortbildung aus den vorhandenen Zeichen werden in Kap. 4.4 erläutert. Die Kapitel 3.2.1 und 3.2.2 behandeln, wie die Standardisierung der Sinogramme bei Zeichenform, Aussprache und Ver-

[124] Die Schriftzeichen werden in Song-Schriftart nach der Reihenfolge Langzeichen/Kurzzeichen angegeben, falls sich das Zeichen in beiden Formen unterscheidet.

[125] Die Zeichenaussprache nach dem modernen Standardchinesisch in Pinyin.

[126] Vgl. Xu S 100 & das Editorial-Komitee von „Shuowen-Jiezi“ 2012.

[127] Das Schriftzeichen war vor der kleinen Siegelschrift das Ersatzzeichen für das gleich ausgesprochene *Negativwort* <非>. Zur graphischen Unterscheidung wurde das Konstruktionsprinzip des Zeichens <飛> mit der kleinen Siegelschrift geändert, welches einen nach oben startenden Vogel symbolisiert.

[128] Vgl. Shuowen-Jiezi Bianwei Hui 2012.

wendung funktioniert (Orthographie) und wie sich die Schrift in den vier Regionen des CJKV-Kreises von der frühen Neuzeit bis heute entwickelt hat, vor allem bezüglich den Schriftreformen, der Zeicheneinstufung nach Häufigkeit und der vereinheitlichten Codierung im Unicode.

3.2.1 Orthographie und Standardisierung in verschiedenen Regionen

Im Vergleich zu den alphabetischen Schriften, deren Orthographie durch Wörterbücher verbreitet wird, gelten so genannte Zeichenlexika bei der chinesischen Schrift als wichtigstes kodifizierendes Werkzeug. Ein Zeichenlexikon (字典 /zìdiǎn/; eng.: character dictionary) ist eine Sammlung von Schriftzeichen. Zu einem Zeicheneintrag gehören die standardisierte Form, Aussprache und Bedeutungen (inklusive Original-, erweiterte sowie angelehnte Bedeutung), unter dem auch die mit ihm als erstes Zeichen gebildeten Kompositionswörter unterordnet werden. Als eine Sonderart des Wörterbuchs wird das Zeichenlexikon hauptsächlich nur für die chinesische Schrift (Hanzi, Kanji und Hanja) eingesetzt. Es gibt vor allem drei Methoden, die Zeichen zu indexieren: die Ordnung nach Klassenhaupt (部首 /bùshǒu/, meistens auch als Radikal verstanden), nach Strichzahl und nach Aussprache. Sie sind ebenso die drei wichtigsten Verfahren für das Recherchieren im Zeichenlexikon (vgl. Lu 2008: 45ff). Die Klassenhaupte sind die Komponenten, die „die semantische Kategorie angeben und zu der das Denotat des gesamten Zeichens gehört“ oder die Grundstriche, mit denen sich Schriftzeichen einfachen Aufbaus „graphisch-strukturell gruppieren“ lassen (Li J 1996: 1406f). Jedes Schriftzeichen besitzt ein Haupt, das es „im Wörterbuch auffindbar macht“ (ibid.). So werden Klassenhaupte auch als Indexierungskomponenten (eng.: index component) bezeichnet. Die Zeichenstandardisierung durch Zeichenlexika wurde seit Jahrtausenden für den Bereich der Kultur und Bildung im chinesischen Schriftkreis eingesetzt.

‚Shuowen-Jiezi‘ (说文解字) war das älteste überlieferte Zeichenlexikon, das großen Einfluss auf die historische Schriftforschung genommen hat.[129] Es wurde im Jahr 100 n. Chr. von XU Shen (许慎) in der Späten Han-Dynastie herausgegeben und umfasste 9.353 Einzelzeichen. Stichwörter des Werks waren die in Siegelschrift dargestellten Schriftzeichen. Bei den Beiträgen wurden hauptsächlich die Konstruktionsprinzipien dargestellt und somit erläutert, wie diese auf die Grundbedeutung hinweisen. Xu Shen hat in diesem Werk auch zum ersten Mal die Zeichenordnung nach Radikal verwendet und die sechs Methoden zur Zeichenbildung (die sechs Schriften) exemplifizierend definiert. Das Buch ist eines der bekanntesten und be-

[129] Es ist dabei aber nicht das älteste Zeichenlexikon nach historischer Aufzeichnung.

deutendsten schriftlinguistischen Forschungswerke zur Sinographie in der Geschichte (vgl. Müller-Yokota 1994a: 362, Lu 2008: 45).

Die Orthographie des ‚Kangxi-Lexikons' (康熙字典) gilt als die letzte vereinheitlichte Standardisierung im gesamten chinesischen Schriftkreis. Es erschien im Jahr 1716 und übertrug bis zum Ende des zweiten Weltkriegs die offizielle Norm der chinesischen Schrift in den CJK-Gebieten. Insgesamt 48.641 Einzelzeichen nach 214 Radikalen wurden in dem Lexikon aufgenommen und die Schriftzeichen, die als Stichwort eines Beitrags aufgelistet wurden, galten als ‚orthographisch richtige Zeichen' (正体字 /zhèngtǐzì/ oder 正字 /zhèngzì/).[130] Heute besteht diese Orthographie nur in Taiwan und Hongkong fort. In Festlandchina und Japan wurde die chinesische Schrift getrennt vereinfacht und die verkürzte Form als orthographische Schrift definiert. In Südkorea wurde keine offizielle Regelung dazu veröffentlicht (vgl. Müller-Yokota 1994a: 371). Im heutigen Festlandchina ist das Xinhua-Zeichenlexikon das häufigste Zeichenlexikon, welches als ein grundlegendes Werkzeugbuch der chinesischen Grund- sowie Mittelschüler Verwendung findet.[131]

Der Begriff ‚orthographisch richtiges Zeichen' stammt aus dem Buch ‚Ganlu-Zishu' (干禄字书)[132] von YAN Yuanxun (颜元孙, ?-714 n. Chr.) und kodifiziert die Zeichen, die öffentlich anerkannt als ‚korrekt' gelten und in offiziellen Dokumenten verwendet werden. Den orthographischen Zeichen gegenüber stehen Variantenzeichen (异体字 /yìtǐzì/). Yan hat in seinem Werk zwei Sorten von Varianten definiert, ‚Ersatz-' (通字 /tōngzì/) sowie ‚unorthodoxe Zeichen' (俗字 /súzì/).[133] Ersatzzeichen können in einem bestimmten Kontext ein anderes Schriftzeichen mit derselben oder einer ähnlichen Bedeutung substituieren. ‚Unorthodoxe' sind auf Basis der Orthographie vom Volk verkürzte Zeichen, die nur in inoffiziellen Situationen benutzt wurden. Viele unorthodoxe Zeichen der Schriftgeschichte werden heutzutage als orthographische Zeichen anerkannt. Sie sind eine der wichtigsten Quellen der Schriftvereinfachung in der VR China (vgl. Gao 2000: 70, Müller-Yokota 1994a: 371ff).

Ab dem 20. Jh. hat sich die Standardisierung der chinesischen Schrift in Festlandchina, Taiwan, Hongkong, Japan und Südkorea hauptsächlich unabhängig voneinander entwickelt. In der VR China wurde die Vereinfachung und Verkürzung der chinesischen Schrift ein wichtiges Projekt. Eine Schriftreform der Zeichenverkürzung (1951-1977) wurde auf Initiative der Regierung eingeführt. In Taiwan und Hongkong wird der traditionelle Standard beibehalten.

---

[130] 正 /zhèng/ bedeutet *Standard*. Im Deutschen heißen die standardisierte Schrift sowie Schriftzeichen deswegen *orthographische richtige Schrift* bzw. *Zeichen* (vgl. Müller-Yokota 1994a: 371).

[131] Die aktuellste Auflage: Xinhua Zidian – 11. Version, 2011, herausgegeben von the Commercial Press.

[132] Das erste Veröffentlichungsjahr ist noch unbekannt.

[133] Die deutschen Übersetzungen *Varianten-*, *Ersatz-* und *unorthodoxe Zeichen* stammen von Müller-Yokota (1994a: 371f.)

In Vietnam und Nordkorea wurde eine Alphabetisierung mit dem lateinischen Alphabet sowie dem koreanischen Hangul eingeführt. In Südkorea wurde das nationale Alphabet als Hauptschrift verwendet, während Hanja in wenigen Fällen benutzt werden, um Homophone zu unterscheiden. In Japan wurde einerseits die Sinographie vereinfacht, andererseits wurden die zum Schreiben verwendeten Kanji auf unter 2000 begrenzt (vgl. Atsuji 1994: 445-450).

Die moderne Standardisierung der chinesischen Schrift kann in vier Punkten zusammengefasst werden: 1) die Festlegung des Inventars der häufig gebrauchten Schriftzeichen, 2) die Standardisierung orthographischer Zeichenformen, 3) die Bestimmung der Zeichenaussprache im Putonghua, 4) die Festlegung der Zeichenindexierung in Lexika (vgl. Sheng 2006: 76f). Die Schriftstandardisierung ist ebenso die Grundlage für Zeichencodierung und Textverarbeitung der chinesischen Schrift in CJK-Gebieten. Im nächsten Kapitel werden die verschiedenen Standardisierungen der Sinographie in der VR China, Taiwan, Japan und Südkorea näher vorgestellt.

3.2.2 Zeichengebrauchs- und Zeichencodierungsstandards in CJK-Regionen

Für Wortbildungen in der chinesischen Schrift gibt es zwei Methoden: die Erfindung von Einzelzeichen für neue Begriffe und die Zusammensetzung von zwei oder mehreren, bereits vorhandenen Zeichen, die entweder symbiotisch Bedeutung oder Aussprache des Wortes beschreiben. Obwohl mit der Zeit die zweite Methode immer häufiger verwendet wurde, endete die Kreation neuer Schriftzeichen nie (vgl. Lu 2008: 39f). Bei den Jiagu-Zeichen vor ca. 3.300 Jahren wurden insgesamt 4.500 Einzelzeichen in zehntausenden Schriftstücken verwendet. Das 1039 veröffentlichte Lexikon ‚Ji Yun' (集韻) umfasste insgesamt 53.525 Schriftzeichen. Heutzutage umfasst das größte Lexikon ‚Dictionary of Chinese Character Variants' (异体字字典, die 6. Ausgabe im Jahr 2017) von ‚Academy for Educational Research' aus Taiwane die Gesamtzahl von 106.330.[134] In dem modernen chinesischen Schriftsystem reichen im Normalfall 3.500 Schriftzeichen für das Lesen im Alltag und 7.000 generelle Schriftzeichen für verschiedene Medien in der VR China. Die Einteilung des Zeicheninventars nach Häufigkeit ist bei der Schulbildung im CJK-Kreis entscheidend.

In Festland China hat die Einstufung der orthographisch vereinfachten Schriftzeichen mit der Schriftreform in den 1950er Jahren begonnen und wird bis heute immer wieder verbessert und aktualisiert. Der aktuellste Stand findet sich in der ‚Tongyong Guifan Hanzi Biao' (通用

[134] Vgl. Yitizi-Lexikon Online: *http://dict.variants.moe.edu.tw/variants/rbt/page_content3.do?pageId=2981893* [letzter Abruf: 2019/02/21]; unter den hunderttausenden Zeichen sind 29.921 als orthographische Zeichen indexiert, während die Mehrheit – 74.407 – Variantenzeichen sind und 2.002 zu überprüfende Zeichen gehören.

规范汉字表, wörtl.: die Liste der gemeingebräuchlichen Schriftzeichen) aus dem Jahr 2013. In der Liste werden insgesamt 8.105 Zeichen in drei Stufen aufgenommen. Die erste Stufe enthält 3.500 höherfrequente Zeichen (常用字 /chángyòngzì/), die für Schulbildung und Fremdsprachenlernen obligatorisch gebraucht werden. 2.500 der dazu zählenden Zeichen werden innerhalb der Liste als sehr häufig eingestuft und im Primärschulbereich vermittelt. Die zweite Stufe ist für weitere 3.000 gemeingebräuchliche Zeichen (通用字 /tōngyòngzì/), die in Büchern, von der Presse, zur Textverarbeitung etc. verwendet werden. In der dritten Stufe werden 1.605 Zeichen für Eigennamen, wissenschaftliche Fachbegriffe sowie klassische Texte aus Lehrwerken der Grund- sowie Mittelschule aufgelistet (nach „Tongyong Guifan Hanzi Biao“ 2013).

Die heutige orthographische chinesische Schrift der VR China besteht aus ‚vereinfachten' und ‚vererbten Schriftzeichen' (传承字/chuánchéng zì/). Die vererbten Schriftzeichen sind die Zeichen, deren Standardform seit dem 8. Jh. unverändert blieb und im ganzen chinesischen Kreis gültig ist. Unter den 8.105 Schriftzeichen von ‚Tongyong Guifan Hanzi Biao' 2013 gibt es insgesamt 3.120 orthographische vereinfachte Schriftzeichen (vgl. ibid.: 48). Die Reform der Schriftvereinfachung hat zwar die Verwendung der chinesischen Schrift vielmals versimpelt, gewisse Nachteile wurden im Laufe der Zeit jedoch offenbar. Erstens verliert die vereinfachte Form eines Zeichens in vielen Fällen den durch die Zeichenstruktur erklärbaren Sinninhalt. Zweitens wurden für die Verkleinerung des Zeicheninventars viele Zeichen von einem anderen Zeichen mit gleicher Aussprache oder mit ähnlicher Form ersetzt, so dass sie ambig geworden sind. Außerdem wurde die schriftliche Kommunikation per Internet zwischen verschiedenen Regionen des CJK-Gebiets, besonders zwischen Festlandchina und Taiwan/Hongkong, hierdurch in mancher Hinsicht behindert.

Im Bereich der Informatik galt vor Einführung des vereinheitlichten CJK-Unicodes in der VR China ‚GB 2312' als wichtigste Norm der chinesischen Schrift. Sie enthält insgesamt 6.763 gemeingebräuchliche Sinogramme, die zur allgemeinen Textverarbeitung, Kommunikation usw. dienen. Solche Schriftzeichen werden mit 14 Bits (das zweifach des ASCII-Codes) definiert und in zwei Stufen eingeteilt: 3.755 häufig gebräuchliche Zeichen, die nach Aussprache geordnet werden, und 3.008 sonstige gemeingebräuchlichen Zeichen, die nach Radikal aufgelistet werden. Auf dieser Basis wurden fünf Erweiterungskategorien innerhalb des GB manifestiert, um seltene, traditionelle und Variantenzeichen zu ergänzen. Die sechs Normen umfassen insgesamt jeweils ca. 21.000 Zeichen von der verkürzten sowie traditionellen Orthographie (vgl. Zhang ZC 1999: 11ff).

In Taiwan, wo die ebenso auf Mandarin bezogene Schriftsprache in traditioneller Schriftform genutzt wird, gelten seit 1949 andere Standards bei Zeichenform und Codierung. Das taiwanesische Bildungsministerium hat 1980 die ‚Liste der orthographischen Glyphen der häufig gebräuchlichen Zeichen' (常用國字標準字體表 /chángyòng guózì biāozhǔn zìtǐ biǎo/) veröffentlicht, in der 4.808 Sinogramme angegeben wurden. Die Entwürfe für die ‚Liste der orthographischen Glyphen der weniger häufig gebräuchlichen Zeichen' (次常用國字標準字體表 /cì chángyòng guózì biāozhǔn zìtǐ biǎo/) mit 6.025 Zeichen und die ‚Liste der orthographischen Glyphen der selten gebräuchlichen Zeichen' (罕用國字標準字體表 /hǎnyòng guózì biāozhǔn zìtǐ biǎo/) mit 12.924 Zeichen wurden jeweils im Jahre 1981 und 1987 offiziell eingeführt (vgl. ibid.: 15). Die drei Listen wurden auf der Basis des Kangxi-Lexikons entworfen. Die Austauschcodierungen namens CCCII (veröffentlicht 1980) und CNS 11643 (veröffentlicht 1986) stimmen mit dem Zeichenstandard der drei Listen hochgradig überein. In CNS werden 5.401 häufige und 7.650 gemeingebräuchliche Zeichen festgelegt. Ein Code der beiden Normen beträgt 21 Bits (3×7) (vgl. ibid.: 15f, CNS 11643-1992 Plane 2).

Aufgrund der Einflüsse der westlichen Welt wurde die Anzahl der Kanji im japanischen Schriftsystem nach Ende des Zweiten Weltkriegs stark reduziert. Nach aktuellen Statistiken beinhaltet die 2010 veröffentlichte ‚Jouyou Kanji Hyou'[135] 2.136 Kanji (vgl. „Jouyou Kanji Hyou" 2010). Im japanischen Industriestandard JIS (vergleichbar mit der deutschen DIN-Norm) wurde 1978 die Norm JIS X 0208-1983 für japanische Textverarbeitung eingeführt. Diese Norm enthält 6.353 Kanji mit 14-stelligen Bitfolgen. Davon gibt es 2.935 Zeichen der ersten Stufe nach dem Ordnungsschema der Aussprache und 3.388 Zeichen der zweiten Stufe mit Radikalanordnung (vgl. Zhang ZC 1999: 16).

Auf der koreanischen Halbinsel haben sich Kultur, Sprache und Schrift nach 1945 aus politischen Gründen in verschiedene Richtungen entwickelt. Im Norden wurde die Geschichte der chinesischen Schrift für beendet erklärt, während im Süden ihre Verwendung erheblich eingeschränkt wurde. 1948 wurde das so genannte ‚Hangul-Gesetz' von der südkoreanischen Regierung ratifiziert, in dem geregelt wird, dass in amtlichen Schriftstücken kein chinesisches Schriftzeichen auftreten darf, sondern in Hangul umschrieben werden muss. Aus kulturellen Gründen haben diese politischen Bemühungen auch negative Wirkungen verursacht, vor allem Lesebarrieren bei nationalen klassischen Werken der Literatur und historischen Aufzeichnungen. 1972 wurde eine Zeichenliste ‚essentieller chinesischer Schriftzeichen für den hammun-Unterricht' mit 1.800 sinokoreanischen Zeichen veröffentlicht, die man in der Mittel-

[135] Titel in Original: 常用漢字表, chi. Aussprache: /chángyòng hànzì biǎo/, jap.: /jouyou kanji hyou/, wörtl.: *die Liste der häufig gebräuchlichen Kanji*.

sowie Oberschule erwerben muss. Heute ist das koreanische Schriftsystem zu einem großen Teil in Hangul und zu einem kleinen Teil nach der chinesischen Schrift aufgebaut (vgl. Atsuji 1994: 447, Lu 2008: 368f). Der südkoreanische Standard der Austauschcodierung lautet KS X 1001-1992 mit insgesamt 8.224 Zeichen in 14 Bits. Darunter gibt es 2.350 Hangul-Silbenblöcke und 4.888 Hanja (vgl. KS X 1001: 1992).

Trotz verschiedener Standards in den vier Regionen gibt es wegen desselben Stamms zahlreiche gemeinsam gebrauchte Schriftzeichen. Die Identität der Schriftzeichen ermöglicht die Entstehung der vereinheitlichten Zeichencodierung der chinesischen Schrift im Unicode – ‚CJK Unified Ideographs'.

3.2.3 Vereinheitlichte Austauschcodierung in Unicode

Trotz derselben Schrift galten vor der Unicode-Einführung verschiedene Normen für Datenaustausch in verschiedenen Regionen des CJK-Kreises. Aus Bedürfnissen der schriftlichen Kommunikationen und der Kompatibilität der Software ist ein international gültiger Austauschcode (Unicode) von großer Bedeutung. Die Vereinheitlichung der Sinogramme, die zehntausende Codepunkte belegen, war ein wichtiges Projekt des Unicode-Konsortiums. Das Projekt bezieht sich vor allem auf zwei Kernzielsetzungen: 1) Ein Computersystem oder eine Software kann auf die Sammlung von verschiedenen Codierungsnormen verzichten, so dass sie schneller und störungsloser im CJK-Gebiet nationalisiert werden kann; 2) Mehr Zeichensätze, die den Qualitätsanforderungen der Presse und der Büroarbeit genügen, können angeboten werden (vgl. Zhang ZC 1999: 18).

Im Jahr 1993 wurde der erste Entwurf des ‚CJK Unified Ideographs' (auch CJKV, V für Vietnam steht), der vom Forschungsteam ‚CJK-IRG' (Abkürzung für ‚China, Japan, Korea Ideographic Research Group') herausgegeben wurde, als eine Kategorie des Unicodes vorgestellt. Diese Kategorie enthält insgesamt 20.902 generelle Sinogramme mit 16-stelligen Bitfolgen im 4E00-9FFF. Die wichtigsten Grundlagen für die generell vereinigten Sinogramme sind GB2312-80 (Festlandchina), CNS 11643 (Taiwan), JIS X 0208 (Japan) und KS C 5601-87 (Südkorea), die als die wichtigsten nationalen Normen galten. Die Standards aus Hongkong, Singapur, Nordkorea und den USA wurden auch berücksichtigt. Von 1993 bis heute wurden immer neue Ergänzungskategorien für die CJK-Ideographen veröffentlicht. Unabhängig von der Aussprache werden die Sinogramme primär nach dem Radikal, sekundär nach der Strichanzahl sowie tertiär nach der Strichordnung aufgelistet. Da die Aussprache bei verschiedenen Schriftsystemen stark variiert, ist eine aussprachenbasierte Reihung in dem vereinheitlichten Standard unmöglich (vgl. Unicode 12.0 Chapters: Kap. 18.1: 716f).

Beim Entwurf der vereinheitlichten CJK-Ideographen wurden vor allem die folgenden drei Prinzipien beibehalten: 1) das Zeicheninventar basiert hauptsächlich auf den vier nationalen Normen der VR China, Taiwan, Japan und Südkorea; 2) die Variantenzeichen, die vorher in einem Standard als verschiedene Codepunkte definiert wurden, werden in Unicode ebenfalls verschiedene Zeichen; 3) Variantenglyphen identischer Zeichen, die sich voneinander wenig unterscheiden, werden mit einem Zeichencode codiert (vgl. Zhang ZC 1999: 21). Diese Prinzipien können durch die folgende Abbildung (Abb. 3-1) näher erläutert werden:

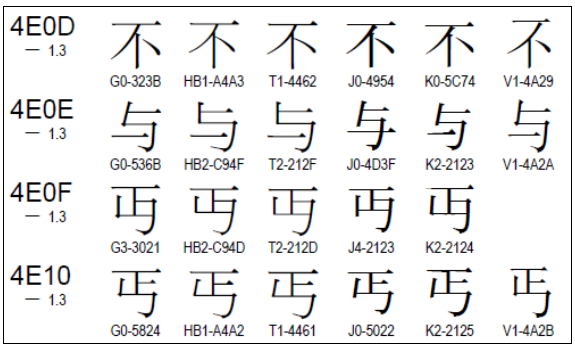


Abb. 3-1: Abschnitt von CJK Unified Ideographs mit Variantenglyphen (Unicode 12.0 Character: U+4E0D-U+4E10)[136]

In der Abbildung sind vier Schriftzeichen mit gleichem Radikal angegeben. Die Zeichenglyphen in verschiedenen Regionen sind zwar nicht hundertprozentig identisch, aber ihre Unterschiede sind relativ gering. Das dritte Zeichen <丐> (/miǎn/, *unsichtbar*) und das vierte <丐> (/gài/, *Bettler*) haben zwar einen sehr geringen Unterschied, sind aber grundsätzlich unterschiedliche Zeichen für verschiedene Begriffe. Variantenzeichen in jedem Schriftsystem müssen berücksichtigt werden. Zum Beispiel ist <幷> (U+4E3C) im chinesischen Schriftsystem das Variantenzeichen von <井> (U+4E95; (/jǐng/ *Brunnen*). Im Japanischen sind die beiden Zeichen aber verschiedene Begriffe (幷 als Schale/Schüssel), weshalb sie im Unicode verschieden codiert werden müssen.

3.3 Graphischer Aufbau und Konstruktionsprinzipien der Schriftzeichen

In Abb. 1-5 (S. 34) wurden die drei Dimensionen des chinesischen Schriftzeichens – Form, Aussprache und Bedeutung – eingeführt. Kap. 3.3 fokussiert nun die Zeichenform, die in den meisten Fällen wegen der allgemeinen Konstruktionsprinzipien im Zusammenhang mit der Zeichenbedeutung steht.

[136] Die nationalen Glyphen eines einzelnen Zeichens und die entsprechenden Codepunkte der nationalen Normen werden auch angegeben. ‚G' steht für die Norm aus der VR China, ‚HB' für Hongkong, ‚T' für Taiwan, ‚J' für Japan, ‚K' für Südkorea und ‚V' für Vietnam.

Der graphische Aufbau der chinesischen Schrift kann nach zwei verschiedenen Systemen analysiert und erforscht werden: 1) das System der Zeichenbildung und graphischen Struktur, das sich auf die Konstruktionsprinzipien der Grapheme bezieht; 2) das System von Strichen (chi.: 笔画 /bǐhuà/, eng.: stroke) und Komponenten (部件 /bùjiàn/, component), bezogen auf jene graphischen Elemente, die die Bausteine der Zeichenform darstellen (vgl. Sun 1988: 282). In Kap. 3.3.2 wird zuerst das System der Bildungsprinzipien (die sechs Schriften) erläutert, bevor in Kap 3.3.2 bis 3.3.4 das zweite System unter die Lupe genommen wird.

3.3.1 Verschiedene Perspektiven über den Aufbau

Die Bildungsprinzipien der chinesischen Schriftzeichen können aus drei verschiedenen Perspektiven betrachtet werden:

1) Vor dem Hintergrund der Erfindung und der Verwendung sind die ‚sechs Schriften' (chi.: 六书 /liùshū/, eng.: the six writings)[137] am bekanntesten, die von Xu Shen in ‚Shuowen-Jiezi' definiert wurde. Es gibt sechs Verfahren für das Schaffen sowie die Anwendung der Schriftzeichen. Dies sind jeweils: Piktogramm, Ideogramm, zusammengesetztes Ideogramm, Determinativphonetikum, Entlehnungs- und Synonymszeichen. Unter den sechs Prinzipien sind die ersten vier Methoden für die Erfindung neuer Zeichen, die anderen zwei für die Zeichenverwendung (vgl. Müller-Yokota 1994a: 362).
2) Unter dem Aspekt der Zeichenstruktur können die chinesischen Schriftzeichen in zwei Subklassen geteilt werden: die simplen Zeichen (独体字 /dútǐ zì/), die nicht mehr in weitere Zeichen geteilt werden; und die komplexen Zeichen (合体字 /hétǐ zì/), die aus zwei oder mehreren Grammwurzeln aufgebaut sind. Eine Grammwurzel (字根 /zìgēn/) ist ein Symbol, das meistens von einem Grundzeichen abgeleitet und als Element zur Zeichenbildung verwendet wird (vgl. Li J 1996: 1407, Lu 2008: 25f).
3) Wie bei den altägyptischen Hieroglyphen kann die chinesische Schrift aufgrund des Typs der Grammwurzel analysiert werden. Die Grammwurzeln unterscheiden sich in phonologische (声符 /shēngfú/, auf die Aussprache bezogen), semiologische (意符/yìfú/, auf die Bedeutung bezogen) und graphologische Symbole (标符 /biāofú/, weder auf Aussprache noch auf Bedeutung bezogen, zur ästhetischen oder graphisch unterscheidenden Funktion fungierende Symbole) (vgl. Lu 2008: 25-28).[138]

[137] ‚Sechs Schriften'bedeutet ‚sechs Klassen der Schriftzeichen' (eng.: six classes of characters) (vgl. Coulmas 1996a: 80).
[138] Auf den englischen Übersetzungen der drei Begriffe (identische Quelle) – *phonological, semological & graphological symbol* – basieren die deutschen Übersetzungen.

3.3.2 Die sechs Schriften

Durch die Erläuterung der Konstruktionsprinzipien lassen sich die drei Aspekte für den Aufbau der chinesischen Schrift näher verdeutlichen. Da die ‚sechs Schriften' vorwiegend im CJK-Kreis erforscht wurden, stelle ich Xu Shens Theorie als Schwerpunkt der Konstruktionsprinzipien vor.

1) Piktogramm (chi.: 象形 /xiàngxíng/; auch: *ursprüngliches Bild*)[139]

„Piktogramm heißt ein Zeichen, das einen Gegenstand nach seiner Gestalt malerisch bezeichnet, wie 日 [rì, *Sonne*] und 月 [/yuè/, *Mond*, archaische Form: ]" (Xu S 100: Vorwort [Übersetzung der Verfasserin]).[140] Die Piktogramme wurden von ursprünglichen Bildern abgeleitet und gehören zu den ältesten und grundlegendsten Sinogrammen. Nach Rechnung von Zhu Junsheng gibt es unter den 9.353 Schriftzeichen von ‚Shuowen-Jiezi' 364 Piktogramme, die somit 3,89% des gesamten Inventars betragen. Trotz des kleinen Anteils gehören die meisten Piktogramme zu den häufigen Zeichen und sind die wichtigsten Grammwurzeln für die Zeichenbildung (vgl. Müller-Yokota 1994a: 362, Sun 1988: 48). Nach anderer Ansicht ist ein Piktogramm ein aus einem semiologischen Symbol bestehendes alleinstehendes Zeichen.

Die Piktogramme können in zwei Subklassen eingeteilt werden: eigenständiges und kombiniertes Piktogramm (vgl. Sun 1988: 47f). In Tab. 3-1 sind z.B. die Zeichen für *Mensch* <人>, *Frau* <女> und *Vogel* <鸟> eigenständige Piktogramme, die nur den Gegenstand gestalten. Das Zeichen <眉> (/méi/, *Augenbraue*) ist dem hingegen ein kombiniertes Piktogramm. Der untere Teil <目> (/mù/, Altform: ) ist das Zeichen für *Auge*. Das Symbol oben ist die Gestalt der *Augenbrauen-Haare*, welches selbst aber Nicht-Zeichen ist. Ohne das Symbol für *Auge* wäre der Begriff *Augenbraue* nur schwer zu symbolisieren.

2) Ideogramm (指示 /zhǐshì/; auch symbolisches Bild)[141]

„Ideogramm heißt ein Zeichen, das beim Lesen seinen Begriff nach Analyse der Zeichenform zu erkennen und zu verstehen gibt, wie 上 [shàng, *oben*] und 下 [xià, *unten*]" (Xu S 100: Vorwort [Übersetzung der Verfasserin]). Die abstrakten Begriffe können unmöglich nach ihrer Gestalt kreiert werden, weshalb die Ideogramme fast so früh wie Piktogramme entstanden,

[139] *Piktogramm* ist von der englischen Bezeichnung *pictographic chracter* abgeleitet (vgl. Coulmas 1996a: 81).
[140] Die angegebene archaische Form ist meist das Jiagu-Zeichen; für die Altform von 日 (*Sonne*) vgl. Tab. 3-1.
[141] *Ideogramm* ist von der englischen Bezeichnung ‚simple ideographic chracter' abgeleitet (vgl. Coulmas 1996a: 81), die den Gegensatz zu ‚zusammengesetztes Ideogramm' bildet.

um die Einschränkung bei der Bezeichnung mancher abstrakten Begriffe zu überwinden (etwa Numerale, Adjektive, unsichtbare Gegenstände usw.). Das Antonympaar <上> und <下> (Altformen:  sowie ) wurde z.B. gleichzeitig sich entsprechend erfunden. Die Horizontale stellt die Bezugslinie dar und die nach oben laufende Vertikale symbolisiert *oben* bzw. umgekehrt *unten*. Ebenso sind die Zahlzeichen <一> (/yī/, *eins*), <二> (/èr/, *zwei*) und <三> (/sān/, *drei*) typische Ideogramme (vgl. Zou 2004: 42f).

Ideogramme können auch in eigenständige und kombinierte Ideogramme subklassifiziert werden. Eigenständige Ideogramme sind rein abstrakte Symbole, wie <上> und <下>. In diesem Fall ist ein Ideogramm ein alleinstehendes Zeichen von einem semiologischen Symbol. Kombinierte Ideogramme sind die Zeichen, in denen ein Symbol auf der Basis eines Piktogramms ergänzt wird. So wird z.B. das Zeichen für Messerklinge <刃> (/rèn/, *Messerklinge*) auf der Basis von <刀> (/dāo/, *Messer*) zur Betonung der scharfen Seite des Messers (*Klinge*) auf der linken Seite durch einen Punkt symbolisiert. In diesem Fall ist ein Ideogramm die Zusammensetzung von einem semiologischen und einem graphologischen Symbol. Ausnahmen von Ideogrammen, die nicht zu den beiden Subklassen gehören, sind <一>, <二> und <三>. In den drei Zeichen stellt der horizontale Strich das graphologische Symbol dar und die Bedeutung wird durch die Häufigkeit des Strichs symbolisiert. Es gab in ‚Shuowen-Jiezi' 125 Ideogramme, ergo 1,34% des Gesamtinventars (vgl. Sun 1988: 49ff).

3) Zusammengesetztes Ideogramm (会意 /huìyì/; auch *zusammengesetztes Bild*; Abk.: ZI)[142]

„Zusammengesetztes Ideogramm heißt ein Zeichen, dessen Bedeutung aus den Sinninhalten der Bestandsymbole auf den Gesamtzusammenhang schließen lässt, wie 武 [wǔ, das *Militär*] und 信 [xìn, die (*moralische*) *Integrität*]" (Xu S 100: Vorwort [Übersetzung der Verfasserin]). Anders formuliert sind solche Zeichen aus zwei oder mehreren Piktogrammen oder Ideogrammen zusammengesetzt, um eine logisch erschließbare Bedeutung wiederzugeben. So besteht das zweite Beispielzeichen <信> aus <亻> (abgeleitet von <人> /rén/, der *Mensch*) und <言> (yán, *sprechen*, das *Wort*) und ist eine Definition des alten chinesischen Volkes für Integrität (in etwa: *Man soll seine versprochenen Worte halten.*). Vor dem Hintergrund der Typologie der Grammwurzel besteht ein ZI aus zwei bis mehreren ‚semiologischen Symbolen' und möglicherweise auch einem ‚graphischen Symbol'. In dem Zeichen <武> sind <止> (*be-*

[142] Das Wort *zusammengesetztes Ideogramm* ist von der englischen Bezeichnung ‚compound ideographic character' abgeleitet (vgl. Coulmas 1996a: 81).

enden) und <戈> (*Speer*) semiologische Symbole und die kurze obere Horizontale ist ein graphisches Symbol.

Es gibt zwei Typen von ZI – die Wiederholung eines Grundzeichens und die Zusammensetzung von verschiedenen Grundzeichen (vgl. Sun 1988: 51ff). Verdoppelung oder Verdreifachung desselben Grundzeichens zur Neuzeichenbildung ist die häufigste Art des ersten Typs. Diese Vorgehensweise kann zugleich eine Stärkung oder Gruppenzugehörigkeit symbolisieren. Ein Beispiel hierfür ist der Zeichenaufbau von *Wald* <林> /lín/, der sich durch die zweimalige Wiederholung von *Baum* <木> /mù/ zusammensetzt. *Drei Bäume* bezeichnen einen *Wald* mit mehr und höheren Bäumen <森> /sēn/. *Volk* oder *große Menge* wird durch die Verdreifachung des Zeichens für *Mensch* geschrieben: <众> /zhòng/. [143]

Die aus verschiedenen simplen Zeichen zusammengesetzten Zeichen sind in der chinesischen Schrift am zweitmeisten (nach Determinativphonetika) vertreten. Solche Zeichen sind häufig rätselhaft und können in vielen Fällen die Philosophie, Weltanschauung usw. der Altchinesen verraten. Zum Beispiel besteht das Zeichen *gut* <好> /hǎo/ aus *Frau* <女> /nǚ/ und *Kind* <子> /zǐ/. Das Zeichen *hell* <明> /míng/ ist die Addierung von *Sonne* <日> /rì/ und *Mond* <月> /yuè/. *Lecker* <鲜> /xiān/ wird durch Komposition von *Fisch* <鱼> /yú/ und *Schaf* <羊> /yáng/ ausgedrückt.

Das Konstruktionsprinzip wird zwar mit wenigen Einschränkungen zur Zeichenerfindung verwendet, hat aber unlösbare Nachteile. In vielen Fällen ist es unmöglich, durch die bestehenden Zeichen den Sinninhalt eines ZI zu erschließen. So mag das Zusammenbild von *Frau* und *Kind* nicht vollends arbiträr auf das Adjektiv *gut* verweisen, es liegt jedoch auf der Hand, dass es eine Vielzahl weiterer Bedeutungen evozieren könnte. Es gibt in ‚Shuowen-Jiezi' insgesamt 1.167 zusammengesetzte Ideogramme (12,48% der Gesamtheit) und mit der Zeit wurden viele neu geschaffene Schriftzeichen dieser Gruppe zugehörig (vgl. Sun 1988: 54).

4) Determinativphonetikum (形声 /xíngshēng/; Abk.: DP)[144]

„Determinativphonetikum heißt ein Zeichen, zu dessen Bildung eine Komponente von zusammenziehendem Begriff und eine Komponente mit gleicher oder ähnlicher Aussprache ausgewählt und kombiniert werden, wie 江 [jiāng, der *Fluss*] und 河 [hé, der *Fluss*]" (Xu S 100: Vorwort [Übersetzung der Verfasserin]). In <江> und <河> ist <氵> das abgeleitete Ra-

[143] In traditioneller Form wird das Zeichen als <眾> geschrieben.

[144] Übersetzung *Determinativphonetikum* stammt von Müller-Yokota (vgl. 1994a: 364), deren Bestandwörter jeweils für 形 /xíng/ und 声/shēng/ stehen.

dikal für *Wasser*, während <工> (/gong/, *Arbeit, Industrie* etc.) und <可> (/kě/, *können, anpassen, erlauben* etc.) wegen der ähnlichen Aussprache zur Zeichenbildung verwendet, die die Bedeutung des gesamten Zeichens nicht beeinflusst.[145] Vor dem Aspekt der Grammwurzel besteht ein DP aus einem semiologischen (wie <氵> im Beispielzeichen) und einem phonologischen Symbol (wie <工/可>). Durch die Komposition von bedeutungs- und phonetikhinweisenden Komponenten in DP entsteht ein stabiler Form-Aussprache-Sinninhalts-Zusammenhang (vgl. Sun 1988: 55, Zou 2004: 45).

DP haben zwei Vorzüge: 1) durch die semiologische und phonologische Angabe können sie einfacher erkannt und gelesen werden; 2) sie können effektiver erschaffen werden (vgl. Sun 1988: 58). Diese Zeichenbildungsmethode spielte bei der Schriftentwicklung eine immer bedeutendere Rolle und die mit DP gebildeten Schriftzeichen werden latent häufiger. Betrugen sie nach statistischen Analysen unter den Jiagu-Zeichen noch 27,27%, war ihr Anteil unter den Jin-Zeichen bereits auf 56,28% gestiegen. Nach der Sammlung in ‚Shuowen-Jiezi' waren in der kleinen Siegelschrift 81,24% und ca. ab dem 10 Jh. über 90% der gesamten Zeichen DP (vgl. Zou 2004: 48, Zhu 1998: 594). Ihre Geschichte zeigt die allgemeine Entwicklungstendenz der chinesischen Schrift auf: die Bezogenheit von Aussprache auf Zeichenform und -sinninhalt wurden immer eindeutiger (vgl. Zhang YJ 1991: 98f). Die in der modernen Zeit neu erfundenen Schriftzeichen sind ebenfalls zumeist DP. Bekannte Beispiele dafür sind die Zeichen der chemischen Elemente. Wegen der Eigenschaften der chinesischen Sprache wären phonetische Übertragungen zu kompliziert für didaktischen Erwerb und Verbreitung der Chemie. So wurden alle Elemente monosyllabisch bezeichnet und für Schriftzeichen mit den Prinzipien der DP erfunden. Mit dem Radikal für *Metall* <钅> (/jīn/, *Gold*/*Metall*), für *Halb-* und *Nichtmetall* <石> (/shí/, *Stein*) und für *Gas* <气> (/qì/, *Gas*) können alle Elemente qua ihrer Schriftform ihre Zugehörigkeit anzeigen, wie z.B. <氢> (/qīng/, *Wasserstoff*)，<钠> (/nà/, *Natrium*) und <硅> (/guī/, *Silicium*).

Die Determinativkomponenten stammen hauptsächlich von Piktogrammen, die im Normalfall als Radikal gelten. Viele davon wurden auf Basis originaler Zeichen erheblich umgeformt, wie <犭> (von dem Piktogramm für *Hund* <犬> für *Säugertier*), <忄> (von <心> *Herz* für *Gefühle* und *Gedanken*), <艹> (von dem Zeichen <艸> *Gras* für *krautige Pflanzen*) oder <

[145] <江> und <河> bezeichnen denselben Begriff *Fluss*, wobei <江> original nur für den *Yangtse-* und <河> für den *Gelben Fluss* steht. Die Aussprache der beiden Zeichen hat sich diachron geändert, weshalb das Phonetikum von <江> undeutlich zu erkennen ist.

灬> (von <火> *Feuer* für Begriffe des zugehörigen Wortfelds). Ca. 200 Komponenten können als Determinativ verwendet werden (vgl. Sun 1988: 57).

Sowohl simples als auch komplexes Zeichen können phonologische Symbole zur Zeichenbildung sein (ibid.), d.h. viele DP können in drei oder mehr Basiszeichen zerlegt werden. Das Phonetikum des Zeichens <影> (/yǐng/ *Schatten*, *Reflektionsbild*) ist z.B. <景> (/jǐng/ *Aussicht*), das selbst ein DP von dem Determinativ <日> (/rì/ *Sonne*) und dem Phonetikum <京> (/jīng/ *Hauptstadt*) ist. Das Beispielzeichen verweist auch darauf, dass in vielen Fällen das Prinzip des zusammengesetzten Ideogramms zusammen mit DP benötigt wird. Anders formuliert wird die phonetische Komponente bevorzugt, deren Sinninhalt näher auf die gesamte Zeichenbedeutung eines DP bezogen ist.

Viele DP sind nach heutigen Ansichten irrational aufgebaut. Das Determinativ passt in manchen Fällen nicht zum Begriff des gebildeten Zeichens und das Phonetikum ist phonetisch abweichend zur Aussprache des Gesamtzeichens. Nach Untersuchungen von Zhou YouGuang an allen orthographisch richtigen Zeichen des 1970 erschienenen Xinhua-Lexikons sind nur 39% der DP bei der Aussprache (ohne Tonberücksichtigung) mit ihrem Phonetikum identisch (vgl. Zhou YG 1980: 4f). Das bedeutet, dass das Phonetikum eines DP – wegen des phonetischen Wandels – in den meisten Fällen nicht korrekt seine Aussprache angeben kann.

5) Entlehnungszeichen (假借 /jiǎjiè/)[146]

„Entlehnung heißt ein Zeichen, das ursprünglich nicht vorhanden war und von einem Zeichen mit derselben Aussprache für seine Bezeichnung entlehnt wurde, wie 我 [wǒ, *ich*] und 足 [zú, *der Fuß*, *genügend*]“ (Xu S 100: Vorwort [Übersetzung der Verfasserin]).[147] Durch Entlehnung wird ein vorhandenes Zeichen mittels phonologischer Einigkeit oder Ähnlichkeit für manche Funktionswörter verwendet, die unmöglich zu gestalten sind. <我> war original ein ZI für *Speer* (Altform: ) und wurde für das Pronomen *ich* entlehnt. Die originale Bedeutung ging mit der Zeit verloren. <足> war ein Piktogramm für *Fuß* (Altform: ) und wurde für den abstrakten Begriff *genügend* entlehnt. Beide Bedeutungen sind bis heute mit dem Schriftzeichen verbunden. Entlehnung folgt dem ähnlichen Prinzip wie die Silbenschrift, wurde aber im Kontrast zur Evolution vieler anderer Schriften der Welt immer seltener verwendet. Im Gegenteil wurden sie in der Schriftentwicklung immer mehr von DP ersetzt.

[146] Diese Übersetzung stammt aus Müller-Yokota (1994a: 370) und basiert auf dem Grundprinzip dieses Verfahrens.

[147] Nach Meinungen der heutigen Schriftlinguisten hat Xu die Beispielzeichen unkorrekt benannt, so habe ich die originalen Beispielzeichen <令> (lìng, *befehlen*) und <长> (/cháng/: *lang*) durch <我> sowie <足> ersetzt.

Vor der kleinen Siegelschrift gab es wegen des begrenzten Zeicheninventars häufiger Entlehnungszeichen. Da diese Methode für die morphemische Funktion der chinesischen Schrift ungeeignet ist, wurden bei dem Entwurf der kleinen Siegelschrift viele neue Determinativphonetika erfunden, um solche Entlehnungszeichen von den originalen Begriffen graphisch zu unterscheiden (vgl. Zhang YJ 1991: 97). Das Zeichen <自> (/zì/, Altform: ) ist ein Piktogramm für *Nase* und wurde für *selbst* abgeleitet. *Nase* wurde später durch das DP <鼻> /bí/ ersetzt, das aus dem Radikal <自> und der Phonetik-hinweisenden Komponente <畀> (/bì/, *geben* [selten gebraucht]) zusammengesetzt ist.

Trotz der Ähnlichkeit sind die Entlehnungszeichen im Grunde unterschiedlich zu Silbenzeichen. Entlehnungszeichen können nicht die Aussprache wiedergeben, sondern übertragen wegen ihres festgelegten Gebrauchs und der Orthographie das entlehnte Morphem. Sie sind deswegen eher vom morphologischen Schrifttyp, als vom phonologischen.

6) Zhuanzhu (转注 /zhuǎnzhù/)[148]

„Zhuanzhu heißt ein Zeichenpaar, das mit demselben Klassenhaupt gebildet wird und sich zur Begriffserklärung untereinander bedienen kann, wie 老 [lǎo, *alter Mann*] und 考 [kǎo, *alter Mann*/*Prüfung*]“ (Xu S 100: Vorwort [Übersetzung der Verfasserin]). Anders als die Definitionen der anderen fünf Schrifttypen wurde Zhuanzhu von Xu Shen nicht eindeutig genug erläutert, so dass dessen Bedeutung bis heute umstritten ist. Aus diesem Grund fallen die deutschen Übersetzungen sehr unterschiedlich aus.

Das am meisten anerkannte Argument lautet, ein Zeichenpaar mit gleichem Radikal, ähnlicher Aussprache (optionale Bedingung) und derselben Bedeutung sei füreinander das ‚Synonymzeichen‘. <老> (die Altform: ) war das Piktogramm eines alten Mannes und <考> war ein Determinativphonetikum mit dem von <老> abgeleiteten Radikal (vgl. Sun 1988: 65f). Dieses Konstruktionsprinzip war vor allem wegen des phonetischen Wandels entstanden. Die Aussprache eines Wortes oder eines Morphems entwickelte sich oder variierte bei Dialekten, aus diesem Grund wurde ein Synonymzeichen auf der Basis vom originalen Zeichen erschaffen, um phonetische Abweichungen zu markieren. Aufgrund dieser Theorie können viele weitere Synonyme genannt werden. <豬> (/zhū/, vereinfachte Form: <猪>, heutige Bezeichnung für *Schwein*) und <豨> (/xī/, veraltete Bezeichnung *Schwein*) waren beide auf Basis des Piktogramms für Schwein <豕> (/shǐ/, veraltete Bezeichnung *Schwein*, Altform: ) gebildete

[148] 转 /zhuǎn/ bedeutet *Wandel, Änderung* und 注 /zhù/ *Markierung, Erklärung*.

DP, die wegen verschiedener dialektaler Aussprachen erfunden wurden. Die originalen Zeichen können entweder Piktogramme oder DP sein, aber die abgeleiteten Zeichen sind fast immer DP. Der wesentliche Unterschied zur Entlehnung wird daran ersichtlich, ob die vorhandenen Zeichen direkt ohne oder mit graphischen Änderungen verwendet werden (ibid.).

Für Zhuanzhu gibt es eine Erläuterung aus gänzlich anderer Perspektive. Dieser Definition zur Folge handelt es sich um „die Verwendung eines und desselben Zeichens in einem von der ursprünglichen Bedeutung abgeleiteten Sinn, wobei fast immer ein gewisser Wandel der Aussprache mit einhergeht“ (Müller-Yokota 1994a: 369). Nach dieser Erklärung könnte Zhuanzhu im Deutschen treffender als ‚Bedeutungswandel‘ übersetzt werden. In diesem Fall können viele Heteronyme mit zwei vergleichbaren, aber verschiedenen Bedeutungen Beispielzeichen sein. Das Zeichen <樂> (Altform:  , heute vereinfacht in <乐>, /yuè/ – *die Musik*; /lè/ – *froh*) war/ist ein Symbol *Musik*. Da Musik den Menschen *Freude* machen kann, wird das Zeichen für die Bedeutung *froh* mit Lautwandel abgeleitet (vgl. ibid.: 369f).

Die zwei Erklärungen für Zhuanzhu drücken zwei verschiedene Verwendungsmethoden der chinesischen Schrift aus. Nach meinem Verständnis ist zwar das erste Argument kompatibler zu Xu Shens Definition. Aber das zweite böte eine hochwertige Ergänzung der *sechs Schriften*, die ebenso ein entscheidendes Prinzip der chinesischen Schrift sind. In der folgenden Tabelle habe ich Zeichen verschiedener Typen als Beispiele herangezogen, von denen die meisten mit dem Zeichen <木> (*Baum*) verwandt sind.

Zeichen	**Typ**	**Pinyin**	**Bedeutung**	**Prinzip der Zeichenstruktur**
木	PG	mù	*Baum*/*Holz*	Altform:  , das Bild von einem Baum mit Ästen und Wurzel
本	IG	běn	*Wurzel*/*Grund*	Altform:  , das Bild von einem Baum mit Betonung der Wurzel
果	PG	guǒ	*Frucht*/*Obst*	Altform:  , das Bild eines Baums mit Früchten
休	ZI	xiū	*ausruhen*	Zusammensetzung von einem Menschen und einem Baum, symbolisiert das Anlehnen und Ausruhen an einem Baum
林	ZI	lín	*Wald*	Zusammensetzung von zwei Bäumen, symbolisiert eine Gruppe von vielen Bäumen
材	DP	cái	*Material*	Determinativ links und Phonetikum <才> (/cái/, die Fähigkeit) rechts
楹	DP & Syn	yíng	*Säule*	Synonym von <柱> /zhù/ (das Phonetikum <主> /zhǔ/, *Besitzer*), Determinativ links und Phonetikum <盈> (/yíng/, *ausfüllen*) rechts
行	PG & BW	háng	*Weg*/*Zeile*/*Bank* etc.	Altform:  , Bild von einer Straße, wurde mit Aussprachewandel auch für den verwandten Begriff *gehen* übernommen
		xíng	*gehen*/*okay* etc.	

Zeichen	Typ	Pinyin	Bedeutung	Prinzip der Zeichenstruktur
來/来	PG & EL	lái	(Original: *Weizen*) *kommen*	Altform: , Bild einer Ähre, wurde für den Begriff *kommen* entlehnt, wobei die originale Bedeutung verloren ging und durch das Zeichen <麥/麦> /mài/ ersetzt wurde
架	DP	jià	*Gerüst/Regal*	Determinativ unten und Phonetikum <加> (/jiā/, *plus*) oben
困	ZI	kùn	*Not*	Bild von einem von vier Wänden umkreisten, allein stehenden Baum; symbolisiert schwierige Situationen ohne Unterstützung

Tab. 3-2: Beispielzeichen für sechs Schriften und ihre Konstruktionsprinzipien[149]

3.3.3 Strich, Strichmerkmal, -ordnung und -zahl

Im letzten Kapitel wurde vorgestellt, wie die quadratische Zeichenform für einen bestimmten Sinninhalt anhand der sechs Schriften bestimmt wird. In diesem und den nächsten zwei Kapiteln wird weiter analysiert, wie sie graphisch aufgebaut wird. Konkreter ist der Zeichenaufbau auf drei Grundelemente bezogen: Strich, Komponente und die geometrische Konstellation der Komponenten (vgl. GB/T 12200.2-94: Kap. 4.1.2.21).

Das wesentliche, unzerlegbare graphische Baumaterial der Schriftzeichen ist der Strich. Strich bedeutet dabei – als Beispiel das Handschreiben in Regelschrift anlegend – „das bewegungssensitive Führen der Stiftspitze von ihrer Niederlegung bis zum Absetzen“ (Peng 1994: 34 [Übersetzung der Verfasserin]). Die Form eines Strichs heißt Strichmerkmal (笔形 /bǐxíng/, eng.: stroke feature; vgl. GF 3001-1997: Kap. 3.2). Es gibt acht Grund- und ca. 30 Kombinationsstrichmerkmale. Solche Grundstriche der modernen chinesischen Schrift sind <一> (Bezeichnung: 横 Héng, Bedeutung: *Horizontal*), <丨> (竖 Shù, *Vertikal*), <丿> (撇 Piě, *Links-Senkung*), <𠂉> (提 Tí, *Steigung*), <丶> (点 Diǎn, *Punkt*), <㇏> (捺 Nà, *Rechts-Senkung*) und zwei Merkmale, die nur mit anderen Strichen verbunden sein können: Zhé (折, *Drehung*) und Gōu (钩, *Haken*), z.B. in dem Kompositionsstrich <㇆> (横折钩, Héng-zhé-gōu) (vgl. Coulmas 1996a: 80). Standardgemäß gibt es insgesamt 37 Arten der chinesischen Striche, die im Unicode zwischen 31C0 bis 31E3 festgelegt werden (vgl. Unicode 12.0 Character: U+31C0-U+31EF). Fünf von den acht Grundstrichen werden als Hauptstrichmerkmale bezeichnet, unter denen die anderen Strichmerkmale untergeordnet werden können. Nach der Reihenfolge der Zeichenrecherche sind diese Héng, Shù, Piě, Diǎn und Zhé. Tí ist beispielsweise ein Nebenstrich unter Héng und Nà unter Diǎn. Die chinesischen Striche sind mit dem Sinninhalt und der Aussprache eines Zeichens in keiner Weise verquickt. Zwischen zwei Stri-

[149] vgl. Xu 100 & Shuowen-Jiezi Bianwei Hui 2012; im Fall der orthographischen Unterscheidung eines Zeichens wird es in der Lang- und Kurzform nachfolgend angegeben; Abkürzungen: PG – Piktogramm, IG – Ideogramm, ZI – Zusammengesetztes Ideogramm, DP – Determinativphonetikum, EL – Entlehnung, Syn – Synonym, BW – Bedeutungswandel.

chen können drei Beziehungsverhältnisse bestehen: Separation (wie die zwei Striche im Zeichen <八> /bā/, *acht*), Verbindung (wie im <人> /rén/, der *Mensch*) und Kreuzung (wie im <十> /shí/, *zehn*) (vgl. Lu 2008: 16ff, Sun 1988: 282).

Die Reihenfolge der Striche beim Zeichenschreiben heißt Strichordnung (笔顺 /bǐshùn/, eng.: stroke order; vgl. GF 3001-1997: Kap. 3.3). Sie ist bei jedem Schriftzeichen einzigartig, aber von bestimmten Prinzipien beeinflusst. Die wichtigsten Prinzipien lauten (hierarchisch geordnet): Héng, Shù, Piě, Nà – von oben nach unten, von links nach rechts und von außen nach innen. Normalerweise ist die Strichordnung eines Zeichens parallel auf mehrere Prinzipien bezogen (vgl. Sun 1988: 286f). Die Anzahl der Striche für die Bildung eines Zeichens oder einer Komponente heißt Strichzahl (笔数 /bǐshù/, eng.: stroke count; vgl.: GF 3001-1997: Kap. 3.4). Sie ist eine der wichtigsten Grundlagen des Recherchierens und der Anordnung der Sinogramme im Lexikon. Ein Beispiel soll dies verdeutlichen helfen: Das Grundzeichen <王> (/wáng/, *König*) setzt sich aus dreimal Héng und einmal Shù zusammen. Seine Strichordnung lautet: Héng oben - Héng mitte - Shù – und zuletzt Héng unten. Seine Strichzahl beträgt dementsprechend vier. Das exemplarische Radikal, welches in diesem Fall semantisch eine Gruppe von auf *Edelsteine* bezogener Begriffe taxiert[150], wird deswegen in dem Index der Zeichenlexika zu den vierstrichigen Radikalen geordnet. Das Zeichen <王> ist sodann erster Strichzeicheneintrag des Indexes, wohingegen z.B. das Zeichen <理> (/lǐ/, *Prinzip*) nach dieser Kategorie aus dem identischen Radikal und sieben weiteren Strichen zusammengesetzt wird. Es ist somit der Oberklasse <王> untergeordnet und befindet sich aufgrund seiner Strichzahl bei Zeichen, die ebenfalls zusätzlichen aus sieben Strichen und dem Radikal bestehen (‚王 +7'-Kategorie).

Die Strichzahl eines chinesischen Zeichens kann zwischen eins bis über dreißig liegen. Das Zeichen mit den meisten Strichen aus dem großen Xihua-Lexikon hat insgesamt 68 Striche. Der Durchschnitt liegt bei elf bis achtzehn (vgl. Peng 1994: 34). Aus diesem Grund wäre die Inputcodierung nach Strichen für Textverarbeitungen uneffektiv, obwohl diese Eingabemethode relativ einfach entworfen und erworben werden könnte. Eine Strichinputcodierung basiert entweder auf den acht Grund- oder den fünf Hauptstrichen. Für Handys wird die strichbasierte Eingabemethode wegen der Mini-Tastatur mit nur neun zeichentragenden Tasten beliebter.

[150] Als Radikal sind die Zeichen <王> und *Jade* <玉> (/yù/, Jade) graphisch identisch.

3.3.4 Komponenten der Sinogramme

Die alleinstehenden simplen Schriftzeichen, die graphisch direkt aus Strichen aufgebaut sind, betragen weniger als 10% des Gesamtinventars. D.h. die meisten Sinogramme sind aus zwei oder mehreren Komponenten in bestimmter Konstellation zusammengesetzte komplexe Zeichen. Ein solches Zeichen, das entweder nach ZI oder DP gebildet wird, kann auf Basis seines Kontexts oder/und seiner referenziellen Quelle zum Zweck der Zeichenstruktur mit Hilfe seiner Komponenten entschlüsselt werden. Jedes komplexes Zeichen hat deswegen ein ‚Strukturoriginal' (结构理据 /jiégoù lǐjù/).[151] Die Zeichenzerlegung in Komponenten kann als ‚Komponentenabbau' (部件拆分 /bùjiàn chāifēn/) bezeichnet werden.[152]

Das nach dem DP-Prinzip gebaute Zeichen <理> /lǐ/ bspw. kann zuerst ins Radikal <王> (/wáng/, für mit *König*/*Jade* bezogene Begriffe) links und der phonetischen Komponente <里> /lǐ/ rechts geteilt werden. Die originale Bedeutung ist übertragen in etwa *Bearbeitung der Jade* und wurde zu dem Begriff *Prinzip* erweitert. Das Phonetikum <里> ist selbst zerlegbar in <田> (/tián/, *Feld*) und <土> (/tǔ/, *Erde*), dessen originale Bedeutung die Messeinheit für die Länge ([heute] 1 Lǐ = ca. 0,5 km) wiedergibt. Eine weiter zerlegbare Komponente heißt ‚Kompositionskomponente' (合成部件 /héchéng bùjiàn/) wie <里>. Im Gegensatz dazu stehen unzerlegbare ‚Grundkomponenten' (基础部件 /jīchǔ bùjiàn/, eng.: basic component), wie <王>, <田> und <土> (vgl. GF 3001-1997: Kap. 3.8).[153] Grundkomponenten können auch als Grammwurzel (字根/zìgēn/), Grammelement (字素 /zìsù/), Grammbasis (字元 /zìyuán/) oder Grundsymbol (基本符号 /jīběn zìfú/) bezeichnet werden, deren Inventar meistens die Grundlage einer zeichenformbasierten Inputcodierung darstellt (vgl. Peng 1994: 34).

Eine Grundkomponente kann sowohl ein simples Zeichen als auch ein von keinem Zeichen gestaltetes Symbol sein. Die letzteren umfassen graphische Symbole (einzelne Striche oder Strichfolgen, wie <丶>, <丿>, <乚>, <𠂇> etc.) und die von Zeichen abgeleiteten Symbole, die auf Basis der originalen Form graphisch umgeformt wurden (wie <氵> von *Wasser* <水> /shuǐ/ und <刂> von *Messer* <刀> /dāo/) (vgl. Sun 1988: 289ff).

Wenn die Zeichenzerlegung auf dem Strukturoriginal basiert, heißt sie ‚originaler Abbau' (有理拆分 /yǒulǐ chāifēn/; eng.: original disassembly). Es gibt jedoch auch viele Zeichen, de-

[151] Wörtlich übersetzt heißt es *der Grund und die Basierung der Struktur*; die hier verwendete deutsche Übersetzung basiert auf dem Englischen *structure origin* (vgl. GF 3001-1997: Kap. 3.10).
[152] Diese Übersetzung basiert auf dem Englischen *component disassembly* (ibid.: Kap. 3.11).
[153] Diese Übersetzungen basieren auf den englischen Äquivalenten *compound component* und *basic component* (vgl.: GF 3001-1997).

ren Strukturoriginal nicht analysiert werden kann oder Konflikte mit der Zeichenform aufwirft. Wenn die Zerlegung nach dem graphischen Abbau erfolgt, wird der sog. ‚unoriginale Abbau' (无理拆分 /wúlǐ chāifēn/; eng: unoriginal disassembly) durchgeführt (vgl. GF 3001-1997: Kap. 3.12). Als Beispiel taugt das Zeichen <旗> (/qí/, *Flagge*). Das ursprüngliche Determinativ <方𠂉> symbolisierte als Piktogramm eine Fahnenstange mit flatternder Flagge, die im Laufe der Schriftentwicklung heute graphisch getrennt geschrieben wird. So wird das Zeichenstrukturoriginal in das Radikal <方𠂉> und das Phonetikum <其> (/qí/, ein Pronomen) segmentiert. Im Gegenteil dazu steht die graphische Realisierung des Zeichens, das sich zuerst in einen linken und rechten Teil zerlegen lässt, wobei der rechte Teil in einem weiteren Schritt in ein graphisches und phonologisches Symbol segmentiert werden kann. Ein weiteres Beispielzeichen 衷 (/zhōng/, *vom inneren*) wird aus dem Determinativ <衣> (/yī/, *Wäsche*) und dem Phonetikum <中> (/zhōng/, *Mitte*) kombiniert. Aber <衣> wird graphisch getrennt geschrieben und umgeht das Phonetikum. Die Grundkomponenten sind aus prinzipiellen und graphischen Ansichten in solchen Schriftzeichen nicht übereinstimmend. Für linguistische und pädagogische Zwecke muss das Strukturoriginal berücksichtigt werden, damit die Zeichen besser zu verstehen, erwerben und analysieren sind. Beim Entwurf der Inputcodierung ist eine Zerlegung nach graphischen Aspekten vorzuschlagen, da die meisten PC-Benutzer keine Schriftlinguisten sind, die die Konstruktionsprinzipien der chinesischen Schrift gut genug beherrschen.

Wie viele Komponenten es in der chinesischen Schrift insgesamt gibt, kann noch nicht festgelegt werden. Nach Angaben verschiedener Experten variiert die Anzahl der Komponenten erheblich zwischen ca. 100 bis über 600. Die großen Unterschiede sind vor allem vom Umfang der erforschten Schriftzeichen und den verschiedenen Kriterien bedingt. Bezüglich des Umfangs ist zuerst die Auswahl des konkreten Forschungsobjekts entscheidend. Diese sind unterschiedlich, können einmal die vereinfachten, die traditionellen oder die Gesamtheit beider Schriftzeichen fokussieren, genauso jedoch auch die generellen Schriftzeichen oder ein breites Spektrum an Sinogrammen. In der Forschungsmethode geht es auch darum, welche Komponenten bei der Analyse gezählt werden. Im Komponentenstandard von GB 13000.1 werden insgesamt 560 Grundkomponenten aufgelistet, anhand dessen die 20.902 Schriftzeichen (gleich wie die Grundkategorie der CJK-Ideogramme von Unicode) in Komponenten segmentiert und analysiert werden können (vgl. Sun 1988: 292, GF 3001-1997: Kap. 4.1).

In der folgenden Graphik (Abb. 3-2 & 3-3) wird das chinesische Wort für *Deutschland* <德國> /déguó/ als Beispiel für Komponentenabbau, Aufbaustriche und Strichordnung ge-

nommen – 德 (*Moral*) wird phonetisch von der ersten Silbe des Wortes Deutsch abgeleitet, 國 (heute in der VR China zu <国> vereinfacht) ist das Zeichen für *Staat*.

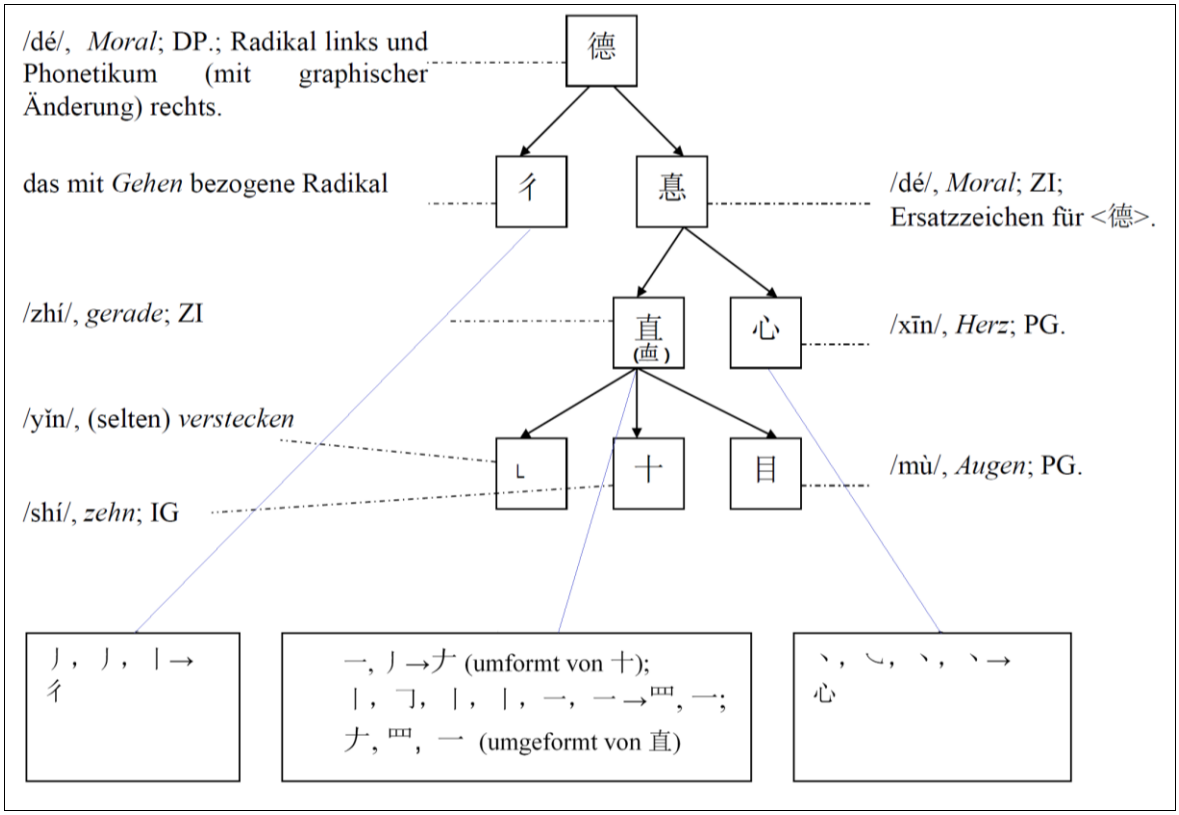


Abb. 3-2: Segmentierung des Beispielzeichens <德> /dé/ in Komponenten

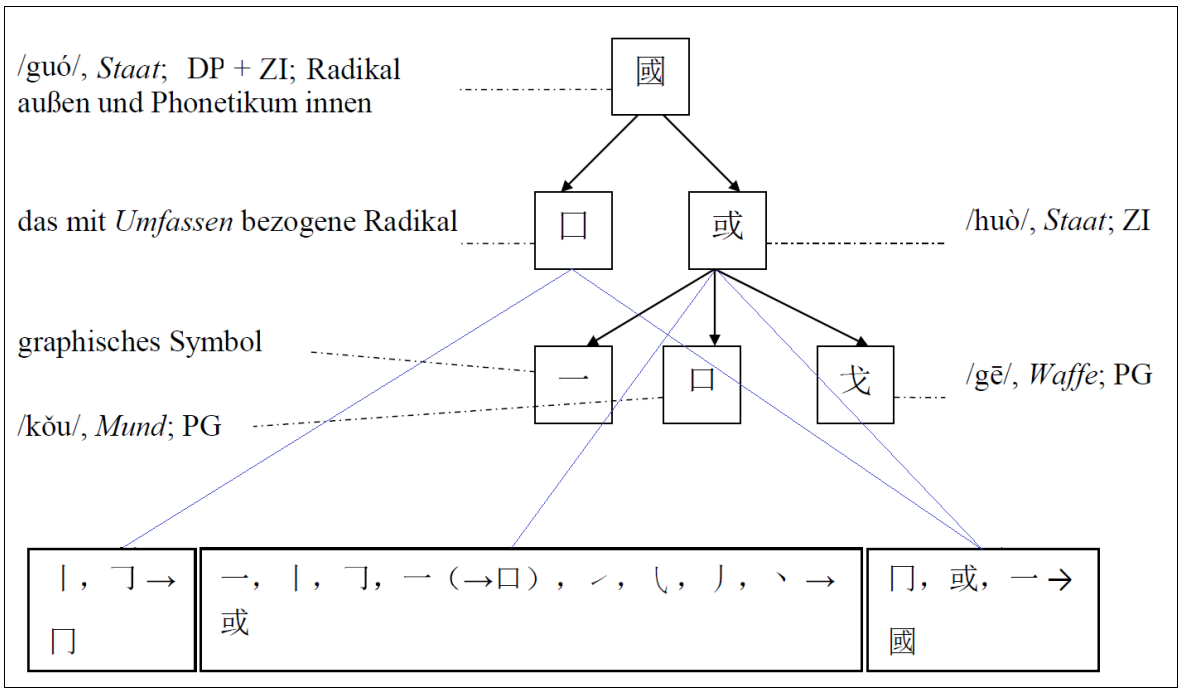


Abb. 3-3: Segmentierung des Beispielzeichens <國> /guó/ in Komponenten

Unter den beiden Zeichen gibt es in <德> Konkurrenzen zwischen den graphischen Bausteinen und dem Strukturoriginal. Der rechte obere Teil, der ursprünglich <直> /zhí/ hieß, wurde

graphisch bei der Entwicklung umgeformt. <直> ist zudem graphisch intransparent aufgebaut. Bei beiden Zeichen muss deswegen in vielen Fällen ein replizierter Abbau durchgeführt werden, statt das Strukturoriginal zu behalten. Im Bereich der Informationsverarbeitung bspw. muss die graphische Segmentierung ausgewählt werden, wenn ein Schriftzeichen graphisch mit dem Strukturoriginal nicht übereinstimmend ist.

Aus linguistischen und pädagogischen Ansichten muss die Zeichensegmentierung auf den authentischen Bausteinen und -prinzipien bezogen bleiben. Die Komponenten auf verschiedenen Ebenen zur Zeichenstruktur sind zu berücksichtigen. Für den Informatikbereich muss die graphische Zerlegung leicht verständlich und sichtbar sein, damit auch PC-Benutzer mit wenigen Kenntnissen der Konstruktionsprinzipien die Inputcodierung anwenden können. Der Entwurf der Inputcodierung nach Zeichenaufbau muss auch vor dem Hintergrund der begrenzten Tastenanzahl und der Schwierigkeit des Auswendiglernens beachtet werden (vgl. Fei 1996: 443f). Die beliebtesten zeichenformbasierten Eingabemethoden sind die Wubi- in Festlandchina und die Cangjie-Eingabemethode in Taiwan und Hongkong.

Im Wubi-Schema der dritten Generation (auch Wang-Schema genannt) sind z.B. insgesamt 226 Grammwurzeln verzeichnet, die auf 25 buchstabenübertragenden Tasten verteilt werden.[154] Solche Grammwurzeln enthalten häufige Grundkomponenten, einzelne Striche und Strichfolgen (vgl. Wubi-Zixing-2000+). <德> lautet nach Wubi-Schema TFLN (T für 彳, F – 犬, L – 罒 und N – 心). Im Cangjie-Schema hingegen werden 24 Haupt- und zwei spezielle Elemente sowie von den Hauptelementen verformte oder abgeleitete Nebenelemente, die insgesamt 140 Zeichenelemente betragen, zur Inputcodierung angewendet (vgl. Cangjie-5). Dasselbe Beispielzeichen wird nach der Cangjie-Eingabemethode mit dem Code HOJWP (H: 竹 → 丿, O: 人 → 亻, J: 十 → 犬, W: 田 → 罒 und P: 心) definiert. Viele Grundkomponenten müssen bei den beiden Schemata als Zusammensetzung von zwei oder mehreren Elementen ausgedrückt werden, die aber mit dem Strukturoriginal konkurrierend sind. Das Zeichen und Radikal <王> etwa muss im Cangjie-Schema in die Grammwurzeln <一> und <土> (/tǔ/, *Erde*) zerlegt werden. Das von dem Piktogramm abstammende Zeichen <自> muss in Wubi auch anhand der Symbole <丿> und <目> (/mù/, *Auge*) kombiniert werden (vgl. Wubi-Zixing-2000+, Cangjie-5). Zusammengefasst könnte die linguistische Segmentierung unpraktisch für die Eingabe sein. Eine praxistaugliche Inputcodierung (wie Wubi oder Cangjie) aber muss in vielen Fällen unlogisch bezüglich der wichtigsten Theorien des Zeichenaufbaus sein.

[154] Die Taste #46 (‚Z' im internationalen Tastaturlayout) kann universal alle Grammwurzeln ersetzen, falls ein Grammwurzelcode vergessen wird. Vgl. auch Kapitel 4.1.2.

Außerdem gibt es bei allen Eingabeschemata nach dem Zeichenaufbau keine bestimmten Regeln, welche Komponente für welche Taste zuständig ist. Die Benutzung einer solchen Eingabemethode erfordert deswegen spezielle Qualifikation und monotones Auswendiglernen.

3.3.5 Geometrische Konstellationen

Striche und Komponenten können bei der Zeichenformbildung verschiedene Beziehungen zueinander aufweisen. Im Vergleich zu den drei Arten der Konstellationen der Striche, nämlich Auseinandersetzung, Verbindung und Kreuzung, gibt es mehr Varianten für die Beziehungen der Komponenten. Zusammengefasst können alle Varianten insgesamt vier Kategorien zugeordnet werden: alleinstehender, Links-Rechts-, Oben-Unten- und Außen-Innen-Aufbau. Außer bei wenigen Grundzeichen des alleinstehenden Aufbaus finden sich in den meisten Sinogrammen eine oder mehrere geometrische Konstellation. Nach statistischen Analysen von zehntausend Sinogrammen betragen die alleinstehenden Schriftzeichen 10% der Gesamtmenge. Die Mehrheit bilden Zeichen des Links-Rechts-Aufbaus mit 60%; auf die übrigen Teile entfallen stets 30% (vgl. Peng 1994: 36f). Der Außen-Innen-Aufbau hat die meisten Varianten: vierseitige, dreiseitige und halbe Einschließungen.

Die Konstellation der Komponenten spielt in manchen Fällen auch eine bedeutungsunterscheidende Rolle. Viele Zeichenpaare, die sich grundsätzlich aus denselben Komponenten zusammensetzen, sind bei Konstellation voneinander abweichend, wie z.B. <吟> (/yín/, *singen*) und <含> (/hán/, *enthalten*), deren Bestandteile gleich <口> (/kǒu/, *Mund* als Radikal) und <今> (/jīn/, *jetzt* als Phonetikum) sind. Dieselben Komponenten <忄/心> (/xīn/, *Herz*) und <亡> (wáng, *sterben*) können ebenso Zeichen für *beschäftigt* (忙 /máng/ in Links-Rechts-Aufbau) und *vergessen* (忘 /wàng/, in Oben-Unten-Aufbau) bilden. In vielen Zeichen existieren hierarchisch zwei oder mehrere Konstellationen. Das Zeichen von Abb. 3-2 als Beispiel nehmend, ist <德> primär links-rechts-aufgebaut. In seiner Kompositionskomponente rechts herrscht ein sekundärer Oben-Unten-Aufbau. Abbildung 3-4 zeigt Symbole und Beispielzeichen der verschiedenen Konstellationen (vgl. ibid.).

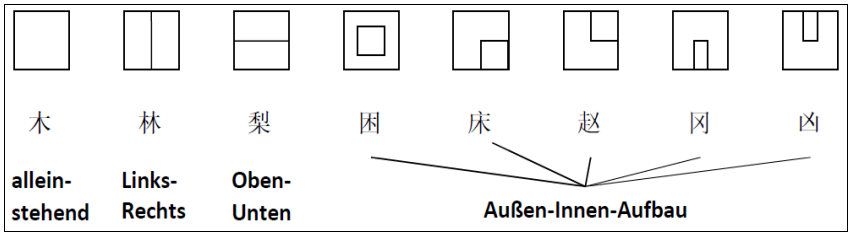


Abb. 3-4: Die Varianten der proportionalen Konstellation der chinesischen Schrift

3.4 Einführung in Inputcodierungsentwurf und Phonetik des Chinesischen

In Kap. 3.3 wurden Grundzeichenattribute der Zeichenform vorgestellt. Wie dargelegt (siehe Abb. 1-5, S. 34) sind für jedes chinesische Schriftzeichen Zeichenform, Aussprache und Sinninhalt unentbehrliche Elemente. Das Beispielzeichen in Abb. 3-2 bspw. wird phonetisch als /dé/ definiert, tritt schriftlich als ‚德' auf und wird durch das Konzept *Moral* repräsentiert. Die Vereinheitlichung der drei Elemente in einem Zeichen hat sowohl arbiträre als auch rationale Eigenschaften. Die arbiträren Attribute verweisen auf de Saussures Zeichentheorie. Die rationalen Attribute beziehen sich hauptsächlich auf die Zeichenkonstruktionstheorien, konkret die ‚sechs Schriften' (vgl. hierzu Kap. 3.3.2, S. 148-155). Diese Regularität bedingt, dass der Sinninhalt eines Schriftzeichens in den meisten Fällen durch die Segmentierung in Komponenten analysiert werden kann, wenn es nach dem Strukturoriginal aufgebaut ist. In Kap. 3.3.4 wurden linguistische Grundlagen für die auf Zeichenform basierten Inputcodierungen vorgestellt. Ziel von Kap. 3.4 ist es nun, linguistische Wissensgrundlagen für die Inputcodierung abzubilden, die auf den drei Grundzeichenattributen basieren müssen.

3.4.1 Allgemeine Zusammenhänge zwischen Zeichenform, -sinninhalt und -aussprache

Wie auf S. 160 vorgestellt wurde, basieren die Wubi- und Cangjie-Inputcodierung auf Zerlegung in Komponenten, die hauptsächlich auf die Dimension der Zeichenform fokussiert sind. Wegen der drei-attributiven Eigenschaft der chinesischen Schrift kann es zahlreiche verschiedene Inputcodierungen geben, aber keine kann vollständig alle Zeichenattribute repräsentieren. Die Beziehungen der drei Grundattribute werden in Abb. 3-5 beschrieben.

Eine Inputcodierung basiert auf einer oder zwei Grundattributen. Da es zwischen zwei Elementen häufig keine Eins-zu-eins-Entsprechung gibt, betrifft ein Inputcode häufig Kandidatenüberschneidungen. Die linguistischen Aspekte, die Ambiguitäten zwischen zwei Elementen verursachen können, sind hellgrau markiert. Die Zusammenhänge miteinander werden nachfolgend von der Abbildung in drei Punkten erklärt.

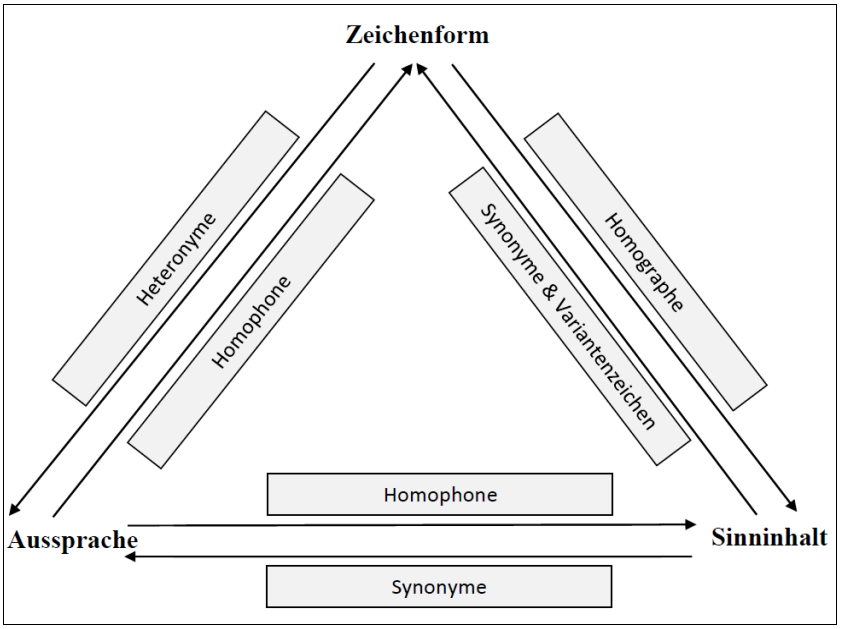


Abb. 3-5: Die Dreieckbeziehungen zwischen Zeichenform, -aussprache und -sinninhalt eines chinesischen Schriftzeichens[155]

1. Die Aussprache-Form-Beziehungen

Die Aussprache variiert sowohl diachron als auch diatopisch, deswegen müssen die Aussprache-Form-Beziehungen meistens nur innerhalb eines Sprach-/Schriftsystems erforscht werden. Wenn eine Silbe mehreren verschiedenen Schriftzeichen entspricht, sind solche Zeichen füreinander Homophone. Der Reichtum an Homophonen ist ein generelles Phänomen im Chinesischen, Japanischen und Koreanischen. Die Homophone der beiden letztgenannten Sprachen werden in Kap. 3.5.1 vorgestellt. Mit der Homophoniewahrscheinlichkeit des modernen Chinesischen beschäftigt sich das Anschlusskapitel.

Im modernen Standardchinesisch hat ein Schriftzeichen normalerweise einen einzigen phonetischen Wert, es gibt jedoch Ausnahmen. Solche mit mehreren Lesarten realisierbaren Zeichen heißen Heteronyme, wie etwa das Zeichen <乐> (ausgesprochen als /lè/ *froh* und als /yuè/ *Musik*). Ein Heteronym kann sowohl verschiedene als auch identische Bedeutung(-en) übertragen. Im Japanischen sind Heteronyme noch häufiger.

2. Die Form-Sinninhalt-Beziehungen

Im Vergleich zu den Aussprache-Form-Beziehungen zeigen die Form-Sinninhalt-Beziehungen übersprachliche Eigenschaften. Wie in Kap. 3.3.2 vorgestellt wurde, versuchte man bei der Schrifterfindung die Zeichenform nach dem Konzept darzustellen oder aufzubauen. Zeichenform und -sinninhalt hängen heutzutage trotz der graphischen sowie sprachlichen Entwick-

[155] Die Darstellungsform orientiert sich vage am semiotischen Dreieck von Ogden/Richards (vgl. z.B. Busch/Stenschke 2008: 28), passt sich dabei aber den linguistischen Gegebenheiten der chinesischen Schriftzeichen an.

lung immer noch eng zusammen. Die Stabilität der Form-Sinninhalt-Beziehungen ist der wesentliche Grund, warum die schriftliche Kommunikation auch ohne fremdsprachliche Kenntnisse in manchen Fällen innerhalb der CJK-Gebiete möglich ist.

Die allgemeinen Entwicklungsphasen und die Zeichenorthographie wurden in Kap. 3.1.3 (S. 137ff) und Kap. 3.2.1 (S. 140ff) vorgestellt. Seit ca. 800 n. Chr. bis zum Ende des Zweiten Weltkriegs veränderten sich die meisten (in Regelschrift geschriebenen) Zeichenformen kaum. Auch nach der Vereinfachung in Festlandchina sowie Japan haben die Sinogramme in den meisten Fällen graphisch große Ähnlichkeiten mit ihrer traditionellen Form und hohe Übereinstimmungen miteinander. Dies ist der Hauptgrund, warum Hanzi, Kanji und Hanja im selben Block von Unicode definiert werden können (vgl. Kap. 3.2.3, S. 145f).

Für eine Zeichenform kann es mehrere Homographe geben, die verschiedene Sinninhalte und manchmal auch unterschiedliche Aussprachen haben. Die Entwicklung des Zeichensinninhalts ist zeitlich, regional und sprachlich abhängig. Einerseits wurden immer neue erweiterte und abgeleitete Bedeutungen auf Basis der ursprünglichen Bedeutung hinzugefügt. Andererseits kann ein Zeichen durch Wortentlehnung irrelevante Bedeutungen aufnehmen. Die Grundbedeutung eines Sinogramms ist allgemein aber in verschiedenen Zeiten, Regionen und Schriftsystemen stabil. Anders formuliert kann ein Schriftzeichen viele verschiedene relevanten oder irrelevanten Bedeutungen übertragen. Bspw. hat das Piktogramm <水> /shuǐ/ im Xinhua-Lexikon Online sieben Bedeutungen. Außer *Wasser* kann es auch *Fluss*, *Gewässer*, *Flüssigkeit*, *zusätzliches Einkommen*, *Waschgang* und einen Nachnamen ausdrücken (nach Xinhua-Lexikon Online). Homographische Morpheme sind Morphempaare, die trotz derselben Zeichenform unterschiedliche Bedeutungen tragen oder phonetisch ausgesprochen werden, wie dasselbe Zeichen in den Wörtern <渭水> (/wèishuǐ/; *Wei-Fluss*) und <水银> (/shuǐyín/; *Quecksilber*). Der Zeichensinninhalt ist alleinstehend so meist unpräzise und ambig und muss in bestimmten sprachlichen Situationen interpretiert werden.

Ein Zeichenpaar, das unterschiedliche Zeichenformen bei identischen Bedeutungen hat, ist füreinander synonym. Dazu gehören vor allem Zhuanzhu- und Variantenzeichen. Die erste Alternative ist eine ‚sechs Schriften'-Methode und bezieht sich auf das Zeichenpaar mit unterschiedlicher Aussprache. Die zweite Alternative stellt im Grunde genommen eine variierte Zeichenform dar, wie z.B. sind <國> und <国> (/guó/, *Land*) im Prinzip dasselbe Zeichen.

3. Aussprache-Sinninhalt-Beziehungen

Wie bereits dargelegt, bezieht sich eine Silbe auf mehrere homophonetische Zeichen, während sich ein Zeichen wiederum auf mehrere Sinninhalte beziehen kann. Aussprache und Sinnin-

halt stehen deswegen meistens im Verhältnis von eins zu vielen. Umgekehrt hat ein Sinninhalt normalerweise mehrere Synonyme, die sich phonetisch und graphisch unterscheiden.

In den drei Sprachen Chinesisch, Japanisch und Koreanisch ist die Aussprache das instabilste und variantenreichste Element. Zeichenform und -sinninhalt haben im CJK-Schriftkreis demhingegen viel höhere Übereinstimmungen. Nachstehend werden sieben Schriftzeichen als Beispiel für verschiedene phonetische Realisierungen desselben Zeichens angegeben.

Zeichen-form	**Phonetik**						**Sinninhalt**
	Chinesisch		**Japanisch**		**Koreanisch**		
	Pinyin	MPS	Romanji	Hiragana	RRK	Hangul	
天	tiān	ㄊㄧㄢ	ten, ame	てん, あめ	cheon	천	*Himmel*
地	dì	ㄉㄧˋ	ti, tuti	ち, つち	ci	지	*Erde*
人	rén	ㄖㄣˊ	zin, hito	じん, ひと	in	인	*Mensch*
山	shān	ㄕㄢ	san, yama	さん, やま	san	산	*Berg*
水	shuǐ	ㄕㄨㄟˇ	sui, mizu	すい, みず	su	수	*Wasser*
上	shàng	ㄕㄤˋ	zyou, ue	じょう, うえ	sang	상	*oben*
下	xià	ㄒㄧㄚˋ	ka, sita	か, した	ha	하	*unten*

Tab. 3-3: Beispiele für gemeinsame Zeichen in Chinesisch, Japanisch und Koreanisch (Unicode 12.0 Chapters: Kap. 18.1: 708)[156]

Durch die Beispielzeichen wird ersichtlich, warum eine auf Phonetik basierende Inputcodierung nur für eine bestimmte Sprache gültig sein kann, aber eine zeichenformbasierte Inputcodierung im ganzen chinesischen Schriftkreis möglich ist. Wegen großer Herausforderungen beim Erwerb, der nicht standardisierten Zeichenzerlegung usw. sind auf Zeichenform basierte Inputcodierungen nicht generell verbreitet. Wegen der semantischen Mehrdeutigkeit ist die Codierung eines einzelnen Zeichens nach Sinninhalt zudem schwer durchzuführen (siehe Form-Sinninhalt-Beziehungen in diesem Kapitel). Eine phonetische Inputcodierung, die den sprachlichen Signalen angemessen und mit wenigen Vorkenntnissen erlernbar ist, ist trotz der sprachlichen Einschränkung in der Praxis gebräuchlicher. Aus diesem Grund werden Erforschungen der phonetischen Dimension in Kap. 3.4.2 und 3.4.3 intensiviert.

3.4.2 Grundlegende Eigenschaften der Phonetik des Chinesischen

Allgemein betrachtet ist einer der größten phonetischen Unterschiede zwischen den drei in der chinesischen Schrift geschriebenen und den indogermanischen Sprachen, dass die ostasiatischen Sprachen eher auf Silben und die europäischen eher auf Phonemen basieren. Außer der phonetischen Entsprechung von einem Schriftzeichen zu einer betonten Silbe im Chinesischen ist auch die syllabische Grundeinheit in Japanisch und Koreanisch auffällig, zu denen

[156] Die zwei angegebenen japanischen Aussprachen sind nachfolgend in häufigster On- und der nativen Kun-Lesart. In der Spalte ‚Sinninhalt' wird nur die Grundzeichenbedeutung angegeben.

jeweils Silbenschriften sowie gruppierten Silbenblöcke zum Schreiben benutzt werden (vgl. Müller-Yokota 1994b: 389f & 399). Nähere Ausführungen folgen in Kapitel 3.5.

Seit der Entstehung der chinesischen Schrift gibt es hauptsächlich vier Entwicklungsperioden der chinesischen Sprache: das frühantike (ungefähr vom 17. Jh. v. Chr. bis zum 2. Jh. n. Chr.), das mittelantike (3. - 13. Jh.), das spätantike (14. - 18. Jh.) und das moderne Chinesisch (seit 1840). Prinzipiell gibt es im Standardchinesischen immer weniger Töne sowie Silbenvarianten. In manchen heutigen Zweigsprachen (wie Kantonesisch) wurden viele Merkmale der archaischen chinesischen Aussprache jedoch beibehalten (vgl. Sun 1988: 209). In dieser Arbeit wird meist nur das moderne Standardchinesisch berücksichtigt.[157]

Zusammengefasst zeichnet sich die chinesische Phonetik vor allem durch vier Eigenschaften aus: 1) es gibt klare Grenzen zwischen zwei Silben, weshalb die nachfolgenden Silben normalerweise nicht kombiniert ausgesprochen werden; 2) die meisten Silben enden mit einem Vokal; 3) eine Silbe enthält ein bis vier Phoneme; 4) ein bestimmter Ton gilt für die ganze Silbe und ist distinktiv für die Bedeutung eines Morphems oder eines Wortes (vgl. Lu 2008: 59). Die Aussprache eines chinesischen Schriftzeichens kann in erster Linie nach einem Segment (eine von Phonemen kombinierte Silbe) und nach einem Ton klassifiziert werden. Ein Segment wird grundsätzlich von zwei Teilen zusammengesetzt: Initial (auch Anlaut, chi.: 声母 /shēngmǔ/; eng.: initial) und Final (auch Auslaut, 韵母 /yùnmǔ/; final). Das Initial kann ein Konsonant oder leer sein.[158] Das Final in der chinesischen Sprache besteht aus Reimkopf (韵头 /yùntóu/), Reimkern (韵腹 /yùnfù/) und Reimende (韵尾 /yùnwěi/).[159] Zu jeder Stelle des Segments – Initial, Reimkopf, Reimkern und Reimende – passt höchstens ein Phonem. Als Reimkopf (auch Mediallaut genannt) kann ein Halbvokal (<i> [i], <u> [u] oder <ü> [y]) eingesetzt werden. Der Reimkern ist das einzige unentbehrliche Element des Finals, an dessen Stelle allein vokalische Phonemen auftreten können. Gibt es einen Reimkopf, so muss ein öffnender Vokal den Reimkern darstellen, wie <a> [ɑ/a], <o> [o/ɔ] oder <e> [ə]. Als Reimende können ausschließlich ein geschlossener Vokal (<u> oder <i>) oder ein Nasalkonsonant (<n> [n] oder <ng> [ŋ]) eingesetzt werden (vgl. ibid.: 60f). Die Silbenstruktur kann mit der folgenden Formel beschrieben werden:

(Konsonant) + (Halbvokal) + Vokal + u/i/n/ng

[157] Putonghua (*Standardsprache*/*Mandarin*) ist die Bezeichnung dafür in der VR China. In Taiwan und unter den amerikanischen Chinesen heißt die Sprache 国语 /guóyǔ/ (*Sprache des Staates*) und in Südostasien 华语 /huáyǔ/ (*Sprache Chinas*). Beim Wortschatz gibt es einige Unterschiede in den verschiedenen Regionen.

[158] Die innerhalb eines Schrägstrichpaars angegebene Transkription ist in Pinyin und die in eckiger Klammer IPA; selbiges gilt bei allen parallelen lautlichen Markierungen der chinesischen Sprache in dieser Arbeit.

[159] Die vier Bezeichnungen basieren auf der Übersetzung der Verfasserin.

Ton (声调 /shēngdiào/) bedeutet im Bereich der Linguistik „[d]istinctive pitch at the syllable level. In tone language, tone is a sense-discriminative feature which distinguishes lexical meanings“ (Coulmas 1996a: 509). In Putonghua gibt es vier Arten des Tonhöhenverlaufs und den unbetonten schwachen Ton, der als ‚neutraler Ton‘ (轻声 /qīngshēng/) bezeichnet werden kann (vgl. GB/T 12200.2-94). Bei der Erläuterung des Tonwerts kann ein fünfstufiges Bezugssystem von ZHAO Yuanren eingesetzt werden. Der erste Ton ‚Yinping‘ (阴平) bleibt beim Aussprechen gleichstehend auf Stufe Fünf. So kann sein Tonwert als *55* angegeben und mit dem Zeichen <¯> symbolisiert werden. Der zweite Ton ‚Yangping‘ (阳平) steigt von Stufe Drei auf Fünf (Tonwert *35* und Zeichen <ˊ>). Der dritte Ton Shangsheng (上声) fällt zuerst von Zwei auf Eins und steigt dann auf Vier (Tonwert *214* und Zeichen <ˇ>). Der vierte Ton ‚Qusheng‘ (去声) fällt von Fünf auf Eins (Tonwert *51* und Zeichen <ˋ>). Die Tonhöhe des neutralisierten Tons kann mit diesem Bezugssystem nicht dargestellt werden. Es ist deswegen umstritten, ob er als Ton definiert werden kann (vgl. Bodmer 1997: 241, Lu 2008: 62f). Zu argumentieren ist, dass es vier Töne, aber fünf Toneme im Standardchinesischen gibt.

Nach verschiedenen Berechnungen gibt es 415 divergierende Silben in Putonghua. Nach Berücksichtigung des Tons (inklusive 38 Silbenwerte mit schwachem Ton) gibt es insgesamt 1.332 Varianten (vgl. Cai 2005: 42). Wenn nur die 8.105 Schriftzeichen von ‚Tongyong Guifan Hanzi Biao‘ 2013 (siehe Kap. 3.2.2, S. 142f) angerechnet werden, gibt es für eine Silbe durchschnittlich 19,53 und nach Einschränkung des Tons 6,08 Homophone. Im Höchstfall kann eine Silbe über hundert Homophone unter den allgemeingebräuchlichen Zeichen verfügen. Für zahlreiche Homophone und die graphisch unterscheidende Funktion der chinesischen Schrift ist ein in klassischer Schriftsprache geschriebener Text mit 96 Schriftzeichen von ZHAO Yuanren ein bekanntes Beispiel. Der Text erzählt eine vollständige Geschichte inkl. dem Titel, wobei alle auftretenden Schriftzeichen im Standardchinesischen als /shi/ ausgesprochen werden. Beim Lesen können die Informationen durch den Sinninhalt eines einzelnen Graphems entschlüsselt werden. Durch reines Zuhören jedoch kann die Geschichte auch kein Muttersprachler verstehen.

3.4.3 Transkriptionen des chinesischen Schriftsystems

Wegen der Eigenschaften der chinesischen Sprache ist es zwar uneffektiv, die chinesische Schrift durch eine phonologische Schrift zu ersetzen, aber eine national anerkannte Transkription ist zum Zweck des Zeichenerwerbs, der Sprachstandardisierung und der international standardisierten Schreibung der chinesischen Eigennamen obligatorisch. Die phonetische Umschrift in China hat sich hauptsächlich in drei Perioden entwickelt: Fanqie (反切) mit vor-

handenen Schriftzeichen, Zhuyin (注音) mit einheimischen phonetischen Symbolen und Pinyin (拼音) mit dem lateinischen Alphabet.

Vor der Sui-Dynastie (581-619) wurde die Zeichenaussprache durch gleich oder ähnlich ausgesprochene Zeichen angegeben. Mit der Verbreitung des Buddhismus aus Indien ist der Begriff ‚Alphabet' zum ersten Mal in China bekannt geworden. Dergestalt beeinflusst wurde im ca. 7. Jh. die Methode ‚Fanqie' begründet und bis zum Anfang des 20. Jh. verwendet. Fanqie meint die phonetische Angabe durch zwei Zeichen, wobei der Initial des Oberzeichens und der Final sowie Ton des Unterzeichens den Silbenwert des anzugebenden Zeichens zusammen bilden (vgl. Cai 2005: 48). Anhand der phonetischen Angabe in Fanqie am Beispiel des Zeichens <德> (/dé/, *die Moral*) sind die Vor- sowie Nachteile dieser Methode beispielhaft zu erkennen. Seine Aussprache wird in <多> (/duō/, *viel*) als Oberzeichen und <则> (/zé/, *Regel*) als Unterzeichen angegeben (nach ‚Yitizi-Lexikon' Online). Vorteil von Fanqie ist, dass es kaum durch regionale und historische Varianten begrenzt wird. So kann die Aussprachemarkierung trotz der phonetischen Änderung des Standardchinesischen im modernen Mandarin immer noch funktionieren. Diese Fanqie-Angabe gilt auf Rundung auch bei anderen chinesischen Sprachen wie Kantonesisch. Die drei Schriftzeichen werden in Kantonesisch nach der Reihenfolge /dak^{1}/ (德), /do^{1}/ (多) und /zak^{1}/ (则) gebildet (nach ‚Yueyu Fanyin Peici Ziku').[160] Nachteile von Fanqie sind, dass es einerseits unpräzise bei der phonetischen Angabe ist und andererseits von den individuellen Kenntnissen des erworbenen Zeichens abhängt (vgl. Cai 2005: 48).

1913 wurde die erste national standardisierte Transkription publiziert, die auch unter dem Terminus ‚Bopomofo' oder ‚Zhuyin-Alphabet' (注音字母, auch *Mandarin Phonetic Symbols*, kurz MPS) bekannt ist. In dem System werden 21 Zeichen für Initial, drei für Medialaute (chi.: 介音 /jièyīn/, eng.: prenuclear glide; ergo: Reimkopf/Halbvokal) und 13 Symbole für Final erfunden, die anhand der Gestalt und Striche der sinographischen Regelschrift entworfen wurden (vgl. Coulmas 1996a: 577). Bis heute ist Zhuyin die offizielle Transkription Taiwans, wohingegen sie in Festland China durch Pinyin ersetzt wurde. Zhuyin stellt deswegen in Taiwan eine wichtige Inputcodierung dar. Ein Tonzeichen wird optional isoliert rechts von den Zhuyin-Symbolen geschrieben, wobei es bei dem ersten Ton weggelassen werden muss und bei dem schwachen Ton mit einem unterstehenden Punkt anzugeben ist (siehe dritte Spalte von Tab. 3-3).

[160] Die hochgestellten Zahlzeichen stellen den Ton des Kantonesischen (insgesamt acht Töne) dar.

Der Entwurf einer nationalen standardisierten phonetischen Transkription im lateinischen Alphabet war eine der wichtigsten Aufgaben der Schriftreformen der Volksrepublik China in den 1950er Jahren, um internationalen Austausch zu gewährleisten und die standardchinesische Aussprache zu kodifizieren. Zu diesem Zweck wurde Pinyin entworfen. Im weiteren Sinne ist das Wort die chinesische Bezeichnung für Alphabet. Im engeren Sinne heißt Pinyin ‚phonetisches Alphabet für Standardchinesisch im lateinischen Alphabet'. In Pinyin werden alle 26 Grundbuchstaben außer ‚v' verwendet. Der Sonderbuchstabe <ê> (in IPA: [ə]) wird in den meisten Fällen und <ü> (IPA: [y]) in vielen Fällen durch <e> sowie <u> ersetzt, die im Allgemeinen für den Laut [ɤ] sowie [u] stehen. Wenn die geschlossenen Vokale [i], [u] oder [y] als Anlaut einer Silbe auftreten, werden <y> (für [i] und [y]) sowie <w> (für [u]) statt <i>, <u> und <ü>) eingesetzt. Das Tonzeichen wird auf dem Vokal des Reimkerns eingesetzt (herzu vgl. die zweite Spalte von Tab. 3-3).

Trotz großer Unterschiede bei der phonetischen Umschreibung behalten Fanqie, Zhuyin und Pinyin dasselbe Grundprinzip bei: die Silbe als eigenständige Einheit und die Kombination von einem Initial und einem Final. Die 21 Konsonanten für Initial (inklusive Nullinitial 22) und die 39 Varianten für Final in Mandarin werden in der Tab. 3-4 parallel in Pinyin, Zhuyin und IPA analog aufgelistet.

Initial								
Pinyin	**MPS**	**IPA**	**Pinyin**	**MPS**	**IPA**	**Pinyin**	**MPS**	**IPA**
b	ㄅ	[ḅ]	l	ㄌ	[l]	zh	ㄓ	[ḍʂ]
p	ㄆ	[pʻ]	g	ㄍ	[g̊]	ch	ㄔ	[tʂʻ]
m	ㄇ	[m]	k	ㄎ	[kʻ]	sh	ㄕ	[ʂ]
f	ㄈ	[f]	h	ㄏ	[x]	r	ㄖ	[ʐ̩]
d	ㄉ	[ḍ]	j	ㄐ	[ḍɕ]	z	ㄗ	[ḍs]
t	ㄊ	[tʻ]	q	ㄑ	[tɕʻ]	c	ㄘ	[tsʻ]
n	ㄋ	[n]	x	ㄒ	[ɕ]	s	ㄙ	[s]
Nullinitial (Vertretungsinitial) [161]**:**								
y	ㄧ	[i]	w	ㄨ	[u]	yu	ㄩ	[y]
Final (Pinyin, MPS, IPA, Beispielzeichen)								
Medialaut[162] / Reimkern+(-ende)		i, ㄧ, [i], 衣[163]			u, ㄨ, [u], 乌			ü, ㄩ, [y], 迂[164]
eigenständig als Final		**Medialaut + Reimkern + (Reimende)**						
a, ㄚ, [a], 啊		ia, ㄧㄚ, [ia], 呀			ua, ㄨㄚ, [ua], 蛙			
o, ㄛ, [o], 喔					uo, ㄨㄛ, [uo], 窝			
e, ㄜ, [ɣ]/[ə], 鹅		ie, ㄧㄝ, [iɛ], 耶						üe, ㄩㄝ, [yɛ], 约
ai, ㄞ, [ai], 哀					uai, ㄨㄞ, [uai], 歪			

[161] Nullinitial sind die Silben gemeint, die mit einem Vokal (als Reimkopf oder Reimkern) anfangen; die Buchstabierung des Finals entspricht dem mit demselben Phonem führenden Fall, wobei der Anfangslaut /i/, /u/ und /ü/ in dem Vertretungsinitial ‚y', ‚w' und ‚yu' geschrieben werden muss, wie /wang/, /yi/, /yuan/ usw.; für die mit /a/, /o/ und /e/ führenden Silben wird kein Vertretunngsinitial gebraucht, wie /an/.

[162] Die drei Medialaute/Halbvokale können sowohl eigenständig als Final als auch als Reimkopf eines Finals auftreten.

[163] Der Pinyin-Wert /i/ wird als alleinstehendes Final in der Silbe <zhi>, <chi>, <shi> und <ri> als [ɻ] und in der Silbe <zi>, <ci> und <si> als [ɯ] ausgesprochen. In den anderen Fällen wird er als [i] ausgesprochen.

[164] Nach dem Initial /j/, /q/ und /x/ wird der Laut in Pinyin als /u/ geschrieben.

ei, ㄟ, [ei], 诶		uei, ㄨㄟ, [uei], 威	
ao, ㄠ, [au], 熬	iao, ㄧㄠ, [iau], 腰		
ou, ㄡ, [ɣu], 欧	iou, ㄧㄡ, [iɣu], 忧		
an, ㄢ, [an], 安	ian, ㄧㄢ, [ian], 烟	uan, ㄨㄢ, [uan], 弯	üan, ㄩㄢ, [yɛn], 冤
en, ㄣ, [ən], 恩	in, ㄧㄣ, [in], 因	un/wen, ㄨㄣ, [uən], 温[165]	ün, ㄩㄣ, [yn], 晕
ang, ㄤ, [aŋ], 昂	iang, ㄧㄤ, [iaŋ], 央	uang, ㄨㄤ, [uaŋ], 汪	
eng, ㄥ, [əŋ], Ø	ing, ㄧㄥ, [iŋ], 英	ung/weng, ㄨㄥ, [uəŋ], 翁	
ong, ㄨㄥ, [uŋ], Ø	iong, ㄩㄥ, [yŋ], 雍		
er, ㄦ, [ɑɻ], 二			

Tab. 3-4: Initiale sowie Finale des Standardchinesischen und der Vergleich mit den drei Transkriptionssystemen[166]

Seit der offiziellen Einführung im Jahr 1958 ist Pinyin ein wichtiges Hilfsmittel zum Zeichenerwerb der Grundschüler und Fremdsprachler und zur Verbreitung des Standardchinesischen geworden. 1979 wurde Pinyin von ISO als internationaler Standard zur Transkription des Chinesischen festgelegt (vgl. DIN-Taschenbuch 343: 410). Zur Vereinfachung des Recherchierens in Lexika wurde die Indexierung nach alphabetischer Reihenfolge in Pinyin seither parallel zur Anordnung nach Klassenhaupt und Strichanzahl eingesetzt. Im Bereich der Textverarbeitung gewann Pinyin auch für die Eingabe der chinesischen Schrift an Bedeutung (vgl. Sun 1988: 341ff).

Die auf Pinyin und Zhuyin basierten Eingabemethoden werden im chinesischsprachigen Gebiet am häufigsten benutzt. Ihr größter Vorteil ist der leichtere Erwerb und die phonologische Wiedergabe des Gesprochenen. Als wichtige Hilfsmittel werden Pinyin oder Zhuyin im ersten Grundschuljahr bzw. am Anfang des Fremdsprachenunterrichts vermittelt. Ein dementsprechend gestaltetes Eingabeverfahren fordert daher kein zusätzliches Auswendiglernen von einem Eingabeschema. Die Entsprechung zu akustischen Informationen verursacht, dass die PC-Benutzer schneller und mit weniger geistiger Belastung die einzugebenden Schriftzeichen in Pinyin sowie Zhuyin decodieren können.

Phonetische Eingabeverfahren sind dabei keineswegs perfekt. Die Nachteile sind vor allem bei den zahlreichen Homophonen, regionalen Dialekten und individuellen Idiolekten zu finden. Auf die unvermeidbaren Nachteile und mögliche Lösungsansätze mit künstlicher Intelligenz wird in Kap. 4.1.3 eingegangen.

[165] Wenn dieser Auslaut allein eine Silbe bildet, wird er in Pinyin als /wen/ geschrieben. Ebenso ist es bei /ung/ als /weng/.
[166] Vgl. Hanyu Pinyin Fang'an, Coulmas 1996a: 409 & 577, Xu YC 2008: 18, DIN-Taschenbuch 343: 425.

3.5 Einführung und weitere Entwicklung im japanischen und koreanischen Schriftsystem

Mit der chinesischen Schrift ist ein großer Schriftkulturkreis in Ostasien entstanden, weshalb sie auch für viele andere Sprachen eingesetzt wurde; insbesondere für die Sprachen der Nachbarländer Japan, Korea sowie Vietnam und der historischen Nationalminderheiten Chinas. Die aufgrund solcher Sprachen notwendig gewordenen Reformen und neuen Erfindungen haben neue Perspektiven der Sinographie aufgeschlagen. In der ca. zweitausendjährigen Weiterentwicklung in nicht-chinesischsprachigen Gebieten wurden einerseits viele neue morphemtragende Schriftzeichen geschaffen, um lexikalische Lücken der nationalsprachlichen Wortschätze zu schließen. Zusätzlich wurden manche phonologische Schriften (wie Katakana, Hiragana, Hangul, die Frauenschrift), Hybriden von Phonographen und Logographien (wie Jurchen-, Kitan-Schrift) sowie logographische Schriften (wie Yi-, Xixia-Schrift, Chữ Nôm) erfunden (vgl. Chen QG 1993: 27-34). Viele solcher abgeleiteten Schriften sind bereits ausgestorben, wurden (z.B. politisch motiviert) abgeschafft oder finden nur noch sehr selten Verwendung. So werde ich in Kap. 3.5 die weitere Entwicklung des japanischen sowie koreanischen Schriftsystems vorstellen, in denen die chinesische Schrift und die davon abgeleiteten Schriften in der modernen Zeit immer noch eine wichtige literarisierende Rolle einnehmen.

3.5.1 Grundlagen der weiteren Entwicklung

In Kapitel 3.1.1 (S. 134f) wurde kurz eingeführt, dass die Entwicklung der chinesischen Schrift in Japan und Korea hauptsächlich in drei Prozesse eingeteilt werden kann: 1) Übernahme der chinesischen Schriftsprache; 2) Reformierung der Schrift sowie Neuschaffung von Schriftzeichen nach den sechs Schriften (abhängig vom Wortschatz der nativen Sprache); 3) die Erfindung der nationalen phonologischen Schriften, konkret die japanischen Silbenschriften Katakana und Hiragana sowie das koreanische Alphabet Hangul.

Katakana und Hiragana werden zusammen als Kana bezeichnet und wurden ca. im 9. Jh. erfunden. Jede dieser Schriften enthält 48 Grundsilbenzeichen. Teils können sie sich durch die Einsetzung des diakritischen Zeichens ‚Dakuten' oder ‚Handakuten' zu erweiterten Zeichen wandeln. Die Eins-zu-eins-Entsprechung zwischen Zeichen und Syllabar herrscht in beiden Kana-Schriften vor. Im japanischen Schriftsystem nehmen Katakana und Hiragana unterschiedliche Funktionen wahr: ersteres ist so für fremdsprachige Eigennamen und Fremdwörter (außer chinesischen), Dialekte, Interjektionen und Fachwörter zuständig, wohingegen Hiragana die Schreibung von Synsemantika ermöglicht (vgl. Dürscheid 2006: 83f, Coulmas

1996a: 252-256). Kanji werden hauptsächlich bei Niederlegung der Autosemantika verwendet, wie Nomen, Verben, Adjektive und manche Adverbien (vgl. Smith 1996: 209).

Das koreanische Alphabet Hangul wurde „1446 als 訓民正音 [sk] Hunmi-chŏngŭm (wörtl. ‚das Volk in den richtigen Lautungen unterweisen' […] veröffentlicht […]" (Müller-Yokota 1994b: 398). Es beträgt historisch 28 und heutzutage 24 Jamo (siehe Kap. 2.4.2, S. 103f). Die Schriftform der Jamo und die Bildungsarten im syllabischen Block wurden hauptsächlich auf Basis der Striche sowie der Zeichengestalt der chinesischen Schrift entworfen. Seit dem Ende des Zweiten Weltkriegs ist Hangul offizielle Hauptschrift der koreanischen Halbinsel. Im Gegenzug wurde die Verwendung der Hanja mit politischer Forderung im Norden abgeschafft sowie im Süden erheblich eingeschränkt.

3.5.2 Kontrastive Analysen von Chinesisch, Japanisch und Koreanisch

Nachfolgend wird ein Vergleich zwischen Chinesisch, Japanisch und Koreanisch auf grammatischer, phonetischer und lexikalischer Ebene durchgeführt. Anhand dieses Vergleichs wird gezeigt, warum ein Hybridschriftsystem von Phonographie und Sinographie heute für die Aufzeichnung des Japanischen und Koreanischen bevorzugt wird.

A) Grammatische Struktur

Vom grammatischen Standpunkt aus gehört Chinesisch zum isolierenden Typus, während Japanisch und Koreanisch zu den agglutinierenden Sprachen zählen. Im Chinesischen werden die syntaktischen Beziehungen im Satz durch Synsemantika oder die Reihenfolge der Wörter ausgedrückt. Der Sprache entsprechend wurden die chinesischen Schriftzeichen original für unflektierte Morpheme sowie Wörter geschaffen (vgl. Bußmann 2002: 321). Im Gegensatz dazu werden Affixe im Japanischen und Koreanischen an eine Wortwurzel gehängt, um die grammatischen Beziehungen zwischen einzelnen Wörtern und Veränderungen in der Bedeutung auszudrücken (vgl. Bodmer 1997: 184). Ein unflektierbares morphemtragendes Sinogramm wäre unpraktisch und uneffektiv für die Markierung eines agglutinierenden Affixes des Japanischen und Koreanischen.

Im modernen japanischen Schriftsystem sind die aus Kanji und Hiragana zusammengesetzten Nomen, Verben, Adjektiven sowie Adverbien häufig zu finden. Das in Hiragana geschriebene Affix wird an die unflektierbare Kanji-Wortwurzel angehängt, wie z.B. <行く> (/iku/, *gehen*) oder <行かす> (/ikasu/, *gehen lassen*). Mit dieser Methode können einerseits Homophone graphisch unterschieden werden, andererseits werden die sprachlichen Informationen mit möglichst wenigen Zeichen dargestellt (vgl. Lu 2008: 19f). Auch in der koreanischen Geschichte (seit Erfindung des Hangul 1446 bis zum Ende des Zweiten Weltkriegs) war

diese Verwendungsform im koreanischen Schriftsystem beliebt (vgl. Sohn 2001: 13). So wurde *chinesisch*/*china-zugehörig* in Koreanisch als <中國의> /cwung kwu kuy/ geschrieben (heute meistens vollständig durch Hangul ersetzt <중국의>) (vgl. King 1996: 225).

B) Phonetik

Phonetisch gewendet unterscheiden sich die drei Sprachen vornehmlich in zwei Punkten: 1) ob es Toneme gibt und 2) wie komplex die Silbenstruktur ist. Die chinesische Sprache ist eine Tonsprache mit relativ einfacher Silbenstruktur, in der die Homophone sprachlich erst durch Toneme distinktiv werden (siehe Kap. 3.4.2, S. 167). Wegen des großen Anteils an chinesischen Lehnwörtern ist die Menge der Homophone im Japanischen und Koreanischen ebenfalls sehr umfangreich; sie sind jedoch nicht durch Ton distinktiv.

Gänzlich gegenteilig sind die japanischen und koreanischen Silbenstrukturen im Vergleich. Eine japanische Silbe wird höchstens aus zwei Phonemen, einem Konsonant (16 Phoneme zur Auswahl, inklusive Nullkonsonant) und einem Vokal (acht Möglichkeiten: a, i, u, e sowie o als einzelner Vokal und ya, yu sowie yo als Diphtong) zusammengesetzt, weshalb nur ca. 100 Varianten existieren (vgl. Dürscheid 2006: 85). Wegen der begrenzten Anzahl von Silben wäre es praktikabel, Syllabar parallel zu Kanji zum Schreiben zu verwenden. Wie in Kap. 2.4.2 eingeführt wurde, besitzt die koreanische Sprache einen relativ komplexen Silbenaufbau. Eine Silbe wird von einem Anlaut, einem Inlaut und (optional) einem Auslaut zusammengesetzt und die Silbenmöglichkeiten betragen insgesamt 11.172 (vgl. Unicode 12.0 Chapters: Kap. 18.6: 737). Anhand der komplizierten Silbenstruktur ist die Graphem-Phonem-Repräsentation definitiv besser für das Koreanische geeignet.

C) Wortschatz

Durch die Verbreitung der chinesischen Schrift und den tausendjährigen Kulturaustausch machen die aus dem Chinesischen entlehnten Wörter einen beträchtlichen Teil aus und sind tief verwurzelt. Im japanischen Wortschatz sind 47,5% der Wörter nach einer 1956 durchgeführten empirischen Forschung des Nationalen Japanischen Sprachinstituts sinojapanisch; die nativen Wörter lassen sich auf rund 37% beziffern (vgl. Gao 1990: 7). Im Koreanischen [des modernen Süden] betragen die sinokoreanischen Wörter ca. 60%, während die nativen Wörter 35% ausmachen (vgl. Sohn 2001: 12f).

Wie erwähnt ist die Homophonie-Wahrscheinlichkeit des Japanischen/Koreanischen ebenfalls hoch. Kanji und Hanja sind so in manchen Fällen erforderlich, um Homophone schriftlich präzise aufzeichnen zu können. Ein japanisches Beispiel liefern die mit dem syllabischen Wert /kō/ produzierten Kanji, auf die 64 von ca. zweitausend häufigen Kanji entfallen

(vgl. Cheng 2002: 130f). Homophonetische Ambiguität ist einer der wichtigsten Gründe, warum auf die chinesische Schrift in Südkorea trotz des ‚Hangul-Gesetzes' (wonach bei offiziellen Schriftstücken nur Hangul erlaubt ist) im alltäglichen Schriftverkehr nur schwer zu verzichten ist. Im Koreanischen haben 473 Silben Korrespondenten zu Hanja, d.h. unter den 4.888 Hanja gibt es im Durchschnitt 10,3 Homophone (vgl. Huang/Bae/Choi 2004: 1).

Aus den Analysen der drei Standpunkte ist zu schlussfolgern, dass ein parallel auf der chinesischen und phonologischen Schriften basiertes Schriftsystem theoretisch am effektivsten für Japanisch sowie Koreanisch wäre. Ausgehend von der Silbenstruktur und der Anzahl der Silbenvarianten wurden in den beiden Nationen geeignete phonologische Schriften erfunden: die Syllabar Hiragana sowie Katakana in Japan und das Alphabet Hangul in Korea.

3.5.3 Methoden zur Anwendung der Sinogramme und das Schaffen neuer morphologischen Schriftzeichen

Es gibt vor allem drei Methoden, um die vorhandenen chinesischen Schriftzeichen für das japanische sowie koreanische Schriftsystem anzuwenden:

1) Die bedeutungsbezogene Strategie – Kun-Lesart (chi.: 訓讀/训读; jap.: 訓読み; kor.: 훈독): Mit dieser Strategie wird ein Sinogramm mitsamt seiner Bedeutung übernommen und mit der Lautung in der nationalen Sprache definiert (vgl. Dürscheid 2006: 81). Die Kun-Lesart <水> (*das Wasser*) z.B. lautet in Japanisch /mizu/, die mit der Aussprache des nativen Wortes verbunden ist.

2) Die lautbezogene Strategie – On-Lesart (音讀/音读; 音読み; 한자음): Diese Methode bedeutet, dass ein Sinogramm sowohl mitsamt seiner Bedeutung als auch mit dem phonetischen Wert entlehnt wurde. Von den Schriftzeichen mit On-Lesart haben sich die sinojapanischen sowie -koreanischen Wörter entwickelt (vgl. ibid.). So wird das Zeichen <水> in Japanisch auch als /sui/ ausgesprochen und stammt daher von der phonetischen Realisation aus dem Altchinesischen (/shuǐ/ in Putonghua). Abhängig von der Zeit und den Dialekten der Wortentlehnung gibt es bei vielen Kanji unterschiedliche On-Lesarten. So hat z.B. das Zeichen <和> mindestens drei verschiedene On-Lesarten: Go-on – /wa/ (aus Wu im 6 Jh.), Kan-on – /ka/ (aus dem Standardchinesischen im 7-8 Jh.) und Tō-on – /o/ (hauptsächlich aus einem südchinesischen Dialekt im 19. Jh.) (vgl. Coulmas 1996a: 240f).

3) Die Rebusstrategie: Diese Strategie besagt, dass ein Sinogramm wegen seiner phonetischen Ähnlichkeit zu einem nativen syllabischen Wert als ein Silbenzeichen übernommen und zur sprachlichen Aufzeichnung verwendet wurde. In diesem Fall verlieren die Sinogramme

ihre morphemtragende Funktion und fungieren als Silbenzeichen (vgl. Dürscheid 2006: 81). Ein bekanntes Beispiel dafür sind die Schriftzeichen des historischen japanischen Syllabarsystems Manyōgana (chi.: 萬葉假名/万叶假名, jap.: 万葉仮名) aus dem 8. Jh., in dem 87 chinesische Schriftzeichen als Silbenzeichen ausgewählt wurden. Sie ist ein Ursprung von Hiragana und Katakana (vgl. Coulmas 1996a: 252f).

Außer den drei Verwendungsmethoden von vorhandenen Sinogrammen wurden auch viele logographische Schriftzeichen anhand der chinesischen Konstruktionsprinzipien (sechs Schriften) für native Wörter erfunden. Die Verfahren des Schaffens solcher Zeichen sind meistens zusammengesetzte Ideogramme und Determinativphonetika. Es handelt sich um reinjapanische und -koreanische Schriftzeichen wie <畑> (jap. /hata/; chi.: /tián/; *Trockenfeld*; ZI aus *Feuer* und *Feld*), <鱈> (/tara/; /xuě/; *Gadus*; DP mit dem Radikal *Fisch* und dem Phonetikum *Schnee*), <畓> (kor. /tap/; chi.: /dā/; *Reisfeld*; ZI von *Wasser* und *Feld*) (vgl. Chen QG 1993: 27f). Manche Schriftzeichen wurden später auch im chinesischen Schriftsystem übernommen und haben die Sprachen CJK-Schriftkreises beeinflusst. Besonders in der frühen Neuzeit haben die Japaner intensiv Wissenschaft und Technologie der westlichen Welt importiert und viele neue Wörter sowie Schriftzeichen erfunden.

Unter den reinkoreanischen Schriftzeichen gibt es viele, die außerhalb der ‚sechs Schrift-Prinzipien' erfunden wurden: die zusammengesetzten Phonogramme. Diese Methode bezieht sich auf den Zeichenaufbau von zwei Grundzeichen, deren phonetische Werte verschmelzen. Solche Zeichen variieren weiterhin in zwei Arten: die Zusammensetzung von zwei Sinogrammen sowie von einem Sinogramm und einem Jamo. Ein Beispielzeichen im ersten Fall ist <갈> (kor.: /kal/; chi.: /jiā/), das aus dem Oberzeichen <加> mit dem phonetischen Wert /ka/ und dem Unterzeichen <乙> mit der Silbenkoda seiner Aussprache /il/ aufgebaut ist. Nach der Erfindung der Hangul sind viele hybride Schriftzeichen mit Sinogramm und Hangul aus dem Prinzip entstanden, um Unterzeichen durch ein phonemtragendes Jamo zu ersetzen, etwa <틱> /kek/ (aus Hanja /ke/ und Jamo /k/) (vgl. Li DC/Jin 1997: 45f, Müller-Yokota 1994: 398).

Wegen des Mischschriftsystems basiert die japanische Inputcodierung im Allgemeinen auf Hiragana oder Romanji. Wegen der großen Menge homophonetischer Wörter werden bei den Eingabemethoden meist auch kontextabhängige Eingaben durchgeführt, wie bei den Pinyin- sowie Zhuyin-Eingabemethoden des Chinesischen. Die erforderlichen technischen Voraussetzungen und computerlinguistischen Anwendungen sind ebenso im Prinzip identisch und werden neben den intelligenten Pinyin-Eingabemethoden erforscht. Wie in Kap. 2.4.3 ge-

zeigt wurde, basiert eine koreanische Eingabemethode hauptsächlich auf der Konversion von Jamo zu Silbenblock (siehe S. 107-111; vgl. Lunde 2009: 306ff). Daher liegen ihre Schwerpunkte bei der Zeichen-Tasten-Repräsentation, der Codierung des Silbenblocks sowie Jamo und der Äquivalenz zwischen einer Silbe und einer Jamofolge. Die technischen sowie computerlinguistischen Schwierigkeiten der koreanischen Textverarbeitung sind deswegen im Prinzip niedriger als die der chinesischen sowie japanischen.

4 Methodik, einbezogene linguistische Erkenntnisse und computerlinguistische Anwendungen der chinesischen Eingabemethoden

In Kapitel 3 wurden die grammatologischen Desiderate eingeführt, auf deren Basis verschiedene Arten von Inputcodierung der Eingabemethoden entworfen werden und die aufzeigen, wie die Textverarbeitung der chinesischen Schrift funktioniert. In diesem Kapitel werden die Arbeitsprinzipien der verschiedenen chinesischen Eingabemethoden erforscht, die verschiedene Forschungsbereiche – vor allem die Informatik, Computerlinguistik und Linguistik – betreffen.

Wegen der großen Menge an verschiedenen chinesischen Eingabemethoden beziehen sich die Analysen auf vier Vorgehensweisen. Zuerst werden die generellen Grundprinzipien der chinesischen Eingabemethoden betrachtet und solche Methoden je nach ihrer Art klassifiziert (siehe Kap. 4.1.1). Selbiges gilt für die Erforschung der Pinyin-Eingabemethode, die zunächst überblickend vorgestellt (siehe Kap. 4.1.3 & 4.1.4) und später im Detail erforscht wird (Kap. 4.2 bis 4.5). Zweitens wird die intelligente Pinyin-Eingabemethode als Schwerpunkt intensiv durchleuchtet, wohingegen die Wubi-Eingabemethode knapp skizziert wird (Kap. 4.1.2). Die sonstigen Eingabemöglichkeiten werden ausgelassen. Drittens verfahren die Analysen intelligenter Pinyin-Eingabemethoden von kleineren bis zu größeren sprachlichen Einheiten, ergo die Reihenfolge: Zeichen (Kap. 4.3) – Wort (Kap. 4.4) – Satz (Kap. 4.5). Dies entspricht einerseits der Entwicklung von der Pinyin-Eingabemethode, andererseits stimmt dieses Schema auch mit der allgemeinen Verarbeitungsreihenfolge der intelligenten Software überein. Viertens sind die linguistischen Erkenntnisse die Voraussetzungen für technische Realisierungen, weshalb sie immer vor den technischen Verarbeitungsverfahren eingeführt werden. So werden in Kap. 4.2 das allgemein benötigte linguistische Fachwissen, die darauf basierte Wissensdatenbank und die Anwendungstechnologien derselben (von der intelligenten Eingabesoftware) vorgestellt. Hierfür notwendig sind Erkenntnisse über Zeichen, Wort, Phrase und Satz, die jeweils zu Beginn von Kap. 4.2 bis 4.5 dargelegt werden.

4.1 Verschiedene Eingabemethoden der chinesischen Schrift

Grundziel der Eingabe der chinesischen Schrift ist es, durch Inputcode den digitalen Code abzurufen und die entsprechende Glyphe und Font auszugeben. Dies ist der Schwerpunkt der ‚Informationsverarbeitung der chinesischen Schrift' (fokussieret einzelne Schriftzeichen) und der ‚Informationsverarbeitung des Chinesischen' (fokussieret die Sprache) (vgl. auch Kap.

1.4.5, S. 57f; Sheng 2006: 78). Solche Eingabeverfahren werden im Normalfall von der PC-Standardtastatur unterstützt und als ‚tastaturbasierte Eingabemethode der chinesischen Zeichencodierung' (eng.: Chinese character coding keyboard input method, chi.: 汉字编码键盘输入法) bezeichnet (vgl. zu einer konkreten Definition in Einleitung, S. 2). Dank der Besonderheiten der Schrift stehen viele technische Anwendungen, zahlreiche Inputcodierungsmöglichkeiten und eine große Auswahl von Eingabesoftwares für das Chinesischschreiben bereit. Ziel von Kap. 4.1 ist es im Grundsatz, die verschiedenen Eingabemethoden und die Grundarbeitsprinzipien von zwei häufig gebrauchten Methoden – Wubi und Pinyin – vorzustellen.

4.1.1 Überblick über die chinesischen Eingabemethoden

Die Textverarbeitung der chinesischen Schrift ist zwar im Verarbeitungsverfahren wesentlich komplizierter als die der alphabetischen Schriften. Der allgemeine Arbeitsprozess und die verwendete Hardware sind aber grundsätzlich identisch. Der allgemeine Arbeitsprozess wurde in Kap. 1.2.3 (S. 17f) und die vier Arten der zur Textverarbeitung benötigten Codierungen in Kap. 1.4.1 (S. 42) vorgestellt. Den allgemeinen Prozess der chinesischen Eingabemethoden hat Wu (1999: 7) wie folgt zusammengefasst.

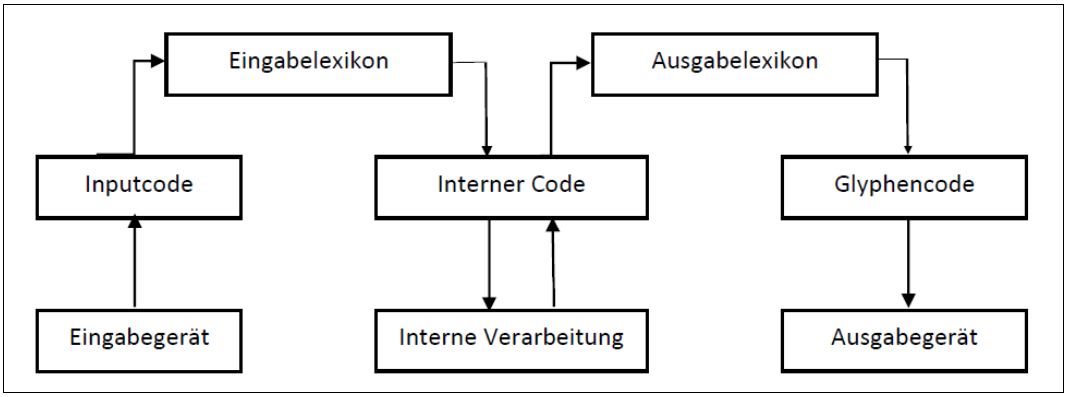


Abb. 4-1: Der allgemeine Arbeitsprozess der chinesischen Eingabemethoden (Wu 1999: 7 [Übersetzung der Verfasserin])

Solche Eingabemethoden basieren auf der Umwandlung zwischen Input-, internem und Glyphencode (Ausgabecode). Das unterscheidende Element, welches zur Eingabe der alphabetischen Schrift normalerweise nicht gebraucht wird, ist das Eingabelexikon. Nachdem ein Inputcode für das einzugebende Schriftzeichen via Tastatur (oder andere Eingabegeräte) eingegeben wurde, wird er anhand des Eingabelexikons einer Eingabemethode zu einem oder mehreren zum Inputcode passenden Schriftzeichen konvertiert. Der interne Code lässt sich weiter mithilfe des Ausgabelexikons mit bestimmten Zeichenglyphen verbinden und auf dem Bildschirm anzeigen. Wenn es mehrere Kandidaten für einen Inputcode gibt, wird eine Wahlliste mit allen möglichen Kandidaten nach bestimmter Reihenfolge zuerst in einer Wahlliste ange-

boten und das einzugebende Schriftzeichen weiter durch die manuelle Auswahl der PC-Benutzer bestimmt (vgl. Wu 1999: 6f).

Wie in Kap. 1.4.3 (S. 53f) zusammengefasst wurde, gibt es zwischen der Eingabe der chinesischen und einer alphabetischen Schrift vor allem drei große Unterschiede: das Mehr an erforderlicher geistiger Arbeit des PC-Benutzers, die größeren Herausforderungen bei der künstlichen Intelligenz des Computers und die Einschränkung der Zeichenmenge wegen des begrenzten Ein- sowie Ausgabelexikons und der Zeichendatenbank. Zeichendatenbank (Datenbank des Hanzi-Fonts, eng.: character library, chi.: 字库) meint die einem Computersystem zugehörige Datenbank mit dem Font der chinesischen Schriftzeichen. Die Anzahl der gemeinsam gespeicherten Zeichen in dem Eingabe- sowie Ausgabelexikon und in der Zeichendatenbank entscheidet, wie viele und welche Schriftzeichen mit der angewendeten Eingabesoftware in diesem Computer verarbeitet werden können (vgl.: ibid.: 4, GB 12200.1-90: Kap. 4.1).

Die Eingabemethoden der chinesischen Schrift unterscheiden sich hauptsächlich in tastaturbasierte und nicht-tastaturbasierte Eingabemethoden. Erstere sind die Forschungsobjekte dieser Arbeit und die allgemein angewendeten Methoden für PC-Benutzer. Zweitgenannte unterscheiden sich weiter in die drei folgenden Unterarten (vgl. Tsai 2005: 9):

1) Optische Zeichenerkennung (eng.: optical character recognition, Abk.: OCR; chi.: 光电汉字识别 /guāngdiàn hànzì shíbié/).

Bei dieser Methode werden die auf Papier handgeschriebenen oder gedruckten Schriftzeichen gescannt, anerkannt und dadurch in Computer eingegeben. Das dazugehörige Eingabegerät ist ein hoch-qualitativer Scanner, mit dem Schriftstücke zu optischen Signalen transformatiert werden können. Durch das Matching der Zeichenform auf Papier mit der Zeichendatenbank wird der interne Code des wahrscheinlichsten Zeichens vom System abgerufen und in den elektronischen Texten hinzugefügt. Dadurch wird der auf Papier stehende Text ‚digitalisiert' (vgl. Peng 1994: 150, Zhang ZC 1991: 183). OCR-basierende Eingabe kann im Normalfall auch Offline-Handschrifteingabe genannt werden, die im Gegensatz zur Online-Handschrift steht. Das Verfahren wird eher in fachlichen Bereichen verwendet, wie Korporabegründungen, dem Druckereiwesen usw. Obwohl die Korrektheit der chinesischen OCR in den vergangen Jahrzehnten erheblich gestiegen ist, kann sie grundsätzlich nicht fehlerfrei sein. In vielen Fällen müssen die digitalisierten Texte manuell korrigiert werden. Zum Alltagsleben und zur Büroarbeit ist diese Eingabemethode unökonomisch und ineffizient (vgl. Peng 1994: 151).

2) Online-Handschrifteingabe (online handwriting; 联机手写识别 /liánjī shǒuxiě shíbié/).

In diesem Eingabeprozess laufen Zeichenschreiben und -erkennung im einzelnem gleichzeitig ab. Das Handschreiben geschieht auf einem speziellen Eingabegerät, entweder mit speziellem Stift auf einem Grafiktablett, mit der Maus in einem bestimmten Feld des PC-Bildschirms oder mit dem Finger auf dem Berührungsbildschirm eines Tablet-PCs oder Smartphones. Das System bietet die möglichen Kandidaten an, die ähnlich wie das handgeschriebene Zeichen aussehen, und der Benutzer trifft die Wahl des einzugebenden Schriftzeichens. Im Vergleich zu der OCR-basierenden Methode, die ebenso auf Zeichenerkennung basiert, ist die Online-Handschriftmethode alltagstauglicher und mit weniger technischen Schwierigkeiten behaftet (vgl. ibid., Sheng 2006: 79). In vielen intelligenten chinesischen Eingabesoftwares wird die Online-Handschriftmethode als Eingabealternative angeboten. In elektronischen Medien mit Berührungsbildschirm wie Tablet-PC und Smartphone funktioniert die chinesische Online-Handschrifteingabe effektiver und wird viel häufiger verwendet.

3) Sprachbasierte Eingabe (speech based input; 语音识别 /yǔyīn shíbié/).

Sprachbasierte Eingabe bedeutet die computergestützte Eingabe durch Spracherkennung. Das Hauptprinzip und die allgemeine Funktionalität umfasst: 1) die Segmentierung eines Satzes in Silben sowie Wörter; 2) Erkennung einzelner Silben nach den phonetischen Regeln und deren Initial, Final und Ton; 3) weitere Analyse und Bestimmung der Wörter sowie Sätze nach Kontext mit linguistischen Erkenntnissen (vgl. Sheng 2006: 79). Im Vergleich zu den indogermanischen Sprachen gibt es bei der Spracherkennung der chinesischen Sprache zusammengefasst vor allem drei Unterschiede: 1) Für die indogermanischen Sprachen sind die zu erkennenden Einheiten Phonem, Diphthong, Halbsilbe, Silbe und Wort, wohingegen für Chinesisch Initial, Final, Ton, betonte Silbe und Wort maßgeblich sind. 2) Die maschinell lesbaren Merkmale der verschiedenen Töne müssen beschrieben werden, um den Ton automatisch zu unterscheiden. 3) Wegen der vergleichsweise häufigen homophonetischen Schriftzeichen und Wörter im Chinesischen müssen die möglichen Kandidaten mit linguistischem Wissen (inkl. Grammatik, Semantik, Pragmatik usw.) ausgefiltert werden, damit der Computer automatisch die richtige Auswahl treffen kann. Wegen der technischen Schwierigkeiten ist ein großer Teil der sprachbasierten Softwares nur für die Konversion von Sprachsignalen zu Pinyin zuständig und die Konversion von Pinyin zu Schriftzeichen wird von einer anderen Software durchgeführt (vgl. Möbius/Haiber 2010: 217f, Peng 1994: 165ff). In den 80er sowie 90er Jahren hat die Korrektheit der Silbenerkennung des Standardchinesischen (Umwandlung in Pinyin)

schon über 90% erreicht; die Korrektheit der Tonunterscheidung lag bei über 99% (vgl. Zhang ZC 1991: 187).

Die tastaturbasierten Eingabemethoden der chinesischen Schrift können aus verschiedenen Perspektiven kategorisiert werden. Vor dem Hintergrund des Tastaturaufbaus variieren sie sich in internationale PC-Standard- und ‚sinisierte' Sondertastaturen, die weiter in drei Arten unterschieden werden: 1) die große Tastatur mit tausenden Tasten von eigenständigen Schriftzeichen, die von der chinesischen Schreibmaschine stammt; 2) die mittelgroße mit ca. 500 Tasten, in der die Komponenten der Schriftzeichen verteilt werden; 3) die kleine mit besonderer Inputcodierung mit Tasten für Striche oder Initial, Final sowie Ton. Die internationale Standardtastatur ist die verbreitetste Variante im Alltag, während die Sondertastaturen kaum Verwendung finden.

In den meisten Fällen wird eine chinesische tastaturbasierte Eingabemethode nach ihrer Inputcodierung benannt und klassifiziert. Wie in Kap. 3.4.1 erwähnt, bezieht sich die sinographische Inputcodierung auf die drei Grundattribute der Schrift: Zeichenform, -aussprache und -sinninhalt (vgl. S. 162f). Für das japanische und koreanische Schriftsystem, in denen die chinesische Schrift und die nativen phonologischen Schriften gemischt verwendet werden, wurden (und werden) hauptsächlich Eingabemethoden mit phonetischen Inputcodierungen entwickelt und von den PC-Benutzern in der Praxis verwendet. In den meisten Fällen basiert die Eingabe des Japanischen auf der Kana-Kanji- oder Romanji-Kana-Kanji- und des Koreanischen auf der Jamo-Hangul-Hanja- oder RRK-Hangul-Hanja-Konversion (vgl. Chen AW 1986: 71; siehe Kap. 2.4.3, S. 108f & Kap. 3.5.3, S. 175f).[167] Für das chinesische Schriftsystem, das im Allgemeinen nach dem morphologischen Schrifttyp funktioniert, könnten zahlreiche Eingabeschemata entworfen werden. Keines davon wäre jedoch unumstritten. Vor 1994 wurden bereits ca. 700 verschiedene Inputcodierungen entworfen, darunter rund 570 aus Festlandchina und 100 aus Taiwan. Unter diesen gibt es vierzig bis fünfzig Schemata, die nach der Eingabemethoden entwickelt und in der Praxis angewendet wurden (vgl. Peng 1994: 102). Solche Schemata können nach dem/den basierten Attribut(en) der chinesischen Schrift in vier Klassen gegliedert werden:

1) Codierung nach der Zeichenform (chi.: 字形编码 /zìxíng biānmǎ/).

In diesem Fall wird ein Schriftzeichen in seine graphischen Bestandteile zerlegt und in den Codes der Bestandteile eingegeben. Die codierten Objekte solcher Inputcodierungen können

[167] Romanji ist die japanische Bezeichnung für lateinische Buchstaben und umschreibt auch die Latinisierung des Japanischen; RRK (Abk. für Revised Romanization of Korean) ist die südkoreanische-standardisierte Transliteration von Hangul.

Striche, Strichfolgen, Komponenten/Grammwurzeln oder simple Zeichen sein, die mit graphischen ASCII-Zeichen repräsentiert werden (vgl. Zhang ZC 1991: 191, Peng 1994: 102). Eingabemethoden nach der Zeichenform haben in der praktischen Verwendung unvermeidbare Einschränkungen.

Wie in Kap. 3.3.3 vorgestellt wurde, beträgt die Strichanzahl eines Schriftzeichens zwischen eins bis mehr als 30 und im Durchschnitt 11 bis 18. In einer strichbasierten Inputcodierung wird ein Zeichen somit durchschnittlich mit zehnfachem Tastenschlag repräsentiert, was sehr uneffektiv ist (siehe S. 156). Da es nur acht Grund- bzw. fünf Hauptstrichmerkmale gibt, ist bspw. die Handytastatur mit zwölf Tasten für die Stricheingabemethode geeigneter. Aus solchen Gründen wird dieses Verfahren oft bei Handys, aber selten bei PCs verwendet.

In Kap. 3.3.4 (S. 158) wurde der Zeichen-Komponenten-Zusammenhang vorgestellt, von dem es zwei Aspekte für komponentenbasierte Inputcodierungen zu rekapitulieren gilt: A) Es gibt keinen festgelegten, allgemein anerkannten Standard für die Grundkomponenten. Zudem variiert ihre Gesamtanzahl meinungsabhängig zwischen 100 bis über 600. B) Ein Schriftzeichen kann mit verschiedenen rationalen oder irrationalen Methoden zerlegt werden. Insbesondere gibt es manchmal Konkurrenz zwischen der Zerlegung nach dem graphischen Aufbau und dem Bildungsprinzip. Verweisend auf die zwei Punkte des Zeichen-Komponenten-Zusammenhangs und der Anwendungssituationen solcher Eingabemethoden können die Nachteile der Komponenteninputcodierung wie folgt zusammengefasst werden: 1) Da hunderte Komponenten mit großer Subjektivität vom Erfinder codiert werden, lässt sich eine solche Inputcodierung schwer auswendig lernen. Zugleich wird sie nach einer längeren Verwendungspause leicht vergessen. 2) In vielen Fällen ist die Zerlegungsart unterschiedlich zu den erworbenen Zeichenkenntnissen der PC-Benutzer. Es erschwert einerseits den Umgang mit der Inputcodierung und verwirrt andererseits wegen der Zeichenkenntnisse. 3) Der Gedankengang der Zeichenzerlegung steht konträr zum Schreiben, bei dem die sprachlichen Signale kurzfristig im menschlichen Kopf erzeugt und gespeichert werden (vgl. Hou 1999: 89).

Die Menge der codierten Komponenten eines Eingabeschemas muss sich auch an der möglichen Effizienz, den Schwierigkeiten für das Gehirn und der PC-Standardtastatur orientieren. Wenn alle 600 Komponenten codiert würden, wäre dies untauglich zum Auswendiglernen und somit für die Verwendung von PC-Benutzern. Wenn hingegen nur hundert Grammwurzeln codiert würden, müssten zahlreiche Grundkomponenten irrational in zwei oder mehrere Grammwurzeln zerlegt werden, was die Eingabe deutlich verlängert (vgl. Chen AW 1986: 10-13). Da es verschiedene Arten zur Zeichenzerlegung gibt, wurden zahlreiche Komponentcodierungen erfunden und danach entworfene Eingabemethoden entwickelt. Außer der

in Festlandchina beliebten Wubi-Eingabemethode von WANG Yongmin und der in Taiwan und Hongkong verbreiteten Cangjie-Eingabemethode von CHU Bong-foo, die beide in Kap. 3.3.4 mit Beispielzeichen vorgestellt wurden, gibt es noch viele andere gebräuchliche Verfahren. Beispiele dafür sind die Dazhong-Pinxing- (大众拼形输入法), Qian-San-Mo-Yi- (前三末一输入法), Sanbi- (三笔输入法) und Zhengma-Eingabemethoden (郑码输入法) (vgl. Peng 1994: 102). Ihre Funktionsprinzipien werden im nächsten Kapitel am Beispiel der Wubi-Eingabemethoden vorgestellt.

2) Codierung nach der Zeichenaussprache (chi.: 字音编码 /zìyīn biānmǎ/).

In diesem Fall wird ein Zeichen nach seiner Phonetik (Initial, Final und Ton) codiert (vgl. Zhang ZC 1991: 191). Im Gegensatz zu den Komponentencodierungen gibt es eine standardisierte Transkription, bzw. Pinyin in Festlandchina oder Zhuyin in Taiwan, die am Anfang der Grundschulbildung und vorm Zeichenlernen erworben werden muss. Die auf ihr basierten Inputcodierungen können deswegen ohne zusätzliches Auswendiglernen erworben und benutzt werden.

Sowohl in Festlandchina als auch in Taiwan werden die phonetischen Eingabemethoden am meisten bevorzugt. Statistisch benutzen mehr als 97% der chinesischen PC-Benutzer Pinyin-Eingabemethoden (vgl. Tsai 2005: 9, Chen Y 1997, nach: Chen Z/Lee 2000: 1). Die Vorzüge von phonetischen Eingabemethoden, die zu ihrer großen Beliebtheit geführt haben, sind einerseits die Leichtigkeit für die standardsprachige Bevölkerung. Andererseits entspricht die phonetische Eingabe den akustischen Informationen und dem Gedankengang des Schreibens. Wegen dieser Vorzüge kann sich ein PC-Benutzer auf den zu schreibenden Text konzentrieren, ohne sich mit Detailüberlegungen zur Zeichenform beschäftigen zu müssen. Funktionsprinzipien und Effizienz werden in Kap. 4.1.3 und 4.1.4 eingeführt und analysiert.

3) Codierung nach der Aussprache-Form-Kombination (chi.: 音形编码 /yīnxíng biānmǎ/).

Um den größten Nachteil (die homophonbedingte Überschneidung vieler Kandidaten) der phonetischen Eingabemethoden zu vermeiden, entstanden die Varianten der Aussprache-Form-Kombination. Das Grundprinzip lautet, dass eine Inputcodierung mit der Aussprache als Hauptteil und der Zeichenstruktur als Nebenteil entworfen wird, um die Homophone mithilfe von Radikal, Komponente oder Strich zu filtern. Zu den Inputcodierungen dieses Typs zählen bspw. die Ziran- (自然码), Initial-Final-Strich- (声韵笔形码) oder Initial-Final-Komponentencodierung (声韵部形码) (vgl. Peng 1994: 103).

4) Codierung nach der Form-Aussprache-Kombination (chi.: 形音编码 /xíngyīn biānmǎ/).

Im Gegensatz zu 3) ist bei der Form-Aussprache-Kombination die Codierung der Zeichenform der Hauptteil, zu der mindestens zwei Alternativen zählen. Die erste Alternative ist die Zeichenzerlegung in Bausteine (wie Komponenten oder Striche), die weiterhin nach ihrer Aussprache codiert werden, z.B. die Codierung vom Klassenhaupt <钅> (abgeleitet vom Zeichen <金> /jīn/, *Metall*) mit dem Initial ihrer Aussprache <J>. Die zweite Alternative ist das Hinzufügen des phonetikhinweisenden Codes nach dem zeichenstrukturrepräsentierten Code. Die dieser Klasse zugehörigen Eingabemethoden sind zahlreich, wie z.B. die AP-Hanzi- (AP 汉字输入法), Jian-Zi-Shi-Ma- (见字识码输入法) oder Yitong-Huawen-Methode (易通华文输入法) (ibid.).[168]

Die zahlreichen Varianten der Inputcodierungen und Eingabemethoden erschwerten den PC-Benutzern die Auswahl und den Umgang mit dem Computer. Es ist deswegen notwendig, ein paar hochwertige Inputcodierungen rational auszuwählen und weitergehend die darauf basierenden Eingabemethoden zu verbessern. Nach Hou gibt es hauptsächlich sechs Kriterien für eine gute chinesische Inputcodierung (vgl. Hou 1999: 94ff):

1. Leichte Erwerbbarkeit. Eine gute Inputcodierung muss leicht erlernbar und dem Kenntnisstand der chinesischen Schrift angemessen sein. Die Kenntnisse der Inputcodierung variieren nach den Grundkenntnissen der chinesischen Sprache und Schrift, die in der Grundschulbildung vermittelt werden müssen, und den speziellen des Codierungssystems, die es zusätzlich zu erwerben gilt. Dabei gilt im Normalfall, dass, je weniger spezielle Codierungskenntnisse zu einer Inputcodierung benötigt werden, diese desto einfacher zu erlernen ist. Das Pinyin-Eingabeschema ist auf dieser Ebene von der größten Simplizität. Die nach der Zeichenform entworfenen Schemata basieren hingegen meistens auf dem unregelmäßigem Auswendiglernen der Codierung.
2. Hohe Effizienz. Um hohe Effizienz zu erreichen, muss der Inputcode für ein Schriftzeichen relativ kurz (= weniger Tastenanschläge), das Tasteneintippen möglichst zeitsparend (setzt vor allem ein rationales Tastaturlayout für eine Inputcodierung voraus) und die möglichen Kandidaten eines Zeichencodes gering sein (weniger Zeitaufwand zur Zeichenauswahl).

[168] Peng hat drei Alternativen der Form-Aussprache-Kombination genannt: neben denen im Text genannten basiert die dritte Alternative auf der Grammwurzelzerlegung, die wiederum nach der Kategorie des Zeichensinninhaltes codiert wird. Demgemäß sollte die Cangjie-Methode nach ihrer Definition zu dieser Kategorie gehören. In dieser Arbeit wird diese Alternative jedoch als auf zeichenformbasierte Inputcodierung kategorisiert.

3. Angemessenheit des Standards der chinesischen Schrift. Um eine Inputcodierung zur Verbreitung zu bringen, muss sie auf den nationalstandardisierten Attributen der Schriftzeichen basieren: die phonetischen Inputcodierungen auf der geregelten Aussprache in Putonghua und die Zeichenformcodierung auf der nationalen Orthographie in der VR China (vereinfachte Schrift) sowie in Taiwan/Hongkong (traditionelle Schrift).
4. Kompatibilität der Tastatur. Im Idealfall kann das internationale Tastaturlayout (das US-amerikanische Tastaturlayout und seine Varianten) für die Ansetzung der Inputcodierung zur Verfügung stehen. So weist das funktionale Tastaturlayout nur wenige Unterschiede zum internationalen Tastaturlayout auf, um so die Eingabe der chinesischen Schrift regional einzuschränken. Pinyin-Eingabemethoden haben binnen dieses Kriteriums einen bestechenden Vorteil im Vergleich zu anderen.
5. Bestimmte und identische Regeln innerhalb eines Codierungssystems. Die Regeln für die Bildung eines chinesischen Schriftzeichens sind nicht festgelegt und haben häufig Ausnahmen. In einer zeichenformbasierten Inputcodierung aber müssen bestimmte und identische Regeln beherrscht werden, damit die PC-Benutzer sie mit wenig Verwirrung erwerben und verwenden können. Sogar die besseren Komponentencodierungen (wie z.B. Wubi) besitzen aus dieser Perspektive bestimmte Konfliktpunkte.
6. Kompetenz für größeres Zeicheninventar. Zum alltäglichen Schreiben in Festlandchina reicht zwar das gemeingebräuchliche Zeicheninventar von dem Standard GB2312-1980 mit 6.763 Schriftzeichen, aber für sämtliche fachlichen Gebiete und den kompletten CJKV-Schriftkreis ist es zu eingeschränkt. Eine für verschiedene Menschengruppen geeignete Inputcodierung muss auf einem größeren Zeicheninventar basieren und zur Verarbeitung von über 20.000 Schriftzeichen bereit sein. Auf beliebten Inputcodierungen basierende Eingabemethoden (wie Pinyin und Wubi) erfuhren mit der Zeit daher eine Vergrößerung ihres Zeicheninventars.

Es ist zwar unmöglich, eine perfekte Inputcodierung der chinesischen Schrift zu erfinden. Mit der steigenden künstlichen Intelligenz ist es aber möglich, die Effizienz einer Eingabemethode mithilfe von computergestützten Sprachanalysen und Kandidatenausfilterung im Kontext zu erhöhen. Die tastaturbasierten Eingabemethoden können nach den Entwicklungsphasen in drei Prozesse eingeteilt werden:

1) Der Prozess der Zeichenverarbeitung mit der Konversion im einzelnen Zeichen.

Dieser Prozess fokussiert die Erforschung der Schriftzeichen und der Schwerpunkt liegt auf dem Inputcodierungsentwurf. Die Entwicklung des Konversionszeichenlexikons, die Proklamation einer standardisierten Austauschcodierung und die Begründung der Zeichendatenbank waren dabei die entscheidenden Teilschritte (vgl. Hou 1999: 90). Die Forschungsaufgaben dieser Entwicklungsphase waren hauptsächlich auf den Teilbereich ‚Informationsverarbeitung der chinesischen Schrift' begrenzt (vgl. Kap. 1.4.5, S. 57).

2) Der Prozess der Wortverarbeitung, in dem die Eingabe in der Einheit des Wortes abläuft.

Dieser Prozess beginnt ungefähr Anfang der 1980er Jahren. Zum Zweck der Wortverarbeitung wurden ‚die Datenbank der Wörter'[169], die Datenbank der Affixe und in manchen Fällen auch Korpora begründet. Eine Wörterdatenbank ist in der Tat eine elektronische Wörterliste, die zu dem Gebrauch mit Computern digitalisiert wird (vgl. Yao 1997: 172). In diesem Prozess wurden viele Forschungen im Bereich der lexikalischen Computerlinguistik durchgeführt, bspw. die Worthäufigkeit, automatische Wortsegmentation, Codierung der Wörter, Struktur der Wörterdatenbank und Wortbildungsregeln des Chinesischen (vgl. Hou 1999: 90f).

3) Der Prozess der Satzverarbeitung, in dem der Inputcode eines Satzes (exklusive des Interpunktionszeichens) gesamt eingegeben und verarbeitet wird.

Wissensdatenbanken mit linguistischen Informationen (Grammatik, Semantik, Pragmatik etc.) wurden einer Eingabemethode beigefügt, damit die satzstufige Konversion aus Inputcode in Zeichen durchgeführt werden konnte. Die Textverarbeitung in der Einheit des Satzes basiert auf ausgereiften Techniken der künstlichen Intelligenz. Dieser Prozess begann in den 1990er Jahren. Zu Beginn des neuen Jahrhunderts verbreitete sich die intelligente Eingabesoftware mit Satzverarbeitungsfunktionen primär im chinesischen Sprachraum. Die Ressourcen (Wissensdatenbank mit maschinenlesbarem Lexikon und linguistischen Informationen) und computerlinguistische Anwendungen (inkl. der automatischen statistischen und sprachlichen Analysen, für die linguistische Informationen nötig sind) sind die Grundlagen solcher intelligenten Eingabemethoden (vgl. Peng 1994: 96, Hou 1999: 91f). Die Satzverarbeitungsfunktionen werden am meisten in Pinyin- sowie Zhuyin-Eingabemethoden benötigt (vgl. Wang XL/Wang YL 1996: 50). In Kap. 4.1.3 werden die drei Prozesse der Pinyin-Eingabemethoden genauer vorgestellt.

[169] Sie ist eine Art von Wissensdatenbank, die Konversionslexikon und maschinenlesbares elektronischens Wörterbuch zusammenfügt. Es gibt zurzeit noch keine festgelegte Bezeichnung auf Deutsch, weshalb im Folgenden der Terminus ‚Wörterdatenbank' verwendet wird (chi.: 词库, eng.: word bank/ word stock).

4.1.2 Vorstellung der Wubi-Eingabemethode

Wie in Kap. 3.3.3 und 3.3.4 ausgeführt, kann ein Schriftzeichen immer in graphische Bestandteile zerlegt werden. Die Zerlegung in Komponenten wird sowohl im Bereich der Didaktik des Chinesischen als auch für den Entwurf der Inputcodierung eingesetzt. Das Bewusstsein für Komponenten besitzt im chinesischen Schriftkulturkreis eine lange Tradition. Wie in Kap. 3.3.2 (S. 148-155) vorgestellt wurde, wird ein komplexes Zeichen nach einem bestimmten Konstruktionsprinzip aus einfachen Zeichen oder verformten Komponenten gebildet. Daher können Zeichenkenntnisse bezüglich Form, Sinninhalt und Aussprache durch die Zeichenzerlegung effektiv erworben werden. Im Vergleich zur phonetischen Inputcodierung respektiert die Komponentencodierung stärker die Kultur der Sinographie, nämlich die Übertragung des Zeichensinninhalts per Zeichenform. Die auf dieser Art der Inputcodierung entwickelten Eingabemethoden erfordern keine Mandarinkenntnisse von den PC-Benutzern und können im Prinzip ohne sprachliche Einschränkung für die Eingabe der Hanzi, Kanji, Hanja und Chũ Nôm zuständig sein (hierzu vgl. Kap. 3.4.1, S. 163ff). Der andere wesentliche Vorteil ist die geringere Kandidatenüberschneidung im Vergleich zur phonetischen Inputcodierung. Anders ausgedrückt kann die Komponentencodierung eine weit höhere Effizienz erreichen, wenn sie fachkundig beherrscht wird (vgl. Chen AW 1986: 8). Die auf dieser Basis entworfene und beliebteste Eingabemethode in der VR China heißt Wubi-Zixing.

Wubi-Zixing (chi.: 五笔字型, wörtlich übersetzt: Fünf-Strich-Zeichenstruktur, auch Wangma genannt) wurde 1983 von WANG Yongmin (王永民) als Inputcodierung erfunden und ab 1986 als Software offiziell in der Computerschreibpraxis (Version-86) eingesetzt und verbreitet. Seitdem wurden die Wubi-Codierung und deren Eingabemethoden mehrmals erneuert und verbessert, insbesondere Version-98 und die Version der dritten Generation des neuen Jahrhunderts. Im Vergleich zu Version-86, mit der 6.763 vereinfachte Schriftzeichen aus dem Standard GB2312-80 zu verarbeiten sind, sind die beiden neuen Versionen jeweils zur Verarbeitung aller 20.902 CJK-vereinheitlichten Ideogramme im Unicode und einer großen Zeichenmenge aus 27.553 Schriftzeichen in der Lage.[170] Trotz der Verbesserungen neuerer Versionen nahm Version-86 großen Einfluss und bietet mehr kompatible Eingabesoftwares, weshalb das ihr zugrundeliegende Schema innerhalb dieses Kapitels beleuchtet werden soll. Im Prinzip gibt es zwischen den drei Versionen (neben dem Hinzufügen und der Umstellung einiger Komponenten) kaum Unterschiede (vgl. Wu 1999: 69f).

[170] Vgl. Wangma-Unternehmen, *http://www.wangma.net.cn/InfoMationDetail.aspx?sm=5&m=41* [Abruf: 2016-05-05].

Die im Wubi-Schema codierten Komponenten heißen Wubi-Grammwurzeln (五笔字根), die in der Version-86 insgesamt 130 betragen.[171] Das Grundprinzip der Wubi-Codierung ist die Kategorisierung aller Grammwurzeln nach ihrem ersten Strich, der immer einem der fünf Hauptstrichmerkmale – Héng (一), Shù (丨), Piě (丿), Diǎn (丶) und Zhé (𠃌) – entspricht (siehe Kap. 3.3.3, S. 155). Dementsprechend werden die 25 alphabetischen Tasten (außer ‚Z') auf fünf Zonen nach den Hauptstrichmerkmalen verteilt. Jede Zone enthält fünf Tasten, denen jeweils eine Serie von mehreren Grammwurzeln zugeordnet wird (vgl. Wang XY/Mao 2000: 11; siehe Abb. 4-2). Durchschnittlich ist eine Taste nach Version-86 für 5,2 Grammwurzeln zuständig. Bei der Eingabe wird ein Zeichen in ein bis vier Grammwurzeln zerlegt, wodurch mit maximal vier Tastenschlägen ein Zeichen geschrieben werden kann. Um die Effizienz möglichst zu erhöhen, wurde das Wubi-Tastaturlayout auf Basis des Zehnfingersystems und statistischer Analysen entworfen. Der Entwurf des Wubi-Tastaturlayouts folgt vier Prinzipien (vgl. Wang YM 2005: 25f, Yin 1998: 173f):

1) Häufiger Handwechsel. In Kap. 2.1.2 (S. 65) wurde erwähnt, dass der häufige Handwechsel die Eingabeeffizienz erhöhen und die Beanspruchung der Finger reduzieren kann. Daher wurden zwei Grammwurzeln, die häufig nacheinander auftreten, nach Möglichkeit in verschiedene Hälften verteilt. Statistiken zufolge ist die Kombination von Héng und Shù, Diǎn und Héng, sowie Piě und Diǎn (Nà) häufiger als die sonstigen Strichkombinationen. Bei dem Layout werden deswegen die Zone für Héng und Piě in der linken Hälfte aufgeteilt, wohingegen Shù und Diǎn in der rechten Hälfte zu finden sind.
2) Rationale Aufgabenverteilung der Finger nach Fingerstärke. Die Stärke vom Zeige- bis zum kleinen Finger nimmt kontinuierlich ab. Aus diesem Grund wurden beim Tastaturentwurf die Zeige- (die erste und zweite Stelle jeder Zone) und Mittelfinger (die dritte Stelle jeder Zone) häufiger beansprucht, wohingegen die beiden kleinen Finger insgesamt nur für vier grammwurzelrepräsentierte Tasten zuständig sind.
3) Belegung relativ häufiger Grammwurzeln in möglichst ‚guter' Lage. Der Grundreihe sind die Héng- und Shù-Zone zugehörig, deren belegten Grammwurzeln statistisch am häufigsten vorkommen. Im Gegenteil dazu befindet sich in der Unterreihe ausschließlich die Zhé-Zone.
4) Verzicht auf Umschalttasten beim Eintippen. Bei dem Eingabeverfahren eines alphabetischen und alphasyllabischen Schriftsystems wird das Tastaturlayout zwei- bis fünfmal be-

[171] Die codierten Grammwurzeln verdoppelten sich in den neuen Versionen beinahe und können in Haupt- und Nebenwurzeln (die von einer Hauptwurzel abgeleiteten Wurzeln) unterschieden werden. In Version-98 gibt es insgesamt 150 Haupt- und 90 Nebenwurzeln (vgl. Wu 1999: 69).

legt, was mithilfe von Umschalttasten funktioniert. Bei der Wubi-Tastatur repräsentiert eine Taste zwar mehrere Grammwurzeln (durchschnittlich 5,2 in Version-86), braucht aber keine Umschalttaste. Einerseits kann so das Eingabetempo vielmals erhöht werden. Andererseits werden mögliche Kandidaten aufgrund anderer Grammwurzeln sowie Erkennungscodes deutlich reduziert.

Beim Wubi-Tastaturlayout besitzt eine Taste drei Kennzeichnungen: 1) die Nummer (wie 11 für die erste Taste, wobei die erste Eins für die Zone und die hintere Eins für die Tastenstelle steht); 2) der entsprechende Buchstabe im US-amerikanischen Layout (wie ‚G' für Taste 11); 3) der Tastenname, der selbst eine von der Taste repräsentierte Grammwurzel ist (wie 王 /wáng/ für die Taste 11-G). Die Belegung der Tastatur und die wichtigsten Grammwurzeln jeder Taste werden in der folgenden Abbildung geschildert. Die Taste <Z> (nach dem US-amerikanischen Tastaturlayout) wird leer belegt und kann für eine beliebige Grammwurzel stehen, d.h. wenn es bei der Zeichenzerlegung eine unbewusste oder unsichere Grammwurzel gibt, kann sie durch <Z> ersetzt werden (vgl. Wu 1999: 50).

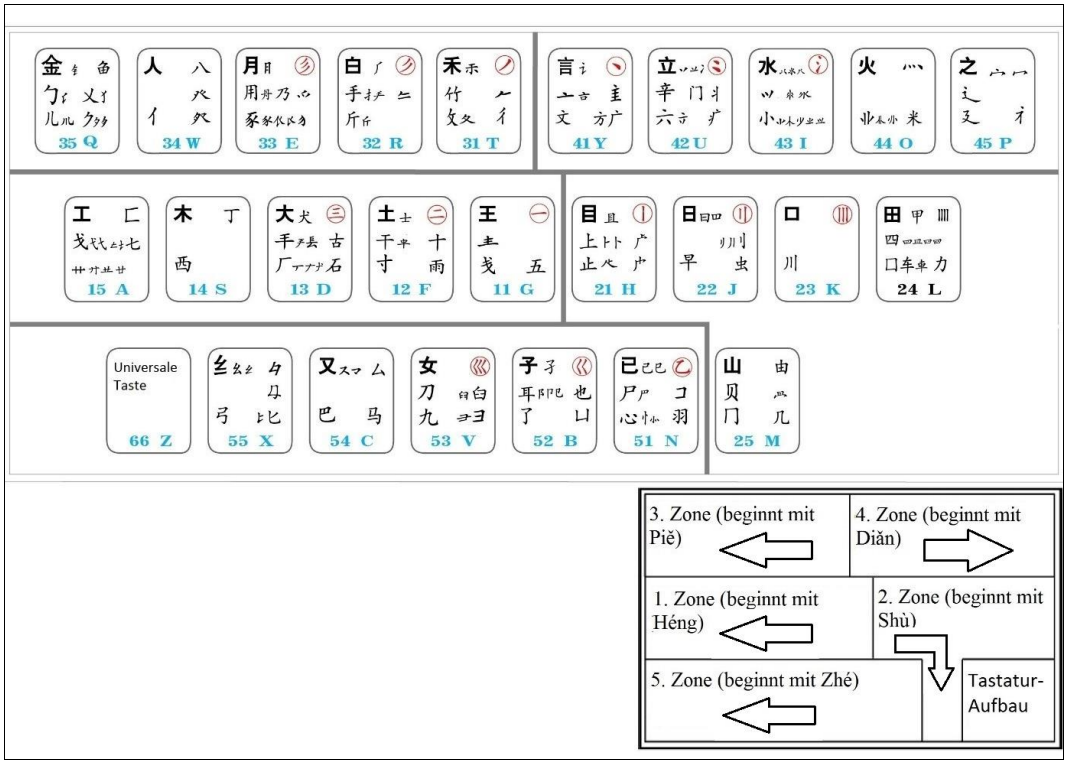


Abb. 4-2: Das Wubi-Tastaturlayout, Version-86[172]

Unter den 130 Grammwurzeln gibt es Bausteine der chinesischen Schrift auf verschiedenen Niveaus. Zunächst gibt es simple Zeichen, die aus dem Prinzip des Piktogramms oder Ideogramms gebildet wurden, selbst häufig gebraucht werden und oft als Grundkomponenten zur

[172] Quelle: *http://www.wangma.net.cn/InfoMationDetail.aspx?sm=5&m=41* [Abruf: 2016-05-05].

Bildung komplexerer Zeichen fungieren. Das Namenzeichen jeder Taste (außer dem der Taste ‚X') kann als Beispiel dafür dienen. Ergänzt werden sie durch Radikale, die selbst keine eigenständigen Zeichen sind, aber bestimmte Bedeutungen übertragen. Beispiele dafür sind 艹 (*Grasradikal*, für krautige Pflanzen, Taste ‚A'), 钅 (*Metallradikal*, für Metall, Taste ‚Q') und 氵 (*Drei-Punkt-Wasserradikal*, für Flüssigkeit, Taste ‚I'). Drittens werden alle Grundstriche und die Strichfolgen durch zwei-/dreimalige Wiederholung eines Grundstrichs ins Schema aufgenommen. Die Belegung solcher Grammwurzeln ist regelhaft, ein einzelner Strich gehört immer zu der ersten Stelle seiner zugehörigen Zone, wie <一> mit der Taste ‚G'. Die Strichfolgen mit Verdopplung eines Strichs befinden sich an der zweiten Stelle der Zone, wie <二> mit der Taste ‚F' und die mit Verdreifachung an der dritten Stelle, wie <三> mit der Taste ‚D'. Mit Ausnahme der drei Arten graphischer Elemente, die aus linguistischer Perspektive in den meisten Fällen rational sind (außer manchen Strichfolgen mit Verdopplung sowie Verdreifachung eines Strichs), gibt es auch einige besondere und irrationale Grammwurzeln (vgl. Wu 1999: 36). Diese können als ‚künstliche Grammwurzeln' bezeichnet werden, die im Gegensatz zu den ‚natürlichen Grammwurzeln' (die rationalen, graphischen Elemente) stehen. Sie wurden subjektiv auf der Basis von natürlichen Grammwurzeln erfunden, damit die nichtcodierten natürlichen Grundkomponenten wiederum zum Ziel der Eingabe in kleinere Bausteine zerlegt werden können (vgl. Chen AW 1986: 13). Im Wubi-Schema werden auf fast jeder Taste eine bis mehrere künstliche Grammwurzeln angeordnet. Bspw. wurde die Komponente 亻 (Taste ‚E') von dem simplen Zeichen 衣 (/yī/, *Wäsche*) abgeleitet. Da das Radikal bei der Zeichenbildung immer geometrisch getrennt geschrieben wird (wie im Zeichen 衷 /zhōng/, *vom Inneren*), wird diese Grammwurzel künstlich gebildet. Das Piktogramm sowie die Grundkomponente 羊 (/yáng/, *Schaf*) muss ebenfalls zweigeteilt zerlegt werden: in Doppelpunkt oben (Taste ‚U') und die künstliche Grammwurzel 羊 (Taste ‚D').

Kennzeichnungen, Grammwurzeln und das häufigste mit dieser Taste eingegebene Schriftzeichen werden in der folgenden Tabelle aufgelistet.

Zone	**Stelle**	**Nr.**	**Taste/ Bs.**	**Name der Taste**	**Strichmerkmal**	**Grammwurzel**	**häufigstes Zeichen**
1. Beginnt mit Héng <一>	1	11	G	王 /wáng/	一	王^丰戈五一	一 /yī/
	2	12	F	土 /tǔ/	二	土士二干十寸雨[⺧]	地 /dì/
	3	13	D	大 /dà/	三	大犬三^彐古石^厂厂^ナ^𠂆	在 /zài/
	4	14	S	木 /mù/		木丁西^覀	要 /yào/
	5	15	A	工 /gōng/		工戈^弋^艹^廿[⺾]匚七	工 /gōng/
2. Beginnt	1	21	H	目 /mù/	丨	目[⺭]上止丨卜^卜[⺊]^卢^卢	上 /shàng/
	2	22	J	日 /rì/	刂	日曰^罒早^刂^刂刂^川虫	是 /shì/

Zone	Stelle	Nr.	Taste/ Bs.	Name der Taste	Strich- merkmal	Grammwurzel	häufigstes Zeichen
mit Shù <丨>	3	23	K	口 /kǒu/	川	口川〢	中 /zhōng/
	4	24	L	田 /tián/		田甲口四皿罒罒车力〣	国 /guó/
	5	25	M	山 /shān/		山由贝冂皿	同 /tóng/
3. Beginnt mit Piě <丿>	1	31	T	禾 /hé/	丿	禾禾竹𠂉丿彳攵夂	和 /hé/
	2	32	R	白 /bái/	〃	白手扌手厂斤𠂉斤〃	的 /dè/
	3	33	E	月 /yuè/	彡	月月丹彡乃用豕豕𧘇𠂎彡	有 /yǒu/
	4	34	W	人 /rén/		人亻八癶癶	人 /rén/
	5	35	Q	金 /jīn/		金钅勹鱼犭乂儿儿夕厂𠂊	我 /wǒ/
4. Beginnt mit Diǎn <丶>	1	41	Y (US)	言 /yán/	丶	言讠文方广亠亠乀主	主 /zhǔ/
	2	42	U	立 /lì/	冫	立辛冫丷丷丬疒立	产 /chǎn/
	3	43	I	水 /shuǐ/	氵	水氺氵业业业业小氺氺	不 /bù/
	4	44	O	火 /huǒ/		火业小灬米	为 /wéi/
	5	45	P	之 /zhī/		之廴辶冖宀礻	这 /zhè/
5. Beginnt mit Zhé <乛>	1	51	N	已 /yǐ/	乙	已己巳コ乙尸尸心忄小羽	民 /mín/
	2	52	B	子 /zǐ/	巜	子孑了耳阝卩也㔾凵巜	了 /le/
	3	53	V	女 /nǚ/	巛	女刀九臼彐巛	发 /fā/
	4	54	C	又 /yòu/		又ㄡマ巴马厶	以 /yǐ/
	5	55	X	纟 /sī/		口弓匕七纟幺幺	经 /jīng/

Tab. 4-1: Die Komponenten der Wubi-Inputcodierung, Version-86[173]

Es gibt drei Arten an Beziehungen von zwei nachfolgend eingetippten Grammwurzeln:

1. Separation (wie z.B. 人 + 人 [/rén/, Mensch] → 从 /cóng/, befolgen)

2. Verbindung (wie z.B. 一 [yī, eins] + 大 [/dà/, groß] → 天 /tiān/, Himmel)

3. Kreuzung (wie z.B. 一 + 米 [/mǐ/, Reis] →来 /lái/, kommen)

Bei der Zeichenzerlegung der Wubi-Eingabe wiederum gibt es vier Hauptregeln (vgl. Wu 1999: 39f, Wang XY/Mao 2000: 30f):

1) Die Reihenfolge der Grammwurzeln ist mit der Reihenfolge beim Handschreiben identisch (generell von oben nach unten, von links nach rechts und von außen nach innen).

2) Die separierte Beziehung wird vor der verbindenden Beziehung bei der Zerlegung bevorzugt. 美 (/měi/, *schön*) z.B. wird in die Wurzeln 丷, 王 (/wáng/, *König*) und 大 zerlegt, obwohl es nach dem Strukturoriginal aus 羊 (/yáng/, *Schaf*) und 大 zusammengesetzt ist.

3) Die verbindende Beziehung wird der kreuzenden Beziehung bevorzugt. 天 z.B. wird in 一 und 大 zerlegt, aber nicht in 二 (/ér/, *zwei*) und 人.

[173] Vgl. Wang XY/Mao 2000: 12-26.

4) Größere Grammwurzeln werden bevorzugt, so dass ein Schriftzeichen mit möglichst wenigen Tasten repräsentiert werden kann. 来 etwa wird in 一 und 米 zerlegt, aber nicht in 一, 丶丿 und 木 (/mù/, *Baum*).

Die Länge eines Inputcodes für ein Schriftzeichen beträgt höchstens vier Buchstaben, wobei jedes Schriftzeichen – sowohl simple als auch komplexe – nach dem Wubi-Schema mit vier Buchstaben repräsentiert wird. Überschneidungen desselben Inputcodes sind selten, eine manuelle Kandidatenauswahl bleibt so erspart (vgl. Wu 1999: 43). Statistisch betrachtet beträgt die Überschneidungswahrscheinlichkeit unter gemeingebräuchlichen Schriftzeichen 1,3%. Nach Korporaanalysen mit 21.600.000 Zeichen liegt die Überschneidung eines Wubi-Inputcodes bei 0,5% (vgl. Peng 1994: 121). Im Vergleich zum Vollpinyin-Inputcode eines Zeichens, dessen Länge unregelmäßig zwischen eins (wie A) und sechs (wie ZHUANG) changiert und höchstens bis zu einhundert Kandidaten haben kann, hat die Wubi-Eingabemethode bei der Inputcode-Zeichen-Konversion erhebliche Vorteile. Damit ein chinesisches Schriftzeichen exakt mit vier Buchstaben repräsentiert werden kann, wurden einige Regeln festgelegt.

Wenn ein Schriftzeichen in mehr als vier Grammwurzeln zerlegt werden muss, wird es bei der Eingabe in die ersten drei sowie die letzte Grammwurzel codiert (vgl. Wang XY/Mao 2000: 32). Bspw. wird 德 mit 彳 (Taste: T), 十 (F) 罒 (L) und 心 (N) eingegeben und der Horizontalstrich mitterechts (die vorletzte Grammwurzel) wird nicht berücksichtigt.

Wenn ein Schriftzeichen in zwei oder drei Grammwurzeln zu zerlegen ist, wird ein so genannter Erkennungscode (chi.: 识别码) gebraucht. Im Wubi-Schema ist der Erkennungscode im Normalfall bedingt vom letzten Strich und dem geometrischen Aufbau des einzugebenden Schriftzeichens. Der letzte Strich bestimmt, in welcher Zone sich der Erkennungscode befindet, während der Aufbau entscheidet, an welcher Stelle die Taste ist. Der geometrische Aufbau lässt sich bei der Wubi-Eingabe in drei Fälle unterscheiden: 1) Links-Rechts-Aufbau (repräsentiert mit der ersten Stelle einer Zone), 2) Oben-Unten-Aufbau (zweite Stelle) und 3) sonstiger Aufbau (alle anderen Arten der geometrischen Beziehungen, dritte Stelle) (vgl. Chen AW 1986: 100; vgl. auch Kap. 3.3.5, S. 161). Bspw. wird 铠 (/kǎi/, *Rüstung*) mit dem Inputcode QMNN (Q→钅, M→山, N→己) eingegeben, wobei das letzte N als Erkennungscode fungiert. Da das Zeichen durch den Links-Rechts-Aufbau gebildet wird und der letzte Strich ein Zhé ist, stellt die erste Stelle der Zhé-Zone (Taste 51 [N]) den Erkennungscode dar. Wenn ein Zeichen in zwei Grammwurzeln zerlegbar ist, setzt sich sein Inputcode aus zwei Grammwurzel- und einem Erkennungscode zusammen und nach dem Eintippen der drei Buchstaben

wird das Leerzeichen als Ersatz gedrückt (ibid.). Zur Eingabe von 好 (/hǎo/, *gut*) bspw. wird nachfolgend V (für 女), B (für 子), G (Erkennungscode für Héng als letzter Strich vom Links-Rechts-Zeichenaufbau) und Space eingetippt.

Die Eingabe von Schriftzeichen, die eigenständige Grammwurzeln sind, funktioniert von Fall zu Fall unterschiedlich. Wenn das Zeichen der Name einer Taste ist, entspricht sein Code der viermaligen Wiederholung der Taste, etwa der Code von GGGG für 王 (/wáng/, *König*). Der Inputcode der sonstigen aus einzelnen Grammwurzeln bestehenden Schriftzeichen wird zuerst gesamt, dann im Einzelnen bestimmt. Die das Schriftzeichen repräsentierende Taste und dann die Tasten für den ersten, zweiten und letzten Strich des Schriftzeichens werden nach der Reihenfolge eingetippt (vgl. Wu 1999: 45). Bspw. lautet der Inputcode von 雨 (/yǔ/, *Regen*) FGHY: F für das gesamte Zeichen: G für den ersten Strich *Héng*, H für den zweiten Strich *Shù* und Y für den letzten Strich *Diǎn*). Falls das einzugebende Zeichen nur aus einem einzelnen Strich besteht, wie 一 (/yī/, *eins*) und 乙 (/yǐ/, *zweite*), wird die entsprechende Taste zweimal gedrückt, gefolgt von zwei weiteren Tastenschlägen auf L (vgl. ibid.). Der Code für das Zeichen für *Eins* ist deswegen GGLL.

Zur Effizienzerhöhung werden abgekürzte Inputcodes für häufige Schriftzeichen des ersten (mit einmaligen Tippen), zweiten (zweimaliges Tippen) und des dritten Grads angeboten. Das häufigste Schriftzeichen, das mit einem Tastenanschlag oder einer Folge von zwei sowie drei Tastenanschlägen abgerufen werden kann, wird durch statistische Analysen festgelegt. Solche Schriftzeichen sind je nach zugehörigem Grad im abgekürzten Inputcode plus einmal Space einzugeben. Die 25 Schriftzeichen vom ersten Grad werden in der letzten Spalte von Tab. 4-1 angegeben. Es gibt ca. 600 Schriftzeichen des zweiten Grads, (prinzipiell 25×25=625-25, manche Zusammensetzungen haben keine entsprechenden Schriftzeichen), deren Inputcodes in zwei Buchstaben abgekürzt werden können. Auf dem dritten Grad gibt es ca. 4.400 Schriftenzeichen, deren Eingabe mit drei Buchstaben durchführbar ist (vgl. ibid.: 47f). Zur Erklärung wird nachfolgend für jeden Grad ein Beispielzeichen angenommen: Das Zeichen auf dem ersten Grad 和 (/hé/ – *und*; /huò/ – *mischen* usw.; Wubi-Code: THG) wird nach dem Eintippen von dem ersten Buchstaben T in der Wahlliste an vorderster Stelle angezeigt, durch das Drücken der Spacetaste kann schnell darauf zugegriffen werden. Mit demselben Prinzip kann das dem zweiten Grad zugehörige Zeichen 行 (/xíng/ – *gehen*; /háng/ – *Zeile, Bank*; Wubi-Code: TFHH) mit TF eingegeben werden. Zur Eingabe von 德 (/dé/, *Moral*)

auf dem dritten Grad (Code: TFLN) werden ebenfalls nur die ersten drei Buchstaben gebraucht.[174]

Um die einzutippenden Inputcodes weiter zu reduzieren und die Effizienz zu erhöhen, wird die Möglichkeit der Eingabe im Wort angeboten, dessen Inputcode auch mit vier Buchstaben repräsentierbar ist. Alle Wörter aus der Wörterdatenbank der Wubi-Eingabesoftware können mit dieser Möglichkeit codiert und mit hohem Tempo eingegeben werden. Wenn ein Wort aus zwei Zeichen zusammengesetzt ist, werden die ersten zwei Buchstaben von dem Inputcode der beiden Zeichen zu dem Inputcode des Wortes kombiniert. Der Inputcode des Wortes 德国 (/déguó/, *Deutschland*) bspw. lautet TFLG (德: TFLN und 国: LGYI). Wenn ein Wort aus drei Zeichen besteht, leitet sich der Wortcode des ersten Buchstaben von dem Inputcode der ersten zwei Schriftzeichen plus der ersten zwei Buchstaben des Zeichencodes des letzten Schriftzeichens ab. Der Ortsname 黑龙江 (/hēilóngjiāng/, *Heilongjiang* [eine Provinz sowie ein Fluss Chinas]) kann deswegen mit dem Inputcode LDIA eingegeben werden (黑: LFOU, 龙: DXV und 江: IAG). Regel zur Bestimmung des Codes für ein Vier-Zeichen-Wort ist, den ersten Buchstaben jedes Zeichencodes nachfolgend zusammenzusetzen. Das Idiom 掩耳盗铃 (/yǎn'ěr dàolíng/, *mit bedeckten Ohren eine Glocke stehlen ~ sich selbstbetrügen*) ist mit dem Code RBUQ (掩: RDJN, 耳: BGHG, 盗: UQWL, 铃: QWYC) einzugeben. Bei einem Wort mit mehr als vier Zeichen werden nur die ersten drei und das letzte Schriftzeichen berücksichtigt, d.h. der Wortcode besteht aus dem ersten Buchstaben der berücksichtigten vier Zeichencodes (vgl. Wu 1999: 48). Die offizielle, vollständige Bezeichnung China aus sieben Zeichen – 中华人民共和国 /zhōnghuá rénmín gònghé guó/ – wird mit KWWL (中: KHK, 华: WXFJ, 人: WWWW, [民/共/和 unberücksichtigt] 国: LGYI) codiert.[175]

Wegen Einschränkungen der Komponentencodierung hat die Wubi-Eingabemethode unvermeidbare Nachteile. Unter den nach Zeichenstruktur entworfenen Eingabemethoden aber ist Wubi-Zixing eine der beliebtesten Varianten. Ihre hohe Effizienz ist einer der wichtigsten Faktoren. Vor allem ermöglicht die Wubi-Eingabemethode eine hohe Schreibgeschwindigkeit, und das aus fünf Gründen: Erstens wird das Tastaturlayout rational nach dem Zehnfingersystem entworfen, damit man schneller und angenehmer tippen kann. Zweitens erfordert der Inputcode für ein Schriftzeichen im Höchstfall nur ein vier Tastenanschläge. Drittens gibt es kaum Code-Überschneidungen, so dass nur selten eine manuelle Auswahl getroffen werden muss. Viertens gibt es für häufige Schriftzeichen abgekürzte Inputcodes, so dass sie mit ein

[174] Nach Eingabetest mit der Sogou-Wubi-Eingabemethode.
[175] Nach Eingabetest mit der Sogou-Wubi-Eingabemethode.

bis drei Tippbewegungen eingegeben werden können. Zuletzt kann ein aus mehreren Zeichen zusammengesetztes häufiges Wort ebenso mit vier Buchstaben (viermaliges Eintippen) verarbeitet werden. Nach fachkundigem Erwerb und ausgiebigem Training können mit der Wubi-Eingabemethode 120 bis 160 Zeichen pro Minute eingegeben werden (vgl. Yin 1998: 167). Solche an höherer Effizienz orientierten Designideen bedingen, dass die Wubi-Eingabemethode unter den beruflichen Schreibkräften in Festlandchina besonders beliebt ist: Ihre Aufgabe ist das Nachschreiben vorhandener Texte per Computer, weshalb ein hohes Tempo das wichtigste Kriterium für die Auswahl der angewendeten Eingabemethode ist. Zudem ist die Eingabe solcher Schriftzeichen, deren Aussprache und Sinninhalt zwar unbekannt ist, deren Zeichenform aber auf Papier geschrieben steht, per Wubi-Eingabemethode kein Problem (vgl. Hou 1999: 81). Für Leute, die mit Computern Texte verfassen, ist die Methode hingegen suboptimal. In derlei Situationen ruht der Fokus auf dem Inhalt eines Textes, weshalb das Schreibtempo vom Verfassungstempo abhängt. In diesen Fällen wird eine Eingabemethode bevorzugt, deren Inputcodierung leichter, umgangsfreundlicher und schwerer zu vergessen ist und zur sprechsprachlichen Gewohnheit passt. Die nach der Aussprache entworfenen Pinyin- oder Zhuyin-Eingabemethoden sind für solche PC-Benutzer die bessere Wahl. Im nächsten Kapitel werden die für standardchinesische Aussprachen designten Eingabemethoden am Beispiel von Pinyin erforscht.

4.1.3 Vorstellung der Pinyin-Eingabemethoden: Allgemeines, Mängel, Codeformen und Entwicklungsphasen

Da sich die Aussprache der chinesischen Schrift in den verschiedenen Sprachen und Dialekten unterscheidet, ist eine nach der Aussprache entworfene Eingabemethode regional und sprachlich eingeschränkt. Im chinesischen Kulturkreis basieren die meisten phonetischen Eingabemethoden auf dem modernen Standardchinesisch Putonghua, welches die offizielle Sprache in Festlandchina, Taiwan und vieler Auslandchinesen ist. Für manch andere chinesische Zweigsprache werden spezielle Eingabemethoden eingesetzt, wie die des Kantonesischen in Hongkong und Guangdong und die Wu-Eingabemethode in Jiangsu, Zhejiang und Shanghai. Die auf Putonghua bezogenen Inputcodierungen gliedern sich hauptsächlich in Pinyin in Festlandchina und Zhuyin in Taiwan (siehe Kap. 3.4.3, S. 168ff).[176] Die Pinyin-Eingabemethoden werden fürs Schreiben am häufigsten gebraucht. Das verwendete Tastaturlayout ist – außer bei manchen Interpunktionszeichen – überwiegend identisch mit der US-amerikanischen Tas-

[176] Pinyin-Eingabemethoden umfassen im weiteren Sinne alle chinesischen phonetischen und im engeren Sinne nur die Pinyin-Eingabemethoden; in dieser Arbeit ist der Begriff *Pinyin-Eingabemethode* hauptsächlich im engeren Sinne gemeint.

tatur, so dass zur Verwendung der Pinyin-Eingabemethoden keine spezielle maschinelle Tastatur vonnöten ist. Die phonetischen Eingabemethoden am Beispiel von Pinyin können wie folgt definiert werden:

> Die phonetischen Eingabemethoden sind die Eingabeverfahren, die aufgrund von chinesischen phonetischen Attributen, nämlich Initial, Final und Ton, entworfen werden und bei denen die chinesischen Schriftzeichen mittels Voll-, Doppel-, Halb- und Misch-Pinyin eingegeben werden. In dem Verarbeitungsprozess lassen sich die homophonetischen Schriftzeichen sowie Wörter durch die Erkenntnisse der Wörter, der Phrasen, der Grammatik, der Semantik und der Pragmatik computergestützt unterscheiden.
>
> (Zhang ZC 1991: 191 [Übersetzung der Verfasserin)

Das Hauptprinzip der Pinyin-Eingabemethode ist die Konversion von Pinyin zu Schriftzeichen in der Einheit von Zeichen, Wort, Phrase oder Satz. Dieser Prozess wird Laut-Zeichen-Konversion (auch Silben-Zeichen-Konversion, chi.: 音字转换, eng.: syllable-to-word/STW conversion) genannt. Mit Unterstützung von Konversionslexika, Wissensdatenbanken usw. werden mögliche Kandidaten angeboten und die einzugebenden Schriftzeichen schließlich vom PC-Benutzer manuell unter den Kandidaten ausgewählt. Die Entwicklungsphasen seit der ersten Pinyin-Eingabemethode in den 1970er Jahren bis heute lassen sich anhand der einzugebenden Einheiten ablesen: beginnend mit der traditionellen Variante in einzelnen Schriftzeichen bis zu der modernen intelligenten Variante im Satz (vgl. Peng 1994: 96ff).

Wegen der großen Vorzüge von Pinyin als Inputcodierung waren und sind Erforschungen und Verbesserungen der Pinyin-Eingabemethoden Schwerpunkte für die Computerunternehmen inner- und außerhalb Chinas. Die älteste Version, die hauptsächlich in der Einheit von einzelnen Schriftzeichen verarbeitet werden konnte, wurde ca. bis in die 1990er Jahre verwendet. Sie hatte vor allem die folgenden vier Beeinträchtigungen (vgl. Hou 1999: 89):

1) Zu lange Buchstabenfolge in vollständiger Pinyin-Form für manche Schriftzeichen. Die längste Pinyin-Kette einer Silbe besteht aus sechs Buchstaben. So muss man z.B. für die Eingabe <双> (/shuāng/, *Doppel*) außer den Buchstaben für Initial und Final zusätzlich einmal Tontaste sowie Enter- oder Spacetaste als Endmarkierung eintippen, was insgesamt acht Tastenanschläge notwendig macht.
2) Sehr häufige Überschneidung der möglichen Kandidaten. Wie voranstehend skizziert wurde, lassen sich die Homophone einer unbetonten Silbe unter den gemein gebräuchlichen Zeichen auf durchschnittlich 19,53 und maximal über hundert beziffern (siehe Kap. 3.4.2, S. 167). Für ein vergleichsweise selten benötigtes Zeichen muss man sich häufig durch die

Auswahlliste scrollen. Dies verlangsamt einerseits die Eingabeeffizienz und stellt andererseits eine kognitive Zusatzbelastung im Schreibprozess dar.

3) Verwendungsschwierigkeiten wegen nichtstandardisierter Aussprache. Wie in Kap. 3.1.1 (S. 134) gezeigt wurde, gibt es unter den chinesischen Sprachen verschiedene regionale Sprachen und binnen einer Zweigsprache zudem verschiedene Dialekte. Trotz der schulischen und medialen Verbreitung des Standardchinesischen im gesamten chinesischen Kreis können große Teile der Chinesen nicht dialekt- sowie akzentfrei Putonghua sprechen, insbesondere ältere Leute, Muttersprachler anderer regionaler Sprachen in Südchina, die Völker mancher Nationalminderheiten usw. Ohne perfekte Mandarinkenntnisse ist es unmöglich, in Pinyin immer korrekt zu codieren.
4) Einschränkungen der unbekannten Schriftzeichen mittels Pinyin. Wegen des immensen Zeicheninventars und zahlreicher selten gebrauchter Zeichen gibt es für jeden Chinesen eine große Menge fremder Schriftzeichen. Ohne phonetische Kenntnisse ist ein Schriftzeichen unmöglich per Pinyin-Eingabemethode einzugeben.

Die Verbesserung und Weiterentwicklung von Pinyin-Eingabemethoden orientiert sich hauptsächlich auf die Verringerung dieser Nachteile. Zur Lösung des ersten Nachteils (lange Pinyin-Ketten) sind verschiedene Formen der Pinyin-Codierung entstanden: Voll- (chi.: 全拼, eng.: full pinyin), Doppel- (双拼, double pinyin), Halb- (简拼, half pinyin) und Mischpinyin (混拼, mix pinyin) (vgl. Zhang ZC 1991: 191).

Vollpinyin bedeutet die vollständige Form für eine sprachliche Silbe mit oder ohne Tonmarkierung. *Doppelpinyin* ist im Prinzip ähnlich zu Fanqie, nämlich der Initial-Final-Kombination, die in Kap. 3.4.3 (S. 168) angerissen wurde. In der dort zu sehenden Tab. 3-4 wurden die 22 Möglichkeiten für Initial (inklusive 21 Konsonanten und der leeres Initial) und 36 Möglichkeiten für Final (Varianten mit derselben Pinyinform werden als eine Möglichkeit definiert) angegeben. Beim Doppelpinyin-Tastaturlayout wird eine Taste zwei- bis dreimal mit Initial- sowie Finallauten belegt, so dass alle Silben durch zweimaliges Tastentippen eingegeben werden können (vgl. Lunde 2009: 301). Die Umschalttaste wird nicht benötigt und die eingegebene Silbe wird vom System automatisch anerkannt. Durch Anwendung des Doppelpinyins kann einerseits die Effizienz erhöht werden, da weniger Eintippbewegungen vonnöten sind. Zudem werden Silbensegmentierungsfehler vermieden, weil jede Silbe durch zwei Tasten repräsentiert wird, wenn mehr als eine Silbe (nämlich ein Zeichen) auf einmal eingegeben wird. Das Tastaturlayout des Doppelpinyins wurde aufgrund der ca. 400 Silben im Putonghua entworfen, d.h. zwei Finallaute, die mit demselben Initial kombinierbar sind, müssen

auf verschiedene Tasten belegt werden, so dass die Folge zweier Tasten einer einzigen Silbe entsprechen kann. Eine verbreitete Tastaturbelegung für Doppelpinyin wird in Abb. 4-3 geschildert.[177] *Halbpinyin* meint die abgekürzte Form einer Silbe und ist im Normalfall der Initial oder der erste Buchstabe des Pinyins. Das Halbpinyin des Zeichens <双> (/shuāng/, *Doppel*) ist deswegen ‚SH' oder ‚S'. Die starke Verkürzung verursachte eine viel häufigere Kandidatenüberschneidung. Deswegen wird Halbpinyin im Normalfall nur für die Eingabe von längeren Wörtern (von mehr als drei Zeichen) bis zu einem Satz verwendet, damit die Kandidaten per Kontextanalyse gefiltert werden können. *Mischpinyin* bedeutet die gemischte Form von Voll- und Halbpinyin bzw. von Doppel- und Halbpinyin. Da für Voll- und Halbform verschiedene Tastaturlayouts gebraucht werden, ist eine Mischform in diesem Fall unmöglich.

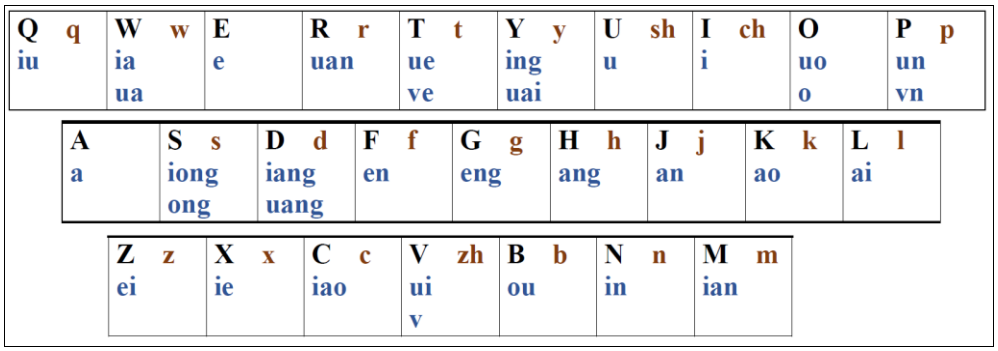


Abb. 4-3: Tastaturbelegung für Doppelpinyin[178]

Die Entscheidung für die Pinyinform ist von Schreibsituationen und -gewohnheiten abhängig. Da im Handel kaum mechanische Doppelpinyintastaturen angeboten werden, muss man die Tastenbelegung auswendig lernen, um mit Doppelform umgehen zu können. Bei den meisten heutigen Pinyin-Eingabemethoden kann der Status der Voll- und der Doppelform nach individuellen Bedürfnissen umgeschaltet werden. Zur Verkürzung der einzugebenden Pinyin-Kette werden sie in den meisten Pinyin-Eingabemethoden ohne Ton verarbeitet. Die Eingabe mit oder ohne Ton hat sowohl Vor- als auch Nachteile. Bei Eingabemethoden mit Toneingabe wird der Ton entweder mit einer Nummerntaste plus Shift (wie ‚Ziguanghuayu-Pinyin-Eingabemethode') oder einem Buchstaben, der nicht im Final auftreten kann (wie ‚Xinhua-Pinyin-Eingabemethode'), repräsentiert. Die erste Variante verlangsamt das Eingabetempo erheblich, während die zweite häufigere Ambiguität bei der Silbensegmentation verursacht, so dass man in vielen Fällen bei der einzugebenden Pinyin-Kette manuell die Silben trennen muss. Die Miteingabe des Tons verringert effektiv die möglichen Kandidaten. Im Durch-

[177] Für verschiedene Pinyin-Softwares werden unterschiedliche funktionale Doppelpinyin-Tastaturlayouts gebraucht.

[178] Quelle: *http://img10.3lian.com/edu201303/f/f103/201303/554c88fa4286f54d62b047d95ab63fee.gif* [Abruf: 2016-05-13]; das leere Initial wird bei der Eingabe mit der Taste ‚E' oder ‚O' repräsentiert. ‚W' und ‚Y' werden wie in Pinyin als Vertretungsinitial für die mit [u], [i] und [y] angefangenen Silben gebraucht. ‚V' steht für ‚Ü'.

schnitt betragen die Kandidaten mit Ton 31,16% von den unbetonten Kandidaten (insgesamt 415 Silbenvarianten ohne Ton und 1.332 mit Ton in Mandarin. Unter den 6.763 Schriftzeichen des nationalen Standards GB2312-1980 hat eine betonte Silbe in Vollpinyin durchschnittlich 5,08 Kandidaten. Aus pädagogischen Ansichten, besonders für Grundschüler und Chinesisch-als-Fremdsprachlerner, ist es sinnvoll, den Ton bei der Eingabe zu berücksichtigen, da der Ton einer der wichtigen Attribute des Sprach- sowie Zeichenerwerbs ist.

Die Hauptlösung des zweiten Nachteils, nämlich die Überschneidung zahlreicher Homophone, liegt in computergestützten Analysen im sprachlichen Kontext. D.h. der Pinyin-Code wird in der Einheit von Wort, Phrase oder Satz eingegeben und mithilfe von Wissensdatenbanken mit linguistischen Informationen sowie statistischen Analysen verarbeitet. Homophone werden dadurch ausgefiltert (vgl. Zhang ZC 1991: 191). Solche von künstlicher Intelligenz unterstützten Pinyin-Eingabemethoden sind heute die allgemein gebräuchlichsten Verfahren zur Eingabe des Chinesischen.

Wie in Kap. 4.1.1 (S. 186) erwähnt wurde, wurden die intelligenten satzstufigen Pinyin-Eingabemethoden auf der Basis von zeichen- und wortstufige Entwicklungsphasen weiterentwickelt. Ein Konversionszeichenlexikon, in dem der Pinyin- und interne Code einzelner Schriftzeichen festgelegt wird, ist die Basis für den Ablauf der Laut-Zeichen-Konversion. In der zeichenstufigen Phase stand außer dem Lexikon kaum eine weitere Ressource zur Verfügung, so dass die Eingabemethoden im alltäglichen Schreiben ineffizient waren. Die Funktionsweise wird in Abb. 4-4 mit dem Beispielzeichen <是> /shì/ abgebildet.

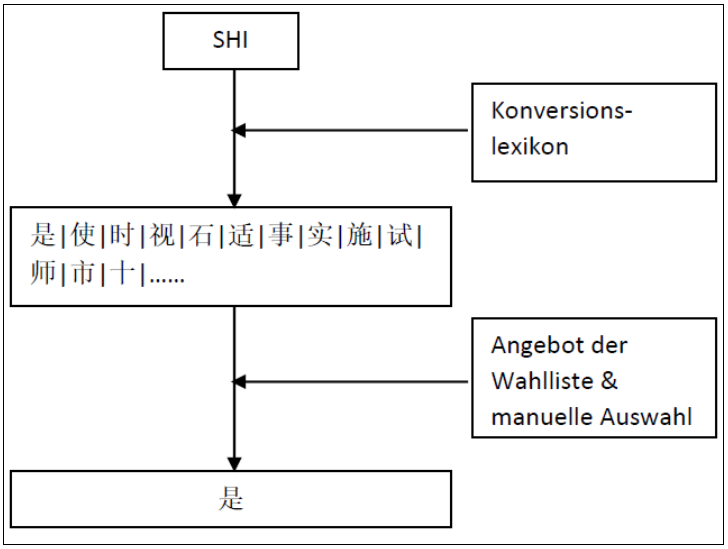


Abb. 4-4: Eingabe in einzelnen Zeichen am Beispiel <是> (/shì/, *sein*)[179]

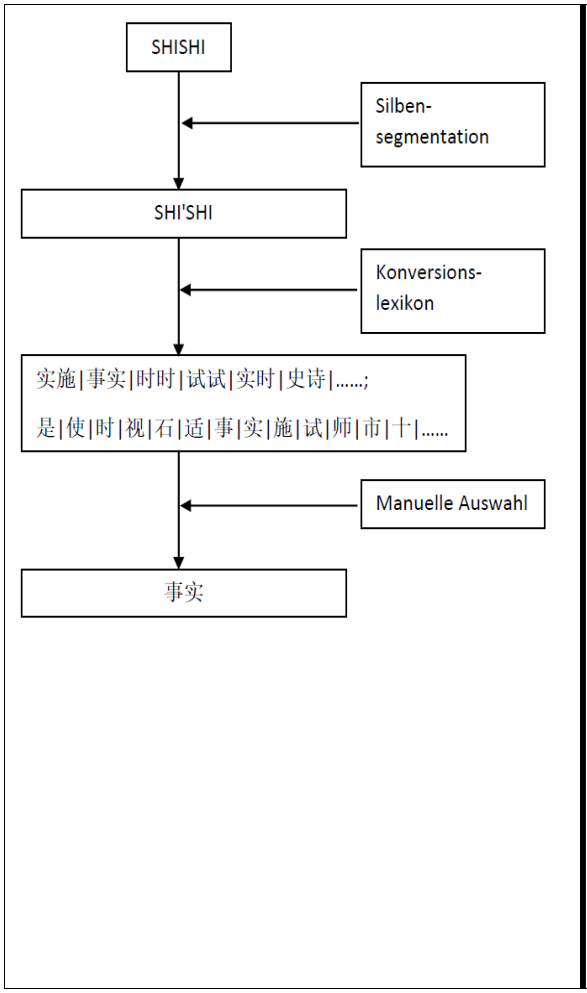
[179] Die Auswahl des ersten Kandidaten der Wahlliste wird mit Leertaste oder Taste ‚1' getroffen; die weiteren Kandidaten können ebenso nach ihrer Nummerierung ausgewählt werden. Mit ‚+' und ‚-, lassen sich die Seiten der Wahlliste umblättern.

Auf der Basis der entwickelten Informationsverarbeitung einzelner Zeichen wurde versucht, Wörter ins Konversionslexikon aufzunehmen. Die Pinyin-Eingabesoftware hatte sich damit zur Phase der Wortverarbeitung entwickelt. Die Sammlung der aufgenommen Wörter wird Wörterdatenbank genannt (vgl. Zhang ZC 1991: 155f). Die Funktionen beinhalten vor allem die Codierung der allgemeinen, speziellen und individuellen Wörter, die Rangliste nach Worthäufigkeit, die Liste der häufigen Wörter und die automatische Speicherung von neuen, nicht-eingetragenen Wörtern. D.h. die allgemein und idiolektal häufiger gebrauchten Wörter werden unter den homophonetischen Wörtern weiter vorne in der Wahlliste angezeigt, so dass sich die Auswahlzeit verkürzt (vgl. Peng 1994: 97). Die Funktionsweise der Eingabe in Wort wird in Abb. 4-5 mit dem Beispielwort <事实> /shìshí/ geschildert. Da es zahlreiche aus einzelnen Zeichen bestehende Wörter gibt und die homophonetischen Wörter zudem in großer Anzahl vorliegen, konnte die Eingabe in Worteinheit noch immer kein hohes Schreibtempo erreichen. Nach Erforschungen jener Wörter, die im Wörterbuch Xiandai Hanyu Cidian (4. Aufl.)[180] aufgenommenen wurden, haben monosyllabische Wörter, die 14,4% des gesamten Wortschatzes ausmachen, ohne Tonunterscheidung mit einer 99,9prozentigen Wahrscheinlichkeit Homophone. Die Wahrscheinlichkeit der homophonetischen bisyllabischen Wörter (68,1% des gesamten Wortinventars) macht ohne Tonunterscheidung 68,1% aus. Mit Tonunterscheidung entsprechen die beiden Wörtergruppen einem Anteil von jeweils 97,6% sowie 16,1% (vgl. Ding/Huang 2009: 136). Diese Statistiken weisen darauf hin, dass die Eingabe in der Einheit des Wortes ebenfalls nicht den Nachteil der vielen homophonetischen Kandidaten effizient vermeiden kann.

Die Entwicklung der Informatik und Computerlinguistik hat die hochintelligente Laut-Zeichen-Konversion in langen Wortketten oder im Satz ermöglicht. Eine Eingabesoftware mit der Möglichkeit zur satzstufigen Laut-Zeichen-Konversion (chi.: 语句级音字转换) wird als intelligente Eingabesoftware (智能输入软件) bezeichnet. Wie in Kap. 4.1.1 (S. 186) erwähnt, ist eine Wissensdatenbank mit sprachlichen Kenntnissen die obligatorische Ressource der satzstufigen Laut-Zeichen-Konversion. Die Wörterliste, die zur wortstufigen Konversion benötigt wird, muss in dieser Entwicklungsphase zu einem maschinenlesbaren Wörterbuch upgedatet werden, in dem die Attribute einzelner Wörter/Zeichen (inkl. Wortart, Wortbedeutung usw.) eingeschrieben werden. Das Wörterbuch und eine linguistische Datenbank bilden zusammen die Wissensdatenbank, um automatische sprachliche Analysen unterstützen zu können. In der linguistischen Wissensbank werden die syntaktischen, semantischen sowie prag-

[180] Originaltitel: 现代汉语词典 /xiàndài hànyǔ cídiǎn/, wörtl.: *Das Wörterbuch des modernen Chinesischen*, 2002 herausgegeben von ‚Commercial Press'.

matischen Regeln über Phrasen- sowie Satzbildung in maschinenlesbare Formeln verdatet (vgl. Hou 1999: 131, Zhang S et al. 1997: 38ff). Der Sinninhalt eines einzelnen Zeichens ist häufig abstrakt und ambig, aber im sprachlichen Kontext eindeutig (vgl. Kap. 3.4.1, S. 164f). Die Transkription in Phrase- sowie Satzeinheit kann deswegen als kombinierte Inputcodierung nach phonetischem und begrifflichem Aspekt verstanden werden. Die von einem einzelnen Zeichenattributaspekt verursachten Ambiguitäten – Homophone und Synonyme – können gegenseitig disambiguiert werden. Die Eingabe im Satz mit dem Beispielsatz <这是事实> /zhè shì shìshí/ wird in Abb. 4-6 bildhaft erklärt.



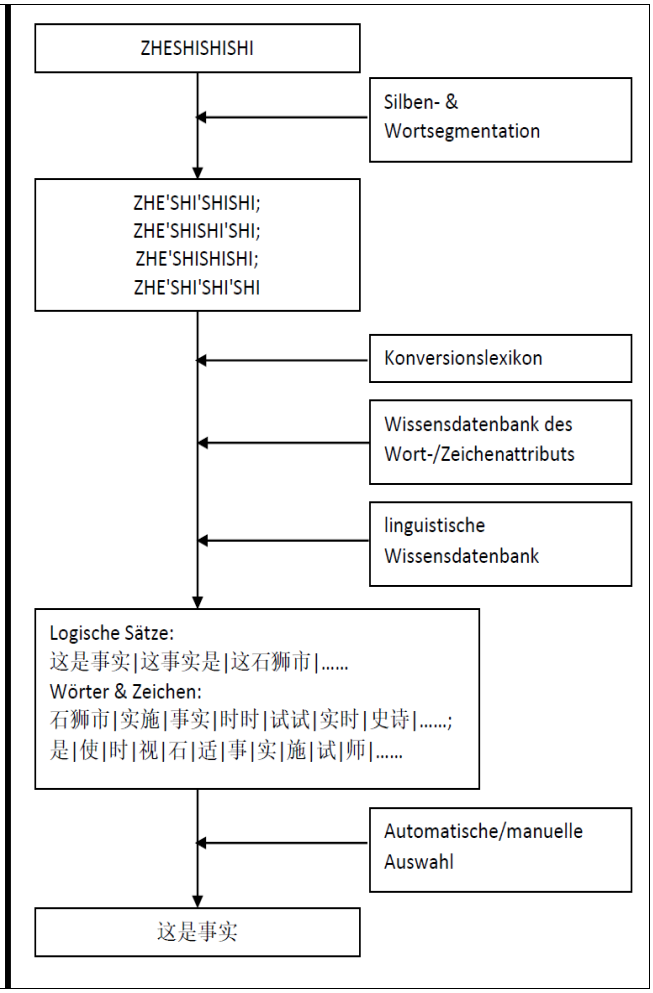


Abb. 4-5 [links]: Eingabe in Wort am Beispiel von <事实> (/shìshí/, *Tatsache*)[181]

Abb. 4-6 [rechts]: Eingabe im Satz am Beispiel <这是事实> (/zhè shì shìshí/, *Das ist Tatsache*)[182]

[181] Die Kandidaten werden von den längsten zu den kürzesten und von den häufigeren zu den seltenen Wörtern angeordnet. In diesem Fall wird das statistisch allgemein am häufigsten oder zuletzt in den PC eingegebene Wort zuvorderst angezeigt. Wenn kein Wort der Wahlliste korrekt wäre, würde Zeichen für Zeichen einzeln ausgewählt.

[182] Vgl. Chen/Zhu 2002b: 13. Diese Abbildung bezieht sich auf eine auf dem sprachlichen Verstehen basierende Methode (vgl. Kap. 4.2.2). Die Reihenfolge der Kandidaten behält dasselbe Prinzip wie in Abb. 4-5.

Welche sprachlichen Kenntnisse bei dem Entwurf der intelligenten Eingabesoftwares gebraucht und welche Laut-Zeichen-Konversionsmethoden im Satz entwickelt werden, wird in Kap. 4.2 erläutert. Neben der Realisierung der Satzverarbeitung wurden Pinyin-Eingabesoftwares auch in vielen anderen Details verbessert. Solche Sonderfunktionen der intelligenten Eingabesoftwares werden im nächsten Kapitel weiter vorgestellt.

4.1.4 Vor- sowie Nachteile und auf künstlicher Intelligenz basierende Funktionen der intelligenten Eingabesoftwares

Die drei Entwicklungsphasen und die heutige Situation der Pinyin-Eingabemethode lassen sich aufbauend auf dem voranstehenden Kapitel insofern zusammenfassen, als dass die Zeichenverarbeitung die Basis darstellt, die Wortverarbeitung eine dominierende Rolle einnimmt und die Satzverarbeitung das Ziel ist (vgl. Yuan 2010: 11). Im Vergleich zu der Eingabe in Zeichen-/Worteinheit hat die satzstufige Eingabe – neben dem geringen Ausmaß möglicher Überschneidungen – vor allem zwei Vorteile (vgl. Peng 1994: 98f):

1. Natürlicher kognitiver Zugang und leichteres Schreiben. Ein Satz wird beim Schreiben von Texten häufig als syntaktisch zusammenhängende Wortfolge in phonetischer Art im Kurzgedächtnis vorformuliert. Schreibt man auf Chinesisch, so denkt man nicht an die Zeichenstruktur, sondern an Konzept und Aussprache des gemeinten Wortes. Mit der satzstufigen Verarbeitung ist es möglich, den Inputcode für den gesamten Satz auf einmal einzutippen, ohne einen Gedanken auf die Zeichenform zu verschwenden. Die sprachlichen Analysen anhand Pinyin-Satz ermöglichen eine hohe Korrektheit der Laut-Zeichen-Konversion, so dass ein PC-Benutzer die Zeit für manuelle Auswahlen sparen und sich auf den Inhalt des Textes konzentrieren kann.
2. Weniger Eintippen und hohe Effizienz. Zeichen und Wort müssen bei der Eingabe normalerweise in vollständige Silben (Voll- oder Doppelpinyin) codiert werden, damit die Überschneidung unter den Homophonen begrenzt bleibt. Durch automatische Analysen im Satzkontext könnte die Überschneidung unter den Schriftzeichen mit demselben Initial gering sein, so dass die Eingabe in Halbpinyin bei der Satzeingabe effizient sein kann. Das Schreibtempo wird dementsprechend dank der Reduzierung von Eintippbewegungen und Einsparungen bei der Kandidatenauswahl erhöht.

Am Anfang von Kap. 4.1.3 (S. 196f) wurden die vier größten Mängel von Pinyin-Eingabemethoden genannt. Mit der satzorientierten intelligenten Eingabemethode sind der erste und zweite Nachteil im Normalfall marginalisiert. Um den dritten und vierten Nachteil

zu verringen (die Erforderlichkeit von Mandarinkenntnissen und die Begrenzung beim Schreiben von unbekannten Schriftzeichen), müssen weitere Funktionen entwickelt werden.

Ohne verbesserte Funktionen kann ein Tippfehler bei Pinyin negative Folgen bei der Satzeingabe verursachen, denn die Korrektur innerhalb einer Pinyin-Kette für einen Satz dauert länger als die Eingabe eines Zeicheninputcodes, besonders wenn sich der Tippfehler am Anfang der Pinyin-Kette befindet. Normalerweise muss die Schreibmarke Buchstabe für Buchstabe rückwärts bewegt werden, um den Fehler zu korrigieren. Bei manchen intelligenten Eingabesoftwares ist die Rückwärtssetzung zuerst in gesamtzerlegte Silben möglich, um die Zeit für Korrekturen zu sparen. Trotzdem kostet die Korrektur in vielen Fällen kaum weniger Zeit als erneutes Eintippen. Der folgende Fehler ist hierfür exemplarisch. Er resultiert durch die Unterschiede des deutschen Tastaturlayouts; das Eingabeziel entspricht Abb. 4-7.

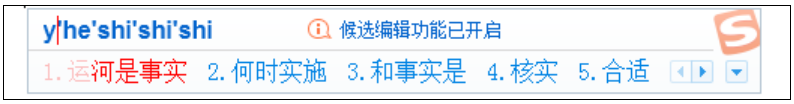


Abb. 4-7: Korrektur des Tippfehlers von ‚Y' zu ‚Z' bei Satzeingabe

Da die Tippfehler unter den phonetischen Eingabemethoden oft passieren, wurden Techniken der automatischen Korrektur und das sog. verschwommene Pinyin entwickelt. Fehlergründe sind hauptsächlich Achtlosigkeit, Schreibgewohnheit und Aussprachedifferenzen in verschiedenen Regionen. Da der schriftliche Verkehr der chinesischen Sprache grundsätzlich nur in der chinesischen Schrift funktioniert, wird Pinyin im Alltag (außer beim Computerschreiben und zur phonetischen Angabe von unbekannten Zeichen) wenig benutzt. Zugleich konzentrieren sich Chinesen beim Schreiben hauptsächlich auf die Korrektheit der Schriftzeichen. Im Gegenteil fehlt es ihnen an Bewusstsein für die ‚Pinyin-Rechtschreibung'. Die sprachlichen sowie dialektalen Varianten der chinesischen Sprachen sind ein weiterer Hauptverursacher für Tippfehler beim Umgang mit Pinyin-Eingabesoftwares. Wie in Kap. 3.1.1 (S. 134) gezeigt, könnte dasselbe Schriftzeichen bei verschiedenen Zweigsprachen sowie Dialekten verschieden ausgesprochen werden. Bei einer phonetischen Eingabesoftware hingegen muss in einer standardisierten Transkription inputcodiert werden (vgl. Wang XL 2005: 108).

Die Funktion ‚verschwommenes Pinyin' (模糊音, eng.: fuzzy Pinyin/ fuzzy sounds) ist eines der wichtigsten Hilfsmittel für PC-Benutzer mit unstandardisierter dialektaler Aussprache bei Tippfehlern. Es bedeutet, dass zwei Phoneme, die im Allgemeinen häufig verwechselt werden, von dem System ohne Unterscheidung behandelt werden können. So kann das richtige Schriftzeichen trotz der falsch eingetippten Pinyin-Kette in der Wahlliste angeboten werden (vgl. Wu 1999: 24). Die Entwicklung dieses Hilfsmittels basiert auf der Erforschung typi-

scher Fehler spezifischer Sprechergemeinschaften, die aus Einflüssen des gesprochenen Dialekts oder der regionalen Sprachen resultieren. In Wu-Gebieten hat man z.B. Schwierigkeiten, /z/, /c/ und /s/ von /zh/, /ch/ und /sh/ zu unterscheiden. In Sichuan, Hunan, Guizhou usw. diskriminieren und produzieren viele /l/ und /n/ mit kaum Abweichung. In Fujian hat man Probleme bei der Unterscheidung zwischen /f/ und /h/ sowie /r/ und /l/. In großen Teilen Südchinas ist die Abweichung von den beiden Finallauten /n/ und /ng/ kaum zu bemerken. Solche PC-Benutzer haben bei der Verwendung der modernen Pinyin-Eingabemethode die Möglichkeit, die verschwommenen Laute nach ihren Bedürfnissen einzustellen. Abb. 4-8 zeigt jenen Einstellungsabschnitt der Sogou-Eingabesoftware, in dem die PC-Benutzer akzentabhängig die gewünschten verschwommenen Laute auswählen können.

Abb. 4-8: Einstellung des verschwommenen Pinyin der Sogou-Pinyin-Eingabesoftware

Die Funktion ‚verschwommenes Pinyin' ist zwar hilfreich für die durch Dialekte verursachten Tippfehler, bei sonstigen Fehlermöglichkeiten gilt dies jedoch nur eingeschränkt. Die automatische Pinyin-Korrektur-Funktion wird deswegen in die meisten intelligenten Eingabesoftwares programmiert. Der Kern dieser Technik ist die Kombination des Sprach- und Tippmodells. Sprachmodell meint die korrekte Pinyin-Umschrift von chinesischen Texten, die auf Korporaanalysen begründet wird. Tippmodell basiert auf Analysen von Tippfehlern bei dem Eingabeprozess mit Pinyin-Eingabesoftwares. Findet sich ein typischer Fehler im Tippmodell, so wird er automatisch anhand des Sprachmodells korrigiert, bevor passende Zeichenkandidaten angeboten werden (vgl. Wang XL 2005: 108).

Um die vierte Einschränkung der Pinyin-Eingabemethoden – die Schwierigkeiten bei der Eingabe unbekannter Schriftzeichen – zu ergänzen und zu überwinden, können die Eingabemöglichkeiten per Handschreiben und per Zeichenform in einer intelligenten Eingabesoftware designt werden. Handschriftliche Eingaben der intelligenten Eingabesoftware werden unter dem Terminus der Online-Handschrift-Eingabemethode subsumiert (siehe Kap. 4.1.1, S.

179f). Die manuelle Eingabe des Beispielzeichens <你> (/nǐ/, *du*) mit der Sogou-Pinyin-Eingabesoftware zeigt die folgende Abbildung:

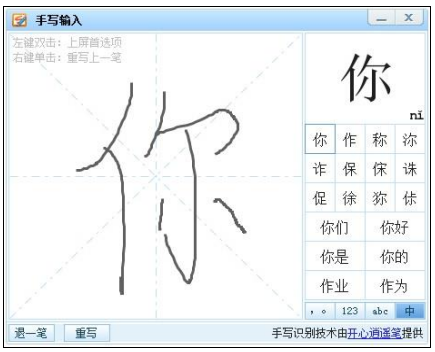


Abb. 4-9: Eingabe per Online-Handschreiben mit Sogou-Pinyin-Eingabesoftware[183]

Die zeichenformbasierten Eingabealternativen einer intelligenten Pinyin-Eingabesoftware funktionieren entweder nach der Codierung der Striche (meistens codiert mit dem ersten Buchstabe der Bezeichnung des Strichs) oder der Aussprache der bestehenden Komponenten/Zeichen (vgl. Yuan 2008: 279). Bei der Sogou-Eingabesoftware wird diese alternative Codierung von dem Buchstaben ‚U' eingeleitet, der nicht Anfangsbuchstabe einer Pinyin-Kette sein kann. Dasselbe Beispielzeichen kann durch Codierung in Hauptstrichmerkmale als ‚UPSPZSDD' codiert werden (U als Einleitung, P: 丿 /piě/, S: 丨/亅 /shù/ und /shùgōu/, Z: 乛 /zhé/, D: 丶 /diǎn/, nach der standardisierten Strichfolge des Schreibens). Das komplexe Schriftzeichen kann ebenso in die Komponenten 亻 (von <人> /rén/, *Mensch*) und 尔 (/ěr/, veraltetes Wort für *du*) zerlegt und in deren Aussprache codiert werden. So lautet die Inputcodierung dieser Alternativeingabemethode ‚URENER' (U als Einleitung, REN für die erste und ER für die zweite Komponente).[184]

Zusammengefasst wurden im vorderen Text hauptsächlich vier Funktionen der intelligenten Pinyin-Eingabesoftware vorgestellt: 1) es können verschiedene Einheiten wie Satz, Phrase, Wort oder Zeichen verarbeitet werden; 2) die verschiedenen Inputcodierungen, Voll-, Doppel-, Halb-, Mischpinyin, Strich- und Zeichenform-Phonetik-Kombination können eingegeben werden; 3) eine Wissensdatenbank wird begründet, welche Zeichenkandidaten durch linguistische Erkenntnisse ausfiltert; 4) für mögliche Tippfehler bei der Pinyin-Kette werden die Funktionen verschwommenes Pinyin sowie die automatische Pinyin-Korrektur angewendet (vgl. Zhou/Zhu 2002: 42). Weiterhin verfügt eine intelligente Pinyin-Eingabesoftware über verschiedene wichtige Funktionen, deren beide bedeutsamsten nachfolgend dargelegt werden.

[183] Nach Eingabetest mit Sogou-Pinyin-Eingabemethode.
[184] Nach Eingabetest mit Sogou-Pinyin-Eingabemethode.

1. Prognose-Funktion (eng.: prediction function).

Ein vollständiges langes Wort oder eine satzwertige volkstümliche Aussage (Phraseologismus) können als Kandidaten vom Programm vorhergesagt werden, wenn der Pinyin-Code nur teilweise eingetippt worden ist (ibid.: 43). Wie in Abb. 4-10 für die Eingabe der Vollbezeichnung Chinas (<中华人民共和国> /zhōnghuá rénmín gònghé guó/) aus sieben Schriftzeichen zu sehen ist, müssen nur dreieinhalb Silben[185] eingetippt werden, damit das Programm den richtigen Vorschlag vorgibt. Dasselbe Prinzip gilt auch für das Schreiben des Sprichwortes <说曹操曹操到> (/shuō cáocāo cáocāo dào/, wörtlich: *wenn man von Cao Cao redet, kommt Cao Cao/* Bedeutung: *jemand kommt zufällig dann, sobald die anderen von ihm reden/wenn man vom Teufel spricht*; Abb. 4-11). Zu dessen Eingabe werden die ersten vier Zeichen in Buchstaben codiert und der vollständige Satz in zwei Varianten (der zweite und der dritte Kandidat) wird blitzschnell ohne weiteres Eintippen angezeigt. Mit dieser Funktion kann einerseits Zeit eingespart werden, andererseits können Tippfehler früher bemerkt und korrigiert werden, wenn sie auftreten.

Abb. 4-10: Wortprognose

Abb. 4-11: Satzprognose

2. Automatische Speicherung von neuen Wörtern und Aktualisierung sowie Begründung der individuellen Wörterdatenbanken (vgl. Zhou/Zhu 2002: 42).

Die zugehörige Wörterdatenbank einer Eingabesoftware kann unmöglich alle Wörter umfassen. Eigennamen aus dem individuellen Bekanntenkreis, Ad-hoc-Bildungen in Medien oder die im Alltag selten gebrauchten fachlichen Wörter werden in den meisten Fällen nicht aufgenommen, obwohl sie individuell, spontan oder für fachliche Arbeiten bestimmter Bereiche häufig gebraucht werden. Zum Gebrauch solcher Wörter ist eine Wörterdatenbank mit weiteren Funktionen obligatorisch: a) Selbstlernfunktion: häufig eingegebene Zeichenfolgen werden automatisch als Wörter gespeichert, wie z.B. <王铠> (der Name der Verfasserin im PC der Verfasserin); b) Aktualisierung via Internet: sie kann sich aktualisieren und aktuelle, medial verbreitete Wörter aufnehmen; c) Angebot der Zellwörterdatenbanken online: ein PC-

[185] Vollform für die ersten drei Zeichen und Halbform für das vierte Zeichen, die sog. ‚halbe Silbe'.

Benutzer kann individuell orientiert Zellwörterdatenbanken von bestimmten Fachgebieten runterladen, um fachliche Wörter eingeben zu können; d) dynamische Umstellung der Wortfrequenz; e) Verwaltung, Importieren, Exportieren, Datensicherung, Korrektur und Kompatibilität der individuellen Wörterdatenbank, so dass sie immer individuell orientiert und auf anderen PCs weiterverwendet werden kann (vgl. Yuan 2008: 279f, Yuan 2010: 11).

3. Automatische Erkennung von unbekannten Wörtern und Phrasen.

Nach Wortbildungsgrammatiken und bekannten Bestandteilen (Zeichen oder Wortteile) kann das System in vielen Fällen ein unregistriertes Wort automatisch bilden und als Kandidaten anbieten. Mit demselben Verfahren kann ebenso eine Phrase gebildet und erkannt werden (vgl. Zhou/Zhu 2002: 42). Eine ausführlichere Erläuterung über das Verfahren der Wort- sowie Phrasenbildung erfolgt in Kap. 4.4.4 und 4.4.5.

Mit der Verwendung solcher Eingabemethoden von hoher Intelligenz kann das künstliche Gehirn viele Aufgaben des menschlichen Gehirns übernehmen, besonders den Gebrauch der korrekten Zeichen im Kontext. Die Chinesischsprecher profitieren einerseits von der erhöhten Effizienz und Erleichterungen beim Schreiben, verlieren andererseits aber langsam ihre Fähigkeit des Zeichenschreibens und der -verwendung. Nach Statistiken sind Schreibfehler mit der Pinyin-Eingabemethode neunmal häufiger als mit einer zeichenformbasierten Methode (vgl. Wang YM/Yang 2005). Zudem ist Textverfassung per se kreativ und vielfältig, im Gegenteil dazu können nur begrenzte Zeichen, Wörter und grammatische Regeln von dem Computer erkannt werden. Diese auf Wörterbüchern und Wissensdatenbanken basierende künstliche Intelligenz kann in manchen Fällen die Schreibfertigkeit und -kreativität begrenzen. Chinesen, die in Chinesisch lesen, aber Probleme beim Schreiben haben, werden als ‚neue Analphabeten' (新文盲 /xīn wénmáng/) bezeichnet und ihre Anzahl steigt mit der Entwicklung der intelligenten Eingabemethoden (vgl. ibid.). Diese rasante Entwicklung stellt eine neue Krise der chinesischen Schriftkultur dar. Es ist deswegen nicht immer sinnvoll, das Künstlich-Intelligenz-Niveau der Eingabemethoden zu erhöhen und mehr geistige Arbeit der PC-Benutzer von dem Computer übernehmen zu lassen. Sowohl bei der Entwicklung als auch bei dem Gebrauch der intelligenten Pinyin-Eingabesoftware muss ein Gleichgewicht zwischen dem künstlichen und dem menschlichen Gehirn eingehalten werden.

4.2 Methoden und linguistische sowie technische Unterstützungen zur Laut-Zeichen-Konversion der intelligenten Pinyin-Eingabemethoden

Wie in Kap. 4.1.3 erwähnt wurde, kann das Grundprinzip der intelligenten Pinyin-Eingabemethoden wie folgt beschrieben werden: Ein Pinyin-Code wird eingetippt und die dem Code entsprechenden Schriftzeichen werden von Datenbanken abgerufen und nach den sprachlichen Erkenntnissen in der Wissensdatenbank automatisch ausgefiltert (vgl. Zhou YG 1996: 389). Welche linguistischen Informationen vonnöten sind und wie sie in der Wissensdatenbank begründet sowie zur Konversion angewendet werden können, sind deswegen die wesentlichen Aspekte zur Erforschung der chinesischen modernen Eingabemethoden. Kap. 4.2 zielt hauptsächlich darauf ab, eine Übersicht dieser drei Aspekte aufzufächern.

4.2.1 Überblick über die chinesische Grammatik und Grundlagen der intelligenten satzstufigen Pinyin-Eingabemethoden

Wie in Kap. 3 vorgestellt wurde, unterscheidet sich das chinesische Schriftsystem aufgrund des isolierenden Sprachbaus und des morphologischen Schrifttyps (Schriftzeichen als sprachliche und schriftliche Grundeinheit) erheblich von den meisten Sprachen der Welt. Die sprachlichen sowie schriftlichen Einheiten der chinesischen Sprache innerhalb eines Satzes sind (vom unteren bis zum höheren Niveau) innerhalb eines Satzes: Silbe / Schriftzeichen / Morphem ≤ Wort < Phrase < Satz (vgl. Xu YC 2008: 165).[186] Die wesentliche Funktionsweise von intelligenten Pinyin-Eingabesoftwares heißt Laut-Zeichen-Konversion im sprachlichen Kontext, die maximal in der Einheit ‚Satz' ohne Interpunktionszeichen möglich ist. Zusammen mit den klassifizierten sprachlichen Einheiten lassen sich die benötigten linguistischen Informationen in vier Punkten zusammenfassen: 1) der Silben-Zeichen-Morphem-Zusammenhang, der zum linguistischen Teilbereich der Grammatologie gehört (siehe Kap. 3.3 & 3.4); 2) die Wortbildung von einem bis zu mehreren Schriftzeichen/Morphemen, zu deren Analysen die Erkenntnisse aus den Bereichen Wortbildung und Lexikologie nötig sind; 3) die Phrasenbildung von Wörtern, die in den Teilbereichen Lexikologie, Syntax und Semantik untersucht werden; 4) die Satzbildung aus Wörtern sowie Phrasen, für die Syntax, Semantik und Pragmatik benötigt werden (vgl. Lu 2008: 3f, 149, 203 & 233).

Wegen grundsätzlicher sprachlicher Unterschiede hat sich die Sprachwissenschaft in China und Europa in entgegengesetzte Richtungen entwickelt. In indogermanischen Sprachen

[186] Silbe, Schriftzeichen und Morphem bilden aus verschiedenen Perspektiven die Grundeinheiten: betonte Silbe – gesprochene Sprache, Schriftzeichen – verschriftete Sprache, Morphem – Grammatik und Semantik; im Normalfall sind die drei Grundeinheiten miteinander identisch, es gibt jedoch Ausnahmen (vgl. Kap. 3.4.1, S. 163ff).

ist das Wort die kleinste, bedeutungstragende Einheit der Sprache, das in flektierender Form auftreten kann, um grammatische Funktionen zu äußern. Schwerpunkte der Forschungen sind folglich Wort und Satz: Morphologie und Syntax sind zwei der wichtigsten Disziplinen (vgl. Bußmann 2002: 259). Im Gegensatz zu den indogermanischen Sprachen ist im Chinesischen die sprachliche Grundeinheit die nicht-flektierbare betonte Silbe sowie morphologische Schriftzeichen, so dass die historische Sprachwissenschaft in China hauptsächlich auf Grammatologie konzentriert war. Die meisten sprachhistorischen Forschungswerke fokussieren so die Schrift, wobei das bedeutendste Schriftwerk *Shuowen-Jiezi* von Xu Shen ca. 100 n. Chr. entstand. Morphologische und syntaktische Regularitäten wurde hingegen kaum erforscht, bis die modernen linguistischen Theorien im 19. Jh. in China verbreitet wurden (vgl. Xu TQ 2008: 2-5). Nach der traditionellen chinesischen Grammatik unterschieden sich die sprachlichen Einheiten in Zeichen (inkl. Autosemantikum- und Synsemantikum-Zeichen) und Satz (vgl. Lü 1979: 14). Auf Basis der aus dem Westen übernommenen sprachwissenschaftlichen Theorien wurden nun die zwei sprachlichen Einheiten Wort und Phrase der sprachlichen Hierarchie zugeordnet (vgl. ibid.: 14-20).

Wegen der grundsätzlichen Unterschiede bei Sprache sowie Schrift ist es unmöglich, die vorhandenen linguistischen Theorien der indogermanischen Sprachen für die chinesischen linguistischen Forschungen zu imitieren. Das an den Grundeigenschaften der chinesischen Sprache orientierte Theoriensystem ist erforderlich und spielt eine grundlegende Rolle für Fachbereiche wie Sprachdidaktik, Literaturwissenschaft, Computerlinguistik usw. Im Grunde genommen können grammatische Forschungen des Chinesischen zeichenbasiert (字本位; character based) durchgeführt werden, während die indogermanischen Theoriensysteme wortbasiert (chi.: 词本位; eng.: word based) sind (vgl. Lu 2008: 3), da ein chinesisches polysyllabisches Wort meistens aus zwei oder mehreren morphologischen Zeichen besteht.

Grammatologie, Grammatik (inklusive der Wort-, Phrasen-, Satzbildung und Wortart-Klassifikation usw.), Semantik und Pragmatik sind vier wichtige Teildisziplinen der Sprachwissenschaft. Die Grammatologie des Chinesischen wurde in Kapitel 3 vorgestellt. Wie bei der Schriftlinguistik steht die chinesische Sprache bei der Grammatik den indogermanischen Sprachen diametral entgegen. Die Besonderheiten der chinesischen Grammatik können in sechs Punkten zusammengefasst werden (vgl. Lan 2002: 8-13):

1. Einer formale Flexion ist gefehlt. Die Form des chinesischen Wortes bleibt im Sprachgebrauch unverändert. Die grammatischen Funktionen, die in den indogermanischen Sprachen durch Flexion geäußert werden, werden im Chinesischen mit Funktionswörtern,

durch Wortstellung und Affixe ausgedrückt. Anders als in vielen indogermanischen Sprachen gibt es kein offensichtliches Merkmal für die Wortart, so dass die Bestimmung nur durch die syntaktische Funktion und die Wortbedeutung markiert werden kann (vgl. auch Lü 1979: 33f). Kapitel 4.4.1 und 4.4.3 beleuchten die Attribute der Wörter näher.

2. Die Reihenfolge der Wörter/Morpheme ist für die Semantik entscheidend. Im Chinesischen kann eine variierte Reihenfolge derselben Wörter/Morpheme deutlich unterschiedliche Bedeutungen ausdrücken. Bspw. können aus den drei Zeichen <不> (/bù/, *nicht*), <很> (/hěn/, *sehr*) und <好> (/hǎo/, *gut*) drei verschiedene Phrasen erzeugt werden: <不很好> (*nicht sehr gut*), <很不好> (*sehr schlecht*) und <很好不> (*Ist es sehr gut?*).

3. Synsemantika spielen in grammatischen Zusammenhängen eine wichtige Rolle. Sie stehen den Autosemantika gegenüber und sind jene Wörter, die keine lexikalische Bedeutung tragen und nicht eigenständig als Satzglied auftreten können. Da im Chinesischen die Wörter unflektierbar sind, werden zumeist Synsemantika gebraucht, um die grammatischen Zusammenhänge herzustellen. Möchte man bspw. eine in der Vergangenheit liegende Aktion ausdrücken, wird die Partikel <过> /guò/ oder <了> /le/ hinter dem Verb eingesetzt.

4. Es gibt zahlreiche Zahleinheitswörter (ZEW). ZEW sind Wörter, die die Zahleinheit von Substantiven und auch Verben bezeichnen. In den indogermanischen Sprachen werden solche Wörter kaum gebraucht, da ein Numeral oder Artikel direkt Gegenstände modifizieren kann und ein Adverb aus den für Numerale sowie Zahleinheiten stehenden Morphemen gebildet wird (z.B. *vier Bücher*, *das Buch*, *zweimal lesen*). Da ein Nomen im Chinesischen keinen Numerus unterscheidet und es keinen Artikel aber zahlreiche homophonetische Wörter gibt, ist der Gebrauch von ZEW in den meisten Fällen obligatorisch zur Mengenangabe und Determination. Rekurrierend auf die obigen deutschen Beispiele werden diese auf Chinesisch mit <四本书> /sì běn shū/, <这本书> /zhè běn shū/ und <读两遍> /dú liǎng biàn/[187] ausgedrückt. Es gibt im Chinesischen eine große Menge an ZEW. Fast jedes davon kann nur eine eingeschränkte Anzahl an Nomen oder Verben modifizieren.

5. Es gibt eine große Variabilität zwischen Wortart und syntaktischer Funktion. Im Chinesischen können die syntaktischen Beziehungen der Wortarten mit zahlreichen Ausnahmen beschrieben werden. Besonders die drei Hauptwortarten – Nomen, Verb und Adjektiv – können im Chinesischen vielfältige syntaktische Funktionen übernehmen. Kap. 4.5.2 geht ausführlich auf diese Aspekte ein.

[187] <四> vier; <本> ist ZEW für Buch; <书>: *Buch*; <这> Determinativpronomen; <读>: *lesen*; <两>: *zwei*; <遍> *Mal*

6. Es herrscht eine hohe Übereinstimmung der Bildungsprinzipien in verschiedenen sprachlichen Einheiten. Die Bildungsprinzipien der Wörter und Phrasen und die Wortstellung in den indogermanischen Sprachen sind unterschiedlich und können meistens nicht zu einem Regelsystem zusammengefasst werden. Dem hingegen ist die Wort- sowie Phrasenstruktur des Chinesischen mit denselben Formeln darstellbar. Arten und Beispiele der grammatischen Strukturen in den drei Einheiten werden in Kap. 4.4.1, 4.4.5 und 4.5.2 erläutert.

Wie die grammatischen Regeln der chinesischen Sprache in formalen Grammatiken ausgedrückt werden können, um computergestützte sprachliche Analysen durchzuführen und darauf basierend die Eingabe im Satz zu unterstützen, ist eine der größten Forschungsaufgaben bei der Entwicklung intelligenter Pinyin-Eingabesoftwares. Dies stellt den Schwerpunkt von Kap. 4.4. sowie 4.5 dar. Zusammenfassend können folgende sprachliche Erkenntnisse angewendet werden (vgl. Zhou YG 1996: 390-393):

1. Silbenstruktur und -varianten.

Das chinesische Syllabar hat eine festgelegte Struktur und begrenzte Varianten mit ca. 415 Möglichkeiten im Putonghua (vgl. Kap. 3.4.2, S.167). Nach diesen Erkenntnissen kann eine Pinyin-Kette in einzelne Silben segmentiert werden, anhand sich die möglichen Zeichen/Wörter recherchieren lassen. Wie in Abb. 4-6 (S. 201) skizziert wurde, ist die Silbensegmentation der erste Schritt und somit Grundlage für die Eingabe im Wort und im Satz mit intelligenten Pinyin-Eingabemethoden. Der Silben-Zeichen-Zusammenhang und die Methoden der automatischen Silbensegmentation werden in Kap. 4.3.1 bis Kap. 4.3.3 erläutert.

2. Die lexikalischen Informationen sowie Wortbildungsgrammatiken und -algorithmen.

Statistisch sind ca. 94% der Wörter des modernen Chinesischen polysyllabisch (vgl. ‚Xiandai Hanyu Changyong Ci Biao' 2008: 2). Die Konversion in der Einheit des Wortes kann deswegen im Normalfall oftmals effektiver funktionieren als in der Einheit des einzelnen Zeichens. Bei einer Pinyin-Eingabesoftware ist eine Wörterdatenbank fast immer die obligatorische Ressource, die die Laut-Zeichen-Konversion der gemeingebräuchlichen Wörter direkt unterstützen kann. Weiterhin gibt es bestimmte grammatische Regeln sowie Algorithmen für die Wortbildung, die in Wissensdatenbanken eingeschrieben werden, um die unbekannten Wörter erkennen, verarbeiten und in der Wörterbank automatisch aufnehmen zu können. Dies wird in Kap. 4.4.1 und 4.4.4 näher ausgeführt.

3. Statistisch-pragmatische Analysen der Zeichen- sowie Worthäufigkeit.

Die Frequenz der häufigen und seltenen Zeichen sowie Wörter ist sehr ungleichmäßig. Die Anordnung der homophonetischen Zeichen sowie Wörter nach Häufigkeit ist deswegen sinn-

voll für die Erhöhung des Schreibtempos. Abgesehen vom allgemeinen Gebrauch ist die Zeichen- und Wörterhäufigkeit abhängig von individuellen und diachronen, kurz idiolektalen Parametern. Aus diesem Grund wird einerseits die allgemeine Häufigkeit der homophonetischen Zeichen/Wörter in der Wörterdatenbank in einer Rangliste erfasst. Gleichzeitig orientiert sich das Programm am individuellen Schreibstil und Verwendungsgebrauch des Nutzers, wie bereits ausgeführt wurde. Dabei werden hochfrequente Kandidaten zuvorderst angeboten, während Kandidaten, die z.B. aus idiolektal-individuellen Gründen vom Nutzer häufig eingegeben werden, ebenfalls nach vorne gezogen werden (vgl. auch Kap. 4.3.4).

4. Die Regel der festgelegten Kollokationen.

Im Chinesischen gibt es zahlreiche festgelegte kollokationäre Verbindungen. Mit der grammatischen sowie semantischen Analyse und der Wahrscheinlichkeitserrechnung solcher Kollokationen können die korrekten Zeichenketten mit viel höherer Wahrscheinlichkeit vom Computer bestimmt werden. Bspw. sind die ZEW ein spezifisches Phänomen der chinesischen Sprache (siehe Punkt 4 der Besonderheiten der chinesischen Grammatik). <书> (/shū/, *Buch*), <树> (/shù/, *Baum*) und <数> (/shù/, *Zahl*) sind z.B. homophonetische Substantive, die bei der Mengenangabe mit verschiedenen ZEW kombiniert werden. So lauten *ein Buch*, *ein Baum* und *eine Zahl* nacheinander: <一本书> /běn/, <一棵树> /kē/ und <一个数> /gè/.[188] Durch die Kombination von ZEW und Substantiv in Pinyin ist es in den meisten Fällen kein Problem, die korrekten Schriftzeichen an den beiden Stellen zu bestimmen. Die festgelegten kollokationären Verbindungen sind in vielen Fällen Phrasen, deren Erkennungs- und Verarbeitungsverfahren in Kap. 4.4.5 vorgestellt werden.

5. Die grammatischen Regeln zur Wort-, Phrasen- und Satzbildung.

In Chinesischen gibt es allgemeine Regelgrammatiken für Kompositionswörter, Phrasen und Sätze (siehe Punkt 6 von den Besonderheiten der chinesischen Grammatik). Sie sind mit Formeln in der Wissensdatenbank beschreibbar und zur intelligenten Eingabe als Satzeinheit verfügbar (vgl. auch Klabunde 2010: 68, Wang XL/Wang YL 1996: 54).

6. Semantische Erkenntnisse über Zeichen, Wörter und Phrasen.

Durch die kombinierten grammatischen und semantischen Kenntnisse kann das zu schreibende Zeichen in den meisten Fällen unter den homophonetischen Zeichen sowie Wörtern ausgewählt werden. Bspw. werden die Pronomen der dritten Person alle als /tā/ ausgesprochen,

[188] Der Ton von <一> (ursprüngliche Aussprache: /yī/, *eins*) muss anhand des folgenden Wortes gewechselt werden: wenn das ZEW dem ersten, zweiten oder dritten Ton angehört, wird es mit /yì/ ausgesprochen (wie <一本书> /yì běn shū/); wenn das ZEW dem vierten Ton angehört, wird es mit /yí/ ausgesprochen (wie <一个数> /yí gè shù/.

unterscheiden sich aber in ihrer Zeichenform: <他> (*er*), <她> (*sie*) und <它> (*es*). Die Pinyin-Kette ‚TASHIWODEDIDI' kann nach grammatischen Analysen in die Schriftzeichenkette , TA 是我的弟弟' (*Er/Sie/Es ist mein jüngerer Bruder*) umgewandelt werden. Mithilfe der lexikalischen und semantischen Kenntnisse kann bestimmt werden, dass das Subjekt dieses Satzes wie das Objekt *Bruder* menschlich und männlich sein muss. Das Schriftzeichen <他> wird sodann abgerufen (vgl. auch Wang XL/Wang YL 1996: 54f).

4.2.2 Vier Hauptmethoden zur satzstufigen Laut-Zeichen-Konversion

Wie in Kap. 4.1.3 und 4.1.4 (S. 200ff) vorgestellt wurde, sind intelligente Eingabesoftwares mit verschiedenen Funktionen in der Mehrheit, deren Kerntechnik eine satzstufige Laut-Zeichen-Konversion ist. Hauptprinzip dieser Konversion ist, bei einer dynamisch eingegebenen Pinyin-Kette linguistische und statistische Analysen zusammenhängend durchzuführen, um automatisch den korrekten Satz in Schriftzeichen auszugeben, ohne vom PC-Benutzer Wort für Wort oder Zeichen für Zeichen ausgewählt werden zu müssen (vgl. Wang XL/Wang YL 1996: 52, Zhang S et al. 1997: 37). Zusammengefasst besteht die Kerntechnik einer chinesischen intelligenten Eingabesoftware aus drei Teilen: 1) Pinyin-Segmentation, die Silben- und Wortsegmentation beinhaltet (ausführlicher erläutert in Kap. 4.3.3 sowie 4.4.2); 2) Abruf der Zeichenkandidaten aus der Wörterdatenbank (Recherche von homophonetischen Wörtern mittels Pinyin) und 3) Satzgeneration von möglichen Kandidaten, die mithilfe der linguistischen Wissensdatenbank unterstützt werden muss (vgl. Chen SY/Zhao/Wang 2015: 2).

Bevor auf die drei Teile detailliert eingegangen wird, soll zunächst unter die Lupe genommen werden, wie die Wissensdatenbank (inkl. Wörter- und linguistischer Wissensdatenbank) begründet wird und wie die Laut-Zeichen-Konversion anhand der eingespeicherten Wissensdaten abläuft. Behandelt werden die vier Hauptkonversionsmethoden.

Die auf sprachlichem Verstehen basierende intelligente Eingabe (基于理解的智能输入) ist die Imitation des menschlichen sprachlichen Verstehens in Computerprogrammen, um die wahrscheinlichste Zeichenkette anhand grammatischer sowie semantischer Regeln anzubieten. In ihrer linguistischen Wissensdatenbank sind hauptsächlich die maschinenlesbaren formalen sprachlichen Regeln eingespeichert. In der Phase der Satzverarbeitung (siehe Abb. 4-12) werden die möglicherweise betroffenen Regeln über die homophonetischen Kandidaten (vor allem die syntaktischen, lexikalischen, semantischen und selbstdefinierten Regeln) in den Wissensdatenbanken recherchiert. Die Kandidaten werden demzufolge zusammenhängend maschinell sprachlich analysiert und ausgefiltert. Diese Methode wurde am frühesten erfun-

den, ist die grundlegendste Variante und wird verbreitet eingesetzt (vgl. Wang XL 2005: 109, Chen YF/Zhu 2002b: 13).

Größte Schwierigkeit dieses Verfahrens ist, Ausführlichkeit und Präzision der linguistischen Regeln ins Gleichgewicht zu bringen. Die Wissensdatenbank soll theoretisch möglichst umfangreich sein, damit möglichst viele sprachliche Phänomene berücksichtigt werden können. Selten benötigte sowie pragmatisch ungültige Sätze können in diesem Fall als Ergebnisse erzeugt werden, was sowohl das Computersystem als auch das menschliche Gehirn bei der Kandidatenauswahl belastet.

Die grundsätzlichen Regeln müssen zuerst auf linguistischen Forschungen basieren. Da die sprachlichen Phänomene unbegrenzt zu interpretieren sind, kann die linguistische Wissensdatenbank durch Analyse realistischer Sätze erweitert werden. Sprachanalysen der Korpora können dazu dienen und auch die Selbstlernfunktion ist zu diesem Zweck entstanden. Wenn ein Satz wegen mangelnder Regeln nicht einheitlich konvertiert werden kann, wird er Wort für Wort vom PC-Benutzer bestimmt. Ein so manuell eingegebener Satz wird dann von der Software analysiert und die unbekannten Regeln darauf basierend ‚gelernt'. Diese Funktion wird relativ häufig bei intelligenten Pinyin-Eingabesoftwares eingesetzt.

Die auf pragmatischen Statistiken basierende intelligente Eingabe (基于语用统计的智能输入) ist von Frequenzanalysen bedingt. Auf statistischer Basis werden homophonetische Zeichen sowie Wörter ausgefiltert. Auch diese Konversionsmethode ist weit verbreitet, wobei sie meist in Kombination mit einer intelligenten Eingabesoftware, die auf dem sprachlichem Verstehen basiert, verwendet wird (vgl. Chen YF/Zhu 2002b: 14).

Im Gegenteil zu der auf sprachlichem Verstehen basierenden Technik, für die die linguistischen Kenntnisse als Grundlage angewendet werden, gehört die statistikbasierte Technik zum Bereich *Operations Research*. Die berechneten Objekte umfassen die Zeichen-/Worthäufigkeit und die Kollokationswahrscheinlichkeit zweier oder mehrerer benachbarter Zeichen/Wörter. Die möglichen Kandidaten zu jedem Pinyin-Wort lassen sich vernetzen. Durch algorithmische Analysen wird die Kollokation mit der höchsten Wahrscheinlichkeit ermittelt. Der Algorithmus anhand N-Gramm-Modelle ist eine der wichtigsten Vorgehensweisen. Auch der sog. Viterbi-Algorithmus wird häufig angewendet, um die Zustandsfolge einer Zeichenkette zu errechnen (vgl. ibid, Wang XL 2005: 109f, Zhou G/Zhu 2002: 42). Zur Funktionalität einer intelligenten Eingabesoftware werden häufig die gemischten Techniken von sprachlichem Verstehen sowie Statistiken gebraucht, um die Korrektheit der Konversion zu erhöhen.

Die auf Sprachmodell-Matching basierende intelligente Eingabe (基于模板匹配的智能输入) ist jene Technik, die mithilfe von gespeicherten Phrasenketten (den festen Kollokationen der Wörter) funktioniert, um homophonetische Zeichen sowie Wörter auszufiltern und die Ambiguitäten der Wortsegmentation auszuschließen. Solche Phrasenketten werden Sprachmodellwörter (模板词) genannt, die zwei oder mehrere zusammenhängende Wörter beinhalten. Für eine bessere Qualität muss das Sprachmodell auf Basis eines riesigen Korpus begründet werden, damit die Sprachmodelldatenbank möglichst umfangreich ist. Nach der Silben- und Wortsegmentation des einzugebenden Satzes werden abhängig vom eingegebenen Pinyin-Code möglichst passende Sprachmodellwörter und grammatische Regeln recherchiert und gematcht. Wenn das Ergebnis des Matchings eindeutig ist, wird das Ergebnis ausgegeben. Gibt es mehrere Ergebnisse, werden diese mit grammatischen Regeln oder Wahrscheinlichkeitserrechnungen näher bestimmt. Das wahrscheinlichste Ergebnis wird dann zuvorderst ausgegeben. Diese Technik ermöglicht zwar große Präzision, beansprucht aber einen großen Speicherplatz in CPU (vgl. Wang XL 2005: 110, Chen YF/Zhu 2002b: 15). Die Arbeitsschritte dieser Funktion bei einer Eingabesoftware werden in Abb. 4-13 dargestellt.

Die auf Kontextzusammenhängen basierende Technik (基于上下文关联的智能输入) beruht auf den Theorien der vagen Kontrolle. Mithilfe von Zusammenhängen im pragmatischen Kontext werden Überschneidungen der Zeichen- sowie Wörterkandidaten ausgefiltert und der wahrscheinlichste Satzkandidat intelligent ausgewählt. Anhand der Attributen der Zeichen sowie Wörter, der grammatischen Regeln und der dynamischen pragmatischen Statistiken werden die Zusammenhänge der nacheinander vorkommenden Zeichen sowie Wörter und die abhängige Wahrscheinlichkeit einer Zeichenkette von ihrem Pinyin-Code (charakteristische Funktion des Zeichens/Wortes) errechnet und darauf basierend die wahrscheinlichste Zeichenkette angeboten. Der Algorithmus für diese Funktion kann mit dem Hidden-Markov-Modell (Abk.: HMM) durchgeführt werden. Mit dieser Technik arbeitet bspw. die Qingyueliang-Hanzitong-Eingabesoftware, in der die vier Wörter im Vorkontext und ein Wort im Nachkontext berücksichtigt werden (vgl. Wang XL 2005: 111, Chen YF/Zhu 2002b: 17). Die Funktionsprinzipien skizziert Abb. 4-14.

Die vier Konversionsmethoden bieten Grundarbeitstheorien der intelligenten Eingabe an. In der realen Anwendung wurden und werden die Eingabesoftwares mit den technischen Fortschritten stets verbessert. Bspw. wird in vielen heutigen Softwares Cloud Computing eingesetzt, so dass vielfach größere Wissensdatenbanken für die Laut-Zeichen-Konversion zur Verfügung stehen. Mit dieser neuen Anwendung wurden Schreibeffizienz und Präzision seit ca.

2010 signifikant erhöht. In dieser Arbeit wird die intelligente Software, bei der die auf Sprachmodell-Matching, sprachlichem Verstehen und Statistiken basierenden Methoden kombiniert eingesetzt werden, vorgestellt.

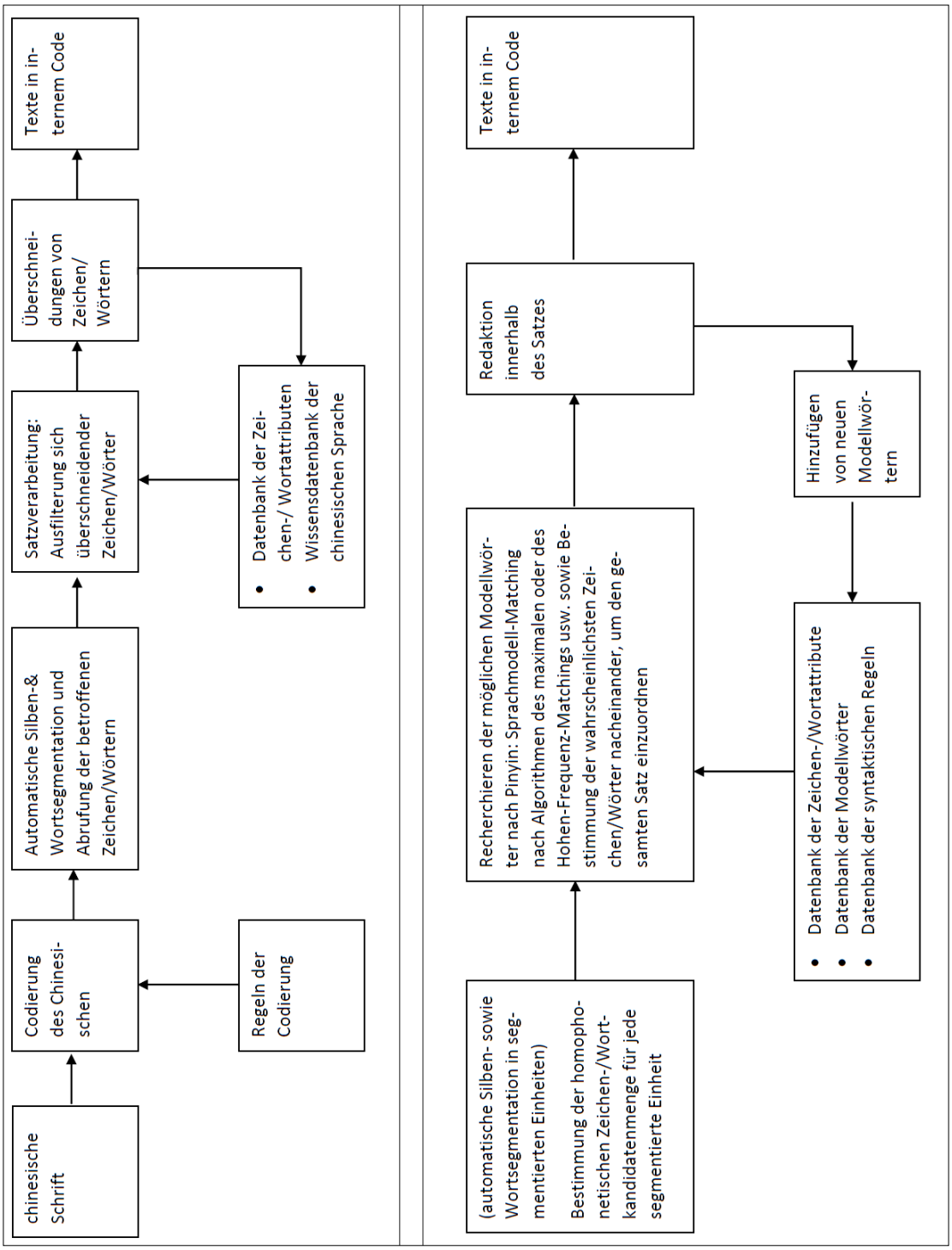


Abb. 4-12 [links]: Struktur der auf sprachlichem Verstehen basierenden Eingabesoftware (Chen YF/Zhu 2002b: 13 [Übersetzung der Verfasserin])

Abb. 4-13 [rechts]: Struktur der auf Sprachmodell-Matching basierenden Eingabesoftware (ibid.: 16)

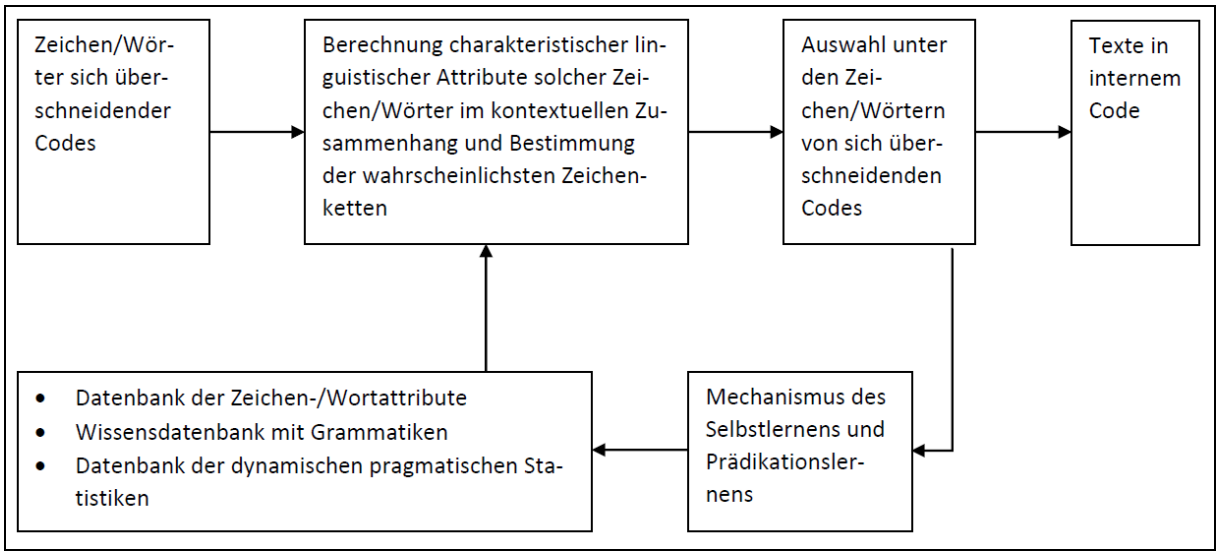


Abb. 4-14: Struktur der auf Zusammenhängen im Kontext basierenden Eingabesoftware (ibid.: 17)

4.2.3 Elektronisches Wörterbuch und linguistische Wissensdatenbank

Aus den Abbildungen zum intelligenten Laut-Zeichen-Konversionsverfahren (Abb. 4-12 bis Abb. 4-14) ist zu ersehen, dass zwei Arten von Erkenntnissen obligatorisch: 1) die in der Wissensdatenbank gespeicherten Wissensdaten, wie die lexikalischen Informationen und die formalen sprachlichen Regeln; 2) die anhand des sprachlichen Kontexts dynamisch erworbenen Informationen, wie die syntaktische Funktionen, die semantischen Beziehungen zweier benachbarter Wörter usw. (vgl. Feng 2001b: 84). Die erste Art, die in der Wissensdatenbank eingeschrieben wird, ist die Grundlage für die Sprachverarbeitung, wodurch die zweite Art akquiriert werden kann. Aufbau und Qualität der Wissensdatenbank sind deswegen entscheidend für die Eingabesoftwareeffizienz.

Wie in Kap. 4.1.3 (S. 200) erwähnt wurde, besitzt eine intelligente Eingabesoftware mindestens zwei Wissensdatenbanken, die jeweils lexikalische Informationen (das elektronische Wörterbuch)[189] und sprachliche Regeln (die linguistischen Wissensdatenbank) anbieten.[190] Zur Effizienzsteigerung werden sowohl elektronisches Wörterbuch als auch linguistische Wissensdatenbank in statische und dynamische Teile getrennt. In den statischen Teilen werden voreingestellte gemeingebräuchliche Wörter sowie sprachliche Regeln eingeschrieben.[191] Das dynamische Lexikon wurde zum Zweck einer schnelleren Datenspeicherung entworfen: Zeichenketten einer bestimmten Länge (im Normalfall zwei Zeichen), die nicht vom statischen Lexikon aufgenommen werden können, werden so dynamisch und einmalig als ein Wort ver-

[189] Entspricht der ‚Datenbank der Zeichen- sowie Wortattribute' in Abb. 4-12, 4-13 und 4-14.
[190] Entspricht der ‚Wissensdatenbank der sonstigen sprachlichen Erkenntnisse' in Abb. 4-12, 4-13 und 4-14.
[191] Für weitere Informationen über Wort und sprachliche Regeln vgl. Kap. 4.4.1 und 4.5.1.

arbeitet. In der Tat sind viele dynamisch als Wörter verarbeitete Zeichenketten grammatisch keine Wörter, sondern festgelegte Phrasen oder Teile eines langen Wortes. Die dynamische linguistische Wissensdatenbank dient dazu, das maschinelle Selbstlernen zu verstärken, individuell-orientiert die sprachlichen Regeln zu erweitern, zu korrigieren und zu verbessern, und so mehr sprachliche Phänomene zu verarbeiten (vgl. Zhang S et al. 1997: 38ff).

Das der intelligenten Eingabesoftware angehörige elektronische Wörterbuch ist ein maschinenlesbares, für Sprachverarbeitungen dienendes Lexikon. Seine Struktur ist im Prinzip nicht viel anders als ein Lexikon für multimodale Systeme. Nach Gibbon (2009: 520f) beinhaltet es vier Substrukturen: 1) Megastruktur, die die Gesamtorganisation behandelt; 2) Makrostruktur, die Lexikoneinträge fokussiert; 3) Mikrostruktur, in der die Wort- sowie Zeichenattribute eingeschrieben werden; 4) Mesostruktur, in der Querverweise eines Eintrages zu einem anderen oder zu einer sprachlichen Regel angegeben werden. Das der intelligenten Eingabesoftware angehörige Wörterbuch wird auch nach diesen vier Substrukturen analysiert.

Über die Megastruktur werden Fragen zu Grundfunktionsweise, Begründung und Indexierung erläutert. In Kap. 4.1.1 wurde der Begriff *Konversionslexikon* eingeführt. Sowohl im Wörterbuch als auch im Konversionslexikon sind Wörter und Zeichen die eingeschriebenen Einheiten. Die Unterschiede liegen darin, dass in dem Konversionslexikon die verschiedenen Codierungen (Input- und interner Code) eines Wortes/Zeichens definiert sind, während in dem Wörterbuch die lexikalischen Informationen (wie Wortart, feste Kollokationen usw.) als Inhalte eingeschrieben werden (vgl. Zhang S et al. 1997: 38). Die Einträge des elektronischen Wörterbuchs können einerseits von traditionellen Wörterbüchern als Basis übernommen, andererseits auch durch Korporaanalysen ergänzt werden (vgl. Hou 1999: 150). Die Indexierung und Anordnung der Einträge sind meistens von der Inputcodierung abhängig und identisch mit dem Konversionslexikon. D.h. das Wörterbuch einer Pinyin-Eingabesoftware ist im Normalfall nach der Aussprache alphabetisch indexiert.

Die Makrostruktur behandelt hauptsächlich den Umfang der Einträge und die Probleme bei homographischen Einträgen. Wie oben vorgestellt, gehören Wörter und Zeichen gleichermaßen zu den Einträgen. Dies findet nicht nur darin Begründung, dass über die Hälfte der Schriftzeichen eigenständige Wörter darstellen können, sondern wird auch durch den isolierenden Sprachbau bedingt. Wie bei der ersten Eigenschaft der chinesischen Sprache in Kap. 4.2.1 (S. 209f) skizziert wurde, werden im Chinesischen Synsemantika, Adverbien sowie Affixe verwendet, um grammatische Phänomene auszudrücken, die in den indogermanischen Sprachen durch Wortflexion realisiert werden. Die idealste Methode für die Lexikalisierung solcher Wörter sowie Affixe ist die Darstellung als eigenständige Einträge, weshalb Wort-

sowie Phrasenbildungsgrammatiken ins elektronische Wörterbuch eingeschrieben werden.[192] Das Partikel <们> /mén/ kann hierfür als Beispiel fungieren. Es wird hinter für Menschen stehende Nomen gesetzt, um eine Pluralbedeutung zu markieren. Wenn <们> in einem Satz einer Personalbezeichnung nachsteht, werden die beiden Wörter durch Unterstützung des elektronischen Wörterbuchs als (nun autosemantische) Einheit betrachtet (vgl. Hou 1999: 138f, vgl. zum selbigen Beispiel auch Kap. 4.4.3).

Im Chinesischen gibt es zahlreiche homographische oder selbe Wörter mit verschiedenen grammatischen Eigenschaften. Wie solche Wörter im Wörterbuch lexikalisiert werden können, ist ebenso ein Schwerpunkt für den Entwurf der Wissensdatenbank. Es gibt im Chinesischen hauptsächlich drei Arten (vgl. Hou 1999: 141f):

1) Homographische Wörter mit unterschiedlicher Phonetik (Heteronyme). Gemeint sind Wortpaare in gleicher Schriftform, die unterschiedlich ausgesprochen werden und sich im Normalfall auch in Bedeutung, Wortart, Pragmatik usw. unterscheiden. Bspw. bedeutet <乐> mit der Aussprache /lè/ *froh*/*freuen* (zumeist prädikatives Wort/Morphem) und mit der Aussprache /yuè/ *Musik* (zumeist nominales Wort/Morphem). In einem der Pinyin-Eingabesoftware angehörigem Wörterbuch müssen Heteronyme getrennt als verschiedene Einträge dargestellt werden, so dass das System nach dem Laut das entsprechende Wort abrufen kann (vgl. zu Heteronymen auch Kap. 3.4.1, S. 163 & Kap. 4.3.1).
2) Homographische Wörter mit gleicher Phonetik aber verschiedenem Sinninhalt. Gemeint sind ein Wort-/Zeichenpaar, das trotz identischer Zeichenform und Aussprache verschiedene semantische Bedeutungen hat. Linguistisch betrachtet werden zwei oder mehrere Wörter in dieser Art als verschiedene Wörter verstanden (vgl. auch Lan 2002: 135). Bspw. ist das Wort <会> /huì/ in dem Satz <他会说汉语> (*Er kann Chinesisch sprechen*) das Modalverb, das *können* bedeutet. In der Phrase <开会> (*Tagung stattfinden*) ist es hingegen ein Nomen, das die Bedeutung *Tagung* überträgt. Bei der Lexikalisierung müssen solche homographischen Wörter getrennt lexikalisiert werden. Wegen der identischen Schriftform und Aussprache muss die Worterkennung, die dem Wort dem richtigen Eintrag zuordnet, anhand des Kontextes überprüft werden.
3) Ein Wort mit Wortart-Ambiguität. Im Vergleich zu den vorherigen Fällen richtet sich dieser Fall an ein Wort, das trotz gleicher Semantik in unterschiedlichen Wortarten funktioniert. Ein Wort, das zu verschiedenen grammatischen Kategorien gehören kann, wird als

[192] Zu Wort- sowie Phrasenbildungsgrammatiken vgl. Kap. 4.4.1, 4.4.4 und 4.4.5.

multi-kategoriales Wort bezeichnet. Bspw. existiert das Wort <忙> /máng/ sowohl als Verb (entspricht im Deutschen *sich beschäftigen*) als auch als Adjektiv (*beschäftigt*). Tritt ein solches Wort in dem zu analysierenden Satz auf, muss ebenso der Kontext überprüft werden (vgl. Lan 2002: 133; zu multi-kategorialen Wörtern und den Methoden des POS-Tagging vgl. Kap. 4.4.3).

Das Ziel der Unterstützung linguistischer Analysen elektronischer Wörterbücher wurde benannt. Welche Attribute über ein einzelnes Wort/Zeichen für die Laut-Zeichen-Konversion nötig sind, muss in der Mikrostruktur eingeschrieben werden. Wie oben erwähnt, sind grammatische und semantische Eigenschaften unentbehrliche Informationen. Bedingt durch die Attribute eines Wortes/Zeichens sind die pragmatischen Eigenschaften in manchen Fällen obligatorisch. Solche Eigenschaften umfassen bspw. feste Kollokationen mit anderen Wörtern/Zeichen, die Wort- sowie Zeichenhäufigkeit usw. (vgl. Hou 1999: 148).

Anhand der Informationen in der Mesostruktur eines einzelnen Eintrags, wo die Querverweise zu anderen Einträgen, bestimmte Wortkategorien und sprachliche Regeln eingeschrieben sind, können einerseits viele festgelegte Kollokationen mit weniger technischen Schwierigkeiten erkannt werden. Andererseits können die Ambiguitäten eines Wortes sowie homographischer Wörter durch Grammatik und Semantik unterstützt werden. Bspw. werden unter dem Affix <们> /mén/ Querverweise mit den für Menschen stehenden Nomen vernetzt, so dass diese Zeichenfolge als die Pluralform des Nomens erkannt werden kann.

Wie betont müssen das Wörterbuch und linguistische Wissensdatenbank kombinierend bei der Laut-Zeichen-Konversion angewendet werden. Die linguistische Wissensdatenbank ist eine aus formalen sprachlichen Regeln bestehende Datenbank, die sich zu verschiedenen Konversionsmethoden unterscheidet (siehe Abb. 4-12, 4-13 und 4-14).

Im Prinzip stehen vier Arten sprachlicher Regeln bei der Satzgeneration mit intelligenten Pinyin-Eingabesoftwares zur Verfügung: 1) die allgemeinen Satz- sowie Phrasenstrukturgrammatiken über unbestimmte Wörter von bestimmten Wortarten (grammatische Regeln); 2) die Satzbildungsregeln über unbestimmte Wörter aus bestimmten semantischen Kategorien (semantische Regeln); 3) die sprachlichen Regeln über zwei oder mehrere bestimmte Wörter (Sprachmodelle); 4) die sprachlichen Regeln über die Kollokation von einem bestimmten Wort zu einem unbestimmten Wort aus bestimmten grammatischen oder semantischen Kategorien (Hybridregeln). Die erste und zweite Variante können manuell in der Wissensdatenbank dargestellt werden, obgleich es in der praktischen Verwendung zu erheblichen Einschränkungen kommen kann. Wegen der großen Anzahl komplizierter sprachlicher Phänome-

ne, der entscheidenden Rolle von Semantik und Pragmatik bei der Satzbildung usw. können diese Einschränkungen derzeit kaum verbessert werden. Aus diesem Grund sind sie zwar die obligatorischen Regeltypen, sollten zum Ziel der hochintelligenten Eingabesoftware aber nicht die einzigen Typen sein. Sprachmodelle (die sprachlichen Regeln über Schlüsselwörter) müssen empirisch sein, auf Korporaanalysen basieren und möglichst umfangreich sein, damit sie bei der Laut-Zeichen-Konversion präzise funktionieren. Dank korpuslinguistischer Fortschritte, der Entwicklung größerer Arbeitsspeicher und der Verbreitung des Cloud Computing ist der Gebrauch dieser Variante bei modernen intelligenten Pinyin-Eingabesoftwares bereits weit verbreitet. D.h. ein Trainingskorpus von bestimmter Größe kann zum Ziel der Akquisition von sprachlichen Regeln analysiert werden. Die Kontexte von einem bestimmten Schlüsselwort können dabei annotiert und mit oder ohne menschliche Korrektur in der Wissensdatenbank eingeschrieben werden. Diese Sprachmodelle können zwar eine hohe Präzision erreichen, aber nur bei einer begrenzten Anzahl definierter Wörter angewendet werden. Ihr Umfang ist außerdem durch den Arbeitsspeicher eingeschränkt. Deswegen kann die Sprachmodelldatenbank des dritten Typs nicht allein Spender der linguistischen Kenntnisse sein (vgl. Luo 1996: 300, Hou 1999: 155f, Zhang S. et al. 1997: 40). Die Akquisition der Regeln des vierten Typs kann z.B. auf der Ableitung von den Regeln des ersten, zweiten sowie dritten Typs basieren (vgl. Wang XL 1993: 374; vgl. dazu ausführlicher auch Kap. 4.5.2 und 4.5.3).

Nach den Eigenschaften des schriftlichen Chinesisch müssen die natürlichen Texte mindestens in den folgenden vier Schritten verarbeitet werden, um die sprachlichen Regeln zu gewinnen: Wortsegmentation → Wortart-Tagging → syntaktische Annotation → semantische und pragmatische Analysen (vgl. Hou 1999: 160). Nach dem Verarbeitungsstand können die Korpora in verschiedene Niveaus klassifiziert werden. Das unverarbeitete Korpus heißt Rohkorpus von Niveau-Null (N-0). Die in Wörter segmentierten Texte gehören zu dem Korpus von Niveau-Eins (N-1). Zum zweiten Niveau gehören die Texte mit Wortart-Tagging. Wenn die Texte mit syntaktischen Funktionen annotiert werden, sind sie den dreistufigen Korpora (N-3) zugehörig (vgl. Luo 1996: 297f). Die vier verschiedenen Korpusniveaus werden in Abb. 4-15 anhand des Satzes <地球所拥有的自然资源也是有限的> (/dìqiú suǒ yōngyǒu de zìránzīyuán yěshì yǒuxiàn de/; *Die verfügbaren Naturressourcen der Erde sind auch begrenzt.*) exemplifiziert. Die syntaktische Annotation wird in dem Beispiel als Grammatikbaum analysiert.

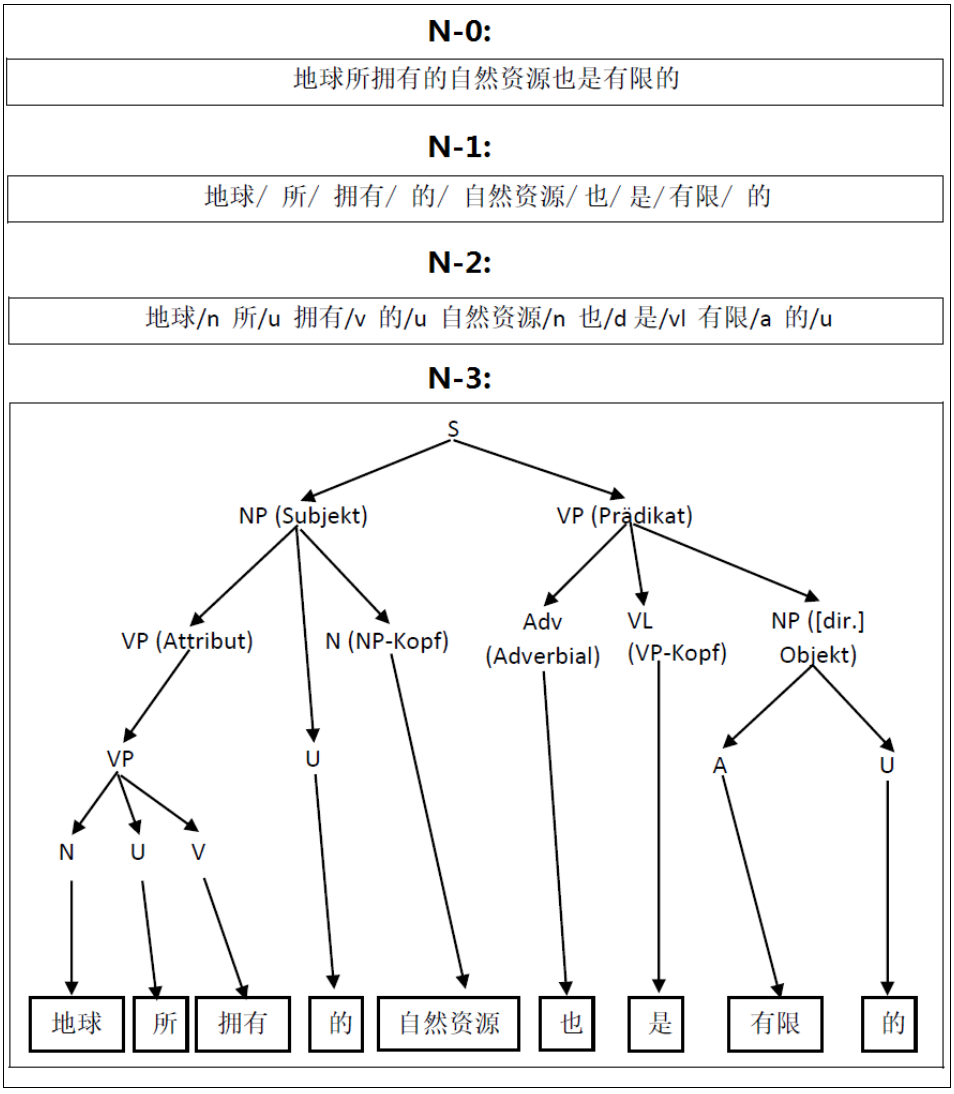


Abb. 4-15: Die vier verschiedenen Verarbeitungsniveaus der chinesischen Korpora[193]

Nach der vierstufigen Verarbeitung der Korpora können neue Regeln über die Schlüsselwörter begründet und die allgemeinen grammatischen sowie semantischen Regeln weiter ergänzt werden, was zur Verbesserung der Wissensdatenbank entscheidend ist. Die Baumdatenbank, also die durch Annotation der syntaktischen Struktur gebildete Datenbank, kann anhand der Korporaanalysen begründet werden (vgl. Hou 1999: 155f, Chen XH/Cui 1996: 303). Die Art und Weise zur Akquisition der sprachlichen Regeln anhand von Korporaanalysen wird in Kap. 4.5.3 beleuchtet.

Nach Informationen über die vier Hauptmethoden zur intelligenten Laut-Zeichen-Konversion im letzten und über die Wissensdatenbank in diesem Kapitel können folgende Punkte zusammengefasst werden:

[193] Für die verwendeten Abkürzungen der Wortarten siehe Kap. 4.4.3, Tab. 4-9. 的 /de/ ist ein spezifischer Partikel für Attributmarkierungen. ‚Attribut+的+Nomen' bildet eine NP; das Nomen kann in dieser Formel weggelassen werden, um etwas Allgemeines oder jemanden auszudrücken, wie z.B. steht 有限的 (*begrenzt* + de) für *Begrenzes*.

Erstens sind die sprachlichen und statistischen Analysen die beiden Grundperspektiven für die intelligente Verarbeitung. Zur Begründung der Wissensdatenbank (vor allem bei Korporaanalysen) und bei dem Verfahren der intelligenten Konversion sind die Ergebnisse von beider Perspektiven entscheidend.

Zweitens variiert die Wissensdatenbank je nach Eingabesoftware in Arten des linguistischen Wissens. Zu der auf sprachlichem Verstehen basierenden (ersten) Methode sind ausführliche grammatische und semantische Regeln für Phrasen- und Satzbildungen der Hauptwissensunterstützer. Im Vergleich dazu wird die statistische Unterstützung (zweite Methode) nebensächlich angewendet, um die syntaktischen Regeln anhand ihrer Wahrscheinlichkeit anzuordnen. Die ebenso hauptsächlich auf sprachliche Analysen konzentrierte dritte, auf Sprachmodellen basierende Methode fokussiert eher die Häufigkeit der Wort- sowie Phrasenkollokationen. Zu statistischen Analysen kann auch die Wahrscheinlichkeit kollokationärer Verbindungen eines Wortes zu anderen (der vierten Methode) errechnet werden.

Drittens spielt Wortverarbeitung die zentrale Rolle bei der satzstufigen Konversion. Silben- und Wortsegmentation und die Recherche der homophonetischen Wörter sind stets die ersten zwei Arbeitsschritte jeder Laut-Zeichen-Konversionsmethode. Danach werden die homophonetischen Wörter nach den sprachlichen Regeln, statistischen Algorithmen oder Sprachmodellen überprüft bzw. gematcht, um die bestmöglichen Kandidaten ausgeben zu können. Anhand der allgemeinen Funktionsweisen der intelligenten Laut-Zeichen-Konversion wird in den folgenden Kapiteln die Reihenfolge des Arbeitsprozesses vorgestellt, konkret die Silbensegmentation, Wortverarbeitung und Satzgenerierung, die jeweils zu Zeichen- (siehe Kap. 4.3.1 - 4.3.3), Wort- (siehe Kap. 4.4.1 - 4.4.5) sowie Satzverarbeitung (siehe Kap. 4.5.1 - 4.5.4) zählen. Zunächst werden dabei linguistische Eigenschaften und mögliche Schwierigkeiten automatischer Analysen skizziert. Darauf basierend werden die computerlinguistischen Lösungen beleuchtet.

4.3 Informationsverarbeitung in der Einheit von Zeichen

Der Silben-Zeichen-Zusammenhang ist die Wissengrundlage für phonetische Inputcodierung und die Laut-Zeichen-Konversion der Pinyin-Eingabesoftware. Die Erkenntnisse darüber sind deswegen die Basis sowohl für die PC-Benutzer als auch für das Design der phonetischen Eingabemethoden. In einer intelligenten Pinyin-Eingabemethode ist der Satz die allgemein eingegebene Einheit, d.h. dass die Silbensegmentation der erste Arbeitsschritt der Konversion sein muss (siehe Kap. 4.2.2). Die zeichenorientierte Verarbeitung konzentiert sich aus diesem

Grund darauf, die davon betroffenen lingusitischen Erkenntnisse und die Verfahren der Silbensegmentation zu erforschen.

4.3.1 Silben-Zeichen-Zusammenhang

Wie in Abb. 3-5 graphisch gezeigt wurde, hat der reziproke Zusammenhang zwischen Aussprache und Zeichenform zwei ambige Fälle: Homophone und Heteronyme. Nach den in Kap. 3.4.2 dargestellten Erkenntnisse gibt es fast zu allen Silben mehrere Homophone und durchschnittlich entsprechen 19,53 Schriftzeichen (innerhalb eines Inventars von 8.105) demselben Laut. Im Gegenteil dazu ist Heteronymie ein selteneres Phänomen. Nach Sun (1988: 210) haben ca. 10% der Schriftzeichen Heteronyme. Homophone, Heteronyme und diffuse Silbengrenzen stellen deswegen jene linguistischen Phänomene dar, die bei der Laut-Zeichen-Konversion als Hindernisse auftreten können.

Die Aussprache eines Schriftzeichens entspricht im chinesischen Sprach- sowie Schriftsystem einer Silbe, die eine klare Grenze zu den benachbarten Silben im sprachlichen Kontext besitzt. In einer sprachlich identischen Situation gleicht die Anzahl der Silben immer die der Schriftzeichen, jedoch mit Ausnahme des sogenannten ‚Erhua-Yin' (儿化音). Erhua-Yin ist der Sonderlaut /er/ (in IPA [ɚ]), der als Suffix mit dem Auslaut der vorderen Silbe zu einem Sonderauslaut kombiniert wird, um einen Diminutiv oder Abstraktion zu bilden. Er wird schriftlich mit dem eigenständigen Schriftzeichen <儿> repräsentiert (vgl. Xinhua-Lexikon Online: Eintrag ‚儿').[194] Z.B. wird die Zeichenkombination von <点> /diǎn/ und <儿> zusammen als /diǎr/ (Bedeutung: *kleiner Punkt*/ *ein bisschen*) ausgesprochen. Transkribiert in Pinyin wird der Erhua-Yin orthographisch unabhängig von dem mit ihm kombinierten Laut immer in Form einer eigenständigen Silbe als /er/ dargestellt. Der Ausdruck *ein bisschen* <一点儿> wird in Pinyin als /yìdiǎn'er/, aber nicht als /yìdiǎr/ codiert. Dieser Fall gilt auch bei manchen chinesischen Ortsnamen mit [ɚ]-Laut, die aus einer anderen Sprache entlehnt sind, wie z.B. der Stadtname <哈尔滨> (/hā'ěrbīn/, *Harbin*), der ursprünglich aus dem Mandschurischen stammt. Bei solchen Namen gibt es zwischen der Pinyin-Angabe und der offiziellen Bezeichnung in dem lateinischen Alphabet Abweichungen. Die Stadt wird z.B. in Pinyin als /hā'ěrbīn/ mittels dreier Silben transkribiert, international aber *Harbin* genannt und geschrieben.

In dem elektronischen Wörterbuch der Wissensdatenbank wird das Suffixmorphem <儿> als ein eigenständiger Eintrag definiert. Gleichzeitig haben die mit dem Morphem gebildeten gemein gebräuchlichen Wörter auch eigene Einträge, wie <一点儿> (/yìdiǎn'er/, *ein bisschen*)

[194] In traditioneller Form ‚兒'; die usprüngliche Aussprache ist /ér/, als Suffix findet ein Wechsel zum schwachen Ton statt.

(siehe Kap. 4.2.3, S. 218f). Sowohl die Wörter mit Erhua-Yin als auch Ortsnamen mit [ɚ]-Laut müssen beim Schreiben mit Pinyin-Eingabemethoden in der orthographischen Form, in der /er/ als eigenständige Silbe auftritt, eingegeben werden. Bspw. kann das Wort <一点儿> mit dem Pinyin-Code ‚YIDIANER', aber nicht mit ‚YIDIAR' eingegeben werden.

Abb. 4-16: Die Eingabe eines Worts mit ER-Laut, in dem /er/ als eigenständige Silbe behandelt werden muss[195]

Bei dem [ɚ]-Laut gibt es zwar Abweichungen zwischen der Umgangssprache und der Pinyin-Transkription, aber nach der Beherrschung der Pinyin-Orthographie, die bereits am Anfang der Grundschulbildung vermittelt wird, ist es normalerweise kein Problem, solche Wörter korrekt in Pinyin zu codieren. Für die Eingabe der Heteronyme sind jedoch weit höhere sprachliche Erkenntnisse vonnöten.

Wie bereits in Kap. 4.2.3 (S. 219) angerissen, werden Heteronyme im elektronischen Wörterbuch der Wissensdatenbank als verschiedene Einträge definiert. Aus Perspektive der drei Grundattribute der chinesischen Schrift – Zeichenform, Aussprache und Sinninhalt – gibt es drei Arten von Heteroymen, die bei der Inputcodierung Probleme bereiten können: 1) Heteronyme mit verschiedenen Sinninhalten und Aussprachen; 2) Heteronyme mit gleichem Sinninhalt aber verschiedener Aussprache; 3) variierte Aussprachen (vgl. Sun 1988: 211f).

Die Heteronyme des ersten Falls sind meistens nach dem Prinzip des Bedeutungswandels (siehe Kap. 3.3.2, S. 153f) oder im Zuge der Vereinfachungsreform entstanden (vgl. ibid: 212). Das Zeichen <长> bspw. bedeutet mit dem Laut /zhǎng/ *wachsen* und mit /cháng/ *lang*. Die Phrase <长长> wird deswegen als /zhǎng cháng/ ausgesprochen, die *länger werden* bedeutet. Bei solchen Zeichen muss beim Lesen im Kontext die korrekte Lesart herausgefunden wer-den. Beim Schreiben mit Pinyin-Eingabemethoden sind sie auch nur durch die dem Kontext angemessene Aussprache eingebbar, wenn das Zeichen in der Einheit vom polysyllabischen Wort oder Satz miteingetippt wird.[196]

Ein Zeichen des zweiten Falls wird trotz derselben Bedeutung in verschiedenen Verwendungssituationen – wie z.B. als eigenständiges Wort sowie als ein Morphem zur Wortbildung in der Schrift- sowie Umgangssprache – unterschiedlich ausgesprochen. Das Phäno-

[195] Nach Eingabetest mit Sogou-Pinyin-Eingabesoftware.
[196] Nach Eingabetest mit Sogou-Pinyin-Eingabemethode.

men ist einerseits durch die Sprachgewohnheit, andererseits wegen der Verschiedenheit zwischen Umgangs- und Schriftsprache bedingt (vgl. ibid.: 219). Z.B. entspricht das Zeichen <血> in der Umgangssprache und als eigenständiges Wort meist dem Laut /xiě/, wird in der Schriftsprache und als Morphem eines polysyllabischen Wortes aber /xuè/ ausgesprochen. Beim Schreiben mithilfe von Pinyin-Eingabemethoden sind solche Schriftzeichen ebenso nur mit der korrekten Aussprache in der Verwendungssituation zu verarbeiten.[197]

Ein Heteronym des dritten Falls meint ein Schriftzeichen, das trotz derselben Bedeutung und Verwendungssituation mit zwei oder mehreren Lesarten im Beijinger Dialekt lesbar ist. Ursache des Phänomens war die Unstandardisierung des Dialektsystems, auf dessen Aussprache Standardchinesisch basiert. In dem Prozess der nationalen Standardisierung der chinesischen Schrift wurde die Aussprache solcher Schriftzeichen im Putonghua standardisiert, weshalb eine standardisierte Aussprache behalten und die restlichen als inoffizielle Varianten definiert werden (vgl. ibid.: 224f). Da die meisten Pinyin-Eingabemethoden nur auf dem Aussprachestandard basieren, sind solche Schriftzeichen meist nur mit der Standardaussprache eingebbar. Aber die Standardisierung wird ab 1950er bis heute stetig in Festlandchina und Taiwan getrennt aktualisiert. Bei den Zeichen mit verwirrender Aussprache bieten manche intelligenten Eingabemethoden wie die Sogou-Pinyin-Eingabesoftware die Funktion des automatischen Matchings zu der korrekten Standardaussprache an. Bspw. wurde das Wort <呆板> (*unflexibel*) nach dem veralteten Standard als /áibǎn/, aber nach dem akutellen als /dāibǎn/ gelesen (nach „Putonghua Yiduci Shenyin Biao“ 1985). Die Korrektur der veralteten Lesart wird in Klammer angezeigt:

Abb. 4-17: Die automatische Korrektur zur aktuellen Standard-Aussprache eines Heteronyms mit variierten Aussprachen[198]

Zusammengefasst sind bei der Eingabe der Heteronyme Zeichen-, Wörter- und Sprachkenntnisse der PC-Benutzer erforderlich, um diese in der korrekten Aussprache zu codieren und mit Pinyin-Eingabemethoden zu schreiben. Trotz Verfahren der künstlichen Intelligenz kann das Computersystem in den meisten Fällen die falsche Aussprache in einer bestimmten Verwendungssituation nicht erkennen. Auf der Seite des Silben-Zeichen-Zusammenhangs, ergo der Entsprechung einer Silbe zu mehreren Schriftzeichen (Homophonie), leistet die künstliche Intelligenz hingegen einen großen Beitrag.

[197] Nach Eingabetest mit Sogou-Pinyin-Eingabemethode.
[198] Nach Eingabetest mit Sogou-Pinyin-Eingabemethode.

Wie die Zeichen- sowie Worthäufigkeit ist die Silbenhäufigkeit der chinesischen Sprache sehr ungleichmäßig verteilt. Unter rund 7.000 Schriftzeichen kann die Menge der zugehörigen homophonetischen Silben (ohne Tonberücksichtigung) zwischen eins bis über hundert divergieren. Unten stehend werden die 45 häufigsten Silben im Putonghua und die Anzahl der Homophone (basierend auf der statistischen Erforschung der 7.376 gemeingebräuchlichen Schriftzeichenangegeben) angegeben (vgl. Li M: Kap. 2.3).[199]

Rang	**Silbe & Zeichenanzahl**	**Zeichenanzahl mit Tonem: Ton/Beispielzeichen/Anzahl**			
1	yi 120	1/衣/16	2/移/27	3/矣/13	4/意/64
2	ji 113	1/肌/36	2/脊/29	3/己/11	4/技/37
3	yu 105	1/迂/6	2/余/35	3/宇/18	4/芋/46
4	fu 89	1/夫/11	2/福/40	3/府/14	4/咐/24
5	zhi 87	1/支/18	2/侄/14	3/旨/16	4/志/39
6	li 84	1/哩/1	2/狸/22	3/李/15	4/粒/46
7	xi 83	1/西/49	2/席/10	3/洗/8	4/系/16
8	qi 77	1/柒/17	2/其/34	3/乞/11	4/契/15
9	jian 73	1/奸/22	---	3/简/23	4/建/28
10	yan 70	1/咽/15	2/言/16	3/眼/17	4/彦/22
11	shi 70	1/虱/16	2/十/14	3/始/7	4/是/33
12	wei 61	1/威/13	2/维/15	3/委/16	4/尉/17
13	bi 59	1/逼/2	2/鼻/2	3/比/9	4/臂/46
14	wu 59	1/污/10	2/无/13	3/伍/15	4/物/21
15	ju 56	1/居/17	2/局/6	3/咀/9	4/拒/24
16	xian 56	1/先/13	2/贤/17	3/蚬/12	4/限/14
17	zhu 53	1/朱/16	2/竹/10	3/嘱/8	4/住/19
18	lu 48	1/噜/2	2/卢/14	3/鲁/8	4/路/24
19	shu 48	1/抒/17	2/熟/7	3/署/9	4/术/15
20	qian 46	1/千/21	2/虔/10	3/浅/5	4/倩/10
21	you 46	1/幽/7	2/油/17	3/友/9	4/右/13
22	gu 45	1/姑/15	---	3/骨/19	4/固/11
23	jing 45	1/精/17	---	3/井/11	4/劲/17
24	jie 44	1/皆/10	2/劫/22	3/姐/2	4/届/10
25	pi 44	1/砒/12	2/皮/16	3/匹/7	4/屁/9
26	hui 43	1/灰/12	2/茴/5	3/虺/4	4/晦/22
27	hu 43	1/呼/8	2/糊/18	3/虎/4	4/护/13
28	jiao 43	1/交/19	---	3/狡/13	4/较/11
29	chi 42	1/吃/14	2/弛/11	3/呎/7	4/叱/10
30	bo 41	1/玻/9	2/薄/26	3/跛/2	4/擘/4
31	yin 41	1/姻/14	2/吟/12	3/引/9	4/印/6
32	di 40	1/低/4	2/狄/12	3/抵/10	4/地/14
33	jia 39	1/加/18	2/夹/9	3/假/7	4/架/5
34	xie 39	1/些/5	2/协/9	3/写/2	4/谢/23
35	ying 39	1/英/13	2/迎/17	3/影/5	4/映/4
36	jin 38	1/津/11	---	3/仅/9	4/近/18

[199] Die 7.376 Schriftzeichen umfassen 7.000 gemeingebräuchliche Schriftzeichen in Festland China, 5.401 häufige sowie 1.719 weniger häufige Standardschriftzeichen in Taiwan; statistisch überdecken sich insgesamt 6.744 Schriftzeichen in Standards beider Regionen.

Rang	Silbe & Zeichenanzahl	Zeichenanzahl mit Tonem: Ton/Beispielzeichen/Anzahl			
37	zi 38	1/兹/22	---	3/姊/11	4/恣/5
38	yao 37	1/吆/7	2/爻/19	3/咬/5	4/要/6
39	ke 36	1/科/17	2/咳/3	3/可/4	4/克/12
40	yuan 36	1/宛/7	2/元/22	3/远/1	4/怨/6
41	xun 36	1/勋/9	2/巡/15	---	4/汛/12
42	si 35	1/丝/16	---	3/死/1	4/寺/18
43	xiao 35	1/削/20	2/淆/2	3/晓/4	4/肖/9
44	qu 35	1/区/13	2/渠/14	3/娶/4	4/去/4
45	xu 35	1/吁/13	2/徐/1	3/酗/6	4/序/15

Tab. 4-2: Die 45 häufigsten Silben im Standardchinesischen (nach Li M: Kap. 2.3)

Die 45 Silben entsprechen 10,84% der insgesamt 415 Silbenmöglichkeiten. Die zu ihnen entsprechenden Schriftzeichen lassen sich auf 2.452 beziffern, was zugleich 33,24% der Gesamheit entspricht. Im Durchschnitt entfallen 17,77 Homophone auf jede Silbe (7.376 Zeichen geteilt durch 415 Silben). Gleichsam entfallen auf die häufigste Silbe /yi/ alleine 120 Homophone und damit fast das siebenfache vom durchschnittlichen Wert. Auch die Initial-, Final- und Tonhäufigkeit kann anhand der Tabelle errechnet werden. Unter den 45 Silben sowie 2.452 Zeichen ist das Vertretungsinitial ‚y' für insgesamt acht Silben bzw. 494 Zeichen (20,15%) am häufigsten, welches sechs mit [i] und zwei mit [y] anfangende Silben in Halbpinyin darstellen kann. Unter den konsonantischen Initialen ist ‚j' für acht Silben bzw. 451 Zeichen (18,39%) am häufigsten. Der wahrscheinlichste Final ist /i/, der für 13 Silben sowie 892 Zeichen benötigt wird. Am zweithäufgisten ist /u/ mit zehn Silben bzw. 511 Zeichen. Die Häufigkeit der Töne kann anhand der Tabelle ermittelt werden: 4. Ton (837 Zeichen, 34,14%) > 1. Ton (630, 25,69%) > 2. Ton (585 Zeichen, 23,86%) > 3. Ton (400 Zeichen, 16,31%).

Je mehr Homophone eine Silbe hat, desto mehr Zeit wird zur Kandidatenauswahl benötigt. Für Schriftzeichen, die qua ihrer Aussprache häufiger produziert werden, ist es deswegen notwendig, sie in der Einheit Wort oder Satz einzugeben. Als die am häufigsten auftretenden Initiale sind ‚y' und ‚j' ebenso die häufigsten Halbpinyin-Codes der Schriftzeichen. In diesem Fall ist es manchmal effizienter, Pinyin in Vollform statt in Halbform einzugeben, damit das System effektiver Kandidaten ausfiltern kann.

4.3.2 Silbenformen und Ambiguitätsfälle von Silbensegmentation

In Kap. 3.4.2 (S. 167) wurde die Struktur der chinesischen Silbe eingeführt und in Kap. 3.4.3 (S. 169f) die Pinyin-Transkription und die Initial- sowie Finallaute vorgestellt. Es gibt insgesamt 22 Varianten für Initial (21 konsonantische und das Null-Initial, das nach dem ersten Phonem einer Silbe leer oder in /y/ oder /w/ dargestellt wird) und 39 für Final. Theoretisch könnte es insgesamt 858 Silben (22×39) geben, aber im modernen Mandarin werden nur ca.

415 davon sprechsprachlich verwendet.[200] Die Daten von den vorhandenen Pinyin-Silben sind Grundlage für die Segmentation der eingegebenen Pinyin-Kette in Silben und der erste Schritt im Arbeitsprozess der Eingabe mit Pinyin-Eingabemethoden. Dieser Prozess wird als Silbensegmentation (eng.: pinyin segmentation, chi.: 音节切分) bezeichnet. Er kann sowohl als eigenständiger Schritt (in Silben für einzelne Zeichen) als auch zusammen mit der Wortsegmentation (als mono- oder polysyllabisches Pinyin-Wort) durchgeführt werden. In Kap. 4.3 wird nur die zeichenorientierte Silbensegmentation berücksichtigt.

In Kap. 4.1.3 (S. 197) wurden die vier Formen des Pinyin-Codes vorgestellt, nämlich Voll-, Doppel-, Halb- und Mischform. Da in der Doppelform alle Silben immer mit zwei Tasten repräsentiert werden, gibt es in diesem Fall keine Ambiguität. In den anderen drei Formen können bei der Zerlegung in Silben zwei oder mehrere Möglichkeiten existieren. Zur Erforschung der Ambiguität der Silbensegmentation wird nachfolgend zuerst die Grundform (Vollpinyin) analysiert. Darauf basierend erfolgt eine erweiterte Betrachtung der Situation bei Halb- sowie Mischform. Zusammengefasst können Ambiguitäten der Segmentation in Vollform in drei Fälle gegliedert werden (vgl. Liu ZY 2007: 56):

1. Silbenkette mit Buchstaben N, G oder R in der Mitte, der sowohl als erster als auch als letzter Buchstabe einer Silbe auftreten kann. Doppeldeutigkeit entsteht, wenn A, O oder E hinter N, NG oder ER folgen. Die drei Vokale können in diesem Fall sowohl mit dem vorderen Konsonant Silben bilden (wie 那 /nà/, 个 /gè/ und 热 /rè/), als auch eigenständig sein (啊 /ā/, 哦 /ò/ und 恶 /è/) oder eine neue Silbe anfangen (wie 安 /ān/, 嗯 /èn/ und 欧 /ōu/). Bspw. kann die Buchstabenfolge ‚FANGAN' sowohl als ‚FANG'AN' als auch ‚FAN'GAN' zerlegt werden, die (jeweils am wahrscheinlichsten) den Wörtern <方案> (/fāng'àn/, *Entwurf*) und <反感> (/fǎn'gǎn, *etw. nicht leiden können*) entsprechen.
2. Bei einer Kette, die sowohl einer Silbe als auch zwei Silben entsprechen kann. Diese Ambiguität kommt vor, wenn ein Final gleichzeitig aus Reimkopf (I, U oder Ü), Reimkern (A, O oder E) und optional Reimende (N/NG oder I/U) zusammengesetzt ist, da die drei Buchstaben für den Reimkern auch Anfangs- oder alleiniger Buchstabe einer Silbe sein können. Ein Beispiel dafür ist XIAN, das entweder aus einer Silbe (Beispielzeichen dazu 鲜 /xiān/, *lecker*) oder aus zwei Silben (Beispielwort 西安 /xī'ān/, *Stadtname*) besteht.
3. Bei dem Diphthong AO, dessen Buchstaben auch zu zwei verschiedenen Silben gehören können. Die Buchstaben A oder O können einerseits Final einer Silbe, andererseits eine ei-

[200] Die genaue Anzahl der Silbenmöglichkeiten ist nach Meinung verschiedener Experten abweichend: ca. 400 (nach Li Mu), ca. 410 (nach Xu YC 2008: 46) und 415 (nach Cai 2005: 42).

genständige Silbe darstellen. So ist es auch bei der direkten Kombination AO (wie 猫 /māo/, *Katze* und 奥 /ào/, *tiefe Bedeutung*). Mithilfe von grammatischen Analysen kann der Pinyin-Code HAOLAO sowohl den Zeichenketten <好老> (/hǎo lǎo/, *sehr alt*) wie auch <好辣哦> (/hǎo là o/, *Wie scharf!*) entsprechen.

Solche Silben können einerseits mit dem Zeichen < ' > manuell getrennt werden, andererseits sind sie bei intelligenten Eingabesoftwares in den meisten Fällen automatisch zerlegbar. Im Normalfall werden Zeichenketten, die nach den verschiedenen Varianten der Silbentrennung gültig sind, weiter nach lexikalischen und grammatischen Kenntnissen verarbeitet und ausgefiltert. Der Verarbeitungsprozess der Pinyin-Kette ‚XIAN' zu Schriftzeichen läuft bspw. ab, wie in Abb. 4-18 dargestellt:

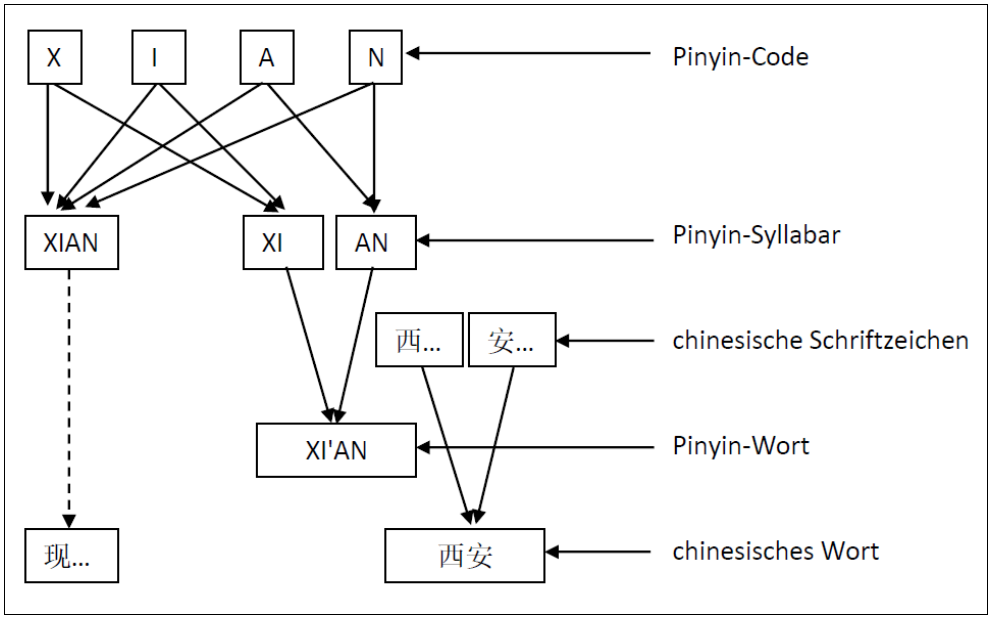


Abb. 4-18: Der Verarbeitungsprozess von ‚XIAN' zu Schriftzeichen

Halbpinyin meint den Initial oder ersten Buchstaben einer Silbe. Aus diesem Grund passieren Ambiguitäten hauptsächlich bei einem Initiallaut, der mit zwei Buchstaben dargestellt wird (wie ZH, CH und SH), und bei den Silben vom Null-Initial, die mit A, O oder E anfangen. Der Code ZH in Halbform kann bspw. einerseits auf eine Silbe mit ZH als Initial verweisen, andererseits zu zwei Silben passen, die jeweils mit Z und H anfangen. Bei A, O und E kann es bei der Silbentrennung zu Ambiguitäten kommen, weil sie in Vollform sowohl mit dem vorderen Initial Silben bilden, als auch als Halbform-Code für mit ihnen angefangene Silben stehen können. In Abb. 4-19 wird die Wahlliste von dem Code ZHE als Beispiel angezeigt. Die Wörter aus der Wörterdatenbank und die optimalen Phrasen sowie Sätze, die zu den verschiedenen Silbentrennungsmöglichkeiten passen, werden angeboten. So wird diese Pinyin-Kette bei

der Umwandlung sowohl ohne Segmentation in einzelne Zeichen (bspw. 蛰 /zhé/, *Winterschlaf;* der erste Kandidat in Abb. 4-19) verarbeitet, als auch in Halbform in drei Silben (Z'H'E) zerlegt, die der Phrase <坐会儿> (/zuò huì er/, *sich kurz setzen;* der zweite Kandidat in Abb. 4-19) entsprechen.

Abb. 4-19: Beispiel für die Ambiguität der Segmentation einer Pinyin-Kette in Halbform[201]

Alle Ambiguitätsfälle in Voll- sowie Halbpinyin können ebenso in Mischpinyin auftreten. Zwischen zwei benachbarten Silben, die jeweils in Voll- und Halbpinyin eingegeben werden, finden häufiger Mehrdeutigkeiten der Silbensegmentation statt. Die Buchstaben, die zur Ambiguität führen, sind ebenfalls N, G, R sowie A, O und E. Sie können sowohl eine Silbe beenden als auch eine Silbe anfangen (nämlich als Halbform-Code einer Silbe). Die Ambiguitätsfälle lassen sich mit exemplarischen Buchstabenfolgen erläutern. So kann ‚YIN' bspw. sowohl eine Silbe repräsentieren (wie das Zeichen <因> /yīn/, *Ursache*) als auch als Mischpinyin wie zwei Silben behandelt werden (/yi'n/ z.B. für <一年> /yì nián/, *ein Jahr*).

Wie auf S. 196f erwähnt wurde, sind Pinyin-Eingabemethoden zwischen Voll- und Doppelformstatus umschaltbar. Im Vollformstatus sind Voll-, Halb und Mischform möglich, weshalb vor der Silbensegmentation meist die Form des Pinyin-Codes erkannt werden muss. Betrachtet man die drei Formen zusammen, können mehrmalige Ambiguitätsfälle binnen einer Pinyin-Kette entstehen. Zum Beispiel kann der Code ‚ZHANG' eine Silbe (alle mit /zhang/ ausgesprochenen Schriftzeichen) darstellen, aber auch als Mischpinyin in zwei (/zh'ang/, /zhan'g/ oder /z'hang/) oder drei Silben (/zha'n'g/, /zh'an'g/ oder /z'han'g/) zerlegt werden.

Manuelle und maschinelle Silbensegmentationen haben beide unvermeidbare Nachteile. Wenn das Silbentrennungszeichen mit eingetippt wird, kann es den Schreibgedankengang stören und daher das Schreibtempo verlangsamen. Wenn das Silbentrennungszeichen immer weggelassen wird, könnte es mehrere Segmentationsmöglichkeiten geben, wodurch sich die Zeichenkettenkandidaten vermehren. Dies belastet das künstliche Gehirn erheblich und auch die Zeit zur Zeichenauswahl kann länger dauern. Es ist PC-Benutzern deswegen zu empfehlen, sich abhängig vom einzugebenden Inhalt zunächst für die eingestellte Pinyin-Form und danach für oder gegen eine manuelle automatische Silbentrennung zu entscheiden (vgl. Wang XL 2005: 106).

[201] Nach Eingabetest mit Sogou-Pinyin-Eingabesoftware.

4.3.3 Automatische Segmentation von Pinyin-Ketten in Voll-, Halb- und Mischform

Soweit kein Silbentrennungszeichen manuell eingesetzt wird, können Ambiguitäten bei der Segmentation existieren. Ohne die zum Eingabeziel geeignete Silbensegmentation ist die korrekte Zeicheneingabe ebenfalls unmöglich. Für die intelligenten Pinyin-Eingabemethoden ist es notwendig, nach einem bestimmten Prinzip die verschiedenen Segmentationsmöglichkeiten zu verarbeiten und durch Analysen mit lexikalischen sowie grammatischen Erkenntnissen die wahrscheinlichste Silbensegmentation und die entsprechenden Zeichenketten herauszufinden. Zur Erforschung der Silbensegmentation muss an dieser Stelle zuerst der Begriff *legales Pinyin* (eng.: legal Pinyin, chi.: 合法拼音) eingeführt werden. Legales Pinyin wird auch als rationale Silbe bezeichnet und meint einen Buchstaben oder eine Buchstabenfolge, der/die für eine Silbe des Chinesischen in Pinyin stehen kann. Für das Computersystem stellt sich bei der Silbensegmentation die Aufgabe, legales Pinyin herauszufinden und Begrenzungen zwischen zwei Silben einzusetzen (vgl. Liu ZY 2007: 57). Eine Pinyin-Kette, aus der ein Zeichen oder eine Zeichenfolge erzeugt werden kann, wird als legale Pinyin-Kette bezeichnet. Sie kann auf der Einheit von einem Zeichen, einem Wort oder einem Satz basieren (vgl. ibid.: 61).

Zur Silbensegmentation können verschiedene Methoden angewendet werden. Eine relativ präzise Methode der vollautomatischen Segmentation basiert auf den statistischen Analysen eines großen Pinyin-Korpus. Damit werden Korpora mit zahlreichen in Pinyin transkribierten chinesischen Texten bezeichnet, auf deren Grundlage Pinyin-Sprachmodelle herausgefunden werden. Durch diese Methode wird jene Segmentationsmöglichkeit einer Pinyin-Kette fest angelegt, die nach statistischen Analysen in diesem Kontext unter allen Segmentationsvarianten die größte Wahrscheinlichkeit besitzt. Für die Praktikabilität einer Pinyin-Eingabesoftware ist diese Methode unzweckmäßig, denn sie kostet sowohl mehr Arbeitsspeicher als auch zusätzliche Zeit für Pinyin-Korrekturen, wenn die wahrscheinlichste Segmentationsmöglichkeit für das Eingabeziel ungeeignet ist (vgl. Wang XL 2005: 106f).

In der Praxis kann die Silbensegmentation bei den Pinyin-Eingabesoftwares mithilfe von Regeln der Silbengrenzmerkmale durchgeführt werden, ergo der auf Silbenregeln basierenden Methode. In Fällen, in denen Ambiguität in großem Maße vorherrscht, ist eine manuelle Silbenzerlegung obligatorisch. Die Daten, von denen diese Anwendung unterstützt wird, bestehen aus Listen von Initiallauten und von gültigen Silben im Standardchinesischen in Pinyin (vgl. ibid.: 107). Mit den verschiedenen Pinyin-Formen zusammen betrachtet bestimmt die Liste der gültigen Silben die Silbenmöglichkeiten in Vollform und die Liste von Initiallauten die Möglichkeiten in Halbform.

Anhand der Informationen des letzten Kapitels und Tab. 3-4 (S. 168f) können die Möglichkeiten für legales Pinyin aufgelistet werden. In Vollform kann eine Silbe aus einem (wie /a/) bis zu sechs Buchstaben (wie /zhuang/) bestehen. Solche Silben enden entweder mit einem Vokal oder einem der drei Konsonanten N, G oder R. In Halbform ist der Code einer Silbe ihr erster Buchstabe oder Initial. Daher haben die in Halbform geschriebenen legalen Silben folgende Varianten: alle für konsonantische Phoneme stehenden Buchstaben (einzelne Buchstaben und ZH, CH und SH), die Buchstaben als Vertretungsinitial W (für [u]) und Y (für [i] und [y]) und die als erster Buchstabe auftretenden Vokale A, O und E.

Die Silbensegmentation läuft meistens synchron zum Code-Eintippen. Wenn die bisher eingegebenen Buchstaben eine gültige Silbe oder einen Teil einer gültigen Silbe bilden können, wird die Segmentation nicht gebraucht. Ein Silbentrennungszeichen wird erst eingesetzt, wenn zwei Buchstaben nacheinander folgen, die definitiv zu verschiedenen Silben zählen. Zum Beispiel kann die Silbentrennung zwischen I und H in der Pinyin-Kette NIHAO (entspricht dem Wort <你好> /nǐhǎo/, *Hallo* [bei normaler Anrede]) nach dem Eintippen von H festgelegt werden, denn innerhalb derselben Silbe kann hinter I nur N oder NG auftreten. Mit dieser Grenzregel zwischen Vokal und Konsonant kann Silbensegmentation in den meisten Fällen korrekt durchgeführt werden. Ambiguitätsverursachende Buchstaben wie N, G oder R müssen als Sonderfall weiter verarbeitet werden. In solchen Ketten muss näher bestimmt werden, ob die hinteren und vorderen Buchstaben mit oder ohne N, G oder R legale Silben bilden können (vgl. Wang XL 2005: 107). Z.B. muss sich ein Silbentrennungszeichen in ‚NINHAO' (entspricht dem Wort <您好> /nínhǎo/, *Hallo* [bei höflicher Anrede]) zwischen N und H befinden, da NHAO ungültig ist, NIN und HAO isoliert jedoch jeweils gültig sind.

Diese Methode ist zwar bei der Programmierung von Eingabesoftware und der computergestützten Anwendung am einfachsten, stößt aber häufig auf Ambiguitätsfälle. Eine Silbe wie XIAN, die selbst in zwei Silben zerlegt werden kann, wird mit dieser Methode meistens als alleinmögliche Gesamtsilbe behandelt. So muss beim Wort <西安> /xī'ān/ bspw. bei der Eingabe das Silbentrennungszeichen <'> manuell hinzugefügt werden.

Um die Korrektheit der Silbensegmentation zu erhöhen sowie die manuelle Silbentrennung zu ersparen, wurde in der jüngeren Vergangenheit das ‚Zustandsraummodell' (chi.: 状态空间模型) bei der Silbensegmentation mancher Pinyin-Eingabemethoden eingeführt.[202] Mit dieser Methode können alle Optionen der Silbensegmentation behalten und die davon wahrscheinlichste nach dem sprachlichen Kontext ausgewählt werden (vgl. Liu/Wu/Liu 2008: 36).

[202] Zuerst veröffentlicht in Liu Zhengyi 2007 (Doktorarbeit an der Universität Anhui).

Hauptidee dieser Methode ist die Traversierung: Man nimmt imaginäre Knoten an, die nach einer bestimmten Reihenfolge jeden Buchstaben der Pinyin-Kette durchlaufen. Nach jedem eingetippten Buchstaben wird ein Pinyin-Phonem-Knoten automatisch zwischen ihm und seinen vorderen Buchstaben in das Zustandsraummodell eingesetzt. Weiterhin wird überprüft, ob er mit einem bis mehreren vorderen Buchstaben zusammen eine legale Silbe bilden kann. Wenn ja, so wird diese Silbe im Modell gespeichert und ein Silbenknoten dahinter eingesetzt. Diese gespeicherte legale Silbe richtet sich weiter an den nach ihr eingetippten Buchstaben und kann korrigiert werden, wenn die Silbenbildung weiter geht. Am Ende richten sich die von den Silbenknoten zerlegten benachbarten Silben aneinander aus, um nach Erkenntnissen der Wörter und/oder der Grammatik die wahrscheinlichste Silbensegmentation weiter zu bestimmen (vgl. ibid.: 36ff). Dieser Prozess kann anhand des folgenden Schaubilds beschrieben werden.

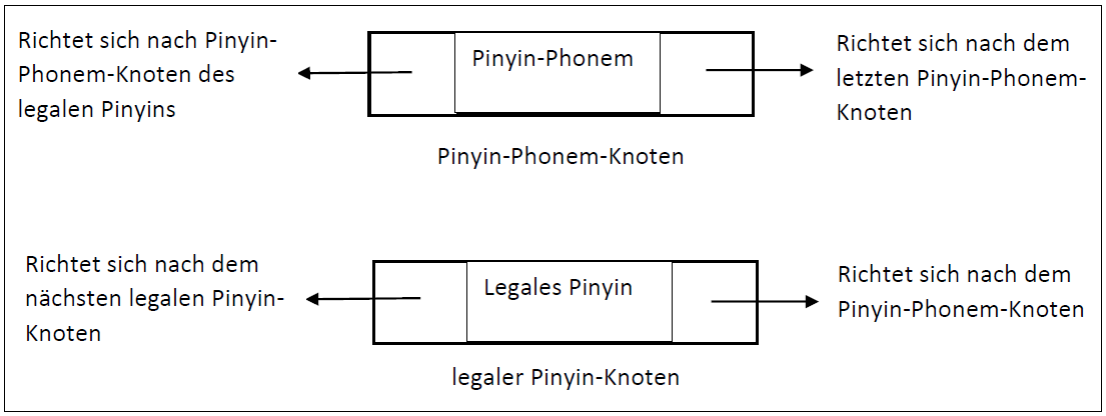


Abb. 4-20: Definition der Knoten bei Silbensegmentation (nach Liu ZY/Wu/Liu 2008: 36 [Übersetzung der Verfasserin])

Zur Veranschaulichung der Ambiguitäten der Silbensegmentation einer Pinyin-Kette wird nachstehend ‚XIANGANG' als Beispiel genommen. Zuerst werden alle Möglichkeiten der Silbensegmentation und der den zerlegten Silben entsprechenden Schriftzeichen vermutet, die in der folgenden Tabelle (Tab. 4-3) angegeben werden. Für ein detailliertes Ergebnis werden die acht Buchstaben der Pinyin-Kette bei der Analyse nacheinander hinzugefügt und die Silbenmöglichkeiten analysiert (siehe Spalte x bis Spalte g). Ein legales Pinyin in Vollform wird mit Nummerierung im runden Kreis markiert. Hinter einer Variante der Silbentrennung wird auch die Anzahl der gesamten zerlegten Silben angegeben, bspw. ‚xian'gang; 2' in der ersten Variante und der vorletzten Spalte. Nach Auflistung aller möglichen Varianten – ohne Unterscheidung von Voll-, Halb- oder Mischformen – wird diese Pinyin-Kette nach den beiden erwähnten Silbensegmentationsmethoden analysiert, konkreter also das Ergebnis mit der allgemeinen Methode samt Regeln der Silbengrenzmerkmale und der Methode mit Zustands-

raummodell. Anschließend werden die Silbensegmentationsvarianten durch den Eingabetest mit Sogou- sowie Microsoft-Pinyin-Eingabesoftware erläutert.

Code	**x**	**i**	**a**	**n**	**g**	**a**	**n**	**g**	**Mögliche Kandidaten**
legales Pinyin / Var.	x	xi ①	xia ②, a③	xian ④, an⑤, n	xiang ⑥, ang ⑦, g	ga⑧, a	gan⑨, an, n	gang⑩, ang, g	
1	x	xi	xia	xian	xiang	xian'ga; 2	xian'gan; 2	xian'gang; 2	岘港 (Ortsname); 蚬岗 (Orts-name); 宪刚 (Vor-name)
2			xi'a	xi'an	xi'ang	xia'n'ga; 3	xia'n'gan; 3	xia'n'gang; 3	夏南刚 (Eigenname)
3						xi'an'ga; 3	xi'an'gan; 3	xi'an'gang; 3	西安港 (Ortsname); 西安刚 (Satzglied)
4				xi'a'n	xian'g	xiang'a	xiang'an; 3	xiang'ang; 2	向昂 (sinnlos)
5					xia'n'g	xian'g'a	xian'g'an, 3	xian'g'ang; 3	西安高昂 (unlogisches Satzglied)
6					xi'an'g	xia'n'g'a	xia'n'g'an; 4	xia'n'g'ang; 4	---
7					xi'a'n'g	xi'a'n'g'a	xi'a'n'g'an; 5	xi'a'n'g'ang; 5	喜爱那个昂 (unlogisches Satzglied)
8							xian'ga'n; 3	xian'ga'n'g; 4	---
9							xia'n'ga'n; 4	xia'n'ga'n'g; 5	下年戛纳个 (sinnlos)
10							xia'n'g'a'n; 5	xia'n'g'a'n'g; 6	下能够按那个 (Satzglied)
11							xi'ang'an; 3	xi'ang'ang; 3	西昂昂 (sinnlos)
12							xi'an'ga'n; 4	xi'an'ga'n'g; 5	西安噶那个 (sinnlos)
13							xi'an'g'a'n; 5	xi'an'g'a'n'g; 6	西安关爱那个 (un-logisches Satzglied)
14							xi'a'n'ga'g; 5	xi'a'n'ga'n'g; 6	喜爱你噶那个 (sinnlos)
15							xi'a'n'g'a'n; 6	xi'a'n'g'a'n'g; 7	喜爱那个爱那个 (un-logischer Satz)

Tab. 4-3: Möglichkeiten zur Silbensegmentation und entsprechende Zeichenketten

In der Tabelle werden fünfzehn Varianten der Segmentation angegeben. Davon sind die erste, die dritte, die vierte und die elfte Variante in Voll- und die restlichen in Mischform. Nach Liu

sind mit der Methode des Zustandsraummodells vier Varianten als gültig anzuerkennen, nämlich ‚XIANG'ANG' (Var.-4), ‚XIAN'GANG' (Var.-1), ‚XI'ANG'ANG' (Var.-11) und ‚XI'AN'-GANG' (Var.-3). Die ersten zwei bisyllabischen Möglichkeiten werden bevorzugt berücksichtigt. Die anderen Varianten außerhalb von den vier werden automatisch ausgelassen (vgl. Liu / Wu / Liu 2008: 38).

Abb. 4-21: Die erste (1), die zehnte (2) und die 87ste Seite (3) der Wahlliste bei der Eingabe der Pinyin-Kette ‚XIANGANG' der Sogou-Pinyin-Eingabesoftware[203]

Abb. 4-22: Die erste Seite der Wahlliste bei der Eingabe der Pinyin-Kette ‚XIANGANG' von Microsoft-Pinyin-IME[204]

Wie die beiden Abbildungen zeigen, werden die logischen Zeichenketten (Wörter oder grammatisch gültige Zeichenkollokationen) mit der geringsten Silbenanzahl (ergo jene Silben mit den längsten Buchstabenfolgen) in der Wahlliste zuvorderst angeboten (siehe ① in Abb. 4-21 und Abb. 4-22). In beiden Eingabesoftwares werden die als /xian'gang/ ausgesprochenen Eigennamen zuerst angezeigt. Der Hauptunterschied bei der Verarbeitung bezieht sich auf die Größe der jeweils angehörigen Wörterdatenbank. In der Wörterdatenbank von Sogou sind beide Ortsnamen (eine vietnamesische Stadt und ein Stadtteil Hongkongs) gespeichert und werden als Kandidaten angeboten. Bei Microsoft-Pinyin-IME ist hingegen der Vorname eines japanischen Fußballspielers das einzige angegebene Wort zu dem phonetischen Wert. Der Name der Provinzhauptstadt Xi'an und die einzelnen Schriftzeichen der Pinyin-Silben ‚XIAN' und ‚XIANG' werden nachfolgend als Kandidaten aufgelistet, bis alle mit Silben ausgesprochenen Schriftzeichen in der Wörterdatenbank recherchiert sind. In dem Beispiel sind alle

[203] Nach Eingabetest mit Sogou-Pinyin-Eingabemethode; in dieser Abbildung wird das Silbentrennungszeichen vom System automatisch hinzugefügt; die hinweisende Silbentrennungsmöglichkeit entspricht immer dem ersten Zeichenkandidaten auf dieser Seite der Wahlliste.
[204] Nach Eingabetest mit Microsoft-Pinyin-IME.

Kandidaten von der zweiten bis zur 85. Seite (außer der Stadt Xi'an) Homophone oder Alternativzeichen mit dem phonetischen Wert /xian/ oder /xiang/ (siehe 2. Teil von Abb. 4-21). Die Schriftzeichen kürzerer Silben werden erst weiter hinten verzeichnet. In dem Beispiel befinden sich die mit /xi/ ausgesprochenen einzelnen Zeichen erst auf der zweiten Hälfte der 86. Seite (siehe 3. Teil Abb. 4-21). Anhand des Eingabetests sowohl mit der Sogou- als auch Microsoft-Pinyin-Eingabesoftware lässt sich konstatieren, dass bei der Silbensegmentation das Prinzip der minimalen Segmentation in Vollform beibehalten wird. Die Möglichkeit der minimalsten segmentierten Silbenanzahl wird vom System als erstes berücksichtigt. Die Pinyinkette am Beispiel kann am seltensten in zwei Silben zerlegt werden (nämlich ‚XIAN'GANG' oder ‚XIANG'ANG'). So werden in der Wahlliste die optimalen Zeichenketten nach den beiden Silbensegmentationsvarianten zuvorderst angezeigt. Weiterhin werden Varianten ab drei bis immer mehr Silben verarbeitet und zu den gültigen Zeichenketten von immer mehr Zeichen angeboten. Nachdem alle polysyllabischen Kandidaten angezeigt wurden, werden die einzelnen Schriftzeichen aufgelistet, von längeren bis zu kürzeren Pinyin-Silben. Zusammen betrachtet mit Tab. 4-3 werden nur die erste, die dritte und die vierte Variante der Silbensegmentation automatisch vom System verarbeitet. Die anderen Varianten in der Tabelle, die hauptsächlich in Mischform dargestellt werden und deren Zeichenkandidaten unnormal sind, werden nicht behandelt. Wenn eine nicht zum allgemeinen Sprachgebrauch gehörende Variante zu dem Eingabeziel passen würde, müsste entweder ein Silbentrennungszeichen manuell beigefügt oder Vollform eingegeben werden. Bei den allgemein gebräuchlichen Wörtern oder Idiomen kann die Eingabe in Halb- oder Mischform meistens problemlos gelingen, wie z.B. die Eingabe von <掩耳盗铃> (/yǎn'ěr dàolíng/, *sich selbstbetrügen*):

Abb. 4-23: Die Eingabe von ‚YEDL' für das Idiom <掩耳盗铃> (der 2. Kandidat) in Halbform[205]

4.3.4 Allgemeine sowie idiolektale Zeichen- und Worthäufigkeit

In Kap. 4.2.2 (S. 214) wurde erwähnt, dass die pragmatisch-statistischen Analysen entscheidende Erkenntnisse für Funktionalität und Effizienz der Eingabesoftware liefern. Bei intelligenten Pinyin-Eingabesoftwares werden hybride Konversionstechniken aus sprachlichem Verstehen und Statistiken am häufigsten angewendet. Überblickt man die verschiedenen chinesischen Eingabemethoden, können Häufigkeitsstatistiken diverse Beiträge leisten: 1) die

[205] Nach Eingabetest mit Sogou-Pinyin-Eingabemethode; wörtliche Übersetzung: *mit bedeckten Ohren eine Glocke stehlen*.

Verkürzung des Inputcodes von den häufigen Schriftzeichen, wie die abgekürzte Inputcodierung des Wubi-Eingabeschemas im ersten, zweiten sowie dritten Grad (siehe Kap. 4.1.2, S. 193); 2) die Anzeige der häufigen Kandidaten an den vordersten Stellen, wie bei der Wahlliste der Pinyin-Eingabemethode; 3) die Ausgabe der wahrscheinlichen Zeichenkandidaten auf Basis ihres kollokationären Kontextes. Wie in Abb. 4-12 bis 4-14 (S. 216f) abgebildet, basieren der zweite und dritte Anwendungsansatz auf dem Entwurf der Wissensdatenbank, genauer der Verwendungshäufigkeit eines einzelnen Wortes/Zeichens in dem Lexikon und der zusammenhängenden Wahrscheinlichkeit einer Zeichenkette in der linguistischen Datenbank der pragmatischen Statistiken. Beide Zelldatenbanken der Wissensdatenbank unterscheiden sich, wie bereits ausgeführt, jeweils in ihrem statischen und dynamischen Teil. Für den statischen Teil des Lexikons sind die Attribute der allgemeinen, voreingestellten Zeichen-/Worthäufigkeit obligatorisch. Für die Begründung der linguistischen statischen Wissensdatenbank werden ebenfalls die voreingestellten sprachlichen Regeln über unbestimmte sowie bestimmte Wörter eingeschrieben (vgl. Bai 1992: 37). In diesem Kapitel wird nur die Zeichen-/Wortfrequenz erforscht. Obwohl Einzelzeichen und Wörter sprachlich sowie schriftlich in den meisten Fällen verschiedene Einheiten sind, da sich die Mehrheit der chinesischen Wörter aus zwei oder mehreren Schriftzeichen zusammensetzt, wird ihre Häufigkeit nach denselben Prinzipien errechnet. Deswegen werden die statistischen Analysen der Zeichen sowie Wörter in diesem Kapitel erläutert. Die Grenze zwischen Zeichen und Wort aus linguistischem Blickwinkel wird Kap. 4.4.1 beleuchtet.

Die Zeichen- sowie Worthäufigkeit gliedert sich in allgemeine und idiolektale Häufigkeit. Die allgemeine Häufigkeit bezieht sich auf den gesamtsprachlichen Gebrauch und wird auf der Basis von Korpora mit Texten verschiedener Bereiche errechnet. Dabei ist die Präzision der allgemeinen Häufigkeit stark von der Größe des Korpus und der gleichmäßigen Verteilung unterschiedlicher Textsortenbereiche abhängig (vgl. Yao 1997: 172f).

Die Formel zur Errechnung der Frequenz lautet dabei ‚Auftrittshäufigkeit eines Zeichens/Wortes geteilt durch die gesamte Anzahl der Zeichen sowie Wörter im Korpus' (vgl. ibid.: 167). Im online verfügbaren ‚CN-Corpus' bspw., der aus 12.842.116 Wörtern und 19.455.328 Schriftzeichen besteht, tritt das Schriftzeichen <德> (/dé/, *Moral*) insgesamt 6.660-mal auf, jedoch in unterschiedlichen semantischen und morphologischen Kontexten. So kann es alleine in seiner Kernbedeutung *Moral* auftreten, aber auch als bedeutungstragendes Morphem, syllabisches Lehnmorphem (bei Verknüpfungen mit auf *Deutschland*/*Déguó* bezogenen Begriffen) und in Eigennamen. Die Zeichenfrequenz insgesamt lässt sich auf 0,342‰

beziffern (6.660/19.455.328). Als alleinstehendes monosyllabisches Wort tritt es 436-mal auf, was einer Wortfrequenz von 0,0456‰ im CN-Corpus Online entspricht.

In Kap. 3.2.2 wurde die Einstufung der 8.105 gemeingebräuchlichen Schriftzeichen in Festlandchina nach dem aktuellen Zeichenlistenstandard vorgestellt, die sich in häufige (3.500, darunter 2.500 sehr häufig), weniger häufige (3.000) und relativ seltene (1.605) Stufen untergliedern (vgl. ‚Tongyong Guifan Hanzi Biao' 2013; siehe auch Kap. 3.2.2, S. 142f). Im Sprachgebrauch ist die Verteilung verschiedener Schriftzeichen extrem asymmetrisch. Quantitative Korporastudien haben ergeben, dass bereits die 200 häufigsten Zeichen 50% aller chinesischen Texte ausmachen. Mit dem ‚Zeichenschatz' der tausend häufigsten Schriftzeichen sind ca. 90% der chinesischen Texte lesbar. Beherrscht man die häufigsten 7.000 ist ein unbekanntes Zeichen mit einer Wahrscheinlichkeit von weniger als 0,01% anzutreffen. Diese Ungleichmäßigkeit der Zeichenhäufigkeit wird in der folgenden Tabelle transparent.

Rg.	Zeichen	Häufigkeit in %	Gesamthäufigkeit (bis dahin) in %	Rg.	Zeichen	Häufigkeit in %	Gesamthäufigkeit (bis dahin) in %
1	的 /dè/	3,8375	3,837	180	或 /huò/	0,1339	52,45
2	一 /yī/	1,2523	5,090	240	很 /hěn/	0,1077	59,67
3	是 /shì/	0,9790	6,069	280	更 /gèng/	0,0877	63,49
4	在 /zài/	0,9449	7,014	360	织 /zhī/	0,0691	69,75
5	了 /le/	0,8194	7,833	400	证 /zhèng/	0,0611	72,35
6	不 /bù/	0,8053	8,638	560	河 /hé/	0,0385	80,02
7	和 /hé/	0,7476	9,386	600	粉 /fěn/	0,0347	81,48
8	有 /yǒu/	0,6933	10,07	1000	贯 /guàn/	0,0152	90,89
9	大 /dà/	0,6863	10,76	1500	亿 /yì/	0,0064	95,96
10	这 /zhè/	0,6398	11,40	1600	丛 /cóng/	0,0054	96,54
20	工 /gōng/	0,4554	16,72	2000	兔 /tù/	0,0029	98,14
26	以 /yǐ/	0,4207	19,37	2500	梗 /gěng/	0,0013	99,15
30	会 /huì/	0,3864	20,97	3000	劈 /pī/	0,0006	99,63
40	义 /yì/	0,3389	24,55	3500	啧 /zé/	0,0002	99,83
60	度 /dù/	0,2611	30,43	3900	赂 /lù/	0,00014	99,91
80	党 /dǎng/	0,2212	35,16	4000	蚬 /xiàn/	0,00012	99,92
100	十 /shí/	0,2024	39,34	5000	幽 /yōu/	0,000027	99,98
160	由 /yóu/	0,1490	49,66	6359	腴 /yú/	0,0000046	100,00 [auf Rundung]

Tab. 4-4: Die Rangliste der Zeichenhäufigkeit und Gesamthäufigkeit (nach Peng 1994: 44 [Appendix-1], nach: Chen ZW/Jin 1988: o.S.)[206]

Bei der Wortfrequenz herrscht ebenso eine starke Asymmetrie vor. In der folgenden Tabelle werden die Statistiken der Wörter-Überdeckungsquoten (ÜDQ) auf verschiedenen Häufigkeitsniveaus nach Chen YF/Zhu (2002a: 10) zitiert, wobei gezeigt wird, wie viel Prozent des Textkorpus von diesen Wörtern überdeckt wird (Anz.).

[206] Unter den aufgeführten Zeichen gibt es einige Heteronyme, für die nur ihre häufigste Aussprache angegeben wird.

ÜDQ %	20	30	40	50	70	90	95	97	98	99	100
Anz.	308	684	1.314	2.405	7.677	26.760	39.870	48.597	54.781	63.678	88.102

Tab. 4-5: Die Überdeckungsquote (ÜDQ) von Wörtern auf verschiedenen Häufigkeitsniveaus (Chen YF/Zhu 2002: 10)

Im Gegensatz zu der allgemeinen Häufigkeit kann die idiolektale Häufigkeit dialektal, regional, diachron und individuell bedingt sein. Das Schriftzeichen <唔> (mit der Aussprache /wú, *nicht*) tritt im Standardchinesischen bspw. nur selten auf (und wird häufiger mit dem Zeichen <不> /bù realisiert). Im Kantonesischen ist <唔> (mit der Aussprache /m4/) hingegen hochfrequent. Diachron verändert sich die Verwendung der Zeichen/Wörter jeden Tag. Bildungshintergrund, Beruf, Fachgebiet, soziale Faktoren der Lebensumgebung usw. führen dazu, dass sich der sprachliche Gebrauch bei jedem Zeichen/Wort erheblich unterscheidet. Nach Beschreibung mit mathematischen Parametern ist die allgemeine Häufigkeit jene Konstante, die den Sprachgebrauch innerhalb des heutigen chinesischen Schriftsystems am ehesten abbildet. Dem hingegen ist die idiolektale Häufigkeit eine Zufallsvariable, die sich bei jedem Individuum ständig verändert.

Aufgrund der starken Abweichung der idiolektalen Zeichen-/Worthäufigkeit sind viele Pinyin-Eingabesoftwares individuell orientiert, d.h. die in einen PC relativ häufig eingegebenen Schriftzeichen und Wörter werden automatisch gespeichert und nach wiederholter Eingabe weiter vorne angezeigt. Die Zusammenhänge der Zeichen-/Worthäufigkeit und der intelligenten Pinyin-Eingabemethode lassen sich insofern zusammenfassen, als dass die allgemeine Häufigkeit als Attribut im Wörterbuch der Wissensdatenbank eingeschrieben wird. Nach dem Installieren wird dann die Reihenfolge der homophonetischen Zeichen/Wörter nach idiolektaler Häufigkeit graduiert und dementsprechend umgestellt (vgl. Yuan 2008: 280). Die auf die idiolektale Zeichen-/Worthäufigkeit bezogenen Funktionen bei den intelligenten Pinyin-Eingabesoftwares sind vor allem folgende (vgl. ibid.: 279):

1) Die automatische und synchronische Umstellung der Kandidatenreihenfolge. Ein Zeichenkandidat (als Einzelzeichen, Wort, Phrase oder Satz), der individuell häufig gebraucht wird oder vor kurzem eingegeben wurde, wird in der Wahlliste nach vorne gezogen.
2) Die Begründung der individuellen Wörterdatenbank. Neben der voreingestellten entsteht nach der Softwareinstallation eine individuelle Wörterdatenbank, deren Daten von dem PC-Benutzer selbst eingeschrieben und verwaltet werden. In der voreingestellten Wörterdatenbank nicht aufgenommene Wörter, die mehr als dreimal in den Computer eingegeben werden, werden automatisch in der individuellen Wörterdatenbank gespeichert. Diese individuelle Wörterbank kann in vielen Fällen nach eigenen Bedürfnissen importiert/ex-

portiert werden, damit sie für eine andere Eingabesoftware oder in einem anderen Computer verfügbar ist.

3) Die automatische Aktualisierung der Wörterdatenbank. Jeden Tag entstehen neue Wörter – Ad-hoc-Bildungen, Termini, Neologismen. Einige verschwinden wie Modeerscheinungen nach kurzer Zeit, andere wandern in den chinesischen Wortschatz.[207] Mit der Funktion der automatischen Aktualisierung ist die Eingabesoftware in der Lage, Schriftmedien zu verfolgen und dadurch Schlagwörter der aktuellen Mediendebatte und beliebte Wörter der heutigen Gesellschaft automatisch aufzunehmen.

4) Das Runterladen und Installieren von fachlichen Zellwörterdatenbanken nach eigenem Gebrauch. Die Verwendungshäufigkeit der fachlichen Wörter ist erheblich von individuellen Faktoren bedingt. Um die intelligente Eingabesoftware zu individualisieren, stehen fachliche Datenbanken zum Herunterladen zur Verfügung.

Die Reihenfolge der homophonetischen Zeichen/Wörter ist deswegen sowohl von der allgemeinen relativen Häufigkeit als auch vom individuellen sowie synchronischen Gebrauch abhängig. Mithilfe des bereits oben verwendeten ‚CN-Corpus' werden die zehn häufigsten homophonetischen Wörter von /shishi/ und ihre Frequenz in der folgenden Tabelle angegeben.

Rang	**Wort & Pinyin**	**Bedeutung**	**Male des Auftretens**	**Frequenz in ‰**
1	事实 /shìshí/	*Tatsache*	2056	0,2151
2	实施 /shíshī/	*durchführen*	666	0,0697
3	时时 /shíshí/	*jede Zeit*	205	0,0215
4	逝世 /shìshì/	*sterben* (euphemistisch*: entschlafen, verscheiden*)	156	0,0163
5	适时 /shìshí/	*richtige Zeit*	94	0,0098
6	史诗 /shǐshī/	*historisches Gedicht*	71	0,0074
7	时事 /shíshì/	*aktuelle Nachricht*	62	0,0065
8	实时 /shíshí/	*Echtzeit*	38	0,0040
9	史实 /shǐshí/	*Tatsache in der Geschichte*	35	0,0037
10	试试 /shìshì/	*versuchen*	27	0,0028

Tab. 4-6: Die zehn häufigsten Homophonen von /shishi/ und ihre Frequenz

In der Praxis verändert sich die Kandidatenreihenfolge jedes Mal, nachdem ein Zeichen oder Wort eingegeben wird. Grundprinzipien sind zusammengefasst somit die automatische Reihenfolgenumstellung je nach idiolektaler Häufigkeit und die Bevorzugung der vor kürzerer Zeit eingegebenen Kandidaten. Exemplarisch wird dazu die erste Seite der Wahlliste von ‚SHISHI' nach zwei verschiedenen Eingaben angegeben. Die zeitliche Distanz betrug dabei nur wenige Minuten:

[207] Beispiele für letztere wären im Deutschen etwa *Public-Viewing*, im Chinesischen <二孩> (/èrhái/, *Zwei-Kind-[Politik]*).

Abb. 4-24: Die erste Seite der Wahlliste von ‚SHISHI' bei zwei verschiedenen Eingaben

4.4 Informationsverarbeitung im Wort

In Kap. 4.2.2 sowie 4.2.3 wurde die zentralisierte Rolle der Wortverarbeitung für die intelligente Eingabesoftware und die Ressource Wörterdatenbank vorgestellt. Da die satzstufige Laut-Zeichen-Konversion auf segmentierten Pinyin-Wörtern beruht, sind Wortsegmentation und Recherche der zur Satzanalyse gebrauchten Wortattribute die Voraussetzungen für die Satzverarbeitung. Kap. 4.4 zielt darauf ab, die linguistischen Eigenschaften des chinesischen Wortes und die darauf basierende Informationsverarbeitung zu erforschen.

4.4.1 Chinesische Wörter aus linguistischer Perspektive

In Kap. 4.2.1 (S. 209ff) wurden linguistische Unterschiede zwischen den indogermanischen Sprachen und der chinesischen Sprache vorgestellt. Die kleinste Grundeinheit der Sprache stellt dabei den wesentlichsten Unterschied dar. In den indogermanischen Sprachen sind Wörter die „kleinste[n], relativ selbstständige[n] Träger von Bedeutung, die im Lexikon kodifiziert sind" (Bußmann 2002: 750). Auf der orthographisch-graphemischen Ebene werden Wörter durch Spatien markiert (vgl. ibid.). Die Wortforschungslehre der Morphologie fokussiert die „Untersuchung von Form, innerer Struktur, Funktion und Vorkommen der Morpheme" (ibid.: 450). Dem hingegen ist im Chinesischen die kleinste, bedeutungstragende Grundeinheit und das im Lexikon kodifizierte Objekt das Schriftzeichen, welches in den meisten Fällen als Morphem zur Wortbildung oder als eigenständiges Wort auftritt. Prinzipiell ist ein chinesisches Wort die Zusammensetzung von einer/einem bis mehreren Silbe(n) / Schriftzeichen / Morphem(en), da sich das moderne Chinesisch aus dem archaischen Chinesisch entwickelte, in welchem die Wörter hauptsächlich betont-monosyllabisch waren. Die chinesische Schrift ist vor diesem Hintergrund als logographischer/morphologischer Schrifttyp entstanden. Mit der Erweiterung und Evolution des Wortschatzes haben sich immer mehr polysyllabische Wörter entwickelt, die aus zwei oder mehreren vorhandenen Schriftzeichen zusammengesetzt

sind. Aus diesem Grund bestehen die Forschungsaufgaben der chinesischen Lexikologie vor allem darin herauszufinden, wie Zeichen zur Wortbildung verwendet werden und wie die bestehenden Zeichen mit dem neu gebildeten Wort zusammenhängen (vgl. Lu 2008: 149).

Laut GB12200.1-90 ist ein Wort „die kleinste, eigenständig verwendbare sprachliche Einheit, wie 大 [/dà/, *groß*], 国家 [/guójiā/, *Staat*] und 奥林匹克 [/àolínpǐkè/, *Olympics*]“ (Kap. 4.1.2.8 [Übersetzung der Verfasserin]). Es gibt kein Spatium zwischen zwei Wörtern, so dass es schriftlich kein Merkmal für die Wortgrenze gibt. Erforschungen der Wortbildungsregeln des Chinesischen unterscheiden sich anhand der zu analysierenden Grundeinheiten, der Perspektive des Morphems und des Schriftzeichens (vgl. Li K 2002: 135f).

Nach dem Form-Aussprache-Sinninhalt-Zusammenhang ist ein Sinogramm im sprachlichen Kontext meist ein monosyllabisches Morphem (vgl. Kap. 3.4.1, S. 163ff). Ein Morphem kann sowohl ein eigenständiges Wort darstellen als auch als Bestandteil eines Mehr-Gramm-Wortes auftreten. Ein aus einzelnen Morphemen bestehendes Wort heißt *simples Wort* (chi.: 单纯词 /dānchún cí/; eng.: simple word). Hingegen ist ein aus mindestens zwei Morphemen zusammengesetztes Wort ein *komplexes Wort* (复合词 /fùhé cí/; compound word) (vgl. GB/T 12200.2-94: Kap. 4.1.4.4 & 4.1.4.5). Statistisch betrachtet sind 95% der chinesischen Morpheme monosyllabisch, ergo schriftlich mit einem Schriftzeichen identisch (vgl. Xu YC 2008: 103).[208] Anders formuliert ist ein monosyllabisches Morphem immer mit einem Schriftzeichen identisch und ein Schriftzeichen hochwahrscheinlich ein bedeutungstragendes Morphem. Den morphemtragenden Schriftzeichen steht eine Minderheit an sog. Nicht-Morphemzeichen (非语素字) gegenüber. Dabei handelt es sich um Zeichen, die keine autosemantische Bedeutung tragen und zur Bildung der polysyllabischen Morpheme benötigt werden, wie <咖> (/kā/, in <咖啡> /kāfēi/, *Kaffee*) (vgl. GB/T 20532-2006: 4).

Abgesehen von den Wortbildungsregeln können die chinesischen Wörter auch auf Basis der Silben- sowie Grammlänge und der Wortart betrachtet und klassifiziert werden. Aus syllabischen Aspekten können die Wörter in monosyllabische (chi.: 单音节词; eng.: monosyllabic word), bisyllabische (双音节词; disyllabic word) und polysyllabische Wörter (多音节词; polysyllabic word; im Normalfall auch inklusive der diasyllabischen) unterschieden werden. Vor dem Hintergrund der schriftlichen Länge gibt es Einzelzeichen-[209], Bigramm-, Trigramm-, Vier-Gramm-Wörter usw. Wie lang ein chinesisches Wort maximal sein kann, changiert von Quelle zu Quelle, in den meisten Fällen jedoch zwischen sechs bis acht Zeichen. Nach Analy-

[208] Ein aus einem monosyllabischen Morphem und <儿> (Erhua-Yin-Träger) bestehendes Wort ist zwar monosyllabisch, wie <花儿> (/huār/, *Blümchen*), aber zugleich ein suffixoides komplexes Wort.

[209] Auch Unigramm genannt; monosyllabische und Ein-Zeichen-Wörter sind (außer bei Wörtern mit Erhua-Yin) identisch.

se der 56.008 am häufigsten gebrauchten Wörter bestehen 3.181 aus Einzelzeichen, was einer Rate von 5,68% entspricht. Beim Rest entfallen 40.351 auf Bigramm-Wörter (72,05%), 6.459 sind trigrammig (11,53%), 5.844 viergrammig (10,43%) und 162 bestehen aus fünf bzw. mehr als fünf Zeichen (0,29%) (vgl. ‚Xiandai Hanyu Changyong Ci Biao' 2008: §4.3).

Wie in allen anderen natürlichen Sprachen ist auch im Chinesisch eine Klassifizierung der Wörter nach Wortarten möglich. Der wesentliche Unterschied liegt darin, dass die Wortart nur nach Bedeutungsmerkmalen und der allgemeinen syntaktischen Funktion bestimmt werden kann (vgl. hierzu auch Kap. 4.2.1). Wie in anderen Sprachen gliedern sich die chinesischen Wörter traditionell und primär in Autosemantika (auch Bedeutungswort; chi.: 实词; eng.: content word) und Synsemantika (auch Funktionswort; 虚词; function word). Autosemantika beinhalten Nomen, Verben, Adjektive, Numerale, Zahleinheitswörter, Pronomen und Adverbien.[210] Synsemantika sind jene Wörter, die eigenständig keine lexikalische Bedeutung aufweisen und grammatische Funktionen übernehmen können. Die zu Synsemantika gehörenden Wortarten der chinesischen Sprache sind Präpositionen, Konjunktionen, Partikel, Onomatopoesien und Interjektionen (vgl. Zhao 1992: 16). Kap. 4.4.3 geht näher auf die chinesischen Wortarten ein. Da die chinesische Schrift zum morphologischen Schrifttypus zählt, wird an dieser Stelle die Wortbildung nach morphologischen Aspekten erforscht.

Wie oben erwähnt wurde, sind die chinesischen Wörter in simple sowie komplexe Wörter zu unterscheiden. Die simplen Wörter sind in den meisten Fällen identisch mit monosyllabischen Morphemen sowie Einzelzeichen. Die meisten als simple Wörter verwendbaren Zeichen können ebenso zur Bildung der Kompositionswörter verwendet werden. Deswegen muss im sprachlichen Kontext festgelegt werden, ob ein solches Zeichen ein eigenständiges Wort ist (vgl. Lan 2002: 85). Bspw. ist das Zeichen <人> in der Zeichenkollokation <一个人> (/yí gè rén/, *eine Person*; 一: *eins*; 个: ZEW; 人: *Person*) ein simples Wort, hingegen ist es ein Morphem in dem Kompositionswort <人民> (/rénmín/, *Volk*). Trotz der Minderheit der monosyllabischen Wörter (5,68% des gebräuchlichen Wortschatzes) werden sie relativ häufig benötigt. Nach der Wortfrequenz des CN-Corpus sind 45 der 50 häufigsten Wörtern monosyllabisch (vgl. CN-Corpus Online: ‚Corpus Word List'). Nach einer anderen Quelle treten die 6.285 diasyllabischen Wörter insgesamt 60-mal, die 2.400 monosyllabischen Wörter 350-mal auf (vgl. Xu YC 2008: 104).

[210] Die Klassifikation der Wortarten in Autosemantikum und Synsemantikum trifft bei verschiedenen Linguisten auf divergierende Meinungen. Adverbien sind bspw. eher eine Wortart zwischen Autosemantikum und Synsemantikum, die daher auch als Semisynsemantikum definiert wird (vgl. Lü 1979: 35).

Die simplen polysyllabischen Morpheme/Wörter können in native und Lehnwörter unterschieden werden. Die in den polysyllabischen Morphemen verwendeten Schriftzeichen sind normalerweise Nicht-Morphemzeichen oder verlieren ihren ursprünglichen Zeichensinninhalt innerhalb der Morphembedeutung. Die polysyllabischen nativen Morpheme/Wörter können näher in vier Arten kategorisiert werden: 1) Verdopplung desselben Zeichens, wie <猩猩> (/xīngxing/, *Gorilla*); 2) Zusammensetzung von zwei Zeichen mit demselben Initial, wie <仿佛> (/fǎngfú/, *es scheint wie…*); 3) Zusammensetzung von zwei Zeichen mit demselben Final, wie <糊涂> (/hútu/, *verwirrt*); 4) Zusammensetzung von zwei Zeichen mit keiner phonetischen Gemeinsamkeit, wie <马虎> (/mǎhū/, *unsorgfältig*; Zusammensetzung von <马> *Pferd* und <虎> *Tiger*) (vgl. Xu YC 2008: 106f, Zhao 1992: 12).[211]

Die simplen Lehnwörter sind meistens phonetische Übertragungen aus einer Fremdsprache. Bei dieser Art der Wortbildung werden Schriftzeichen wegen der ähnlichen Aussprache entlehnt, wohingegen ihr Zeichensinninhalt, ergo die Funktion als Morphem, verblasst, etwa <咖啡> (/kāfēi/, *Kaffee*) (vgl. Cai 2005: 70). Trotz des Verlustes der morphologischen Bedeutung bei der Wortbildung ist es immer noch wichtig, der Verwendungssituation angemessene Schriftzeichen unter den ähnlich ausgesprochenen Zeichen auszuwählen. *Kaffee* etwa zählt zum Inhaltsfeld ‚Getränke' und wird in zwei Schriftzeichen mit der Aussprache /kā/ und /fēi/ mit dem Radikal <口> (für auf *Mund* bezogene Begriffe) beschrieben. Ein weiteres typisches Beispiel ist der chinesische Name von ‚Coca Cola' <可口可乐> /kěkǒu kělè/. Obwohl es viele ähnlich ausgesprochene Schriftzeichen gibt, wurden diese vier Schriftzeichen wegen ihrem Sinninhalt ausgewählt. Zusammen bedeutet diese Kollokation *schmeckt gut und macht Freude*. Mit der ‚Sinisierung' hat sich die Bedeutung von ‚Coca Cola' weit über die phonetische Ähnlichkeit hinaus erweitert; da sie diese aber eben nicht verliert, ist <可口可乐> /kěkǒu kělè/ ein Paradebeispiel für eine kurze, bedeutungsvolle Werbung innerhalb des chinesischen Kulturkreises (vgl. Lu 2008: 129f). Solche Lehnwörter können ebenso als Morpheme zur Wortbildung benutzt werden, wie z.B. <咖啡厅> (/kāfēi tīng/, *Café*), das aus den Wortwurzeln für *Kaffee* und *Halle* zusammengesetzt ist. Da die phonetischen Lehnwörter nicht mit der Wortbildungsgrammatik analysiert werden können, bleiben sie bei der Wortbildungsforschung in diesem Kapitel sowie in Kap. 4.4.4 ausgespart.

[211] In <猩猩>, <糊涂> und vielen anderen diasyllabischen simplen Wörtern wird das letzte Zeichen häufig zum schwachen Ton umgewandelt.

Es gibt drei Bildungstypen der komplexen Wörter: Grundmorphemkombination-, affixoide Komposition und Morphemverdopplungskomposition. Anders formuliert hängt die Wortbedeutung im ersten Fall meistens mit zwei bis mehreren Grundmorphemen zusammen. Hingegen fungiert bei der affixoiden Wortbildung ein Morphem als Wortwurzel und ein Affix als Träger der zusätzlichen Bedeutung. Wort der Morphemverdopplung meint ein Wort, in dem dasselbe Morphem wiederholt wird. Alle Typen der Komposition können mit bestimmten grammatischen Regeln beschrieben werden und sind sowohl zur Wort- als auch Phrasenbildung einsetzbar (vgl. Zhao 1992: 12ff, Xu YC 2008: 109, Feng 2001c: 26).

Die Wörter der Grundmorphemkombination können in fünf Grund- und drei spezielle Formeln klassifiziert werden. Sie werden am Ende dieses Kapitels bei der Unterscheidung zwischen Wort und Phrase vorgestellt und analysiert. An dieser Stelle werden sie mithilfe von zwei Beispielwörtern erläutert. Das Wort <多少> (/duōshǎo/, *wie viel*) besteht aus den Zeichen für *viel* und *wenig*. Durch Zusammensetzung der Antonyme wird erst die Drittbedeutung *wie viel* erzeugt. Im zweiten Beispielwort <读书> /dúshū/ ist das erste Zeichen <读> (*lesen*) ein Verb und das zweite Zeichen <书> (*Buch*) das gelesene Objekt. Das entstehende Wort dieser Verb-Objekt-Kombination bedeutet grundsätzlich *Bücherlesen* sowie erweitert *lernen* und *Schulbildung* (vgl. Xinhua-Lexikon Online).

Das für die chinesische affixoide Wortbildung verwendete Affix ist stets ein Präfix oder Suffix. Präfixe sind bspw. <老> (/lǎo/, eigenständige Bedeutung: *alt*), <小> (/xiǎo/, *klein/jung*) und <第> (/dì/, *Ordnungsreihenfolge*).[212] Als Suffixe fungieren <子> /zi/, <儿> /er/, <者> /zhě/ und <化> /huà/.[213] Der Affixgebrauch wird in den Fußnoten mit einem Beispielwort erklärt. Um den Sinn der Kompositionswörter mit Affixen zu erklären, werden die Unterschiede zwischen <虎> /hǔ/ und <老虎> /lǎohǔ/, die beides Wörter für *Tiger* sind, analysiert. Obwohl das Zeichen <老> /lǎo/ mit dem Sinninhalt *alt* verbunden ist, investiert es bei der Wortbildung als Präfix keinen eigenen Sinninhalt. <老虎> steht für Tiger jeden Alters und jeder Größe. Im Gegenteil zu <老虎> kann das Einzelzeichen <虎> sowohl eigenständig ein

[212] Zur Präfix-Verwendung dieser drei Zeichen: <老> vor Familiennamen, Personalnomen, Verwandtschaftsbezeichnungen und einigen Tiere wie <老虎> (/lǎohǔ/, *Tiger*); <小> vor Familiennamen, Verwandtschaftsbezeichnungen und einigen Tieren wie <小猫> (/xiǎomāo/, *Katze*); <第> vor einer Nummer, um eine Ordnungszahl auszudrücken, wie <第一> (/dìyī/, *der/die/das erste*).

[213] Zur Suffix-Verwendung dieser vier Zeichen: <子> wird als Suffix (unbetont /zi/) hinter einem nominalen Zeichen verwendet, wie <桌子> (/zhuōzi/, *Tisch*); <儿> (Erhua-Yin-Träger) wird als Suffix hinter einem nominalen Zeichen (Diminutiv, Abstraktion oder semantische Umwandlung) oder verbalen/adjektivischen Zeichen verwendet, wie <花儿> (/huar/, *Blümchen*); <者> kann als Suffix mit Nomen, Verb, Adjektiv, Numeral oder Phrase kombiniert werden, um Personen, oder Gegenstände etc. zu bedeuten, wie <读者> (/dúzhě/, *Leser*); <化> wird als Suffix mit Nomen oder Adjektiv kombiniert, um die Umwandlung zu einer Eigenschaft oder einem Zustand auszudrücken, wie <绿化> (/lǜhuà/, [ein Gebiet] *grüner machen* [durch Anbau mehrer Bäume], 绿 *grün*).

Wort sein (hauptsächlich in Schriftsprache), als auch als Grundmorphem eines Kompositionsworts fungieren, wie <虎口> (/hǔkǒu/, *gefährliche Situation*; 口: *Mund*). Vor dem Hintergrund der Wortbedeutung ist der Sinninhalt des Einzelzeichens abstrakter, neben der Bedeutung *Tiger* trägt das Zeichen auch erweiterte Bedeutungen, etwa *wild, tapfer*. Im Gegenteil dazu hat das Wort <老虎> eine relativ stabile Bedeutung und entspricht dem deutschen Wort *Tiger* eins-zu-eins. Wegen der vielen Homophone im Chinesischen kann ein komplexes Wort beim Sprechen viel präziser wirken, als ein simples Wort (vgl. Cai 2005: 104, nach Xinhua-Lexikon Online). Anhand des Beispiels ist zu ersehen, dass eine wichtige Funktion von Affixen in der Wortbildung darin besteht, die Wortbedeutung präziser zu modifizieren.

Wie oben erwähnt wird auch die Verdopplungsform zur Wortbildung im Chinesischen bevorzugt und häufig gebraucht, wie <妈妈> (/māma/, *Mama*). Der wesentliche Unterschied zwischen der Morphemverdopplung im Wort und der Zeichenwiederholung im Morphem (wie 猩猩 /xīngxing/, *Gorilla*) ist, ob der Bestandteil im Einzelnen morphologische Funktionen hat (vgl. Lan 2002: 90f). Die Ziele der Verdopplungsstruktur sind, die Bedeutung der Bestandteile (Morpheme oder Wörter) zu betonen, die Semantik zu verstärken und die sprachlichen Signale klangvoller zu machen (vgl. Lu 2008: 184). Einige häufige Verdopplungsstrukturen können durch die Symbolisierung der Zeichen mit Buchstaben zusammengefasst werden:

- ‚AA', z.B. <妈妈> /māma/, *Mama*; Bedeutung und Wortart gleich wie A (meistens nominal, verbal, adjektivisch oder adverbial).
- ‚ABB', z.B. <喜洋洋> /xǐ yángyáng/, *sehr fröhlich*; Bedeutung: Addierung von A und B (喜: froh, 洋洋: reichlich wie B); Wortart gleich wie A (normalerweise verbal oder adjektivisch).
- ‚ABAB', z.B. <安排安排> /ānpái ānpái/, *etw. ordentlich planen*; Bedeutung und Wortart gleich wie AB <安排> (meistens diasyllabisches Verb).
- ‚AABB', z.B. <马马虎虎> /mǎmǎ hūhū/, *so lala*/ *unsorgfältig*; abgeleitet von AB (马虎 *unsorgfältig*); Bedeutung und Wortart wie AB (meistens diasyllabisches Adjektiv).
- ‚ABAC', z.B. <一心一意> /yìxīn yíyì/, *vom ganzen Herzen*; BC kann meistens ein Wort sein.
- ‚ABCC', z.B. <喜气洋洋> /xǐqìyángyáng/, *sehr fröhlich und glücklich*, (喜气: *glückliches Phänomen*; 洋洋 *reichlich* wie 洋); C ist normalerweise ein Verb oder Adjektiv. (vgl. ibid.: 184, Lü 1979: 37, Xiao 1988: 25ff).

Die Zusammenhänge zwischen einem komplexen Wort und seinen Bestandzeichen aus Sicht der Wortbildungsgrammatik sowie der Bedeutung sind entscheidend dafür, die unregistrierten Wörter automatisch zu erkennen und die Wortart zu bestimmen (vgl. Yao 1997: 181f). Die fünf Arten der Morphembeziehungen zur Wortbildung werden in dem folgenden Abschnitt zusammen mit der Phrasenbildungsgrammatik erläutert. Die Zusammenhänge der Bedeutung von bestehenden Morphemen/Zeichen und der gesamten Bedeutung des Wortes werden in Kap. 4.4.4 angerissen.

Ein Zwischenfazit könnte so lauten, dass ein chinesisches komplexes Wort eine festgelegte Zeichenkollokation und als kleinste Einheit Satzbildungsfunktion hat. Zwischen den Einheiten *Zeichen* und *Wort* sowie *Wort* und *Phrase* gibt es viele Grenzfälle. Die verwendeten grammatischen Regeln zur Phrasenbildung sind in vielen Fällen identisch mit den Regeln der Wortbildung von zwei oder mehreren Grundmorphemen.[214] Aus diesem Grund gibt es im Bereich der Linguistik zahlreiche Zeichenkollokationen, die umstritten als Wort *oder* Phrase definiert werden (vgl. Lu 2008: 151). Bspw. wird die Zeichenkollokation <吃饭> (/chī fàn/, *Essen haben*/*Mahlzeit*; 吃 *essen* und 饭 *gekochter Reis*/*Essen* [zur Mahlzeit]) von vielen Muttersprachlern als Wort angesehen. Nach grammatischen Analysen sollte es aber eher eine Phrase sein. Der Grund besteht darin, dass die beiden Zeichen manchmal getrennt gebraucht werden, z.B. <吃过饭> (*gegessen haben*; 过 ist ein Vergangenheitspartikel) und <吃早饭> (*Frühstück essen*; 早: *morgen*/*früh*; 早饭: *Frühstück*) (vgl. Lü 1979: 23).

Welche Zeichenkollokationen als Wörter definiert werden, steht im Bereich der Linguistik nicht fest. In Zeichenlexika werden die allgemein anerkannten und gebräuchlichen Wörter zu dem Eintrag ihres ersten Zeichens geordnet und erklärt. Da die grundlegende Einheit das Schriftzeichen ist, spielt die Unterscheidung zwischen Wörtern und Phrasen im Spracherwerb und -gebrauch keine entscheidende Rolle. Auf der Ebene der Textverarbeitung hingegen könnte diese linguistische Unklarheit große Schwierigkeiten verursachen, wie z.B. beim Bau der Wörterdatenbanken, der automatischen Wortsegmentation, Part-of-speech Tagging und die auf der Wortsegmentation basierten syntaktischen Forschungen und Sprachgenerierungen. Als Ziel der chinesischen Textverarbeitung muss ein national sowie international gültiger Standard zur Wortsegmentation verkündet werden (der Standard GB/T 13715-92). Die Beziehungen zwischen der segmentierten Einheit nach Standard, den Wörtern der Wörterdatenbank und den grammatisch gültigen Wörtern werden im Anschlusskapitel erläutert.

[214] Die fünf Grundkompositionsarten sind identisch mit denen der Phrasenbildung.

Die drei sprachlichen Ebenen *Morphem*, *Wort* und *Phrase* sind mit bestimmten Bedeutungen und Wortartattributen verbunden. Die Wort- sowie Phrasenbildung von Grundmorphemen sowie Wörtern ist von bestimmten grammatischen Regeln abhängig.[215] Wie in Kap. 4.2.1 (S. 210f) erwähnt, herrscht bei den Bildungsprinzipien der Wörter, Phrasen und Sätze hohe Übereinstimmung. An dieser Stelle werden die fünf Grundbildungsarten, die gemeinsam für Kompositionswort- und Phrasenbildungen gültig sind, vorgestellt (vgl. Zhao 1992: 13f, Lan 2002. 89f):

1) Koordinationsart (auch Lianhe-Art, 联合式)

Sie bezeichnet die Komposition von zwei oder mehreren Morphemen, Wörtern oder Phrasen mit gleicher, ähnlicher, verwandter oder entgegengesetzter Bedeutung von derselben Wortart. Beispielwörter sind das zuvor exemplifizierte <多少> und <父母> (/fùmǔ/, *Eltern*), in dem das erste Zeichen für *Vater* und das zweite für *Mutter* steht. Die Wortkollokation von derselben Bedeutung <爸爸妈妈> (/bàba māma/, Zusammensetzung von *Papa* und *Mama*) ist im Gegenteil eine Koordinationsphrase.

2) Subordinationsart (auch Pianzheng-Art, 偏正式)

In dieser Art wird der letzte Bestandteil (Morphem, Wort oder Phrase) von dem/den ersten Teil(-en) modifiziert oder begrenzt, wobei der letzte Bestandteil die zentrale Rolle bei der gesamten Wortbedeutung spielt (vgl. auch Xu YC 2008: 196). Diese Art kann in zwei Unterarten unterschieden werden: ‚Attribut-Kopf' (der Kopf ist meistens nominal) und ‚Adverbial-Kopf' (der Kopf ist meistens verbal oder adjektivisch). <汽车> (/qìchē/, *Auto*) ist bspw. ein Attribut-Kopf-Wort, in dem <汽> ursprünglich für *Dampf* steht und <车> der Oberbegriff für *Wagen* (Landverkehrsmittel) ist. <好看> (/hǎokàn/, *schön*) ist ein Adverbial-Kopf-Wort mit <好> (*gut*) und <看> (*sehen*), die zusammen die Bedeutung *angenehm zu sehen* das Wort *schön*, erzeugen. Die Wortkollokation <出租汽车> (/chūzū qìchē/, *Taxi*), in der <出租> (*vermieten*) das Wort <汽车> modifiziert, wird in den meisten Fällen als eine Attribut-Kopf-Phrase definiert. Die Zeichenfolge <很好看> (/hěn hǎokàn/, *sehr schön*) ist im Gegensatz eine Adverbial-Kopf-Phrase. Eine Subordinationseinheit kann mehrmals hierarchisch sein. Bspw. wird die Kollokation <中国人民生活> (/zhōngguó rénmín shēnghuó/, *das Leben des chinesischen Volks*) primär in <中国人民> (*das chinesische Volk*) und <生活> (*Leben*) zerlegt. Der erste Bestandteil <中国人民> /zhōngguó rénmín/ ist selbst eine Phrase in Attribut-Kopf-

[215] Exklusive der Schriftzeichen, die nicht allein benutzt werden können.

Subordination aus den Wörtern <中国> (*China*) und <人民> (*Volk*). Das Wort <中国> /zhōngguó/ ist ebenso ein komplexes Wort derselben Art aus <中> (*Mitte*) und <国> (*Land/Staat*) (siehe Abb. 4-25).

3) Prädikat-Komplement-Art (auch Shubu-Art, 述补式)

In dieser Art wird der erste Bestandteil (Morphem, Wort oder Phrase) von dem letzten Bestandteil ergänzend erklärt, wobei der erste Bestandteil einen Prädikatkopf (Verb oder Adjektiv) und der letzte syntaktisch ein Komplement darstellt. In dem Wort <发展> (/fāzhǎn/, *entwickeln*) z.B. kann das Zeichen <发> allein für die Bedeutung *verschicken* und *entwickeln* stehen. Das Zeichen für *entfalten* <展> spielt sodann eine ergänzende Rolle zur Wortbedeutung. In der Prädikat-Komplement-Phrase <发展壮大> (/fāzhǎn zhuàngdà/, *erheblich entwickeln*) ergänzt das Wort <壮大> (*stark und groß*) den Grad der Entwicklung näher.

4) Verb-Objekt-Art (auch Dongbin-Art, 动宾式)

In dieser Art ist der letzte Bestandteil (Morphem, Wort oder Phrase) das dependente Objekt des von dem ersten Bestandteil dargestellten Verbes. Das oben genannte Wort <读书> (/dúshū/, *Bücherlesen / lernen / Schulbildung*) ist Verb-Objekt-Wort. Im Gegenteil dazu ist <读小说> (/dú xiǎoshuō/, *Romane lesen*) eine Verb-Objekt-Phrase, in der *Roman* nicht fest mit dem Verb <读> verbunden ist und keine erweiterte Bedeutung hat im Gegensatz zu <读书>.

5) Subjekt-Prädikat-Art (auch Zhuwei-Art, 主谓式)

In dieser Art stellt der erste Bestandteil (Morphem, Wort oder Phrase) ein Subjekt (Nomen) und das letzte Bestandteil ein Prädikat (Verb oder Adjektiv) dar.[216] Beispielwort hierfür ist <你好> (/nǐhǎo/, *hallo*), das aus den Zeichen für *du* <你> und *gut* <好> zusammengesetzt ist. Wenn in dieser Kollokation <你> durch das Wort <你们> (/nǐmen/, *ihr*) ersetzt wird, nämlich <你们好> (/dàjiā hǎo/, *Hallo [an euch]*), wird eine Subjekt-Prädikat-Phrase erzeugt.

Zur Unterscheidung zwischen Wort und Phrase kann zuerst berücksichtigt werden, ob Morpheme oder Worte die Bestandteile sind und ob ein Sonderpartikel dazwischen hinzugefügt werden kann, ohne die Bedeutung zu wandeln (vgl. Feng 2001c: 27f). Bei den anderen zwei Wortbildungstypen – Affixoide und Verdopplungsstruktur – können vergleichbare Methoden zur Unterscheidung herangezogen werden. <小老虎> (/xiǎo lǎohǔ/, *kleiner Tiger*) ist

[216] Im Chinesischen kann das Adjektiv als Prädikat auftreten.

bspw. ein affixoides Kompositum mit dem Affix <小> (*klein*) und dem bedeutungstragenden Wort <老虎> (*Tiger*). Unter den sechs genannten Arten der Verdopplungsstruktur ist die AA-Form (Wiederholung desselben Morphems) für Wortbildung zuständig, während die ABAB-Form (Wiederholung desselben Wortes) typisch für Phrasenbildung ist (vgl. ibid.: 25f). Die anderen vier Strukturen (ABB, AAB, AABB, ABAC), die zu Ambiguitätsfällen zwischen Wort und Phrase gehören, werden nach dem national-standardisierten Segmentationsstandard als Wort behandelt (nach GB/T 13715-92).

Zur näheren Erklärung wird der Satz <他是德国人> (/tā shì déguó rén/; *Er ist Deutscher*) nachfolgend als Beispiel in Wörter segmentiert. Die beiden Zeichen <他> (/tā/, *er*) und <是> (/shì/, *sein*) können zweifelsohne als eigenständiges Wort erkannt werden. <德国人> (/déguó rén/, *Deutsche /-r /-n*) wird nach dem Standard der Wortsegmentation und des online abrufbaren CN-Corpus als Phrase behandelt (nach GB/T 13715-92: Kap. 5.1, CN-Corpus Online). Sie wird von den zwei Wörtern <德国> (*Deutschland*) und <人> (*Menschen*) mit Subordinationsart gebildet und nominal definiert. Wie eine Phrase vom Computer verarbeitet werden kann, wird in Kap. 4.4.5 dargestellt. Die syntaktische Formel dieses Satzes lautet deswegen: ‚Pronomen + Kopula + nominale Phrase'.

Die Informationsverarbeitung mit intelligenten Pinyin-Eingabesoftwares unterscheidet sich in zwei Perspektiven. Die unmittelbare Perspektive bezieht sich auf die Laut-Zeichen-Konversion, ergo wie eine Pinyin-Kette nach sprachlichen Informationen zu einem gültigen Satz verarbeitet werden kann (siehe Kap. 4.1.3). Die indirekte Perspektive stammt von der Begründung der Wissensdatenbank einer intelligenten Eingabesoftware (siehe Kap. 4.2.3). Um ausreichende sprachliche Informationen für die Wissensdatenbank zu beinhalten, müssen Korpora automatisch vom Computer verarbeitet werden. Die nachfolgend genannten Verarbeitungsteilbereiche Wortsegmentation, Wortart-Tagging, Erkennung der unregistrierten Wörter, Phrasenerkennung und Satzgeneration sind für beide Perspektiven entscheidend. Aus diesem Grund werden diese Teilbereiche im Einzelnen erläutert, wobei jeder Teilbereich zwei verschiedene Perspektiven betrifft, die jeweils von der Verarbeitung in Schriftzeichen sowie in Pinyin handeln (vgl. Kap. 4.4.2 bis 4.4.5 sowie 4.5.1).

4.4.2 Automatische Wortsegmentation

Die Wortsegmentation ist ein spezifischer Arbeitsprozess beim Lesen und der Textverarbeitung des Chinesischen sowie Japanischen, bei dem kein Spatium für die Grenze zwischen Wörtern eingesetzt wird. Im Leseprozess meint es die geistige Segmentation des Textes in einzelne Wörter. Hierfür ist ein gewisses Sprachverstehen erforderlich, um diese kognitive

Leistung zu erzielen. Im Normalfall aber ist mit Wortsegmentation die computergestützte Segmentation gemeint (vgl. Yao 1997: 174). In dem nationalen Standard der Wortsegmentation GB/T 13715-92 wird Wortsegmentation als „Arbeitsprozess der Zerlegung der chinesischen Texte in Wortsegmentationseinheiten [definiert], der auf den Gebrauch der Textverarbeitung abzielt und nach bestimmten Regeln abläuft“ (GB/T 13715-92: Kap. 3.5 [Übersetzung der Verfasserin]). Wie in Kap. 4.4.1 (S. 248) erwähnt, sind Wort und Phrase in vielen Fällen schwer zu unterscheiden. Aus diesem Grund wurden die Wortsegmentationseinheiten national standardisiert, um die chinesische Textverarbeitung zu unterstützen. In diesem Standard werden die sprachlichen Grundeinheiten aufgenommen, die über bestimmte syntaktische und/oder grammatische Funktionen verfügen. Zu den Wortsegmentationseinheiten gehören sowohl Wörter und als auch manche Phrasen, die nach den Regeln von GB/T 13715-92 gültig sind (vgl. ibid.: Kap. 3.4). Die Wortsegmentation hat bei den intelligenten Eingabemethoden zwei verschiedene Funktionen: die Segmentation von eingegebenen Pinyin-Ketten und von Zeichenketten aus Korpora, was zur Begründung der Wissensdatenbank der Eingabesoftware dient. Die Wortsegmentation von Texten (in Zeichen) stellt die Basis derselben in Pinyin dar (vgl. Zhang S et al. 1997: 40). Nachfolgend wird daher zunächst die Wortsegmentation auf Textebene erforscht. Die Schwierigkeiten der chinesischen Wortsegmentation können in drei Punkten zusammengefasst werden:

1. Kein offensichtliches Wortmerkmal.

Es gibt weder Leerstellen zwischen den Wörtern noch Flexionen. Ebenso kann die Länge eines Wortes von einem Zeichen bis zu über fünf Zeichen abweichend sein. Aus diesen Gründen ist die Tokenisierung, die für die in alphabetischen Schriftsystemen verfassten Texte verwendet wird, für die Verarbeitung der chinesischen Texte untauglich. Chinesisch gehört daher zu „nicht-segmentiertes Schriftsystemen“ (Hagenbruch 2010: 264).

2. Unklare Grenze zwischen Wort und Phrase.

Wie im letzten Kapitel vorgestellt wurde, sind sowohl Wort als auch Phrase mit vergleichbaren grammatischen Formen gebildete Kollokationen und die Grenze zwischen ihnen wird von verschiedenen Linguisten unterschiedlich definiert. Diese linguistische Unklarheit bringt Schwierigkeiten bei der chinesischen Textverarbeitung mit sich. Zur erleichterten Nutzbarkeit werden die Wörter in verschiedenen Bereichen zu verschiedenen Zwecken unterschiedlich definiert, vor allem bei dem Wortsegmentationsstandard, der Wörterdatenbank einer Eingabesoftware und dem Wortschatz in der Sprachdidaktik.

Wie oben erwähnt heißt der nationale Wortsegmentationsstandard GB/T 13715-92 und die Zeichenkollokationen nach diesem Standard sind Wortsegmentationseinheiten. Dazu gehören Wörter und manche Phrasen, die nach dem Wortsegmentationsstandard gültig sind. Dieser Standard bietet die Regeln an, wonach eine Zeichenkollokation als eine Segmentationseinheit definiert wird (vgl. Hou 1999: 99). Zum Entwurf des Segmentationsstandards stehen in ihrer Bedeutung vollständige, unzerlegbare sprachliche Einheiten im Fokus. So werden bspw. Vier-Zeichen-Idiome als segmentierte Einheiten definiert, obwohl die meisten davon nach den linguistischen Theorien Phrasen sind. Dem hingegen wird ein Personenname in Familien- und Vorname segmentiert (wie <王/铠> /wáng kǎi/), obwohl er nach dem grammatischen Aspekt eher als Wort definiert wird (vgl. GB/T 13715-92: Kap. 5.1.2).

Die Einträge einer Wörterdatenbank der Eingabesoftware sind immer von der Verwendungsfrequenz, bestimmter Menge und bestimmten Fachbereichen etc. eingeschränkt. Der aufgenommene Wortschatz sollte einerseits weder zu klein noch zu groß sein, denn ein zu großer Umfang kann theoretisch das Verarbeitungstempo verlangsamen und Arbeitsspeicher belasten. Die linguistische Definition für *Wort* ist deswegen in diesem Fall nicht der primäre Faktor für die Einträge, denn ob eine festgelegte Zeichenfolge als Wort oder Phrase definiert wird, ist bei Satzanalysen nicht entscheidend. Anderseits muss das Wörterbuch die Segmentation unterstützen, so dass die Anzahl der segmentierten Wörter möglichst minimal und das sprachliche Verstehen möglichst einfach für den Computer ist (vgl. Wang XL et al. 1993: 372). Die Häufigkeit und die Länge einer Zeichenfolge sind somit bedeutend dafür, ob sie als ein Wort eingetragen wird. Eine häufig verwendete Bigrammphrase wird daher häufig als Wort aufgenommen, wohingegen ein langes und selten gebrauchtes Wort im Normalfall als Wortfolge erkannt wird. Die einzutragenden Wörter sind von verschiedener Art. Über die allgemeinen Wortdefinitionen hinaus gehören zudem Morphemzeichen, Wortaffixe, Vier-Zeichen-Idiome, Sprichwörter, Abkürzungen usw. zu den Einträgen (vgl. Zhou Q/Duan 1999: 2). Wie in Kap. 4.2.3 (S. 217) vorgestellt wurde, dienen die statische und das dynamische Datenbanken kombiniert zur Wort- sowie Satzgeneration, so dass die Laut-Zeichen-Konversion möglichst effektiv abläuft und das Selbstlernmodul gezielt neue Wörter lernen kann.

Die nach den Wortbildungsgrammatiken bestätigten Zeichenkollokationen sind grammatische Wörter, deren Definition im letzten Kapitel (S. 242) angegeben wurde. Grammatische Wörter sind auch im Bereich der Linguistik anerkannte Wörter des modernen Chinesischen. Es gibt jedoch immer noch viele neu entstehende oder selten verwendete grammatische Wörter, die nicht von Wörterdatenbanken aufgenommen werden. Diese sind unregistrierte Wörter (vgl. Hou 1999: 93f). Zusammengefasst basieren sowohl der Wortsegmentationsstandard als

auch die Datenbankbegründung auf der grammatischen Vagheit zwischen Wort und Phrase, damit die automatische Textverarbeitung sowie die Eingabe ohne grammatische Unklarheit besser funktioniert.

3. Ambiguitäten bei Wortsegmentation im sprachlichen Kontext.

Doppeldeutigkeiten bei der Wortsegmentation kann passieren, wenn ein Schriftzeichen sowohl mit dem vorderen als auch mit dem hinteren Zeichen ein Wort bilden kann. Beispielsweise kann in der Zeichenkollokation <七十分> /qīshí fēn/ das Zeichen <十> (*zehn*) sowohl mit dem vorderen Zeichen das Wort <七十> (/qīshí/, *siebzig*) bilden, als auch mit dem hinteren Zeichen das Wort <十分> (/shífēn/, *sehr*). Ambiguitäten der Wortsegmentation vermögen auch bei solchen Zeichenkollokationen aufzutreten, die sowohl als ein Wort als auch als Zusammensetzung von zwei oder mehreren Wörtern verstanden werden können. <马上> /mǎ shàng/ bedeutet als ein Wort *bald*, kann aber auch in die zwei Wörter <马> *Pferd* und <上> *oben* segmentiert werden (vgl. Yao 1997: 177f). In ambigen Kontexten muss die korrekte Wortsegmentation mit syntaktischen sowie semantischen Erkenntnissen analysiert werden. In <七十分> bedeutet das Zeichen <分> *Note*, das mit dem vorderen Wort *siebzig* <七十> eine gültige Phrase bildet, die ihrerseits *Leistung von siebzig Punkten* ausdrückt. Hingegen ist die Verbindung des Wortes <七> *sieben* mit <十分> *sehr* ungültig. Ebenso ist <马上> in der Phrase <马上来> (/mǎshàng lái/, *bald kommen*) ein Wort, während es in der Phrase <在马上> (/zài mǎ shàng/, *auf dem Pferd*) zwei Wörtern entspricht.

Die automatische Wortsegmentation ist die Grundlage für satzstufige Eingabe und Akquisition der sprachlichen Regeln durch Korporaanalysen. Weiterhin ist sie Basis und Voraussetzung für verschiedene Aufgaben der chinesischen Textverarbeitung, wie die maschinelle Übersetzung, Information Retrieval und die Text-to-speech-Techniken etc. (vgl. ibid.: 174f). Die Methoden zur automatischen Wortsegmentation lassen sich in auf dem maschinellen Matching sowie Statistiken basierenden Verfahren (maschinelle Segmentationsmethoden genannt) und intelligenten Segmentationsmethoden unterscheiden (vgl. Liu Q/Jia 2006: 176f). Maschinelle Segmentationsmethode meint die bedingungslose Zerlegung in Wörter durch Matching mit Wörtern des maschinenlesbaren Wörterbuchs oder Wahrscheinlichkeitsanalysen des Zusammenauftretens benachbarter sowie zusammenhängender Wörter/Zeichen. Im Gegenteil dazu läuft die intelligente Segmentation unter sprachlichen Bedingungen ab, konkret mithilfe von lexikalischen, syntaktischen und semantischen Kenntnissen etc. Heutzutage werden fast allen Wortsegmentationssysteme von relativ hoher Präzision gemischt von maschinellen und

intelligenten Methoden verwendet (vgl. ibid., Hou 1999: 101,). Da die intelligenten Segmentationsmethoden immer zusammen mit POS-Tagging kombiniert durchgeführt werden, werden sie im nächsten Kapitel zusammen mit POS-Tagging vorgestellt. Zunächst werden jedoch diverse relativ häufig angewendete, auf maschinellem Matching und Statistiken basierende Methoden vorgestellt.

Die am frühesten erfundene maschinelle Segmentationsmethode heißt ‚maximale Vorwärts-Matchingmethode' (chi.: 正向最大匹配法; eng.: forward maximum matching method; Abk.: FMM). Die Hauptidee des maximalen Matchings ist das Matching maximal langer Zeichenketten mit den Wörtern des Wörterbuchs. Der Arbeitsprozess von FMM kann im Allgemeinen wie folgt beschrieben werden: Es wird angenommen, dass das längste Wort des zum Matching gebrauchten Wörterbuchs von I Schriftzeichen zusammengesetzt wird (im Normalfall entspricht I dabei sechs bis acht Zeichen). Die ersten I-Zeichen nach einem Interpunktionszeichen werden zuerst mit den Wörtern des Wörterbuchs abgeglichen. Gibt es das Wort, ist das Matching erfolgreich und nach dem I-Zeichen-Wort wird ein Worttrennungszeichen eingesetzt. Wenn nicht, wird das letzte Zeichen ausgeklammert und erneut gematcht. Die Reduzierung des letzten Zeichens geht immer weiter, bis die abgezogene Zeichenkette mit einem Wort aus dem Wörterbuch aligniert werden kann. Mit demselben Verfahren werden ebenso alle restlichen Zeichenketten eines Satzes vorwärts segmentiert. Nach offiziellen Statistiken liegt die Fehlerquote der Segmentierung mit FMM bei 1 aus 169 (vgl. Hou 1999: 101, Yao 1997: 176, Wang XL 2005: 35). Sie ist heute eine der am häufigsten angewendeten Methoden.

Komplett gegenteilig funktioniert die ‚maximale Rückwärts-Matchingmethode' (逆向最大匹配法; the opposite directional maximum matching method/ the backward maximal matching method; OMM oder BMM). OMM läuft nach demselben maximalen Prinzip wie FMM ab, jedoch rückwärts. Das Matching mit FMM folgt nach der logischen Ordnung des Textes. Dem hingegen werden mit OMM die letzten I-Zeichen eines Satzes berücksichtigt und das erste Zeichen wird weggelassen, wenn diese Kette mit keinem Wort aligniert. Im Vergleich zu FMM kann OMM mit 1 aus 245 eine noch niedrigere Fehlerquote erreichen, technisch ist sie dafür aber schwerer durchzuführen. Häufig wird sie zusammen mit FMM durchgeführt, um die Präzision zu erhöhen (vgl. ibid.).

Die ‚minimale Matchingmethode' (最小匹配法; the minimal matching method; MM) kehrt das Grundprinzip des maximalen Matchings um. Das erste Zeichen einer Zeichenkette wird zuerst ausgenommen und mit dem Wörterbuch gematcht. Ist es kein Wort, wird das nächste Zeichen hinzugefügt und weiter überprüft. Da die meisten chinesischen Schriftzeichen auch isoliert als Wörter auftreten können, aber die Einzel-Zeichen-Wörter die Minder-

heit des Gesamtwortschatzes darstellen, kann diese Methode unmöglich eine hohe Korrektheit erreichen. Aus diesem Grund wird sie kaum allein verwendet (vgl. Zhai 2012: 258).

Die ‚bidirektionale Matchingmethode' (双向扫描法; the bi-direction matching method; BM) ist für die Überprüfung und die Korrektur der Wortsegmentation zuständig. Die Ergebnisse der Wortsegmentation eines Textes, die getrennt mit FMM und OMM generiert wurden, lassen sich vergleichen. Die Gemeinsamkeiten werden vom System als korrekt erkannt und die Unterschiede müssen näher überprüft werden, und zwar durch manuelle Analyse, nach der Wortfrequenz oder mit intelligenten Wortsegmentationsmethoden. Diese Methode kann zwar präzise Ergebnisse erbringen, kostet aber auch mehr Zeit und fordert höhere technologische Grundlagen (vgl. Hou 1999: 104, Feng 2001a: 6).

Die ‚Methode der Begründung der Segmentationsmerkmale' (设立切分标志法; the method of segmentation marks researching; SMR) wird meist zur Vorbereitung der automatischen Wortsegmentation verwendet. In chinesischen Texten gibt es zwar kein formales Merkmal eines Wortes, aber manche Sonderschriftzeichen sowie Wörter können trotzdem die Rolle der Segmentationsmerkmale spielen. Solche Merkmale beinhalten arabische Zahlzeichen, nur als Präfix oder Suffix auftretende Schriftzeichen, normalerweise als eigenständiges Wort auftretenden Zeichen, polysyllabisch simple Wörter usw. Mit dieser Methode werden diese Segmentationsmerkmale im Korpus recherchiert, wodurch die Sätze in kleinere Teile zerlegt werden. Weitergehend werden diese Teilstücke mit FMM, OMM oder einer anderen Methode segmentiert (vgl. Hou 1999: 103, Yao 1997: 176).

Die ‚optimale Matchingmethode' (最佳匹配法; the optimum matching method; OM) ist eigentlich eine Vorbereitungsphase zur Wortsegmentation. Vor der Durchführung werden die Wörter aus dem elektronischen Wörterbuch nach ihrer allgemeinen Verwendungsfrequenz angeordnet, so dass das Matching bei den Wörtern mit höherer Verwendungsfrequenz anfängt und sich so beschleunigt (vgl. ibid., Feng 2001a: 6).

Die ‚Methode der Wort-für-Wort-Traversierung' (逐词遍历法; the word by word traversing method; WWT) basiert im Gegenteil zu den oben genannten maschinellen Methoden auf dem Matching der Wörter aus dem Wörterbuch mit den Zeichen im Korpus. Die Hauptidee dieses Verfahrens ist, alle Wörter des Wörterbuchs zuerst nach dem ersten Zeichen zu ordnen und anschließend nach ihrer Länge (von lang bis kurz) zu arrangieren. Bei einer zu segmentierenden Zeichenkette im Korpus wird dann Zeichen für Zeichen mit den Wörtern des Wörterbuchs nach der Rangliste gematcht, bis alle möglichen Wörter der Zeichenkette herausgefunden und alle möglichen Wortsegmentationsvariationen entdeckt wurden (vgl. Zhai 2012: 258, Hou 1999: 102).

Die ‚Methode der Priorisierung von hoher Frequenz' (高频优先法; the priorty of high frequency method; PHF) ist ein auf Statistiken basierendes Verfahren zur Wortsegmentation und vor allem für Ambiguitätsfälle derselben zuständig. Angenommen eine Zeichenkette ABC existierte, wobei A, B und C jeweils für ein einzelnes Schriftzeichen stehen, in der sowohl AB als auch BC Wörter sein können, so entschied die analysierte Höhe der Frequenz, ob die Zeichenkette als A/BC oder AB/C zerlegt würde (vgl. Zhai 2012: 258, Hou 1999: 103f). Ohne sprachliche Analysen könnte diese Methode in vielen Fällen Fehlergebnisse anbieten. So ist bspw. bei der Zeichenkollokation <七十分> (siehe S. 253 des Kapitels) die Frequenz von <十分> (*sehr*) viel höher als <七十> (*siebzig*) (3199 vs. 412 Auftritte; nach CN-Corpus Online). ‚七/十分' ist aber sowohl grammatisch als auch semantisch ein falscher Ausdruck.

Die ‚Methode der Wortfrequenzauswahl von minimaler Segmentation' (最少分词词频选择法; the fewest segmentation based method with selecting of word frequency; FWF) ist eine Mischtechnik aus der minimalen Segmentation und den statistischen Analysen. Minimale Segmentation bedeutet die Zerlegung einer Zeichenkette in die wenigsten Wörter. Die Funktionsweise verläuft in zwei Hauptschritten. Zuerst werden alle möglichen Wörter mithilfe des Wörterbuchs herausgefunden. Zwischen den möglichen Wörtern werden Pfade gebildet und die Variante von den am wenigsten segmentierten Wörtern als Ergebnis der Wortsegmentation anerkannt. Die FWF-Methode basiert zuerst auf den Ergebnissen der minimalen Segmentation. Gibt es hier jedoch mehrere Möglichkeiten, wird die optimalste Variante anhand von Wahrscheinlichkeitsanalysen näher bestimmt, konkret die Variante mit der höchsten Frequenz der Wörterkollokationen (vgl. Zhai 2012: 258f, Zhang S et al. 1997: 40).

Die auf ‚Statistiken basierende Wortvernetzungsmethode' (基于统计的词网格分词法; the statistic based word gridding method; SWG) ist später als die oben genannten Methoden entstanden, von relativ hoher Präzision und mit erweiterten Funktionen ausgestattet. Das Verfahren hat zwei Hauptarbeitsphasen. Zuerst werden alle möglichen segmentierten Wörter mithilfe des Wörterbuchs herausgefunden und in Wortnetzform gespeichert. Danach wird die wahrscheinlichste Segmentationsvariante mit einstelligem (die Frequenz des einzelnen Wortes) sowie zweistelligem Algorithmus (die Frequenz von bisyllabischen Wörtern) ausgewählt (vgl. Wang XL 2005: 36). Die Wortsegmentationsvernetzung lässt sich anhand der Phrase <中国人民生活> (/zhōngguó rénmín shēnghuó/, *das Leben des chinesischen Volkes*) erklären, in welcher es ohne syntaktische Analysen 16 Möglichkeiten zur Wortsegmentation gäbe.[217] Nach

[217] Bei dem Knoten nach a, b, c und d gibt es jeweils zwei Pfade (siehe Abb. 4-25), weshalb die gesamten Segmentationsmöglichkeiten 16 (2*2*2*2) betragen.

sprachlichen Kenntnissen jedoch ist alleine die Segmentation ‚中国/人民/生活' [*China* / *Volk* / *Leben*] korrekt, die mithilfe von der algorithmischen Rechnung der Wortvernetzung automatisch bestimmt werden kann.

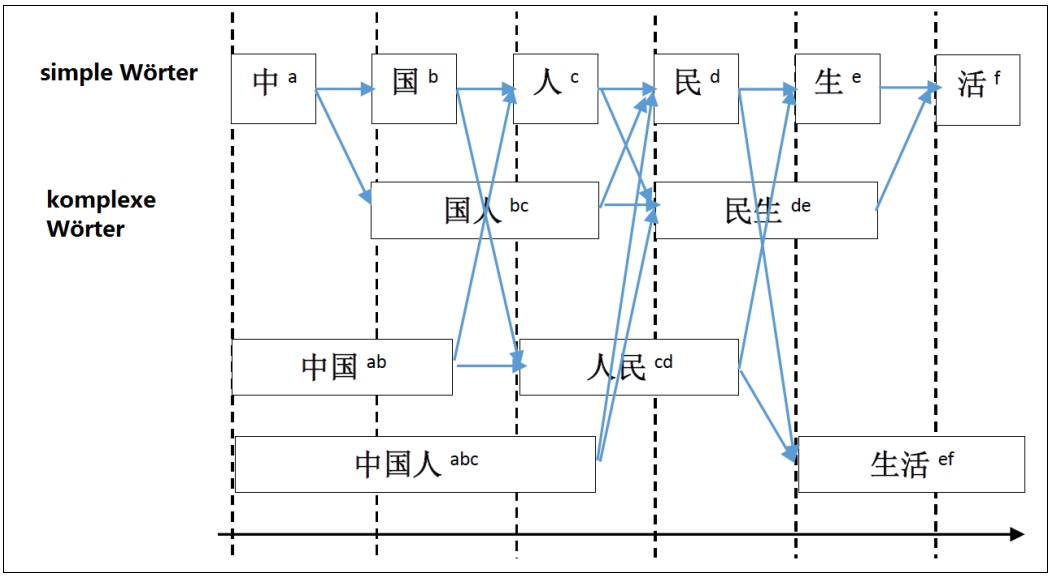


Abb. 4-25: Die Vernetzung der Wortsegmentation von der Zeichenkette ‚中国人民生活' (vgl. Wang XL 2005: 36)[218]

Die auf maschinellem Matching und Statistiken basierenden Methoden können mit weniger Komplexität einprogrammiert und durchgeführt werden. Allerdings basieren die intelligenten Segmentationsmethoden auf elektronischem Wörterbuch und linguistischen Wissensdatenbanken, um den menschlichen Gedankengang beim Lesen zu simulieren. Programmierung und Durchführung der intelligenten Segmentation bringen deswegen große technische Schwierigkeiten mit sich. Heutzutage funktioniert die Wortsegmentation auf der Grundlage der maschinellen Segmentation und die davon erhaltenen Ergebnisse werden weiter mit intelligenten Segmentationsmethoden verarbeitet, um Fehler zu lösen (vgl. Yao 1997: 177).

Bei der Anwendung in intelligenten Pinyin-Eingabesoftwares muss die automatische Wortsegmentation bei der Pinyin-Kette statt im Text durchgeführt werden. Theoretisch werden für die Pinyin-Wortsegmentation dieselben Funktionsprinzipien und -weisen benötigt, wegen homographischer Wörter aber treten Ambiguitätsfälle um ein vielfaches häufiger auf als bei Zeichen (vgl. Chen YF/Zhu 2002b: 13). Wie in Kap. 4.3.2 erwähnt wurde, können die Silben- und Wortsegmentationen kombiniert durchgeführt werden. Diese Verbindung erhöht

[218] Aussprache und Bedeutung der einzelnen Zeichen/Wörter: a - 中 /zhōng/, *mitte*; b - 国 /guó/, *Staat*; c- 人 /rén/, *Mensch*; d - 民 /mín/, *Volk*; e - 生 /shēng/, *Geburt, wachsen, leben*; f - 活 /huó/, *leben*; Aussprache/Bedeutung der Kompositionswörter: ab - 中国 /zhōngguó/, *China*; abc - 中国人 /zhōngguórén/, *Chinese -r/-n*; bc - 国人 /guórén/, *Landesvolk*; cd - 人民 /rénmín/, *Volk*; de - 民生 /mínshēng/, *Volksleben*; ef - 生活 /shēnghuó/, *Leben*.

einerseits die Effizienz und kann andererseits viele Ambiguitätsfälle ausfiltern. Die häufig gebrauchten Methoden sind die maximale Matchingmethode (inkl. FMM und OMM), die minimale Wortfrequenzauswahlmethode (FWF), die Wort-für-Wort-Traversierungsmethode (WWT) und die Wortvernetzungsmethode (SWG) (vgl. Zhang S et al. 1997: 40).

Um die Wortsegmentation bei der intelligenten Pinyin-Eingabemethode zu forschen, wird nachfolgend die Wortsegmentation der Pinyin-Kette ‚TASHIDEGUOREN' (für den Text <他是德国人> /tā shì déguó rén/, *Er ist Deutscher*) mit den vier genannten Methoden analysiert. Nachdem diese Pinyin-Kette in einzelne Silbe (‚TA'SHI'DE'GUO'REN') segmentiert wurde, beginnt die Wortsegmentation. Angenommen wird dabei, dass die Zeichenkollokation <德国人> (/déguó rén/, *Deutsche -r/-n*), die sowohl als Wort als auch als Phrase definiert werden könnte, als zwei Wörter erkannt wird. Die Verarbeitungsprozesse mit jeweils FMM und OMM werden in der folgenden Tabelle angegeben.

	Verarbeitung mit FMM		**Verarbeitung mit OMM**	
Prozess	gematchte Pinyin-Kette	Status	gematchte Pinyin-Kette	Status
1	/ TASHIDEGUOREN/	×	/ TASHIDEGUOREN/	×
2	/ TASHIDEGUO/	×	/ SHIDEGUOREN/	×
3	/ TASHIDE/	×	/ DEGUOREN/	×
4	/ TASHI/	√	/ GUOREN/	√
5	/ DEGUOREN/	×	/ TASHIDE/	×
6	/ DEGUO/	√	/ SHIDE/	√
7	/ REN/	√	/ TA/	√
Ergebnis	/ TASHI/ DEGUO/ REN/		/ TA/ SHIDE/ GUOREN	
Möglicher Text	踏实/德国/人 (/tāshí déguó rén/, die zuverlässigen und beständigen Deutschen)		他/师德/过人 (/tā shīdé guòrén/, *seine Lehrmoral ist sehr gut*)	
Ziel geeignet?	nein		nein	

Tab. 4-7: Die Wortsegmentation mit der maximalen Matching-Methode (FMM und OMM) bei Pinyin-Ketten ‚TA'SHI'DE'GUO'REN'[219]

Von den Ergebnissen mit FMM und OMM ist festzustellen, dass es mit der maximalen Matchingsmethode häufig Fehler bei der Wortsegmentation der Pinyin-Kette geben kann: In Pinyin-Codes gibt es viel mehr Ambiguitätsmöglichkeiten als in Zeichen. Mit der FMM- sowie OMM-Methode kann nur eine Variante für die Wortsegmentation als Ergebnis ausgegeben werden. Trotz vieler möglicher Fehler ist das maximale Matching das einfachste Verfahren in Programmierung und Anwendung, insbesondere FMM. Aus diesem Grund werden FMM und OMM kombiniert angewendet. Wenn die Ergebnisse der beiden Methoden nicht übereinstimmen, läuft die Prüfung weiter mit statistischen Analysen, um das Ergebnis mit höherer Wahrscheinlichkeit zu bestimmen (vgl. Zhang S et al. 1997: 40). Bei diesem Beispiel kann weder

[219] Status ×oder √ steht dafür, ob die Wortsegmentation nach Wörterbuch gültig ist.. Das Zeichen ‚/' in einer zu segmentierenden Pinyin- sowie Zeichenkette symbolisiert Worttrennung.

mit FMM noch OMM die korrekte Segmentation herausgearbeitet werden. Zur besseren Effizienz der intelligenten Eingabe ist es in manchen Fällen notwendig, alle möglichen Segmentationsvarianten zu behalten und darauf basierend die optimalen Zeichenketten nach sprachlichen sowie statistischen Analysen im Kontext zu erzeugen. Dazu ist die Methode WWT oder SWG notwendig (vgl. ibid.: 41, Wang XL 2005: 36).

Bei den Segmentationsmethoden FWF, WWT und SWG werden zuerst alle möglichen Wörter einer Pinyin-Kette recherchiert und der Pfad zwischen benachbarten Wörtern gebaut. Mit derselben Pinyin-Kette wie in Tab. 4-7 als Beispiel können die Ergebnisse in folgender Abbildung angegeben werden.

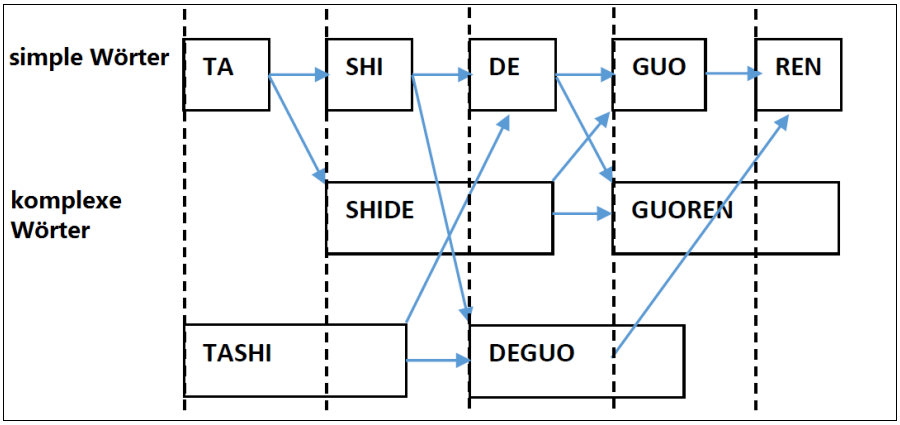


Abb. 4-26: Die möglichen Wörter von der Pinyin-Kette ‚TA'SHI'DE'GUO'REN' und die Pfade der möglichen Segmentation

Anhand der Abbildung kann festgestellt werden, dass es insgesamt acht Möglichkeiten der Segmentation gibt, davon fünf mit ‚TA' und drei mit ‚TASHI' als erstes Wort (siehe Tab. 4-8). Mit FWF werden zuerst nur die Varianten mit der minimalen Segmentation behalten und weiter zur Konversion in Zeichen verarbeitet. In diesem Fall sind drei Varianten von drei Wörtern gültig: TASHI/DE/GUOREN, TASHI/DEGUO/REN und TA/SHIDE/GUOREN. Keine Variante davon ist für das Eingabeziel <他/是/德国/人> geeignet. Somit ist es wiederum problematisch, FWF allein für die Wortsegmentation der Pinyin-Kette zu verwenden. Meistens wird es zusammen mit WWT benutzt (vgl. Zhai 2012: 258f, Zhang S et al. 1997: 40f).

Der größte Vorzug von WWT und SWG im Vergleich zu anderen Segmentationsmethoden ist, alle möglichen Pinyin-Wörter zur weiteren Konversion im Satz einsetzbar zu machen. Die möglichen Pinyin-Wörter können dann in Zeichen umgewandelt werden und die möglichen Zeichenketten lassen sich weiter auf der syntaktischen Ebene nach sprachlichen Regeln sowie Wahrscheinlichkeitserrechnungen überprüfen.

Variante	Wortsegmentation in Pinyin	Mögliche Zeichenkette
1	TASHI/ DEGUO/ REN (AB/ CD/ E)	踏实/ 德国/ 人
2	TASHI/ DE/ GUOREN (AB/ C/ DE)	踏实/ 的…/ 国人…

Variante	Wortsegmentation in Pinyin	Mögliche Zeichenkette
3	TASHI/ DE/ GUO/ REN (AB/ C/ D/ E)	踏实/ 的…/ 国…/ 人…
4	TA/ SHIDE/ GUOREN (A/ BC/ DE)	他…/ 是的…/ 国人…
5	TA/ SHIDE/ GUO/ REN (A/ BC/ D/ E)	他…/ 是的…/ 郭…/ 人…
6	TA/ SHI/ DEGUO/ REN (A/ B/ CD/ E)	他…/是…/德国/人…
7	TA/ SHI/ DE/ GUOREN (A/ B/ C/ DE)	他…/是…/得…/国人…
8	TA/ SHI/ DE/ GUO/ REN (A/ B/ C/ D/ E)	他…/是…/得…/郭…/人…

Tab. 4-8: Alle möglichen Segmentationsvarianten von ‚TA'SHI'DE'GUO'REN'[220]

Wie die in Wörter segmentierten Pinyin-Ketten weiter mit grammatischen sowie semantischen Erkenntnissen verarbeitet werden, um optimale Zeichenketten ausgeben zu können, wird in den nächsten Kapiteln aufgezeigt.

4.4.3 Part-of-Speech Tagging

Das Part-of-Speech Tagging (chi.: 自动标注, eng.: part-of-speech tagging, Abk.: POS-Tagging) wird auch Wortart-Tagging genannt und seine Aufgabe ist, Wörter „gemäß ihrer Wortart zu klassifizieren" (Hagenbruch 2010: 271). Wegen der Besonderheiten des Chinesischen muss das POS-Tagging von der automatischen Wortsegmentation vorausgesetzt sein, sowohl bei der Verarbeitung des Textkorpus als auch der eingegebenen Pinyin-Kette.

In Kap. 4.2.3 wurde bereits eingeführt, dass die Wortart eines Wortes/Zeichens der Mikrostruktur eines Wörterbucheintrags eingeschrieben wird, anhand derer das POS-Tagging im Normalfall ablaufen kann. Die Herausforderungen des chinesischen POS-Taggings sind einerseits von den Schwierigkeiten der Wortsegmentation bedingt, darunter vor allem die Faktoren 1) kein ersichtliches Wortmerkmal, 2) unklare Begrenzung zu Phrasen und 3) Ambiguitätsfälle (vgl. Kap. 4.4.2, S. 252ff). Andererseits führen die linguistischen Wortartprobleme zu weiteren zwei Schwierigkeiten.

Die erste Schwierigkeit liegt an der Klassifikation und Definition der chinesischen Wortarten, die bis heute umstritten sind (vgl. Bai 1992: 41). In Kap. 4.2.1 (S. 209f) wurde skizziert, dass es bei keiner Wortart im Chinesischen flektierende Merkmale gibt und zwischen Wortart und syntaktischer Funktion auch keine Eins-zu-eins-Entsprechung existiert. Nur nach der Semantik und der allgemeinen syntaktischen Rolle kann die Wortart klassifiziert werden. In Kap. 4.4.1 (S. 244) wurde des Weiteren referiert, dass die chinesischen Wörter den zwei Hauptkategorien Autosemantika und Synsemantika zugeordnet werden können. Wegen der linguistischen Schwierigkeiten bei der Wortartdefinition gibt es unterschiedliche Argumenta-

[220] Für eine eindeutige Angabe wird in der zweiten Spalte jede Stelle des Zeichens mit einem lateinischen Buchstaben symbolisiert; in der letzten Spalte werden Wörter/Zeichen, die mit benachbarten Einheiten kein Wort bilden können. Grammatische sowie semantische Erkenntnisse zu den möglichen Zeichenketten werden nicht berücksichtigt.

tionen verschiedener Linguisten. Im selben Kapitel wurden jene Wortarten, die nach der heute anerkannten linguistischen Auffassung jeweils zu Auto- und Synsemantika zählen, entsprechend angegeben (vgl. Zhao 1992: 16). Die Darlegungen orientieren sich dabei an chinesischen Erst- und Fremdsprache-Lehrbüchern.

Das Zahleinheitswort (kurz: ZEW) ist unter den chinesischen Wortarten eine spezifische Wortart, die in den indogermanischen Sprachen so nicht definierbar ist. Auch das chinesische Partikel wird wegen seiner umfassenderen grammatischen Funktionen anders definiert und eingesetzt. Wie vorgestellt markieren im Chinesischen hauptsächlich die Funktionswörter jene grammatischen Phänomene, die in indogermanischen Sprachen mit Wortflexion geäußert werden. Bspw. werden <过> /guò/, <了> /le/ oder <着> /zhe/ nach dem Verb gebraucht, um vergangene Aktionen (wie <吃过> /chī guò/ oder <吃了> /chī le/, *gegessen haben/aß*) oder die gegenwärtige Fortsetzung einer Aktion darzustellen (wie <吃着> /chī zhe/, ähnlich dem Gerundium im Englischen: *be eating*). Das Wort <的> /de/ kann nach einem Personalnomen folgen, um das Besitzverhältnis zu markieren, wie <我的> (/wǒde/, *mein -e/-er/-es/*). Solche Funktionswörter werden bei Wörtern, Phrasen oder Sätzen angehängt, um eine ergänzende Bedeutung zu äußern. Sie werden als chinesische Partikel definiert (vgl. GB/T 20532-2006: 3). Neben ZEW und Partikeln werden auch andere Wortarten – besonders Adjektiv, Adverb, Nomen, Verb usw. – unterschiedlich als in den indogermanischen Sprachen verwendet. In Tab. 4-9 wird die grammatische Rolle jeder Wortart verkürzt angegeben. Ausführlichere Informationen über die Beziehungen zwischen Wortart und syntaktischer Funktion werden in Tab. 4-13 und 4-14 (vgl. Kap. 4.5.2) vorgestellt.

Eine weitere Schwierigkeit des chinesischen POS-Taggings liegt an homographischen sowie multi-kategorialen Wörtern (siehe Kap. 4.2.3, S. 219f). Ein homographisches Wortpaar besteht aus zwei verschiedenen Wörtern, die unterschiedlich oder identisch ausgesprochen werden und sich semantisch in der Regel unterscheiden. Im Gegensatz dazu ist ein multi-kategoriales Wort (chi.: 兼类词, eng.: multi-category words) ein Wort, das mit zwei oder mehreren verschiedenen Wortarten verbunden ist (vgl. Lan 2002: 133). Im Gesamtwortschatz machen solche Wörter ca. 5 bis 11% aus, im alltagssprachlichen Gebrauch aber ist ihre Frequenz mit 40 bis 45% weit höher. Je häufiger ein Wort daher gebraucht wird, desto wahrscheinlicher ist es multi-kategorial. Die Multi-Kategorisierung von Nomen und Verb, Adjektiv und Adverb sowie Verb und Adjektiv sind dabei die dominierenden Fälle (vgl. Wang XL 2005: 87). Betrachtet man homographische sowie multi-kategoriale Wörter zusammen, kann ein Zeichen oder eine Zeichenfolge viele verschiedene Anwendungsfälle haben. Ein Beispiel

dafür ist das Zeichen <是> /shì/ im modernen Chinesischen, das kontextunabhängig mit sieben verschiedenen Wortarten definiert werden kann (nach Xinhua-Lexikon Online). Ohne formale Merkmale kann die Wortart solcher Wörter nur nach semantischen und syntaktischen Gesichtspunkten bestimmt werden. Nach dem POS-Tag-Standard werden die Möglichkeiten der Wortartmarkierung und ihr Taggingcode in der folgenden Tabelle dargestellt.

Nr.	**Tagging-Code**		**POS in DE & ZH**	**Erklärung**[221]
	1. Klasse	2. Klasse		
1	a		Adjektiv 形容词	Als Prädikat, Attribut, Adverbial, Komplement, Subjekt oder Objekt in der Syntax; ohne Unterscheidung bei Komparation.
2		aq	Adjektiv-Eigenschaft 性质形容词	Vom Englischen *adjective-quality*
3		as	Adjektiv-Status 状态形容词	Vom Englischen *adjective-state*
4	c		Konjunktion 连词	Vom Englischen *conjunction*; Kombination von Wörtern, Phrasen oder Sätzen.
5	d		Adverb 副词	Vom Englischen *adverb*; Bestimmung oder Einschränkung von Verb und Adjektiv; als Adverbial in der Syntax.
6	e		Interjektion 叹词	Vom Englischen *exclamation*; eigenständig als ein Satz oder Satzglied darstellbar.
7	f		Unterscheidungswort 区别词	Vom Englischen *difference*; als Attribut oder Komplement in der Syntax; Markierung der unterscheidenden Eigenschaft eines Gegenstandes; Unterart der Adjektive (umstritten), aber nicht eigenständig als Prädikat darstellbar: muss als Attribut zu Nomen und Verb oder mit <的> kombiniert benutzt werden.
8	g		Morphemzeichen 语素词	Vom Pinyin von <根> /gēn/, *Wurzel*; mit g wird kein Wort, sondern ein Morphem in unbekannten Wörtern markiert.
9		ga	adjektivisches Morphem-Zeichen 形容词性语素字	Von ‚gēn' und *adjective*
10		gn	nominales Morphem-Zeichen 名词性语素字	Von ‚gēn' und *noun*
11		gv	verbales Morphem-Zeichen 动词性语素字	Von ‚gēn' und *verb*
12	h		Anfangsteil 前接成分	Vom Englischen *head*; Wortpräfix oder Schriftzeichen, das meistens als das erste Zeichen eines Wortes auftritt.
13	i		Idiom 习用语	Chinesische Idiome, die festgelegte Phrasen, aber eigenständige Segmentationseinheiten sind; zumeist Vier-Zeichen-Idiome.
14		ia	adjektivisches Idiom 形容词性习用语	Von *idiom-adjective*

[221] Wie in Kap. 4.2.1 (S. 210) vorgestellt wurde, kann eine Wortart hochvariabel verschiedene Satzglieder darstellen, besonders bei Nomen, Verben und Adjektiven; wegen der Platzeinschränkung können nicht alle möglichen Fälle genannt werden.

Nr.	Tagging-Code		POS in DE & ZH	Erklärung[221]
	1. Klasse	2. Klasse		
15		ic	konjunktives Idiom 连词性习用语	Von *idiom-conjunction*
16		in	nominales Idiom 名词性习用语	Von *idiom-noun*
17		iv	verbales Idiom 动词性习用语	Von *idiom-verb*
18	j		Abkürzung 缩略语	Von Pinyin von ‚jiǎn' (<简>, *Kürzung*); abkürzende Wörter für Eigennamen sowie häufige Ausdrücke.
19		ja	adjektivische Abkürzung 形容词性缩略语	Von jiǎn und *adjective*
20		jn	nominale Abkürzung 名词性缩略语	Von jiǎn und *noun*
21		jv	verbale Abkürzung 动词性缩略语	Von jiǎn und *verb*
22	k		Endungsteil 后接成分	Durch Ausschlussprinzip; Wortsuffix oder ein Schriftzeichen, das meistens als das letzte Zeichen eines Wortes auftritt.
23	m		Numerale 数词	Von *numeral*; für Nummern und Ordnungszahlen.
24	n		Nomen 名词	Als Subjekt, Objekt oder Attribut in der Syntax; ohne Unterscheidung bei Kasus, Genus und Numerus.
25		nd	direktionales Nomen 方位名词	Von *noun-direction*; Nomen für Richtungen; kann hinter einem Nomen oder Verb eine Direktionalphrase bilden (DP); DP als Adverbial; nd in der Syntax eigenständig verwendbar als Attribut, Adverbial, Subjekt und Objekt.
26		ng	generelles Nomen 普通名词	Von *noun-general*; generelle Nomen.
27		nh	Personalname 人名	Von *noun-human*; Eigennamen der Menschen.
28		ni	Institutsname 机构名	Von *noun-institution*; Eigennamen für Teams, Konsortien und Institutionen.
29		nl	lokales Nomen 处所名词	Von *noun-location*; Nomen für Standort.
30		nn	nationales Nomen 族名	Von *noun* und *nation*; Eigennamen für Nationen oder Volksstämme.
31		ns	Ortsname 地名	Von *noun-space*; Ortseigennamen.
32		nt	zeitliches Nomen 时间名词	Von *noun-time*; als Subjekt, Objekt, Adverbial und Attribut in der Syntax verwendbar; die Zahleinheitswörter für die Zeit sind auch in der Kategorie eingeschlossen.
33		nz	sonstige Eigennamen 其他专有名词	Von *noun* und dem Pinyin von <专> (/zhuān/, *speziell*).
34	o		Onomatopoesie 拟声词	Nachahmung von natürlichen Lauten; nicht eigenständig als Satz verwendbar.
35	p		Präposition 介词	Bildet mit dem nachfolgenden Nomen eine Präpositionalphrase (PP), die als Adverbial, Komplement oder Attribut in der Syntax auftritt.

Nr.	Tagging-Code		POS in DE & ZH	Erklärung[221]
	1. Klasse	2. Klasse		
36	q		Zahleinheitswort 量词	Von *quantity*; dient, um die Zahleinheit von Menschen, Gegenständen oder Akten zu äußern.
37	r		Pronomen 代词	Von *pronoun*; Ersatz für oder Verweis auf ein Nomen.
38	u		Partikel 助词	Von *auxilary*; bei Wort, Phrase oder Satz anhängend, um eine ergänzende Bedeutung zu äußern.
39	v		Verb 动词	Als Prädikat, Attribut, Adverbial, Komplement, Subjekt oder Objekt in der Syntax; ohne Unterscheidung bei Person, Modus, Tempus, Aspekt und Diathese.
40		vd	direktionales Verb 趋向动词	Vom Englischen *verb-direction*; dient, um die Richtung einer Aktion zu äußern; als Prädikat oder Adverbial hinter einem anderen Verb in der Syntax.
41		vi	intransitives Verb 不及物动词	Verb ohne Objekt
42		vl	bindendes Verb 联系动词	Vom Englischen *verb-linking* (Kopula).
43		vt	transitives Verb 及物动词	Verb mit Objekt
44		vu	Modalverb 能愿动词	Von *verb-auxilary*; unklare Grenze zu Adverb; drückt eine Modalität aus.
45	w		sonstige Wörter 其他	Nicht-Chinesische-Zeichen
46		wp	Interpunktionszeichen 标点符号	„w“ und *punctuation*
47		ws	Nicht-Chinesische-Zeichenkette 非汉字字符串	„w“ und *string*
48		wu	unbekannte Zeichen 其他未知符号	„w“ und *unknown*
49	x		Nicht-Morphem-Zeichen 非语素字	Schriftzeichen, die eigenständig keine Bedeutung übertragen.

Tab. 4-9: POS-Tag-Standard der Wortart und deren Erklärung im Sprachgebrauch[222]

Das POS-Tagging muss wörterbuchunterstützt ablaufen und in kontextabhängigen Situationen auch anhand syntaktischer Erkenntnisse durchgeführt werden (vgl. Hou 1999: 160). Bei der Verarbeitung des chinesischen Textkorpus sind Wortsegmentation und POS-Tagging miteinander fest verknüpft. Die kombinierten Arbeitsphasen von den beiden Anwendungen mithilfe von der auf dem sprachlichen Verstehen basierenden Funktion werden hauptsächlich anhand der folgenden Hauptschritte durchgeführt (vgl. Zhou Q/Yu 1993: 2, Yao 1997: 184). Der Satz <工人们忙忙碌碌地生产触屏> (/gōngrén mén mángmánglùlù de shēngchǎn chùpíng/, *Die*

[222] Vgl. GB/T 20532-2006: Kap. 5, Lan 2002: 104-133.

Arbeiter sind sehr beschäftigt mit der Produktion der Berührungsbildschirme) soll als Beispiel für die Erklärung der Arbeitsphasen genommen.

1. Der Einsatz eines maschinenlesbaren Wörterbuchs mit Wortartmarkierung. Nach der Wortsegmentation wird jede Wortsegmentationseinheit mit allen möglichen Wortarten markiert. Die bekannten Wörter/Zeichen werden anhand des Wörterbuchs herausgefunden und mit der möglichen Wortart getaggt, in diesem Fall: ‚工人/$_{n}$' (*Arbeiter*), ‚们/$_{k}$' (Endung für die für *Menschen* stehenden Nomen, um die Pluralform auszudrücken), ‚忙碌/$_{a/v}$' (*beschäftigen*), ‚地/$_{u}$' (das Partikel als Adverbialmarkierung), ‚生产/$_{v}$' (*produzieren*), ‚触/$_{v}$' (*berühren*), und ‚屏/$_{n}$' (*Bildschirm*).
2. Analysen mit Wortbildungsregeln über Wortaffix, Morphemverdopplung usw. bei den Ergebnissen der Wortsegmentation (vgl. Kap. 4.4.1, S. 246f). Den Regeln gemäß kann <们> /mén/ nach <工人> vorkommen, um die Menschengruppe *Arbeiter* darzustellen. Deswegen müssen die beiden zu einem Satzglied gehören. Die Zeichenkollokation <忙忙碌碌> ist dem Wörterbuch zwar unbekannt, kann aber nach der Grundwortbildungsformel als die Verdopplungsform von <忙碌> in der Struktur AABB verstanden werden, deren Wortart und Bedeutung identisch mit AB sind (nämlich 忙忙碌碌/$_{a/v}$).
3. Ausfilterung der Wortartambiguität nach sprachlichen Regeln. Die sprachlichen Regeln über <忙忙碌碌>, <地> und <生产> werden abgerufen: Das Partikel <地> markiert das Satzglied Adverbial und muss mit seinem vorherigen Wort (das generell ein Adjektiv ist) ein Adverbial bilden; hinter dem Adverbial ist das Verb <生产> am wahrscheinlichsten Prädikatkopf des Satzes. So kann die Verboption von <忙忙碌碌> ausgeschlossen werden. Durch die Analysen in den drei Schritten kann der Satz in Wörter segmentiert und getaggt werden: ‚工人/$_{n}$ 们/$_{k}$ 忙忙碌碌/$_{a}$ 地/$_{u}$ 生产/$_{v}$ 触/$_{v}$ 屏/$_{n}$' (nach CN-Corpus Online).
4. Erkennung der unregistrierten Wörter und Bestimmung ihrer Wortart mit Wortbildungsregeln. Das Wort <触屏> (/chùpíng/, *Berührungsbildschirm*; <触>: *berühren*; <屏> *Bildschirm*) im Beispielsatz wurde von den meisten softwareinternen Wörterbüchern nicht aufgenommen. Nach Wortattributen treten die beiden Zeichen <触> und <屏> in den meisten Fällen nicht eigenständig als Wörter auf. Grammatisch und semantisch ist es regelkonform, die beiden Morpheme zu einem Kompositionswort zu verknüpfen, dessen Wortart am wahrscheinlichsten ein Nomen ist. Wenn diese Zeichenfolge im satzstufigen Kontext nach sprachlichen und statistischen Analysen als wahrscheinliches Wort erkannt werden kann

(siehe Kap. 4.4.4), wird sie vom System als Einheit segmentiert, mit der wahrscheinlichsten Wortart getaggt und im Wörterbuch eingetragen.

5. Überprüfung der Wortart mit syntaktischen Erkenntnissen, konkret den grammatischen Modellen der möglichen Reihenfolgen der Wortart im Satz. Durch diese Überprüfung werden mögliche Fehler herausgefunden, woraufhin diese erneut auf ihre Wortart hin getaggt werden.

Unter den fünf Schritten ist der erste die Grundlage und obligatorische Phase des POS-Taggings. Schritt 2 sowie 4 bieten zusätzliche Verfahren zur Worterkennung an. Nur der dritte Schritt konzentriert sich auf die Diskriminierung der Wortartambiguität, die mit Unterstützung der Wissensdatenbank der grammatischen Regeln abläuft. Größtes Problem des auf grammatischen Regeln basierenden Taggings ist jedoch: Je detaillierter die Regeln begründet werden, desto mehr Schwierigkeiten hat das System unter den möglichen Wortarten zu unterscheiden. Aber wenn die Regeln unausführlich eingetragen werden, kann die Wortart in vielen Fällen nicht automatisch korrekt markiert werden (vgl. Wang GZ/Wang XF 2008: 426). Das auf Statistiken basierende POS-Tagging wird deswegen in vielen Fällen für die Wortartdiskriminierung der multi-kategorialen Wörter eingesetzt, um Präzision und Effizienz des POS-Taggings zu erhöhen (vgl. Zhou Q/Yu 1993: 3f). Das auf Statistiken basierende POS-Tagging arbeitet nach folgender Funktionsweise (vgl. Feng 2001a: 7):

1) Eine bestimmte Menge von Texten wird ausgewählt, um ein Trainingsset (chi.: 训练集, eng.: training set) anzufertigen. Die Wortart aller Wörter in dem Trainingsset wird dann manuell markiert. Ein Wort wird dabei zusammen mit seinen beiden benachbarten Wörtern im Kontext mithilfe der Bigrammgrammatik berücksichtigt.
2) Anhand des Trainingssets wird die Frequenz des Zusammenauftretens zweier benachbarter Wörter errechnet und darauf basierend statistische Modelle gebildet.
3) Anhand der statistischen Modelle wird POS-Tagging bei den anderen Texten des Korpus durchgeführt.
4) Die möglichen Wortarten eines Wortes werden bei dem Wort des zur Textverarbeitung benötigten Wörterbuchs eingetragen.

Mithilfe der Wortartmöglichkeiten und der statistischen Berechnung kann ein Wort aus anderen Korpora getaggt werden, die in ähnlichen syntaktischen Kontexten innerhalb des Trainingssets die größte Wahrscheinlichkeit aufweisen. Der Nachteil dieses Verfahrens besteht darin, dass Fälle mit niedriger Wahrscheinlichkeit überwiegend ignoriert werden (vgl. Hou 1999: 160, Wang GZ/Wang XF 2008: 426f). Wegen der Einschränkung der auf grammati-

schen Regeln sowie Statistiken basierenden Verfahren ist es ideal, die beiden für das chinesische POS-Tagging zu verknüpfen. Ein Beispiel für Hybridverfahren ist das auf der Regelpriorität basierte Tagging. Die Regeln werden nach den Statistiken in verschiedene Klassen unterschieden und die prioritären Regeln zuerst berücksichtigt.

Um das Verfahren des POS-Taggings zu realisieren, müssen zuerst eine Wissensdatenbank mit ausreichenden Regeln und die Regelpriorität anhand des Trainingskorpus importiert werden. Die Priorität einer Regel ist durch die Frequenz dieser Regel im Bigrammmodell bestimmt. Die höchste Priorität wird mit ‚0' definiert und je nach Priorität werden die weiteren Regeln addierend mit ‚1', ‚2', ‚3' usw. definiert. Dabei kann eine Regeldatenbank nie vollkommen sein, da in einer natürlichen Sprache immer Ausnahmen vorkommen. Außer den durch die Korpusanalysen automatisch erhaltenen sprachlichen Regeln kann die Regeldatenbank auch manuell korrigiert, modifiziert und ergänzt werden. Unter Voraussetzung der Priorität kann der Algorithmus durchgeführt werden, um die Wortarten der Wörter im Satz zu definieren. Liegt bei einem Wort Wortartambiguität vor, muss es einzeln mit dem davor sowie dahinter stehenden Wort zusammengesetzt werden. Danach lässt sich die Priorität der beiden Seiten vergleichen, wobei mit der wahrscheinlichsten begonnen und der unwahrscheinlichsten Wortart aufgehört wird (vgl. Wang GZ/Wang XF 2008: 427).

Der segmentierte Satz <他/是/德国/人> (/tā shì déguó rén/; *Er ist Deutscher*) kann als Beispiel für POS-Tagging dienen. Das Zeichen <是> besitzt sieben mögliche Wortarten: Kopulaverb, Adjektiv, Adverb, Nomen, Pronomen, Konjunktion und Partikel (nach Xinhua-Lexikon Online). Die Wortart der anderen drei Wörter im Satz ist eindeutig und kann mithilfe des Wörterbuchs entsprechend markiert werden: ‚他/$_{v.}$ 是/$_{?}$ 德国/$_{ns.}$ 人/$_{n.}$' (nach CN-Corpus Online). Nach der Phrasenstrukturgrammatik bilden *Deutschland* und *Mensch* zusammen eine nominale Subordinationsphrase mit der Bedeutung *Deutsche -r/-n*. Nach der statistischen Berechnung können die Regeln zwischen <他> und <是> (Index1 zwischen ‚A' und ‚B') sowie <是> und der nominalen Phrase (Index2 zwischen ‚B' und ‚C') so ausgedrückt werden:

$Index1\text{-}P[0] = A_{r.} + B_{vl.}$,
$Index1\text{-}P[1] = A_{r.} + B_{d.}$,
$Index1\text{-}P[2] = A_{r.} + B_{c.}$,
$Index1\text{-}P[3] = A_{r.} + B_{n.}$,
$Index1\text{-}P[4] = A_{r.} + B_{a.}$;

$Index2\text{-}P[0] = B_{vl.} + C_{n.}$ (Entsprechung zu Index1-P[0]),
$Index2\text{-}p[0] = B_{c.} + C_{n.}$ (Entsprechung zu Index1-P[2]).

Unter der Voraussetzung, dass die Wortart von ‚B' identisch ist, wird eine Regel von Index1 sowie Index2 nach der Priorität angenommen, syntaktisch-kontextuell gültig ist. So schlägt das System zuerst das Ergebnis ‚A: 他/r. B: 是/vl. C: (德国/ns.+人/n.)/n.' vor. Da dieses Resultat grammatisch korrekt ist, wird das POS-Tagging des Satzes beendet. Wäre der Vorschlag syntaktisch ungültig, würden weitere Regeln eingesetzt und getestet, bis alle Wörter in ihrem syntaktischen Zusammenhang als grammatisch anerkannt wären. Der Wortart-Syntax-Zusammenhang wird in Kap. 4.5.2 und 4.5.3 näher erläutert.

Beim Durchlaufen des POS-Taggings in Pinyin-Ketten vervielfachen sich die Ambiguitätsfälle im Vergleich zu Schriftzeichen drastisch. Dies liegt hauptsächlich an den homophonetischen sowie homographischen Wörtern und multi-kategorialen Möglichkeiten. Exemplarisch wird nachfolgend der Arbeitsprozess des POS-Taggings anhand der Pinyin-Kette ‚TASHIDEGUOREN' analysiert, deren Wortsegmentationsergebnisse in Tab. 4-8 angegeben wurden. Basierend auf diesen acht Segmentationsmöglichkeiten wird das POS-Tagging in Tab. 4-10 durchgeführt. Die möglichen Zeichenketten werden nach drei Kriterien bewertet, nämlich ob sie syntaktisch sowie semantisch im modernen Chinesisch gültig sind (durch manuelle Analyse) und ob sie von der Sogou-Pinyin-Eingabesoftware als Kandidaten angeboten werden können (durch Eingabetest).

Varianten	**Pinyin-Kette**	**Kand.-1**; syntaktisch gültig; semantisch gültig; Ausgegeben?	**Kand.-2**; syntaktisch gültig; semantisch gültig; Ausgegeben?	**Kand.3**; syntaktisch gültig; semantisch gültig; Ausgegeben?
V. 1	TASHI/ DEGUO/REN	踏实/a (德国/ns+人/n)/n; ja; ja; nein.		
V. 2	TASHI/DE/ GUOREN	踏实/a 的/u 国人/n; ja; ja; ja.	踏实/a 的/u 果仁/n; ja; nein; nein.	踏实/d 地/u 过人/a; ja; nein; nein.
V. 3	TASHI/DE/ GUO/REN	踏实/d 地/u 裹/vt 人/n ; ja; nein; nein.	踏实/a 的/u 过/vd 任/v; nein; nein; nein.	踏实/a 德/n 过/vd 仁/a; nein; nein; nein.
V. 4	TA/SHIDE/ GUOREN	他/r 是的/u 国人/n ; nein; nein; nein.	他/r 师德/n 过人/a ; ja; ja; nein.	他/r 使得/v 国人/n ; ja; ja; nein.
V. 5	TA/ SHIDE/ GUO/REN	他/r 是的/u 过/vd 忍/a/v ; nein; nein; nein.	他/r 师德/n 果/d 仁 a/n ; nein; nein; nein.	他/r 使得/v 过/vd 韧/a ; nein; nein; nein.
V. 6	TA/SHI/ DEGUO/REN	他/r 是/vl (德国/ns+人 /n)/n ; ja; ja; ja.	他/r 时/nt 德国/ns 任/v ; nein; nein; nein.	她/r 是/vl (德国/ns+人 /n)/n ; ja; ja; ja.
V. 7	TA/SHI/DE/ GUOREN	他/r 使/v 得/u 国人/n ; ja; ja; nein.	他/r 识/v 得/u 国人/n ; ja; ja; nein.	他/r 试/v 地/u 过人/v ; nein; nein; nein.
V. 8	TA/SHI/DE/ GUO/REN	他/r 是/vl 得/u 过/vd 任/v ; nein; nein; nein.	他/r 是/vl 得/u 过/vd 任/v ; nein; nein; nein.	他/r 实/a 德/n 过/vd 仁/a ; nein; nein; nein,

Tab. 4-10: Das POS-Tagging der Pinyin-Kette ‚TA'SHI'DE'GUO'REN'[223]

[223] Die Ergebnisse des POS-Taggings der möglichen Zeichenketten verweisen auf den CN-Corpus Online; für die Abkürzungen der Wortarten siehe Tab. 3-13.

Durch Eingabetest mit der Sogou-Eingabemethode werden drei Kandidaten angeboten, die der Pinyin-Kette im vollen Umfang entsprechen. Abb. 4-27 zeigt nachstehend, wie V. 6-1 und V. 6-3[224] sowie V. 2[225] in Tab. 4-10 mithilfe von sprachlichen Erkenntnissen generiert und als Kandidaten angeboten werden.

Abb. 4-27: Wahlliste von ‚TA'SHI'DE'GUO'REN' mit Sogou-Eingabesoftware

Die Methoden des Pinyin-POS-Taggings sind im Prinzip identisch mit denen der Schriftzeichen, bei denen zuerst die auf Lexik und auf Statistiken basierenden Taggingverfahren, danach die syntaktischen Analysen im Kontext durchgeführt werden. So werden z.B. bei der Wortsegmentationsvariante ‚TA/SHI/DEGUO/REN' zuerst die entsprechenden, relativ häufigen Wörter inklusive ihrer jeweiligen Wortart angenommen. Nach Algorithmen und den grammatischen Regeln werden die wahrscheinlichsten Zeichenkandidaten erzeugt und angeboten. In diesem Beispiel kann die Silbe ‚TA' für eines der Pronomen <他>, <她> oder <它> (/tā/, *er/sie/es*), ein Nomen wie <塔> (/tǎ/, *Turm*) und ein Verb wie <踏> (/tà/, *trampeln*) stehen. Das Pinyin ‚SHI' hat wegen seiner Homophonvielfalt mindestens sechs Wortartmöglichkeiten, von denen Nomen, Verb und Adjektiv von dominierender Wahrscheinlichkeit sind. Die bisyllabische Einheit ‚DEGUO' hat im Gegensatz zu den beiden vorderen monosyllabischen Wörtern in der Wörterdatenbank nur eine Wortmöglichkeit, nämlich <德国> /déguó/ für *Deutschland*. Die Silbe ‚REN' allein hat zwar mehrere Zeichen- sowie Wortartmöglichkeiten, aber das Zeichen <人> (/rén/, *Mensch*) kann mit dem vorderen Wort eine festgelegte, häufig gebrauchte Phrase <德国人> (*Deutsche -/r/n/*; nominale Phrase) bilden. Nach den statistischen und grammatischen Analysen ist so ‚Pronomen+Kopula+NP' am wahrscheinlichsten. Nach den semantischen Kenntnissen sind <他> und <她> als Subjekt gültig. Wie die automatischen syntaktischen sowie semantischen Analysen bei einer intelligenten Pinyin-Eingabesoftware ablaufen, wird in Kap. 4.5.1 und 4.5.4 erläutert.

[224] V. 6-1 und V. 6-3 entsprechen dem ersten sowie dem dritten Kandidaten in der Wahlliste in Abb. 4-27, die *er/sie ist Deutscher/Deutsche* bedeutet; ihre syntaktische Struktur kann mit ‚A_r + B_{vl} + CDE_n' symbolisiert werden. ABCDE stehen für die fünf Zeichen der Zeichenkette nach ihrer Reihenfolge, die kleinen Buchstaben rechtsunten für Wortart (vgl. Tab. 4-9).

[225] V.2 entspricht dem zweiten Kandidaten der Wahlliste von Abb. 4-27, der *zuverlässiges Landesvolk* bedeutet. Seine syntaktische Struktur kann mit ‚AB_a + C_u + DE_n' symbolisiert werden (vgl. Tab. 4-9).

4.4.4 Erkennung, Festlegung und POS-Tagging unregistrierter Wörter

Wie in Kap. 4.2.3, 4.4.2 und 4.4.3 vorgestellt wurde, sind Worterkennung sowie POS-Tagging von der Indexierung im maschinenlesbaren Wörterbuch bedingt. Wegen der Einschränkung des Wörterbuchs kann nicht jedes Wort aufgenommen werden. Ein unregistriertes Wort kann nicht mit allgemeinen Methoden segmentiert sowie mit Wortart getaggt werden. Die Verfahren zur Erkennung solcher Wörter stellen den Schwerpunkt dieses Unterkapitels dar. Diese Techniken sind einerseits für das Selbstlernmodul der Eingabesoftware notwendig, andererseits auch entscheidend für die bessere Qualität automatischer Korporaanalysen (vgl. Wang XL 2005: 31f). Wie in den letzten beiden Kapiteln werde ich zuerst die linguistischen Grundtheorien der Wortbildung erläutern, dann die Worterkennungstechniken bei Korpora skizzieren und zuletzt ihre Anwendung bei Eingabesoftwares analysieren.

Wie in Kap. 4.4.1 (S. 248) eingeführt wurde, sind die Grenzen zwischen den beiden Einheiten undeutlich und die linguistischen Bildungsprinzipien der beiden Einheiten hochgradig übereinstimmend. Das hat Konsequenzen für die Verarbeitungsmethoden. Bei der Textverarbeitung können (vor allem die unregistrierten) Wörter und Phrasen in den meisten Fällen sowohl durch innerliche als auch äußerliche Aspekte analysiert und verarbeitet werden. Der innerliche Aspekt bezieht sich auf die Analysen der Bestandteile, konkret der Bedeutung, der Wortart und den grammatischen Zusammenhängen von Morphemen/Zeichen eines Wortes sowie von Wörtern einer Phrase. Dem hingegen sind mit dem äußerlichen Aspekt die syntaktischen Beziehungen zu den anderen sprachlichen Einheiten im Kontext gemeint (vgl. Feng 1999: 301-304). Wie Wortsegmentation und POS-Tagging kann die Verarbeitung der unregistrierten Wörter und Phrasen ebenso sowohl mit linguistischen Erkenntnisse als auch mit statistischen Analysen durchgeführt werden (vgl. Wang XL 2005: 41, Evert et al. 2010: 114).

Unregistrierte Wörter sind mehrheitlich komplexe Wörter und Eigennamen für Personen, Orte, Organisationen etc. Erstere bestehen aus Zeichen/Morphemen, die dem elektronischen Lexikon bekannt sind. Sie können normalerweise mit Grammatiken und Algorithmen erkannt werden (vgl. Yao 1997: 187). Solche Zeichenfolgen werden aber – ohne ausreichende linguistische und statistische Analysen innerlicher und äußerlicher Aspekte – normalerweise nicht als Segmentationseinheiten, sondern als Phrasen erkannt und verarbeitet (vgl. Wang XL 2005: 40). Bspw. ist das Wort <房奴> (/fángnú/, wörtlich: *Wohnungssklave*) ein in der heutigen chinesischen Gesellschaft neu entstandenes Wort. Es bezeichnet Menschen, die lebenslang hart arbeiten müssen, um den Kauf einer Wohnung (bzw. den dafür notwendigen Kredit) zurückzuzahlen. Wegen der beiden bestehenden nominalen Morpheme kann das System zwar diese Zeichenkollokation als nominale Einheit erkennen, nicht aber als eine Komposition, sondern

als eine Phrase aus zwei Nomen (nach CN-Corpus Online). Wie das System unregistrierte Wörter als eigenständige Einheit erkennen, segmentieren und taggen kann, wird im weiteren Verlauf des Kapitels erläutert.

Zunächst soll jedoch der Erkennungsprozess chinesischer Eigennamen verknappt dargestellt werden. Hauptproblem hierbei ist, dass die für Eigennamen stehenden Zeichen generell keine bestimmte grammatische und semantische Beziehung zueinander haben und statistisch gesehen häufig zufällig zusammen auftreten, weshalb sie nicht durch allgemeine Wortbildungsregeln erkannt werden können. Für die Verarbeitung können die Strukturen der Eigennamen und syntaktische Analysen herangezogen werden (vgl. Yao 1997: 186f). Ein Ortsname endet in offiziellen Schriftwerken häufig mit einer administrativen Einheit wie <康金镇> (/kāngjīn zhèn/; 镇 entspricht *Großgemeinde*). Der Name eines Instituts oder einer Organisation hat häufig ein Institutionsnomen am Ende, etwa <金汉斯餐厅> (/jīnhànsī cāntīng/; 餐厅 entspricht *Restaurant*). Dank der Eigennamenmarker können die beiden Beispielwörter mit wenigen Schwierigkeiten als Eigennamen – *Kangjin-Gemeinde* sowie *Goldener Hans Restaurant*[226] – erkannt werden. In dem Satz <王铠很忙> (/wáng kǎi hěn máng/, *Wang Kai ist sehr beschäftigt*) z.B. ist <王铠> eine dem Computersystem unbekannte Zeichenkollokation. Im syntaktischen Kontext wird an der Stelle von <王铠> ein menschliches Subjekt benötigt. Nach den semantischen Informationen der beiden Zeichen sollte diese Kollokation *Rüstung des Königs* bedeuten, die aber unmöglich *beschäftigt* (忙) sein kann. Nach der Struktur allgemeiner chinesischer Namen (monosyllabischer Familienname + mono-/bisyllabischer Vorname) und der Liste der häufigen chinesischen Familiennamen kann herausgefunden werden, dass es sich am wahrscheinlichsten um einen chinesischen Personalnamen handelt.[227]

Wie erwähnt sind die Schlussfolgerung der Bedeutung und das Tagging der Wortart zwei der wichtigsten Aufgaben bei der Verarbeitung der unregistrierten Wörter. Wie in Kap. 4.2.1 vorgestellt wurde, funktioniert die Wortbildung des Chinesischen zeichenbasiert, d.h. Bedeutung und Wortart eines unbekannten komplexen Wortes können durch das Attribut des bestehenden Zeichen/Morpheme analysiert werden. Damit unbekannte Wörter vom Computer verstanden werden können, müssen zuerst linguistische Erkenntnisse über Wortbildungsgramma-

[226] Dabei handelt es sich um eine Restaurantkette in China, die westliche Küche feilbietet; benannt wurde sie nach dem deutschen Brauer Hans Müller, der zum Aufbau der ketteneigenen Brauerei beigetragen hat.

[227] Dabei darf nicht vergessen werden, dass die Anzahl der chinesischen Familiennamen für europäische Verhältnisse begrenzt ist. Laut der Chinesischen Volkszählung von 2010 machen die hundert häufigsten Familiennamen bereits 82,1% der Gesamtbevölkerung aus. Im Gegenteil zur westlichen Welt kann es jedoch keine bestimmte chinesische Vornamensliste geben, da die Zeichen für Vornamen sehr frei ausgewählt werden.

tik und die Beziehungen der Gesamt-Bestandteil-Semantik in der Wissensdatenbank eingeschrieben werden. Die semantischen Zusammenhänge zwischen einem komplexen Wort und seinen Bestandteilen zeigen sich vor allem auf fünf Weisen.

> ① Gleiche Bedeutung, nämlich A+B = A = B, wie ‚声音' [/shēngyīn/, *Laut*], 声+音=声=音 [*Laut + Laut = Laut = Laut*];
> ② Kombinierende Bedeutung, nämlich A+B = AB, wie ‚品德' [/pǐndé/, *moralischer Charakter*], 品+德=品质道德 [*Charakter + Moral = moralischer Charakter*];
> ③ Additionsbedeutung, nämlich A+B = AB+C, wie ‚景物' [/jǐngwù/, *Sehenswürdigkeit*], 景+物=（可供欣赏的）景致和事物 [*Landschaft + Gegenstand = (sehenswürdige) Landschaft und Gegenstände*];
> ④ Rand- und Kernbedeutung, nämlich A+B = A oder B, wie ‚国家' [/guójiā/, *Staat/Land*], 国+家=国 [*Staat/Land + Familie = Staat/Land*];
> ⑤ Wandelnde Bedeutung, nämlich A+B = C, wie ‚毛病' [/máobìng/, *menschlicher Nachteil*], 毛+病=人的缺点 [*Pferdefell + Krankheit = menschlicher Nachteil*].
> (Yao 1997: 181 [Übersetzung der Verfasserin])

Unter diesen fünf machen die zweite und dritte Art 89,7% der Fälle von komplexen Wörtern aus, d.h. in den meisten Fällen wird die gesamte Wortbedeutung von der Bedeutung der bestehenden Morpheme/Zeichen kombiniert oder mit zusätzlichen Informationen addiert (vgl. ibid.: 181f). Im Gegenteil dazu wird der fünfte Fall relativ selten benötigt, der kaum durch die Bestandteile aufgelöst werden kann. Die semantischen Beziehungen der Bestandzeichen müssen bei der Anwendung gemeinsam mit den grammatischen Zusammenhängen und der Wortbildungsart berücksichtigt werden. In Kap. 4.4.1 wurden die drei Formen der chinesischen komplexen Wörter vorgestellt: Grundmorphemkombination, Affixoid-Wortwurzel und Morphemverdopplung. Die erste – von zwei oder mehreren Grundmorphemen kombinierte – Komposition ist der häufigste Fall, der sich in fünf Grund- und drei Sonderbildungsarten unterscheidet. Die grammatischen sowie semantischen Beziehungen der Bestandteile eines Kompositionsworts lassen sich kombiniert betrachtet so zusammenfassen:

1. Klasse	**2. Klasse**	**Wortart**	**Grammatische Beziehungen**	**Semantische Beziehungen**	**Beispiel**[228]
Grundmorphem-Kombination	Koordination	n/ v/ a/ d	$W_n=A_n+B_n$; $W_v=A_v+B_v$; $W_a=A_a+B_a$; $W_d=A_d+B_d$.	A und B sind identische, aufeinander bezogene oder antonyme Begriffe; die Semantik von W entspricht entweder ① ②③④ oder ⑤.	声音 n① 飞跃 v② 多少 a② 始终 d② 开关 n② 景物 n③ 国家 n④ 东西 n⑤

[228] Nach der Reihenfolge: Wortform, Wortart und Art der semantischen Beziehung des Beispielworts.

1. Klasse	2. Klasse	Wort-art	Grammatische Beziehungen	Semantische Beziehungen	Beispiel[228]
	Subordination	n/ v/ a	$W_n=A_{n/v/a}+B_n$; $W_v=A_{n/a/v/d}+B_v$; $W_a=A_{n/a/v/d}+B_a$.	A modifiziert B; B stellt die Kernbedeutung von W dar (Fall ②④⑤).	开水 n② 重视 v② 飞快 a④ 毛病 n⑤
	Verb-Objekt	v	$W_v=A_v+B_n/B_a/B_v$ [229]	W hat eine kombinierende Bedeutung aus A und B, mit oder ohne Bedeutungswandel (Fall ②⑤).	开会 v② 读书 v⑤
	Prädikat-Komplement	v/ a	$W_v=A_v+B_{v/a/d}$; $W_a=A_a+B_{v/a/d}$	A stellt Kopf und B die Ergänzung von W dar (Fall ②④⑤).	开明 v② 茂密 a④ 吃紧 v⑤
	Subjekt-Prädikat	n/ v/ a	$W_n=A_n+B_{v/a}$; $W_v=A_n+B_v$; $W_a=A_n+B_a$.	W hat eine kombinierende Bedeutung aus A und B oder A als Kernbedeutung (Fall ②④⑤).	月亮 n④ 心开 v② 年轻 a② 牛饮 v⑤
	Verb-Kombination	v	$W_v=A_v+B_v$	A und B sind fortsetzende, aufeinander bezogene Aktionen (Fall ②).	退休 v② 躲开 v②
	Drehpunkt	v	$W_v=A_v+B_v$	Objektivierung des Verbs A und Subjektivierung des Verbs B sind semantisch identisch (Fall ②).	请教 v②
	Nomen-Zahleinheit	n	$W_n=A_n+B_q$	B ist das bestimmte ZEW von A und W ist das Synonym von A bei Mengenangaben (Fall ④).	书本 n②
Affixoid	Mit Präfix	n/ v/ a/ m/ nt	$W_n=A_{h1}+B_n$; $W_v=A_{h2}+B_v$; $W_a=A_{h3}+B_{a/v}$; $W_m=A_{h4}+B_m$; $W_{nt}=A_{h5}+B_m$.[230]	A stellt die Rand- und B die Kernbedeutung von W dar (Fall ④).	老虎 n④ 打听 v④ 可怕 a④ 第五 m④ 初三 nt④
	Mit Suffix	n/ v/ a/ d	$W_n=A_{n/v/a}+B_{k1}$; $W_a=A_{n/a}+B_{k2}$; $W_d=A_{v/a}+B_{k3}$.[231]	A stellt die Kern- und B die Randbedeutung von W dar (Fall ④).	孩子 n④ 绿化 a④ 竟然 d④
Verdopplung	AA-Form	n/ v/ a/ d	$W_n=A_n+A_n$; $W_v=A_v+A_v$; $W_a=A_a+A_a$; $W_d=A_d+A_d$.	W verstärkt durch Verdopplung die Bedeutung von A (Fall ①③).	妈妈 n① 尝尝 v① 悠悠 a① 常常 d① 天天 n③

[229] Im Chinesischen können Verben und Adjektive wie Nomen auch als Subjekt und Objekt verwendet werden.

[230] A_{h1}= {老, 小, 阿…}, A_{h2}= {打}, A_{h3}= {可}, A_{h4}= {第}, A_{h5}= {初}.

[231] B_{k1}= {子, 儿, 者, 家, 手, 气…}, B_{k2}= {化}, B_{k3}= {然}.

1. Klasse	2. Klasse	Wort-art	Grammatische Beziehungen	Semantische Beziehungen	Beispiel[228]
	ABB-Form	a/ v	$W_a=A_{a/n}+BB_{a/o}$; $W_v=A_v+BB_{o/a/v}$.	A und BB sind in einer Prädikat-Komplement-Struktur (Fall ②).	乐呵呵 a② 笑哈哈 v②
	AAB-Form	n/ v/ a	$W_n=AA_n+B_n$; $W_v=AA_{o/n}+B_v$; $W_a=AA_a+B_a$.	AA subordiniert B (Fall ②).	面面观 n② 呱呱叫 v② 麻麻亮 a②
	ABAB-Form[232]	v	$W_v=AB_v+AB_v$	AB wird verdoppelt (Fall ②).	安排安排 v②
	AABB-Form	n/ v/ a	$W_n=AA_n+BB_n$; $W_v=AA_v+BB_v$; $W_a=AA_a+BB_a$.	AA und BB sind koordinativ; dabei muss AB im Kern ein Wort bilden können (Fall ②).	家家户户 n② 吵吵闹闹 v② 浩浩荡荡 a②
	ABAC-Form	n/ v/ a/ d	$W_n=AB_n+AC_n$; $W_v=AB_v+AC_v$; $W_a=AB_a+AC_a$; $W_d=AB_d+AC_d$.	AB und AC sind koordinativ (Fall ②).	挨家挨户 n② 百发百中 v② 不伦不类 a② 不上不下 d②
	AABC-Form	n/ v/ a	$W_n=AA_{n/v/a}+BC_n$; $W_v=AA_{n/a/v/d}+BC_v$; $W_a=AA_{n/a/v/d}+BC_a$.	AA subordiniert BC (Fall ④).	鼎鼎大名 n④ 哈哈大笑 v④ 比比皆是 a④
	ABCC-Form	v/ a	$W_v=AB_n+CC_v$; $W_a=AB_n+CC_a$	AB und CC sind in Subjekt-Prädikat- oder Prädikat-Komplement-Struktur (Fall ②④).	来去匆匆 v② 喜气洋洋 a②

Tab. 4-11: Die Wortbildungsgrammatik und die Semantik der chinesischen Kompositionswörter[233]

Von der formalen Grammatik dieser komplexen Wörter ist es zu ersehen, dass die mit Affix- sowie Verdopplungsform gebildeten Wörter mit relativ wenigen technischen Schwierigkeiten vom Computer erkannt, getaggt und verstanden werden. Sowohl Präfixe als auch Suffixe gehören zu bestimmten Affixlisten und sind nur mit einem Grundmorphem bestimmter Wortarten kombinierbar (siehe Fußnoten der Zeile ‚Affixoid' in Tab. 4-11). Bei den Wörtern mit Verdopplungsform können sich wiederholende Bestandteile herausgefunden und anhand von lexikalischen Regeln die Bildungsart, die Wortart und die vage Semantik bestimmt werden. Im Gegenteil zu der zweiten sowie dritten Klasse sind die Bildungsgrammatiken von der ersten Klasse komplizierter darzustellen. Einerseits kann ein Schriftzeichen in verschiedenen Bildungsarten in unterschiedlicher Position eines komplexen Wortes auftreten. Bspw. kann

[232] In vielen Fällen wird eine Zeichenfolge in dieser Form als Phrase betrachtet.

[233] A, B und C stehen in den meisten Fällen für verschiedene Morpheme; die Codes für die Wortarten orientieren sich an GB/T 20532-2006 und werden in Tab. 4-10 angegeben; die Fälle der semantischen Beziehungen entsprechen den im vorletzten Abschnitt angegeben fünf Fällen.

das Zeichen <开> (/kāi/, Grundbedeutung: *öffnen* [v.], weitere Bedeutungen: *stattfinden*, *steuern* usw. [v.]) mindestens in sechs verschiedenen Wortbildungsformen verwendet werden (siehe Beispielwörter mit Unterstrich in Tab. 4-11). Andererseits ist die Zusammensetzung zweier oder mehrerer Grundmorpheme ebenso stark von semantischen sowie pragmatischen Erkenntnissen bedingt. Analysiert mit dem Morphem <开> ist es nur mit begrenzten Morphemen kombinierbar, bspw. in der Verb-Objekt-Struktur nur als *geöffnetes*, *gesteuertes* oder *veranstaltetes* Morphem. Unter gleich- oder ähnlich bedeutenden Morphemen wird in den meisten Fällen nur ein einziges Morphem als feste Kollokation zur Wortbildung anerkannt. Zum Ausdruck *Tagung stattfinden* etwa ist das richtige Wort <开会> /kāihuì/, wohingegen es sich um kein Wort handelte, wenn <会> (/huì/, *Tagung*, *Konferenz*) durch ein Synonym ersetzt würde (vgl. Lu 2008: 152f).

Wegen der Komplexität der Wortbildungsprinzipien können die auf den Wortbildungsgrammatiken basierenden Techniken nicht allein für die Verarbeitung der unregistrierten Wörter zuständig sein, weshalb algorithmische Methoden unentbehrlich sind. Die Statistik für Worterkennung bezieht sich im Normalfall auf vier Attribute: 1) die Wortbildungskraft der bestehenden Schriftzeichen, 2) das Wortbildungsmodell eines Zeichens, 3) Korrelationen von Zeichenpaaren im Wort, 4) Beziehungen zu anderen Wörtern im Kontext (vgl. Wang XL 2005: 40f). Jedes der vier Attribute kann mit mathematischen Methoden errechnet werden und zur Errechnung der Wahrscheinlichkeit eines unregistrierten Wortes beitragen. Zur Erklärung der vier Attribute wird das Kompositionswort <图书馆> (/túshūguǎn/, *Bibliothek*) als Beispiel analysiert. Das Wort ist in der ersten Hierarchie die Subordinationskomposition von <图书> (Gesamtbezeichnung für *Buch, Lehrwerk* und *Bild*, meistens als ein Synonym für <书> *Buch*) und <馆> (*Räumlichkeit*) mit der zweiten semantischen Beziehung. Das Kurzwort <图书> besteht weiterhin aus den Morphemen <图> (*Bild*) und <书> (*Buch*) in Subordination mit der vierten semantischen Beziehung.

Die Wortbildungskraft (chi.: 构词能力, eng: word formation power, Abk.: WFP) eines Zeichens meint die Wahrscheinlichkeit, dass ein Zeichen zur Wortbildung verwendet wird, deren Wert zwischen Null und Eins liegt. Die Wortbildungskraft eines Schriftzeichens C kann mit der folgenden Formel berechnet werden:

$$WFP(C) = \frac{Anzahl\ der\ mit\ C\ gebildeten\ Mehr-Gramm-Wörter}{Anzahl\ von\ C}$$

Formel 1 (Wang XL 2005: 40)

Das Ergebnis eines Zeichens mit dieser Formel ist die Wahrscheinlichkeit, dass es in einem Mehr-Gramm-Wort (inklusive komplexen und polysyllabischen simplen Wörter) auftritt. Die Wahrscheinlichkeit, dass C eigenständig als Wort auftritt, ist 1−WFP(C). Nach der Häufigkeitsanalyse des CN-Corpus tritt z.B. das Schriftzeichen <馆> (/guǎn/, *Räumlichkeit*) insgesamt 7.662-mal auf, darunter 7.571-mal in einem komplexen Wort. So beträgt die Wortbildungskraft von <馆> 98,81% (7.571/7.662*100%). Dem hingegen beträgt die Wahrscheinlichkeit, dass es als eigenständiges Wort auftritt, 1,19% (1−98,81%). Nach derselben Methode beziffert sich die Wortbildungskraft der beiden anderen Zeichen auf 56,73% für <图> und 75,63% für <书> (errechnet nach CN-Corpus Online).

Das Wortbildungsmodell meint die Wahrscheinlichkeit der Position eines Zeichens in einem Mehr-Gramm-Wort. Die Zeichenposition kann zusammenfassend eine von drei Fällen sein: als Kopf- (head; Abk.: H; Modell H), als Ende- (tail; T; Modell T) und als Mitte-Zeichen (middle; M; Modell M). Die Struktur eines Mehr-Gramm-Worts ist entweder ‚H+T' oder ‚H+M(+M…)+T'. Die Wahrscheinlichkeit eines Schriftzeichens in einem Wortbildungsmodell (als H, T oder M) kann mit der folgenden Formel ermittelt werden:

$$\Pr(pttn(C)|C) = \frac{Anzahl\ von\ pttn\ (C)}{Anzahl\ der\ mit\ C\ gebildeten\ Mehr-Gramm-Wörter}$$

Formel 2 (Wang XL 2005: 40)[234]

Nach Statistiken des CN-Corpus wird <馆> (/guǎn/, *Räumlichkeit*) 52-mal als H, 7.503-mal als T und 13-mal als M in einem komplexen Wort benutzt. Die Häufigkeit von <馆> in Mehr-Gramm-Wörtern in demselben Korpus beträgt 7.571 (vgl. auch das Beispiel für Wortbildungskraft). Nach Anwendung der Formel gehört das Zeichen mit 99,14% (7.503/7.571*100%) zum T-Modell und ist damit weit wahrscheinlicher als das H- (0,69%) und M-Modell (0,17%) ist. Analysiert mit derselben Formel liegt die H-Modell-Wahrscheinlichkeit von <图> bei dem Wert 44,53%, die damit kaum Abstand zu seiner T-Modell-Wahrscheinlichkeit hat. Die M-Modell-Wahrscheinlichkeit <书> liegt bei 11,45% und so unter der H- sowie T-Modell-Wahrscheinlichkeit (nach CN-Corpus Online).

Die Korrelation eines Zeichenpaars ist im Prinzip die bedingte Wahrscheinlichkeit eines Bigramms, die mit Bayes Theorem errechnet werden kann. Angenommen wird, dass es ein unregistriertes Wort AB (wobei A und B jeweils für ein einzelnes Zeichen steht) gibt. So kann

[234] In der Formel steht Pr(pttn(c)|c) für die Wahrscheinlichkeit von c in einem Wortbildungsmodell eines Mehr-Gramm-Wortes, konkret die Wahrscheinlichkeit als H, T oder M in einem Mehr-Gramm-Wort aufzutreten.

die Wahrscheinlichkeit, dass B von A bedingt ist und nach A im Text vorkommt, mit der folgenden Formel ermittelt werden:

$$P\,(B|A) = \frac{P\,(B \cap A)}{P\,(A)}$$

Formel 3 (Evert 2010: 120)[235]

Bestünde ein unregistriertes Wort aus drei oder mehr Zeichen, würde die bedingte Wahrscheinlichkeit von jedem benachbarten Zeichenpaar ausgerechnet. Durch die Differenz zwischen den Wahrscheinlichkeiten ist es in vielen Fällen möglich, das Wort in Morpheme zu strukturieren. Exemplarisch soll die Korrelation des Zeichenpaars in dem Wort <图书馆> (/túshūguǎn/, *Bibliothek*) analysiert werden. Mithilfe von Statistiken des CN-Corpus und nach der Anwendung der obigen Formel ist <书> zu 3,40% von <图> abhängig; die Abhängigkeit von <馆> zu <书> beträgt 2,14%.[236] Das stärker korrelierende Zeichenpaar gehört mit höherer Wahrscheinlichkeit als Morphem oder Wort zusammen. Deswegen wird das Wort vom System hierarchisch in dem Zeichenkombinationsmodell (character join model; Abk.: CJM) in <图书/馆> zerlegt. Im nächsten Schritt kann die bedingte Wahrscheinlichkeit von <馆> zu <图书> ausgerechnet werden, um weiter zu bestimmen, ob diese Kollokation ein Wort ist. Algorithmisch liegt der Wert bei 55,48%, was im Vergleich zu der anderen Modellmöglichkeit um ein zehnfaches wahrscheinlicher ist (nach CN-Corpus Online).[237]

Die Wortbildungskraft sowie das -modell der bestehenden Zeichen und die Korrelation eines Zeichenpaars beziehen sich auf die Analysen der inneren Struktur eines Wortes. Um ein unbekanntes Wort festzulegen, ist auch der äußerliche Aspekt entscheidend, z.B. die Zusammenhänge zu den benachbarten Wörtern. Je häufiger die Zeichenfolge in verschiedenen sprachlichen Situationen auftritt und seltener sie von den benachbarten Wörtern abhängig ist, desto wahrscheinlicher ist sie ein eigenständiges Wort.

Angenommen, es gäbe eine Wortkette $W_L W_{Single} W_R$, wobei in der Kette W_L und W_R bekannte Wörter sind, W_{Single} aber eine Zeichenkollokation aus $C_1C_2\ldots C_n$ darstellt. In diesem Fall wird die Hypothese vom System gesetzt, dass es unter W_{Single} ein unregistriertes Wort $\hat{W}_{new}$ gibt ($\hat{W}_{new} \leq W_{Single}$). $\hat{W}_{new}$ besteht aus einer Zeichenkette $X_1X_2\ldots X_n$, die innerhalb von $C_1C_2\ldots C_n$ ist. So kann die maximale Wahrscheinlichkeit, dass $\hat{W}_{new}$ ein Wort ist, mit der folgenden algorithmischen Formel analysiert werden:

[235] In der Formel steht P(B|A) für die bedingte Wahrscheinlichkeit von B durch A; P(B∩A) repräsentiert die Häufigkeit, dass B nach A vorkommt.

[236] P(书∩图) = 301 und P(图) = 8.863; P(馆∩书) = 203 und P(书) = 9.468.

[237] P(馆 | 图书) = 167 und P(图书) = 301; die Wahrscheinlichkeit von P(书馆 | 图) beträgt entgegen nur 1,88%.

$$\hat{W}_{new} = \underset{W_{new}}{arg\,max} \{P_{WFP}(W_{new}) + P_{pttn}(W_{new}) + P_{CJM}(W_{new}) + P_{bigram}(W_{new})\}$$

$$= \underset{W_{new}}{arg\,max} \{\Sigma_{i=1,\ldots n} (P_{WFP}(X_i) + P_{pttn}(X_i) + P_{CJM}(X_i) + P_r(X_i/X_{i-1}))\}$$

Formel 4 (Wang XL 2005: 41)[238]

Die Verarbeitung mithilfe dieser Formel fasst Wang XL (2005: 41 [Übersetzung der Verfasserin]) in vier Arbeitsschritte zusammen.

1) In der Kette $W = X_1X_2\ldots X_n$ einzelne Zeichen sowie bekannte untergeordnete Wörter herausfinden.
2) Überprüfen, ob es in der Kette Verdopplungsformen gibt, wie die Struktur ABB, AAB, AABB, ABAB, ABAC etc.[239] Wenn ja, folgt unmittelbar Schritt 4). Ansonsten erfolgt der dritte Schritt.
3) Die Hypothese wird aufgestellt, dass W_{new} ein unbekanntes Wort ist. Es wird dann mit der algorithmischen Formel errechnet. Wenn das Ergebnis den für ein Wort definierten Wert überschreitet, wird die Hypothese als wahr festgelegt. W_{new} wird dann als Wort segmentiert.
4) Recherche der restliche(n) Zeichenkette(n) von W_{Single}. Gibt es noch weitere Zeichenketten, werden Schritt zwei und drei wiederholt.

Mit diesem Arbeitsprozess kann ein unregistriertes Wort innerhalb einer Zeichenkette in einem Trainingskorpus erkannt und automatisch in einem Wörterbuch gespeichert werden. Mithilfe von den Zusammenhängen der Bedeutung der bestehenden Morpheme/Zeichen ist es weiterhin möglich, die vage Bedeutung des unregistrierten komplexen Wortes zu mutmaßen (vgl. Hou 1999: 162). Angenommen wird, dass das Wort <图书馆> nicht vom elektronischen Lexikon aufgenommen worden sei: Nach den statistischen Analysen mit den oben genannten Arbeitsphasen wird es als ein Wort aus <图书> (*Buch*, n.) und <馆> (*Räumlichkeit*, n.) verstanden. Nach der Wortbildungsgrammatik ist ein aus zwei Nomen bestehendes komplexes Wort ebenfalls ein Nomen. Nach der semantischen Eigenschaft der beiden Bestandswörter müssen diese sich zu einer Subordinationsform kombinieren. Die Semantik von <图书馆> ist hochwahrscheinlich die semantische Kombination/Addierung von *Buch* und *Räumlichkeit*, was am wahrscheinlichsten *Bibliothek* bedeuten könnte.

Die Erkennung unregistrierter Wörter kann die Satzverarbeitung als Grundlage unterstützen. Sie ist auch eine wichtige Aufgabe des Wortsegmentationssystems und ein entscheidender Arbeitsprozess für die Verbesserung zur Textverarbeitung (vgl. Wang XL 2005: 40).

[238] In dieser Formel steht $P_{CJM}(W_{new})$ für *Zeichenkombinationsmodell* (character join model), während $P_{bigram}(W_{new})$ für die Korrelation zweier Zeichen innerhalb des Wortes steht.
[239] Siehe Zeile ‚Verdopplungsform' in Tab. 4-11.

In intelligenten Eingabesoftwares sind normalerweise relativ einfache Worterkennungstechniken gebräuchlich. Der Umfang des der Eingabesoftware angehörigen Wörterbuchs wurde in Kap. 4.4.2 (S. 253f) eingeführt. Gematcht mit dem Wörterbuch werden die unregistrierten Wörter verschiedener Typen unterschiedlich verarbeitet. Die aus Affixen sowie in Verdopplungsform gebildeten Wörter können nach den bekannten Erkenntnissen als Worteinheit erkannt, weiter verarbeitet und ins Wörterbuch hinzugefügt werden. Die Wörter der Grundmorphemkombination werden generell als Phraseneinheit erkannt. Man kann ein unregistriertes Wort manuell bestimmen und ins individuelle Wörterbuch importieren. Wenn eine Zeichenfolge relativ häufig benutzt wird, wird sie in vielen Fällen automatisch als Wort vom Selbstlernmodul akquiriert und im Wörterbuch eingetragen (vgl. Wang XL 1993: 372).

4.4.5 Erkennung und Verarbeitung der Phrasen

Im letzten Kapitel wurde erläutert, dass Phrasen mit vergleichbaren Methoden wie unregistrierte Wörter verarbeitet werden können, nämlich mit linguistischen Sprachregeln und statistischen Analysen. Nach dem nationalen Standard GB/T 12200.2-94 ist die Phrase des Chinesischen „die sprachliche Einheit, aus der sich zwei oder mehrere Wörter in bestimmter Stufenordnung und nach bestimmten syntaktischen Beziehungen zusammensetzen“ (GB/T 12200.2-94: Kap. 4.1.5.2 [Übersetzung der Verfasserin]).[240] Wegen der komplizierten Komposition der Grundmorpheme können die statistikbasierten Methoden bei der Erkennung der unregistrierten Wörter effektiver wirken und werden deswegen bevorzugt (siehe Kap. 4.4.4). Im Vergleich zur Wortbildung ist die Wortkollokation zur Phrasenbildung flexibler. Wie in Kap. 4.4.1 und 4.4.3 gezeigt wurde, ist die Wahrscheinlichkeit von Ambiguitäten und multikategorialen Wortarten bei Mehr-Gramm-Wörtern viel geringer, als bei monosyllabischen Wörtern. Die Bestandteile der Phrasen hängen pragmatisch seltener zusammen. Aus semantischer Perspektive ist die Semantik der Phrasen meist eine kombinierende Semantik der einzelnen Bestandteile (vgl. Lü 1979: 30). Aufgrund linguistischer Unterschiede (vor allem bei Grammatik, Semantik und Pragmatik) im Kontrast zur Ebene des Wortes sind bei Phrasen solche Methoden beliebter, die auf dem sprachlichen Verstehen basieren (vgl. Yao 1997: 231).

Zur Erläuterung der automatischen Phrasenerkennung und -verarbeitung werden zunächst die chinesischen Phrasen aus linguistischer Perspektive (äußerlicher syntaktischer Kontext und innere Konstruktionsprinzipien) vorgestellt. Von der allgemeinen syntaktischen Funktion aus betrachtet können Phrasen in drei Typen klassifiziert werden:

[240] Im Chinesischen gibt es für den Terminus *Phrase* die Begriffe 短语/duǎnyǔ/, 词组/cízǔ/ und 仂语 /lèyǔ/. Dabei sind 短语 /duǎnyǔ/ und 词组 /cízǔ/ nahezu gleichbedeutend, während 仂语 /lèyǔ/ für solche Mehrworteinheiten steht, die mindestens aus zwei Autosemantika gebildet werden.

- nominale Phrase: entspricht grammatisch dem Nomen; in der Syntax meistens an der Stelle von Subjekt oder Objekt;
- prädikative Phrase: entspricht grammatisch dem Verb oder Adjektiv; meistens an der Stelle des Prädikats;
- ergänzende Phrase: entspricht grammatisch dem Adjektiv oder Adverb; meistens an der Stelle von Attribut, Adverbial oder Komplement (vgl. Xu YC 2008: 197).

Nach der inneren Struktur der Phrase sind fünf Haupt- (vgl. Kap. 4.4.1, S. 249f) und mehrere Nebenarten zu nennen. Die zugehörigen Beziehungen zwischen den Phrasen nach der syntaktischen Funktion und der internen Struktur werden in Tab. 4-12 angegeben. Neben den fünf Hauptarten gibt es noch vier weitere Arten von Phrasen, die mit Nomen, Verb oder Adjektiv als Phrasenkopf gebildet werden: Appositions- (同位短语), Verbkombinations- (联动短语), Drehpunkt- (兼语短语) und Abkürzungsphrase (紧缩短语) (vgl. ibid.: 195, Lan 2002: 154). Funktionswörter sowie sonstige Autosemantika können ebenfalls als Phrasenkopf fungieren. Hierfür sind vier Arten zu nennen: Präpositional- (介词短语), direktionale (方位短语), Zahleinheitswort- (量词短语) und Partikelphrase (介词短语; darunter sind Partikel wie <的>, <得> und <地> /de/ besonders häufig) (vgl. ibid.: 196f, Lan 2002: 154).

	Phrasenart	**Struktur**[241]	**Beziehungen der Bestandteile (BST)**	**Beispiel**	**Wort- und Phrasenbedeutung**
Nominale Phrase (im Normalfall syntaktisch als Subjekt oder Objekt)	Subordinationsphrase (Attribut-Zentral), 偏正短语 (定心)	NP = a/v (+的) + n; NP = n/r (+的) + n	Attribut plus Beziehungswort (nominaler Kopf)	好天气 /hǎo tiānqì/; 德国人 /déguó rén/	*gut + Wetter = gutes Wetter*; *Deutschland + Person = Deutsche -r/-n*
	Koordinationsphrase (nominal), 联合短语（名词）	NP = n (+c) + n	parallele oder alternative Beziehungen der BST	语言文字 /yǔyán wénzì/	*Sprache + Schrift = Umgangs- und Schriftsprache*
	Appositionsphrase 同位短语	NP = n/r + n/r	zwei BST sind auf denselben Begriff bezogen	你自己 /nǐ zìjǐ/	*du + selbst = du selbst*
	direktionale Phrase (nominal) 方位短语 （名词）	NP = n + nd	nd hängt an n, um auf die Direktion hinzudeuten	桌上 /zhuō shàng/	*Tisch + oben = auf dem Tisch*
	Partikelphrase mit 的[de]	NP = n/r + 的	Besitzanzeige von etwas oder jemandem	我的 /wǒ de/	*ich + 的 = mein -e/-er/es*

[241] Die Abkürzungen für die Wortarten orientieren sich an GB/T 20532-2006, die in Tab. 4-9 angegeben sind; an jeder Stelle kann sowohl ein Wort als auch eine Phrase von dieser Wortart auftreten.

	Phrasenart	**Struktur[241]**	**Beziehungen der Bestandteile (BST)**	**Beispiel**	**Wort- und Phrasen-bedeutung**
	“的”字短语	NP = v/a/f + 的	Substantivierung von v/a.; Status-betonung	好吃的 /hǎochī de/	*lecker + 的 = Leckeres*
Prädikative Phrase (im Normalfall syntaktisch als Prädi-kat)	Subordinations-phrase (Adverbial-Zentral), 偏正短语 （状心）	VP = a/d + v; AP = d + a.	Adverbial plus Beziehungswort	很好 /hěn hǎo/	*sehr + gut = sehr gut*
	Koordinations-phrase (prädi-kativ), 联合短语 （谓词）	VP = v + v; AP = a + a.	parallele oder alternative Be-ziehungen der BST	读书写字 /dúshū xiězì/	*bücherlesen + zeichenschrei-ben = lesen und schreiben*
	Subjekt-Prädikat-Phrase, 主谓短语[242]	VP = n + v/a/n[243]	Subjekt plus Prädikat	天气好 /tiānqì hǎo/	*Wetter + gut = das Wetter ist gut*
	Verb-Objekt-Phrase, 述宾短语	VP = v + n	Prädikat plus Objekt	读小说 /dú xiǎoshuō/	*lesen + Roman = Roman lesen*
	Prädikat-Komplement-Phrase, 述补短语	VP = v (+得/不) + v/a/d/n; AP = a (+得) + d/v.	Prädikat plus Komplement	读懂 /dú dǒng/	*lesen + verständlich = klar lesen*
	Verb-kombinations-Phrase, 联动短语	VP = v + v	Prädikat 1 leitet Prädikat 2 ein	爱读 /ài dú/	*lieben + lesen = lesen lieben*
	Drehpunktphrase, 兼语短语	VP = v + n/r + v	n/r als Objekt von dem ersten v und als Subjekt von dem zweiten v	请他吃饭 /qǐng tā chīfàn/	*anfragen + er + essen = ihn zum Essen einladen*
	Abkürzung-sphrase, 紧缩短语	VP = c + v/a + c + v/a; AP = c + a + c + a.	Semantische Be-ziehungen der beiden Bestand-wörter, wie ko-ordinative, pro-gressive, suppo-sitive und kondi-tionale Bezie-hungen	越远越好 /yuè yuǎn yuè hǎo/	*je + weit + je + gut = je weiter desto besser*

[242] Eine Subjekt-Prädikat-Phrase kann sowohl eigenständig einen Satz als auch ein Satzglied darstellen. Als Satzglied ist sie am häufigsten Prädikat und zeigt verbale Eigenschaften. Im Vergleich zu den anderen Arten der verbalen Phrasen, deren Kopf ein Verb ist, sind ihre verbalen Eigenschaften umstritten.

[243] Das Nomen kann in speziellen Fällen direkt ein Prädikat darstellen.

	Phrasenart	**Struktur[241]**	**Beziehungen der Bestandteile (BST)**	**Beispiel**	**Wort- und Phrasen-bedeutung**
Ergänzende Phrase (im Normalfall syntaktisch als Attribut oder Adver-bial)	ZEW-Phrase, 量词短语[244]	ZEWP = m/r_{det} + q[245]	q bei m oder r_{det} anhängend, um das Attribut eines Nomens oder Adverbial eines Verbs darzustel-len	这本 /zhè běn/	*dies* + q für Buch = *dieses (Buch)*
	Präpositional-phrase, 介词短语	PP = p + n/v/r	Präposition und Objekt	在德国 /zài déguó/	*in* + *Deutsch-land* = *in Deutschland*
	Sonstige Partikelphrasen	ParP = n/v/a/d + u	Semantikum plus Partikel	老虎似的 /lǎohǔ shìde/	*Tiger* + u für wie = *wie ein Tiger*

Tab. 4-12: Die Phrasenarten im Chinesischen[246]

In der Tabelle wurden einfache Phrasen (Phrasen von einer Hierarchie) als Beispiele angegeben. Wie die Definition besagt, kann eine Phrase in mehreren Stufen gebildet sowie analysiert werden. Bspw. ist die Phrase <读这本书> (/dú zhè běn shū/, *das Buch lesen*) primär eine Verb-Objekt-Phrase. Das Objekt ist in der zweiten Stufe eine Subordinationsphrase aus <这本> (Attribut) und <书> (Bezugswort; Nomen). <这本> ist näher betrachtet eine Zahleinheits-wortphrase aus *dies* (Determinativpronomen) und dem Zahleinheitswort für *Buch*.

Zur Phrasenerkennung müssen Phrasenstrukturgrammatiken (Abk.: PSG; eng.: phrase structure grammar) vorausgesetzt sein. Aus der Spalte ‚Struktur' (Tab. 4-12) ist zu ersehen, dass es viele strukturelle Ambiguitätsfälle gibt. Bspw. kann die Struktur ‚v+n' sowohl eine NP in Subordination als auch eine VP in Verb-Objekt-Form sein. Damit die Phrasenerkennung effektiv disambiguiert wird und präzise weiter das Satzverstehen unterstützen kann, sind semantische sowie logische Erkenntnisse obligatorisch (vgl. Zhu 2000: 147, Xu ZM et al. 2000: 53). Die Regeln der Verb-Objekt-Phrase mit dem Phrasenkopf <踢> (/tī/, *kicken / treten / spielen*) sowie <打> (/dǎ/, *schlagen / eintippen / spielen* usw.) sind bspw. wie folgt.

Regel-1 (für <踢>): 踢 + mit Füßen berührtes Objekt → gültige VP
Regel-2 (für <打>): 打 + mit Händen berührtes Objekt → gültige VP

Nach Informationen zur Mikrostruktur der beiden Einträge im Wörterbuch stehen <足球> (/zúqiú/, *Fußball*) und <篮球> (/lánqiú/, *Basketball*) jeweils als für mit Füßen bzw. Händen zu spielenden Ball. So können bspw. im Satz <他爱踢足球> (/tā ài tī zúqiú/, *Er liebt das Fuß-*

[244] Eine ZEW-Phrase kann als Attribut, Komplement oder Objekt in der Syntax auftreten.
[245] r_{det} steht für Determinativpronomen, bspw.: <这> (/zhè/, *dies*), <那> (/nà/, *jenes*) und <每> (/měi/, *jedes*).
[246] Vgl. Xu YC 2008: 195-198, Lan 2002: 154, Xinhua-Lexikon Online.

ball spielen) die nachfolgenden Wörter <踢> und <足球> als eine Phrase bestimmt werden. Im Gegensatz dazu ist <打足球> (*schlagen* + *Fußball*) entweder eine falsche Phrase oder wird nur in speziellen Sprachsituationen (wie *Handspiel*) gebraucht.

Durch die PSG und semantische Kenntnisse kann überprüft werden, ob eine in Pinyin geschriebene Wortkette aus zwei oder mehreren Wörtern eine Phrase sein kann. Zudem kann dies anhand der allgemeinen syntaktischen Funktion einer Phrase bestimmt werden. Unter den vier intelligenten Konversionsmethoden (siehe Kap. 4.2.2, S. 213-217) sind die syntaktischen und semantischen Analysen einer Phrase vor allem bei der auf sprachlichem Verstehen basierenden intelligenten Konversion obligatorisch. Bei einer Pinyin-Kette funktioniert die Erkennung der Phrase nach derselben Grundfunktionsweise wie bei Schriftzeichen. Nach der Wortsegmentation einer eingegeben Pinyin-Kette werden alle betroffenen homophonetischen Wörter eines Pinyin-Wortes abgerufen. Anhand der Wortattribute (Wortart, Bedeutung usw.) und sprachlichen Regeln kann dann analysiert werden, ob zwei oder mehrere benachbarte Wörter eine Phrase bilden können, die bestimmte syntaktische Funktionen erfüllt (vgl. Wang XL 2005: 109, Chen YF/Zhu 2002b: 13f). Die Erkennung einer Phrase ist der grundlegende Prozess syntaktischer Analysen. Mit derselben Funktionstheorie können kurze Sätze auf einfacher Stufe verarbeitet werden, wie <天气好> (/tiānqì hǎo/, *das Wetter ist gut*; besteht aus einer Subjekt-Prädikat-Phrase). Wie ein relativ langer Satz auf mehreren Stufen der sprachlichen Einheiten verarbeitet werden kann, wird in Kap. 4.5.1 und Kap. 4.5.3 näher beleuchtet.

4.5 Informationsverarbeitung im Satz

In Kap. 4.3 und 4.4 wurden Vorbereitungsphasen der satzstufigen Laut-Zeichen-Konversion vorgestellt, zu denen die Silben- sowie Wortsegmentation, POS-Tagging und Phrasenbildung zählen. Die Verarbeitungsschritte vom Roh- bis zum getaggtem Korpus, die für die Regelakquisition in linguistischen Wissensdatenbanken entscheidend sind, werden auch an dieser Stelle analysiert. Um die satzstufige Konversion durchführen zu können, liegt der Schwerpunkt auf den Techniken zur Satzverarbeitung, die in den folgenden Unterkapiteln untersucht werden. Diese gliedern sich in vier Teile. In Kap. 4.5.1 werden die Arbeitsschritte der intelligenten Laut-Zeichen-Konversion konkret erläutert, um den roten Faden von den dargestellten Erkenntnissen (in Kap. 4.3 & 4.4) und den neu zu erläuternden Informationen über Satzverarbeitung fortzuführen. Kap. 4.5.2 fokussiert die Syntax, die die Wissensgrundlage zur Satzverarbeitung liefert. Anschließend werden die Methoden zur Akquisition der sprachlichen Regeln von realen Sätzen unter die Lupe genommen, die aus Korpora oder manuell eingegebenen

Sätzen stammen (Kap. 4.5.3). Die Regelanwendung zur Laut-Zeichen-Konversion wird abschließend in 4.5.4 auf Basis jener Analysen erforscht, die in Kap. 4.5.1 erfolgten.

4.5.1 Arbeitsphasen der satzstufigen Laut-Zeichen-Konversion

Wie in Kap. 4.1.3 (S. 200f) dargelegt, können als Eingabeobjekte einer intelligenten Pinyin-Eingabesoftware sowohl einzelne Zeichen, Wörter, Phrasen als auch Sätze fungieren. Eine polysyllabische Pinyin-Kette kann dementsprechend ohne Berücksichtigung des sprachlichen Kontexts sowohl ein einzelnes Wort repräsentieren, als auch einen ganzen Satz bzw. alle Einheiten dazwischen. Ist das Eingabeobjekt ein registriertes Wort, so wird das Pinyin-Wort im Konversionslexikon recherchiert und die möglichen Kandidaten ausgegeben. Wenn mehr als ein Wort eingegeben wird (z.B. eine Phrase oder ein Satz), gehört die Konversion zum Teilbereich der Satzverarbeitung (vgl. Wang XL/Wang YL 1996: 51).

Der Verarbeitungsprozess der Laut-Zeichen-Konversion im Satz kann zusammenfassend wie folgt analysiert werden. Angenommen wird, dass ein Satz S zu schreiben ist, der sich aus der Wortkette $W_1W_2\ldots W_m$ zusammensetzt. Die einzutippende Pinyin-Kette für S ist A. Nach der Silben- sowie Wortsegmentation wird A in die Pinyin-Wortfolge $A_1A_2\ldots A_m$ zerlegt. A_i steht dabei für ein beliebiges Pinyin-Wort in dieser Kette und entspricht (als Zeichen) W_i. Jedes zerlegte Element A_i kann mehreren homophonetischen Wörtern/Zeichen entsprechen, deren Kandidaten mit der Mengensammlung $\{W_{i1}, W_{i2}, \ldots, W_{in}\}$ symbolisiert werden können. Ein Zeichenkandidat des Pinyin-Elements A_i wird an der Stelle mit W_{it} ($1 \leq t \leq n$) versinnbildlicht. Der letzte entscheidende Arbeitsschritt der Pinyin-Eingabesoftware ist die automatische Ausgabe, ergo die optimale Variante der Zeichenkette $W_{1t}W_{2t}\ldots W_{mt}$ (in diesem Fall die Variante $W_1, W_2, \ldots, W_m$) (vgl. Xu ZM et al. 2000: 52). Der Verarbeitungsprozess von A zu S mithilfe einer intelligenten Pinyin-Eingabesoftware kann mit der folgenden Abbildung (Abb. 4-28) veranschaulicht werden.

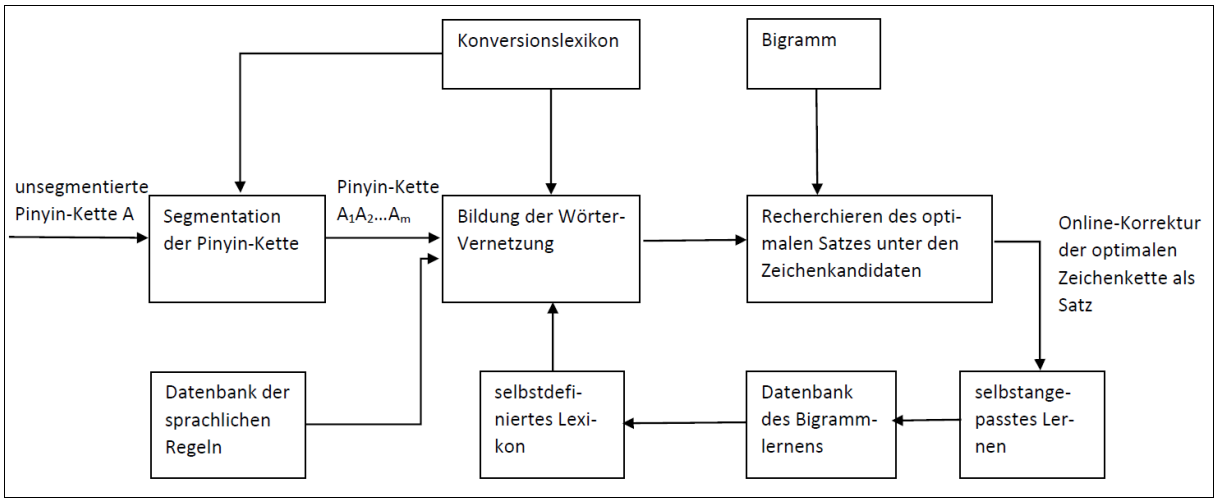


Abb. 4-28: Strukturschaubild einer verbreiteten tastaturbasierten satzstufigen intelligenten chinesischen Eingabesoftware (nach: Xu ZM et al. 2000: 54 [Übersetzung der Verfasserin])[247]

Wie dargestellt arbeiten vier Module in einer intelligenten Eingabesoftware nacheinander folgend zusammen, in denen jeweils Segmentation, Wörtervernetzung, Kandidatenauswahl und Selbstlernfunktion durchgeführt werden. Zur Analyse der Modulverarbeitung wird der Beispielsatz <他是德国人> (/tā shì déguó rén/; *Er ist Deutscher*) unter die Lupe genommen.

1) Modul der automatischen Silben- sowie Wortsegmentation (A → $A_1 \ldots A_i \ldots A_m$ [A_i = ein Pinyin-Wort; $1 \leq i \leq m$]).

Die beiden Segmentationsphasen werden meistens in einem Modul kombiniert ausgeführt. Die Silbensegmentation basiert auf der Erkennung der Silbengrenzmerkmale oder der Anwendung des Zustandsraummodells, wie in Kap. 4.3.3 (S. 232ff) vorgestellt wurde. Zur Eingabe des Beispielsatzes wird die Pinyin-Kette zuerst in fünf Silben, nämlich TA'SHI'DE'GUO'REN, zerlegt. Zur Wortsegmentation sind wie erwähnt oft vier Segmentationsverfahren für eine Eingabesoftware vonnöten: FMM (maximales Matching), FWF (Wortfrequenzauswahl der minimalen Segmentation), WWT (Wort-für-Wort-Traversierung) und SWG (Wortvernetzung) (siehe Kap. 4.4.2, S. 258-261). Nach der Verarbeitung mit dem Modul der Segmentation wird die eingetippte Pinyin-Kette in eine Folge von Pinyin-Wörtern umgewandelt. Nach den Analysen in Tab. 4-8 gibt es für dieses Beispiel acht verschiedene Segmentationsmöglichkeiten. Die Variante mit der minimalen Segmentation oder mit der höchsten Frequenz wird bevorzugt berücksichtigt.

[247] Wie in Kap. 4.2.2 vorgestellt wurde, gibt es mehrere Varianten für intelligente Laut-Zeichen-Konversionen. Dabei wird die auf dem Bigramm-Sprachmodell basierende Technik, die ein Hybrid der zweit- und drittgenannten Konversionsmethode ist, an dieser Stelle als Beispiel genommen.

2) Modul für Wörtervernetzung (A_i → {W_{i1}, …, W_{it}, …, W_{in}} und die Vernetzung von $W_{1t}…W_{it}…W_{mt}$ [W_{it} = ein Zeichenkandidat; $1 \leq t \leq n$; $1 \leq i \leq m$]).

Anhand der Recherche im Konversionslexikon werden alle Zeichen sowie Wörter (jeweils symbolisiert mit W_{it}[248]), denen jedes Element der segmentierten Pinyin-Kette $A_1…A_i… A_m$ entspricht, abgerufen. Danach wird die Vernetzung zwischen einzelnen Zeichenkandidaten gebildet, die mit $W_{1t},…W_{it},…,W_{mt}$ symbolisiert werden kann. Unter den acht Wortsegmentationsmöglichkeiten der Pinyin-Kette (siehe Tab. 4-8) wird nur die für das Ausgabeziel geeignete Variante TA/SHI/DEGUO/REN an dieser Stelle bei der Wörtervernetzungsbildung analysiert (siehe Abb. 4-29). Die Erkenntnisse über die Wortart der Kandidaten und die grundsätzlichen grammatischen Regeln können in dieser Phase auch berücksichtigt werden, um die Vernetzungsmöglichkeiten mehrfach zu reduzieren.

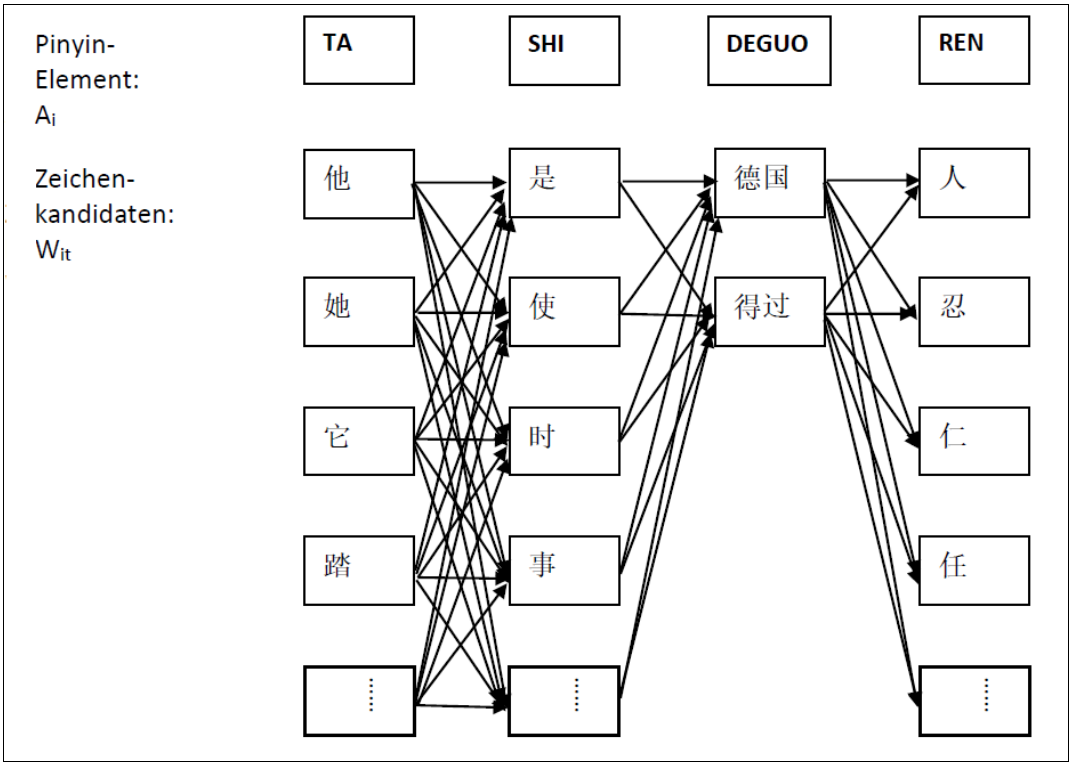


Abb. 4-29: Schaubild der maximalen Zeichenkandidatenmöglichkeiten der segmentierten Pinyin-Kette TA/SHI/DEGUO/REN[249]

3) Modul der Auswahl der optimalen Zeichenkandidaten (automatische Bestimmung von $W_{1t}…W_{it}…W_{mt}$).

In Kap. 4.2.2 (S. 213-217) wurden die vier Hauptmethoden zur Ausfilterung der Zeichenkandidaten auf Satzebene erläutert: die auf sprachlichem Verstehen und auf pragmatisch-

[248] i symbolisiert die Nummerierung der Silben und t die der homophonetischen Kandidaten; die Menge der Zeichenkandidaten für die Silbe A_1 lautet bspw. {W_{11}, W_{12}, …, W_{1n}} ($t \leq n$).
[249] Das vertikal dargestellte Auslassungszeichen steht für viele weiteren Kandidaten.

statistischen Analysen beruhenden sowie die auf Sprachmodell-Matching und Kontextzusammenhängen basierenden Methoden. Diese lassen sich wie folgt zusammenfassen:

A) Nach der auf sprachlichem Verstehen basierenden Technik ist $W_{1t}…W_{it}…W_{mt}$ grammatisch und semantisch gültig.

Hierbei werden die für den sprachlichen Kontext geeigneten Schriftzeichen durch grammatische sowie semantische Regeln bestimmt. Das heißt, dass die Wortattribute und die sprachlichen Regeln von jedem W_{it} in Abb. 4-28 im Kontext von $W_{1t}…W_{it}…W_{mt}$ analysiert werden. Die Satzgenerationsverfahren werden in Kap. 4.5.3 diskutiert.

B) Nach der auf pragmatisch-statistischen Analysen basierenden Technik ist die Wahrscheinlichkeit des Zusammenauftretens von $W_{1t}…W_{it}…W_{mt}$ am größten.

Die auf pragmatischen Statistiken basierende Technik funktioniert nach dem N-Gramm-Modell, welches die Häufigkeit des Zusammenauftretens von N benachbarten Zeichen sowie Wörtern ermittelt. An dieser Stelle wird das Bigramm-Modell als Beispiel herangezogen. Innerhalb des Beispielsatzes wird jedes Paar der benachbarten Pinyin-Wörter analysiert und statistisch errechnet. Beim Bigramm TA-SHI sind 他/她/它-是 am häufigsten (/tā shì/; *er/sie/es ist*).[250] Wenn die Zeichenkandidaten von SHI(是)-DEGUO berücksichtigt werden, ist die Variante 是-德国 (/shì déguó/; *sein* + *Deutschland*) am wahrscheinlichsten.[251] Bei der Pinyin-Kette DEGUO(德国)-REN ist die Zeichenkette 德国人 (/déguó rén/, *Deutsche -r/-n*) der wahrscheinlichste Fall.[252] Zusammenhängend ist die Vernetzungsvariante 他 / 她 / 它 是德国人 (*Er*/*Sie*/*Es ist Deutscher/Deutsche*) die statistisch denkbarste Variante. Nach weiteren statistischen und semantischen Analysen wird dann der Satz <他是德国人> (*Er ist Deutscher*) als der erste Kandidat ausgegeben.

C) Nach der auf Sprachmodell-Matching basierenden Technik gehören W_{it} und $W_{(i-j)t}$ ($1 \leq j \leq i-1$) zu einem Sprachmodell.

Diese Technik ist von Sprachmodellen abhängig, in denen zwei oder mehrere oft zusammen auftretenden Wörter als Modell-Wörter gespeichert werden. Nach den Funktionsprinzipien (siehe Abb. 4-13) kann der Verarbeitungsprozess des Sprachmodell-Matchings in dem Beispielsatz wie folgt zusammengefasst werden. Alle Sprachmodell-Wörter, die der segmentierten Pinyin-Kette TA/SHI/DEGUO/REN entsprechen, werden abgerufen. Diese sind bspw.:

[250] Im CN-Corpus Online treten die drei Varianten jeweils 1.470-, 556- bzw. 1.889-mal auf.
[251] 45 Treffer im CN-Corpus Online.
[252] 55 Treffer im CN-Corpus Online.

TA + SHI: ‚他/她/它 + 是' (*er/sie/es* + *sein*),
DEGUO + REN: ‚德国 + 人' (*Deutschland* + *Mensch -en*) und
SHI + REN: ‚是 + (Attribut) + 人' (*sein* + Attribut + *Mensch -en*).

Auf dieser Basis werden die drei Satzvarianten <他/她/它+是+德国+人> (*Er/Sie/Es ist Deutsch -e/r/en*) generiert. Zusätzlich mit semantischen Erkenntnissen analysiert kann das Wort <它> (*es*), das zudem nicht für Menschen stehen kann, ausgelassen werden. Nach Häufigkeitsanalysen hat <他> (*er*) eine höhere Frequenz als <她> (*sie*), so dass <他是德国人> (*Er ist Deutscher*) als der erste Kandidat ausgegeben wird (vgl. Wang XL 2005: 110).

D) Nach der auf Zusammenhängen im Kontext basierenden Technik ist die bedingte Wahrscheinlichkeit von W_{it} durch den Vorkontext $W_{1t}\ldots W_{(i-1)t}$ und den Nachkontext $W_{(i+1)t}\ldots W_{mt}$ am größten $[2 \leq i \leq m\text{-}1]$.

Wie in Kap. 4.2.2 erwähnt, ist HMM eine der am häufigsten eingesetzten algorithmischen Verfahren dieser Technik. An dieser Stelle wird die auf Zeichen basierende Verarbeitung erklärt, wobei die Vorphase der Wortsegmentation ausgespart werden kann. Analysiert wird der Beispielsatz S_1 (bestehend aus fünf den Zeichen $C_1C_2C_3C_4C_5$; C_i steht für einen Zeichenkandidaten der Silbe A_i), zu dessen Eingabe in der Pinyin-Kette A (bestehend aus $A_1A_2A_3A_4A_5$) TASHIDEGUOREN codiert wird. So lautet die Formel dazu:

$$S = \arg\max P(S/A)$$

Nach dieser Formel wird die Zeichenkandidatenfolge, die unter allen im Korpus auftretenden Zeichenkettenmöglichkeiten mit demselben Pinyin-Code am häufigsten ist, als Ergebnis der Ausgabe erzeugt. Diese Formel beinhaltet zwei Modelle: Sprach-Laut-Modell (die bedingte Wahrscheinlichkeit der Zeichenkette des Pinyin-Codes) und Sprach-Zeichen-Modell (die bedingte Wahrscheinlichkeit eines Zeichen/Wortes von den vorderen i-1 Zeichen oder Wörtern).

Wie in Kap. 4.2.2 erwähnt wurde, werden bei den meisten intelligenten Pinyin-Eingabesoftwares zwei oder drei der vier Laut-Zeichen-Konversionsmethoden hybrid eingesetzt, um die Qualität dieses Moduls nach Möglichkeit zu erhöhen. Wie in Abb. 4-28 zu sehen ist, gehören zum Wissensspender sowohl die vorgegebene Wissensdatenbank (inklusive des Lexikons, der Datenbank der sprachlichen Regeln usw.) als auch die dynamisch von dem Model des selbstangepassten Lernens integrierten Informationen.

4) Das Modul des selbstangepassten Lernens (manuelle Auswahl von $W_{1t}W_{2t}…W_{mt}$ und automatische Speicherung neuer sprachlicher Regeln).

Wie bereits in Kap. 4.2.2 dargelegt, kann die Wissensdatenbank unmöglich alle sprachlichen Regeln umfassen, weshalb in vielen Situationen die gewünschten Sätze nicht automatisch ausgegeben werden können. In diesem Fall muss ein Satz in zerlegten Einheiten verarbeitet und manuell ausgewählt werden. Das Modul des selbstangepassten Lernens dient dazu, neue Regeln von den bis dahin nicht zu verarbeitenden Sätzen zu lernen, in der Wissensdatenbank einzuschreiben und so perspektivisch die intelligente Laut-Zeichen-Konversion besser zu unterstützen.

Aus den vier Modulen kann man schlussfolgern, welche sonstigen Satzverarbeitungstechniken für die Erforschung der intelligenten Eingabesoftware disponiert werden müssen. Zuerst ist es wert zu fragen, wie die syntaktische Annotation auf der Basis von getaggten Korpora weiter durchgeführt wird, was eine wichtige Quelle von sprachlichen Regeln für die Wissensdatenbank ist (siehe Kap. 4.2.3, S. 221f; vgl. Luo 1996: 300). Der zweite Schwerpunkt der Satzverarbeitung liegt darin, wie die sprachlichen Regeln auf die Auswahl der Zeichenkandidaten einwirken können. Zuletzt muss analysiert werden, wie das Modul des selbstangepassten Lernens neue, nicht-eingeschriebene sprachliche Regeln lernen kann. Da das gelernte Objekt für ein Selbstlernmodul ebenso ein Satz ist, sind die Verarbeitungstechniken vergleichbar mit denen für Regelakquisition anhand der Korporaanalysen. Aus diesem Grund werden die erste und die dritte Frage zusammen in Kap. 4.5.3 betrachtet.

4.5.2 Allgemeine Eigenschaften der chinesischen Syntax

Unter den vier Modulen der intelligenten Eingabesoftware ist das dritte Modul technisch am herausforderndsten und stellt den Schwerpunkt der Laut-Zeichen-Konversion dar (siehe Kap. 4.5.1). Syntaktische Informationen sind aus dieser Sicht unentbehrlich. Einerseits sind syntaktische Regeln die Grundlage für die satzstufigen Konversionstechniken. Andererseits sind sie die Wissensgrundlage für die Annotation der Korpora, auf deren Basis die Wissensdatenbank der Sprachmodelle sowie der statistischen Kollokationswahrscheinlichkeit begründet werden. Aus diesem Grund werden in diesem Unterkapitel chinesische Sätze aus linguistischer Perspektive erforscht.

Wie in allen anderen Sprachen besteht ein Satz im Chinesischen aus Wörtern und Phrasen in bestimmten syntaktischen Strukturen.[253] Wegen des isolierenden Sprachbaus, des mor-

[253] Phrasen und Wörter derselben Wortart haben identische syntaktische Funktionen, weshalb die beiden Einheiten bei syntaktischen Problemen in diesem Kapitel gleichrangig berücksichtigt werden.

phologischen Schrifttyps und sonstigen Besonderheiten müssen die allgemeinen syntaktischen Regeln für die chinesische Sprache formuliert und ergänzt werden. Da es innerhalb eines zusammengesetzten Satzes meistens Interpunktionszeichen gibt, können solche Sätze nicht einheitlich per intelligenter Eingabemethode eingegeben werden. In dieser Arbeit werden nur einfache Sätze analysiert, die sich dadurch definieren, dass sie obligatorisch ein einzelnes Hauptprädikat, ein Wort oder eine Phrase beinhalten. Sie können zuerst in Subjekt-Prädikat- und Nicht-Subjekt-Prädikat-Sätze subklassifiziert werden. Ein Nicht-Subjekt-Prädikat-Satz ist ein Satz, der nur aus Subjekt, Prädikat oder einem einzelnen Sonderwort (etwa Interjektionen, Onomatopoesien) besteht (vgl. Lan 2002: 226-229). Die Subjekt-Prädikat-Satzstruktur gilt als Grundart und wird im Folgenden als Schwerpunkt analysiert.

Wie im sechsten Hauptunterschied des Chinesischen zu den indogermanischen Sprachen in Kap. 4.2.1 (S. 209ff) vorgestellt wurde, sind die syntaktischen Grammatiken im Allgemeinen identisch, was Wortbildungsprinzipien und allgemeine Phrasenstrukturgrammatiken (PSG) anbelangt. Die fünf Hauptarten und sonstigen Arten zur Wort- sowie Phrasenbildung werden in Kap. 4.4.1 (S. 249f) und 4.4.5 (S. 281ff) erläutert. Aus diesen fünf wird die allgemeine Reihenfolge der Satzglieder transparent: Subjekt-Prädikat, Prädikatkopf-Objekt, Prädikatkopf-Komplement, Adverbial-Prädikatkopf und Attribut-Subjekt-/Objektkopf. Die sechs genannten Satzglieder sind im Chinesischen gleichsam die Hauptarten des Satzgliedes. Darunter sind Subjekt und Prädikat jene direkten Satzglieder, die ein vollständiger Satz beinhalten muss. Der Kopf des Prädikats im Chinesischen ist meistens ein Verb/Adjektiv oder eine verbale/adjektivische Phrase. Das dem Prädikat untergeordnete Satzglied Objekt hat im Prinzip die gleichen Bildungsprinzipien wie das Subjekt, dessen Kopf im Normalfall nominal ist. Im Gegenteil zu Subjekt/Objekt und Prädikat können Adverbial, Komplement und Attribut nur den Kopf eines direkten Satzgliedes modifizieren (vgl. ibid.: 184).

Aus dieser Perspektive sind Struktur und Reihenfolge der Satzglieder im Chinesischen formal vergleichsweise unkompliziert darzustellen. Die Schwierigkeiten der syntaktischen Formeln liegen eher bei den einzelnen Satzgliedern und konkret den Fragen, welche Wortart mit welchem Satzglied zusammenhängt und wie die innerliche Struktur eines Satzgliedes aussieht. Folgende zwei Eigenschaften der chinesischen Syntax stellen und stellten die Syntaktik vor Herausforderungen.

1) Es gibt eine große Variabilität zwischen Wortart und syntaktischer Funktion und keinen Wortartwechsel eines Wortes.

In indogermanischen Sprachen kann ein Satzglied meistens von Wörtern oder Phrasen einer bestimmten Wortart dargestellt werden. Deswegen ist (bspw. im Deutschen) ein Subjekt oder

Objekt zumeist die nominale Einheit. Das Prädikat eines Satzes entspricht ebenso immer einem Verb oder einer Verbalphrase. Adjektivphrasen müssen im Prinzip ein Attribut, Prädikativum oder eine adverbiale Bestimmung sein und eine Präpositionalphrase entspricht einem Adverbial oder Prädikativum. In indogermanischen Sprachen ist die Transposition der Wortart im grammatischen Kontext möglich, die von Suffixen oder durch Flexion gekennzeichnet wird. Bspw. referieren das Verb *nutzen*, das Adjektiv *nutzbar* und das Nomen *Nutzung* auf denselben Wortstamm (vgl. Bußmann 2002: 713).

Im Gegensatz zu den indogermanischen Sprachen kann eine Wortart im Chinesischen – insbesondere Nomen, Verben und Adjektive – multiple syntaktische Funktionen haben. Anders formuliert gibt es im Chinesischen keinen Wortartwechsel, sondern eine Wortart hat mehrere Möglichkeiten als Satzglied zu fungieren. Die vielfältigen syntaktischen Funktionen einer Wortart unterscheiden sich in Haupt- sowie Nebenfunktionen. Bspw. sind die Hauptfunktionen des Nomens sowie der Nominalphrase Subjekt und Objekt. Außerdem kann ein Nomen bzw. eine NP in bestimmten Situationen als Prädikat auftreten, was als eine Nebenfunktion der Wortart ausgelegt werden kann (vgl. Xu YC 2008: 173; zu syntaktischen Funktionen einer Wortart vgl. Tab. 4-9, Spalte ‚Erklärung'). Eine Wortart kann in verschiedenen Unterarten auch unterschiedliche syntaktische Funktionen erfüllen, woraus die Unterscheidung der ersten sowie zweiten Klasse des Taggingcodes in Tab. 4-9 resultiert. Bspw. können zeitliches, lokales sowie direktionales Nomen (Code: nt/nl/nd) das Adverbial eines Satzes sein, was im Normalfall nicht für andere Unterarten des Nomens gilt. Die Beziehungen zwischen Wortart und Satzglied können anhand der folgenden Tabelle beschrieben werden.

Verb & Adjektiv	Nomen	Unterscheidungswort	Adverb
Komplement			
Prädikat	Prädikat*		
Subjekt & Objekt	Subjekt & Objekt		
Attribut	Attribut	Attribut	
Adverbial	Adverbial**	Adverbial	Adverbial
* Das nominale Prädikat gilt nur für begrenzte Fälle; es kann zudem als Weglassung des Kopulaverbs <是> (*sein*) betrachtet werden. ** Normalerweise kann im modernen Chinesisch nur eine begrenzte Zahl von Nomen als Adverbial auftreten, nämlich die den Kategorien Zeit, Lokal sowie Direktion zugehörigen.			

Tab. 4-13: Die Wortart-Satzglied-Beziehungen im Chinesischen (nach: Lu 2008: 112; ergänzt von der Verfasserin durch Zhao 1992: 73-103)

2) Es gibt keine übereinstimmende Subjekt-Prädikat-Beziehung; Prädikatkopf entspricht nicht immer dem Verb.

In den indogermanischen Sprachen herrscht eine reziproke Subjekt-Prädikat-Übereinstimmung vor, weshalb die Auswahl der Wörter, die vom Prädikat redigiert werden (Valenz des Verbs), und des Subjekts (Numerus und Person) voneinander abhängig sind (vgl. Bußmann

2002: 662). In Satzanalysen ist es bei solchen Sprachen daher möglich, die Form des Prädikats zu bestimmen, wenn das Subjekt bekannt ist, bzw. umgekehrt.[254]

Im Vergleich zu den indogermanischen Sprachen fehlt im Chinesischen die grammatische Relation zwischen Subjekt und Prädikat. Des Weiteren gibt es mehr als eine Wortartmöglichkeit für Prädikate. Grammatisch gesehen können Verb, Adjektiv, Nomen und die meisten Arten von Phrasen als Prädikat auftreten. Treten in einem Satz alle sechs Satzglieder auf, kann die generelle syntaktische Struktur mit der folgenden Abbildung dargestellt werden (die Reihenfolge der Satzglieder im Satz ist im Normalfall von links nach rechts).

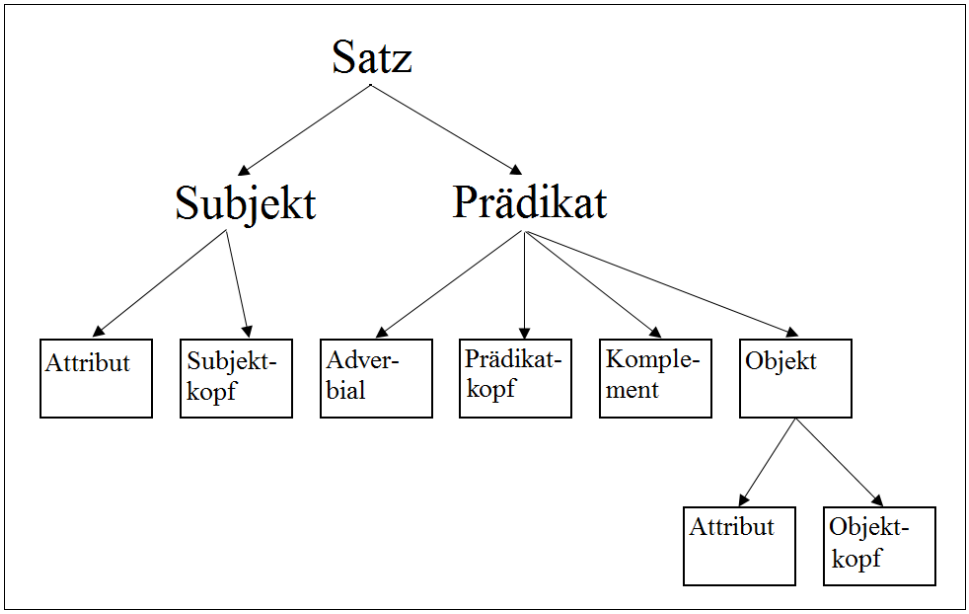


Abb. 4-30: Die allgemeine syntaktische Struktur bei Sätzen mit sechs Satzgliedern

Im Gegensatz zu den indogermanischen Sprachen (wie Deutsch) sind im Chinesischen verschiedene Satztypen (Aussage-, Frage- und Aufforderungssatz) nicht durch Umstellung eines Satzgliedes gekennzeichnet. Das heißt, dass die allgemeine syntaktische Struktur (vgl. Abb. 4-30) für alle drei Satztypen gültig ist. Soll bspw. eine Entscheidungsfrage anhand des Aussagesatzes <我爱你> (/wǒ ài nǐ/, *Ich liebe dich*) (um)formuliert werden, so wird das Fragepartikel <吗> /ma/ am Satzende hinzugefügt: <我爱你吗> ‚*Liebe ich dich*‘. Wenn eine Ergänzungsfrage formuliert werden soll, muss das entsprechende Fragepronomen an der entsprechenden Stelle ersetzt werden. So bedeutet <谁爱你> ‚*Wer liebt dich*‘ (das Fragepronomen ist das Subjekt am Satzanfang), hingegen bedeutet <你爱谁？> ‚*Wen liebst du*‘ (das Fragepronomen ist das Objekt am Satzende). Ähnlich wie beim Fragesatz wird ein Aufforderungssatz ebenfalls

[254] Intelligente Handy-Eingabesoftwares solcher Sprachen können anhand von Subjekt-Prädikat-Übereinstimmungen automatisch das Folgewort antizipieren. Gibt man im deutschen *SwiftKey* bspw. das Wort *bin* ein, so stellt die Software automatisch die Kollokation *ich* zur Verfügung (nach eigenem Eingabetest mit der *SwiftKey*-Eingabemethode für Deutsch).

mit zusätzlichen Partikeln, der Prosodie beim Sprechen und einem Ausrufezeichen in der Schriftsprache markiert (vgl. Xu YC 2008: 229f). Die relativ stabile Satzgliedanordnung bedingt, dass die computergestützte Satzgliederkennung relativ simpel durchgeführt werden kann, obwohl es keine eindeutigen Wortart-Satzglied-Beziehungen im Chinesischen gibt. Welche Wortarten an der Stelle eines Satzgliedes auftreten können, wird in der folgenden Tabelle dargestellt. Analysiert wird der Satz: <勤奋的我们很快地翻译完了这篇文章> (/qínfèn de wǒmén hěn kuài de fānyì wánle zhè piān wénzhāng/; *Wir, die Fleißigen, haben die Übersetzung des Artikels sehr schnell geschafft*).

Satzglied	**Position**	Wortart im **allgemeinen Fall**[255]	Wortart im **Sonderfall**	**Phrasenart**	Im **Beispielsatz**; Worterklärung von markierten Stellen[256]
Subjekt	Satzanfang, vor dem Prädikat	Sub. = n/r	Sub. = v/a/m	Sub. = NP/ VP/AP/ ZewP	勤奋的<u>我们</u> ‖ 很快地翻译完了这篇文章; Subjektkopf: r- <u>*wir*</u>
Prädikat	Nach dem Subjekt	Prä. = v/a	Prä. = n/r	Prä. = VP/AP/ NP/ PP/ZewP/ BkP	勤奋的我们 ‖ 很快地<u>翻译</u>完了这篇文章; Prädikatkopf: v- <u>*übersetzen*</u>
Objekt	Nach dem Prädikat-Kopf, dem Prädikat zugehörig	Obj. = n/r	Obj. = v/a/m	Obj. = NP/ VP/AP/ ZewP	很快地翻译完了 | 这篇<u>文章</u>; Objektkopf: n- <u>*Artikel*</u>
Adverbial	Vor dem Prädikatkopf, kann in vielen Fällen auch am Satzanfang stehen	Adv. = d/a/vu (+地)	Adv. = r/n (nt/nl/nd)/	Adv. = PP/ AP/VP/ NP	[很快地] <u>翻译</u>完了; Adverbial: d+a+u [*sehr schnell*]
Komplement	Nach dem Prädikatkopf	Kom. = (得+) v/a	Kom. = d	Kom. = VP/AP/ PP/ZewP	很快地<u>翻译</u> <完了>; Komplement: v- <*fertig*>
Attribut	Vor dem Subjekt- sowie Objektkopf	Att. = n/v/a/r (+的)		Att. = NP/ VP/AP/ PP/ZewP	(勤奋的) <u>我们</u>; Subjekt-Attribut: a- (*fleißig*) <u>*wir*</u> + (这篇)<u>文章</u> Objekt-Attribut: r+q - (*das Stück*)

Tab. 4-14: Struktur und Erklärung jedes Satzglieds mit Beispiel[257]

[255] Die Wortart im Allgemeinen sowie in Sonderfällen bezieht sich einerseits auf die Wortart einzelner Wörter, falls das Satzglied aus einem Wort besteht. Andererseits geht es auch um die Wortart des Kopfes des Satzgliedes.

[256] Zeichen zur Markierung der Satzglieder: Subjekt und Prädikat werden mit Doppelstrich ‖ getrennt. Ein Einzelstrich | zeigt die Grenze zwischen Prädikatkopf und Objekt. Das Adverbial wird mit eckiger, das Komplement mit spitzer und das Attribut mit runder Klammer markiert. Der Kopf eines direkten Satzgliedes wird via Unterstrich gekennzeichnet. Nur der betroffene und die mit ihm zusammenhängenden Satzteile werden in der Spalte angegeben. Bei der Worterklärung der markierten Stellen steht zunächst die Wortart der jeweiligen Wörter, dann die entsprechende Bedeutung im Deutschen.

[257] Zu Grundlagen der chinesischen Syntax vgl. z.B. auch Zhao 1992: 73-103 und Lan 2002: 184-197.

4.5.3 Syntaktische Annotation und Akquisition sprachlicher Regeln für verschiedene Arten von Wissensdatenbanken

In Kap. 4.2.3 und 4.5.1 wurden die Zusammenhänge zwischen den Korporaanalysen und der Begründung der Wissensdatenbank vorgestellt. Durch Korporaanalysen können realistische Beispielsätze als empirische Wissensgrundlage gesammelt und zur intelligenten Konversion verwendet werden. D.h. die sprachlichen Regeln zur Satzbildung können von annotierten realistischen Sätzen nach grammatischen, semantischen und statistischen Aspekten interpretiert werden, so dass die sprachlichen Regeln in der Wissensdatenbank möglichst umfangreich und vielfältig sind. In Kap. 4.2.3 (S. 221f) wurden die drei wichtigsten Verarbeitungsschritte eines Rohkorpus zur Akquisition der sprachlichen Regeln aufgezeigt: Wortsegmentation (siehe Kap. 4.4.2 & Kap. 4.4.4), POS-Tagging (siehe Kap. 4.4.3 & Kap. 4.4.4) und die syntaktische Annotation. Mit Phrasenstrukturgrammatiken beschäftigt sich Kap. 4.4.5. Die allgemeine Reihenfolge der Satzglieder und Satzformeln, auf der die syntaktische Annotation basiert, wurde im voranstehenden Kapitel 4.5.2 ausgeführt.

Wenn ein Satz syntaktisch annotiert werden soll, kann an ihm zunächst eine Top-down-Analyse durchgeführt werden. Der Satz wird dabei zuerst in Subjekt und Prädikat geteilt, bevor der Kopf der beiden Elemente und sonstige Satzglieder bestimmt werden. Für letztere werden Bottom-up-Analysen eingesetzt, um Phrasen niedrigeren bis höheren Niveaus Schritt für Schritt zu markieren (vgl. Chen XH/Cui 1996: 311f). Dependenzgrammatisch wird im Beispielsatz von Abb. 4-31 der Satzteil <这篇文章> (/zhè piān wénzhāng/, *dieser Artikel*) mit einer Top-down-Untersuchung als Objekt erkannt. Danach wird die Satzannotation durch Bottom-up-Untersuchungen anhand der Wortart jedes bestehenden Wortes als Kombination aus einer Zahleinheitsphrase (Determinativpronomen <这> + Zahleinheitswort <篇>) und einem Nomen <文章> verstanden.

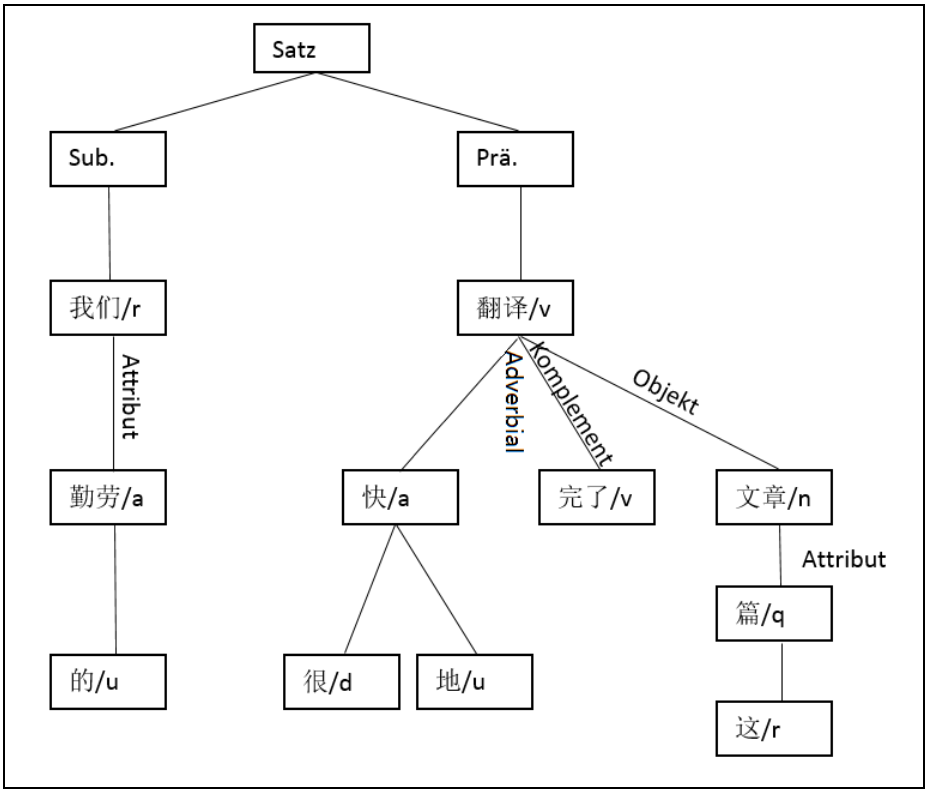


Abb. 4-31: Die Dependenzgrammatik vom Beispielsatz , (勤劳的) 我们 || [很快地] 翻译 <完了> | (这篇) 文章‘

Die größten Schwierigkeiten bei der Erkennung des Satzgliedes mit Top-down-Analysen entstehen (wie in Kap. 4.5.2 angerissen) einerseits durch die multiplen Funktionen der Wortarten für Satzglieder, andererseits durch die vielfältigen Möglichkeiten zur Bildung eines Prädikats.

Die Satzgliederkennung basiert auf der Stellung und der Wortart jedes Wortes im Satz, wie in Tab. 4-14 angegeben wird. Neben den zwei Hauptfaktoren können Adverbial, Komplement und Attribut in vielen Fällen auch mithilfe von Modalpartikeln erkannt werden. Das Modalpartikel <的> ist bspw. die Kennzeichnung zwischen Attribut und Subjekt-/Objektkopf, während <得> Komplement und <地> Adverbial markiert.[258] Im Beispielsatz von Tab. 4-14 sowie Abb. 4-31 kann nach diesen dargestellten Erkenntnissen nun problemlos die Subjekt-Prädikat-Grenze nach <我们> (/wǒmen/, wir) erkannt werden. Bei weiteren Satzgliedannotationen steht die Erkennung des Prädikatkopfs im Fokus, was aufgrund der beiden nachfolgenden Verben <翻译> (/fānyì/, *übersetzen*) und <完了> (/wánle/, *fertig sein*) erschwert wird. Im Prinzip und ohne Berücksichtigung von Semantik und anderen Satzgliedern können zwei aufeinanderfolgende Verben drei Varianten von Beziehungen eingehen: Koordination, Verbkombination und Komplement. Wegen der Intransivität von <完了> kann das Verb kein Objekt

[258] <的>, <得> und <地> werden als Modalpartikel immer /de/ ausgesprochen.

einleiten, wohingegen die Kombination von <翻译> mit dem Objekt (für *Artikel*) nach allgemeinen Sprachkenntnissen sowohl grammatisch als auch semantisch gültig ist. Das Verb <完了> gehört auch zu der als Komplement verwendbaren Verbliste, die das Perfekt eines Verhaltens ausdrückt. Demgemäß wird <翻译> als Prädikatkopf und <完了> als Verbkomplement annotiert.

Wie bei der zweiten Eigenschaft der chinesischen Syntax vorgestellt wurde (Kap. 4.5.2, S. 292f), sind die strukturellen Varianten des Prädikats vielfältig, besonders wenn ein Verb als Prädikatkopf gebraucht wird. Damit die Satzformel ausführlich dargestellt werden kann, werden die Möglichkeiten des Prädikats in der folgenden Tabelle zusammengefasst. Um diese Formeln möglichst klar und deutlich zu beschreiben, werden in der Tabelle nur Subjekt, Prädikatkopf und Objekt berücksichtigt. Alle Satzformeln können aber anhand von Tab. 4-14 um Adverbial, Komplement und Attribut erweitert werden.

Wortart des Prädikats	**Unterart** des Falls	**Grundformel**	**Beispiel**	**Erklärung** des Beispiels	**Anmerkung** zu der Satzformel
Nomen/ Pronomen/ NP		S → Sub. + n/r/NP	明天春节。 /míngtiān chūnjié/; *Morgen ist Frühlingsfest.*	明天/n ‖ 春节/n	Dieser Fall gilt nur bei begrenzten Nomen, Pronomen und NP in bestimmten Situationen.
Verb/ VP	Ohne Objekt	S → Sub. + v/VP	他起床了。 /tā qǐchuáng le/; *Er ist aufgestanden.*	他/r ‖ 起床 /vi 了/u	Gilt bei fast allen Verben (im Fall der Objektweglassung) und verbalen Phrasen.
	Mit einem Objekt	S → Sub. + vt + Obj.	他读小说。 /tā dú xiǎoshuō/; *Er liest Romane.*	他/r ‖ 读/vt | 小说/n.	Gilt bei vt, vd und manchen vu.
	Mit doppelten Objekten	S → Sub. + vt + Obj. (*jn.*) + Obj. (*etw.*)	我借他小说。 /wǒ jiè tā xiǎoshuō/; *Ich leihe ihm Romane.*	我/r ‖ 借/vt | 他/r-obj.1 | 小说/n-obj.2	Gilt nur bei begrenzten Wörtern von vt.
	Kopulaverb	S → Sub. + vl + Obj.	他是德国人。 /tā shì déguó rén/; *Er ist Deutscher.*	他/r ‖ 是/vl | 德国/n 人/n	Gilt nur bei vl. In dieser Satzformel hat das Objekt eine semantische Entsprechung zu dem Subjekt.
	Modalverb	S → Sub. + vu + v/a/VP/AP	他要读小说。 /tā yào dú xiǎoshuō/; *Er will Romane lesen.*	他/r ‖ [要 /vu] 读/vt | 小说/n	vu wird als Adverbial angesehen und nach ihm tritt ein Verb als Prädikatkopf auf.

Wortart des Prädikats	**Unterart** des Falls	**Grundformel**	**Beispiel**	**Erklärung** des Beispiels	**Anmerkung** zu der Satzformel
	Satz mit Bǎ- (把) sowie Bèi- (被)	S → Sub. + 把/被 + $Obj._1$ + vt/VP	电脑被我买了。/diànnǎo bèi wǒ mǎi le/; Der *Computer wurde von mir gekauft.*	电脑/n ‖ [被/p 我/r] 买/vt 了/u	Charakteristika beider Satzformeln sind die Reihenfolge-Umwandlung von Subjekt, Prädikatkopf und Objekt. Der Bǎ-Satz ist Aktiv und der Bèi-Satz Passiv. Im Bǎ-Satz ist das Objekt von Bǎ das Patiens des Prädikatkopfs. Im Bèi-Satz ist das formelle Subjekt das Patiens vom Prädikatkopf, während das Objekt von Bèi das Agens ist.
	Mit einer Drehpunkt-Phrase	S → Sub. + v + Obj. + v/a/VP/AP	我请他吃饭。/wǒqǐng tā chīfàn/; *Ich lade ihn zum Essen ein.*	我/r ‖ 请/v 他/r 吃饭/v	Diese Satzformel gilt nur bei begrenzten Verben als Prädikatkopf. Das Objekt von dem ersten Verb ist gleichzeitig auch das Subjekt von dem/der zweiten prädikativen Wort/Phrase.
	Verb-Kombination	S → Sub. + v_1/VP_1 + v_2/VP_2	他爱读小说。/tā ài dú xiǎoshuō/; *Er liebt es Romane zu lesen.*	他/n ‖ 爱/v 读/v | 小说/n	1. v_2/VP_2 ist der Zweck von v_1/VP_1; 2. v_1/VP_1 ist die Art und Weise von v_2/VP_2; 3. v_1/VP_1 ist die Ursache oder Voraussetzung für v_2/VP_2.
	Subjekt-Prädikat-Phrase	S → Sub. + $Sub._1$ + $Prä._1$	今天天气好。/jīntiān tiānqì hǎo/; *Das Wetter heute ist gut.*	今天/n ‖ 天气/n 好/a	$Sub._1$ kann sich meist auf Sub. beziehen.
Adjektiv/ AP		S → Sub. + a/AP	天气好。/tiānqì hǎo/; *Das Wetter ist gut.*	天气/n ‖ 好/a	Anders als in den indogermanischen Sprachen ist ein Kopulaverb nicht obligatorisch.
Sonstige Phrasen	Präpositionalphrase	S → Sub. +在 + Obj.	我在北京。/wǒ zài běijīng/; *Ich bin in Beijing.*	我/r ‖ 在/p 北京/ns	在 ist ein häufiges Partikel. Mit ihm gebildete Phrasen kann als Adverbial, Komplement, Attribut oder Prädikat verwendet werden. Sein Objekt in der Satzformel kann sowohl einen Ort als auch eine Aktion ausdrücken (Verlaufsform).
	Zahleinheitsphrase	S → Sub. + ZewP	他二十岁。/tā érshí suì/; *Er ist 20 Jahre alt.*	他/r ‖ 二十/m 岁/q	Gilt bei fast allen Zahleinheitswörtern.

Wortart des Prädikats	**Unterart** des Falls	**Grundformel**	**Beispiel**	**Erklärung** des Beispiels	**Anmerkung** zu der Satzformel
	Bikuang-Phrase	S → Sub. + BkP	他老虎似的。 /tā lǎohǔ shide/; *Er ist wie ein Tiger.*	他/r ‖ 老虎/n 似的/u	Bikuang ist eine Partikelart mit einigen bestimmten Wörtern, die die Bedeutung *wie etwas* übertragen.

Tab. 4-15: Die Prädikatvarianten im Chinesischen und ihre Grundsatzformeln[259]

Nach Erkennung der Satzglieder durch Top-down- und der gestuften Phrasenbestimmung nach Bottom-up-Analysen können des Weiteren die sprachlichen Regeln zu verschiedenen Zwecken akquiriert werden. Wörter oder Phrasen, die im Satz abhängig auftreten, können zu einer sprachlichen Regel zusammengefasst werden (vgl. Chen XH/Cui 1996: 304). Wie in Kap. 4.2.2 und Kap. 4.2.3 erwähnt, gibt es verschiedene Arten linguistischer Wissensdatenbanken, die zu unterschiedlichem Zweck gebraucht werden. Wie in den Schaubildern der Konversionsmethoden (vgl. Abb. 4-12, 4-13 & 4-14) gezeigt, gibt es hauptsächlich drei Arten: die Datenbank von Grammatik und Semantik, die von pragmatischen Statistiken und die von Sprachmodellen. An dieser Stelle wird zuerst die grammatisch-semantische Datenbank berücksichtigt. Die sprachlichen Regeln dieser Datenbank orientieren sich an den grammatischen sowie semantischen Zusammenhängen von zwei oder mehr Wörtern sowie Phrasen. Zieht man den in Tab. 4-14 verwendeten Beispielsatz heran (vgl. Abb. 4-31), können folgende sechs sprachliche Regeln begründet werden:

R1. Subjektkopf – Prädikatkopf: 我们/$_{r}$ + 翻译/$_{v}$ (WOMEN + FANYI; *wir* + *übersetzen*);
R2. Prädikatkopf – Objektkopf: 翻译/$_{v}$ + 文章/$_{n}$ (FANYI + WENZHANG; *übersetzen* + *Artikel*);
R3. Attribut – Kopf im Subjekt: (勤劳/$_{a}$ 的)$_{AP}$ + 我们/$_{r}$ (QINLAO DE + WOMEN; *fleißig* + *wir*);
R4. Adverbial – Prädikatkopf: [很快/$_{a}$ 地]$_{AP}$ + 翻译/$_{v}$ (HENKUAI DE + FANYI; *schnell* + *übersetzen*);
R5. Prädikatkopf – Komplement: 翻译/$_{v}$ + 完了/$_{v}$ (FANYI + WANLE; *übersetzen* + *fertig*);
R6. Attribut – Kopf im Objekt: (这篇/$_{q}$)$_{ZewP}$ +文章/$_{n}$ (ZHE PIAN + WENZHANG; *Stück* + *Artikel*).

Die dritte, vierte und sechste Regel behandeln die Beziehungen zwischen einer Phrase und einem Wort. In jeder der drei Phrasen herrscht ein Phrasenkopf (das Wort mit Unterstrich), der

[259] Zu Grundlagen der Prädikatvariante vgl. auch: Lan 2002: 230-241, Zhao 1992: 107-136.

direkt mit dem anderen Wort der sprachlichen Regel zusammenhängt. Derlei sprachliche Regeln können daher auf die Beziehungen zwischen Phrasenköpfen reduziert und bei weiteren Verwendungen ohne Berücksichtigung der modifizierten Teile verwendet werden. Im Prinzip können die sprachlichen Regeln mit dem semantischen Netz oder der paradigmatischen Wörterliste kombiniert werden, wobei jedes Wort derselben Kategorie ein Wort als Variable in einer sprachlichen Regel ersetzen kann (vgl. Wang XL et al. 1993: 373, Lobin 2010: 83f). Da es im Chinesischen keine Flexion gibt, ist mit einer paradigmatischen Wortliste meistens ein semantisches Paradigma gemeint. Z.B. sind bei der ersten Regel alle Nomen und Pronomen für Personalbezeichnungen und Eigennamen (nh) austauschbar, um neue, auf sie bezogene Regeln zu gewinnen. Aus R1 kann deswegen ‚R[1-var.]: *Jemand*/$_{n/r/nh}$ + 翻译/$_{v}$' geschlussfolgert werden. Nach selbigem Prinzip kann R2 auch in ‚R[2-var.]: 翻译/$_{v}$ + Nomen für Schriftwerek/$_{n}$' umgeschrieben werden.

Die sprachlichen Regeln bestimmter Wörter sind Regeln mit Konstanten, während die Regeln über unbestimmte Wörter aus einer Wortliste als Regeln mit Variablen angesehen werden können. Regeln mit Variablen sind ungewiss, weil ihre Korrektheit nicht durch realistische Beispielsätze im Korpus bestätigt wird. Derlei ungewisse Regeln werden deswegen nicht in einer voreingestellten Datenbank, sondern in einer Testdatenbank gespeichert, deren Anwendung unter Beschränkungen abläuft. Nachdem ein variantenreiches Wort vom individuellen PC-Benutzer einer sprachlichen Regel zugeführt bzw. diese manuell bestätigt wurde, wird es offiziell in die Wissensdatenbank aufgenommen (vgl. Wang XL et al. 1993: 373).

Von dem mit syntaktischen Funktionen annotierten Korpus können auch die Regeln aus Sicht der Pragmatik sowie des Sprachmodells gewonnen werden. Zusammengefasst hat ein annotiertes Trainingskorpus für die Begründung der linguistischen Wissensdatenbanken hauptsächlich folgende Möglichkeiten.

1. Die Wahrscheinlichkeit einer sprachlichen Regel:

Solche Regeln gehören zu der Datenbank der pragmatischen Statistiken oder dienen als Attribut einer grammatischen sowie semantischen Regel. Die statistischen Analysen gelten sowohl bei den Regeln mit Konstanten, als auch bei jenen mit Variablen. Auf Statistiken basierende Sprachmodell-Formeln wie das N-Gramm- und Hidden Markov Model (HMM) sowie die Maximum-Entropie-Methode können dazu angewendet werden (vgl. Wang XL 2005: 47). Wenn zu einer Pinyin-Kette mehrere sprachliche Regeln passen, werden die Regeln von der größten bis zur kleinsten Wahrscheinlichkeit überprüft.

2. Die bedingte Wahrscheinlichkeit von zwei Wörtern sowie Phrasen:

Das errechnete Wort- sowie Phrasenpaar betrifft nicht nur den benachbarten, sondern auch die in einem Satz grammatisch zusammenhängenden Einheiten (vgl. Chen XH/Cui 1996: 304). Bspw. sind die beiden in Verb-Objekt-Zusammenhang stehenden Wörter <翻译> *übersetzen* und <文章> *Artikel* jene Wörter, zwischen denen die Satzglieder Komplement (des Prädikats) oder Attribut (des Objekts) auftreten können. Mit der Formel des Bayes-Theorems kann die Wahrscheinlichkeit von A (ein Wort oder eine Phrase) unter der Bedingung B ausgerechnet werden. Wenn der Wert der bedingten Wahrscheinlichkeit den minimalen Wortpaarwert erreicht, wird ‚A + B' als eine festgelegte Kollokation anerkannt und als ein Sprachmodell in der Datenbank mit seinem Wahrscheinlichkeitsattribut eingeschrieben. Solche Wortpaare sind sozusagen die Wissensunterstützung von der auf dem Sprachmodell-Matching basierenden intelligenten Konversionstechnik (vgl. Tsai 2005: 11f, Chen YF/Zhu 2002b: 16).

3. Die Häufigkeit eines Wortes in bestimmten Satzstellungen:

Die Formel für Wortbildungsmodelle eines Zeichens (vgl. Kap. 4.4.4, S. 277) kann für den Algorithmus des Satzbildungsmodells eines Wortes abgeleitet werden. Mit dieser Methode kann ausgerechnet werden, welchem Satzglied Wörter – die z.B. multi-funktionalen Wortarten angehören – am wahrscheinlichsten zugehörig sind. Bspw. kann ein Verb bei der Satzbildung die syntaktische Funktion eines Prädikatkopfs, Attributs, Komplements, Subjekts/Objekts und Adverbiales übernehmen, wobei verschiedene Verben sprachpragmatisch unterschiedlich eingesetzt werden. Anhand dieser statistischen Analysen kann ermittelt werden, welche Verben dominant als Prädikatkopf, häufiger als Adverbial, Komplement usw. verwendet werden. Das Verb <完了> (/wánle/; *fertig*) im zuvor genannten Beispielsatz wird so z.B. am häufigsten als Komplement gebraucht, das hinter dem Prädikatkopf-Verb auftritt. Davon abgeleitet ergibt sich die Regel ‚Verb + 完了'. Solche Ergebnisse können einerseits die automatische syntaktische Annotation besser unterstützen und andererseits die Korrektheit sprachlicher Regeln mit Variablen erhöhen. Das dialektale nordostchinesische Verb <嘚瑟> (/dèse/; *sich aufspielen* [ironisch markiert]) z.B. findet in standardsprachlichen Schriftwerken selten Gebrauch, weshalb es in der Wissensdatenbank wenige sprachliche Regeln gibt. Mit dieser Art der statistischen Rechnung und den Variable-Regeln kann der korrekte Satz <他嘚瑟完了> (/tā dèse wánle/, *Er hat das Aufspielen hinter sich.*) in vielen Fällen trotzdem anhand seines Pinyin-Codes erzeugt werden.

Im vorderen Text wurden die Methoden zur Akquisition der sprachlichen Regeln mittels annotierter Korpora vorgestellt. Wie skizziert, läuft das Selbstlernmodul einer intelligenten

Eingabesoftware auf identische Art und Weise (hierzu vgl. Kap. 4.5.1, S. 290). Das Modul startet, wenn ein zuvor nicht zu verarbeitender Satz manuell eingegeben wird. Der allgemeine Verarbeitungsprozess ist gleich wie bei Sätzen aus Korpora – Wortsegmentation, POS-Tagging, syntaktische Top-down-Analyse, Bottom-up-Analyse nach PSG und zuletzt die Annotation mit der syntaktischen Funktion bei jedem Wort. Aus einem annotierten Satz können sprachliche Regeln grammatischer, semantischer sowie statistischer Natur zielorientiert gewonnen werden. Solche neuen Regeln werden dann in der Wissensdatenbank mit aktualisierter Wahrscheinlichkeit eingeschrieben (vgl. Wang XL et al. 1993: 371).

4.5.4 Anwendung der sprachlichen Regeln zur Laut-Zeichen-Konversion

Wie im letzten Kapitel erwähnt wurde, gibt es Regeln mit Konstanten (bei bestimmten Wörtern) und mit Variablen (bei unbestimmten Wörtern aus bestimmten Kategorien). Variablen-Regeln lassen sich näher in grammatische und semantische Regeln unterteilen. Die grammatischen Regeln werden meistens manuell beschrieben und als Vorlage für Annotationen sowie vage Satzausgaben bei der intelligenten Eingabesoftware gebraucht. Sie beziehen sich auf die Satz- sowie Phrasenformel der Wortarten der Bestandteile (vgl. Tab. 4-12, Tab. 4-14 und Tab. 4-15). Die semantischen Regeln meinen die Darstellung der Beziehungen von zwei oder mehreren Wörtern aus bestimmten semantischen Kategorien, bspw. ‚*jemand* + menschliches Verhalten'. Sie können sowohl auf Basis von grammatischen Regeln manuell umgeschrieben als auch anhand der Konstanten-Regeln abgeleitet werden. Im Allgemeinen sind die Regeln über bestimmte Wörter (K) eindeutiger als die semantischen Regeln (S). Die semantischen Regeln sind wiederum präziser als die rein grammatischen Regeln (G) (vgl. Wang XL et al. 1993: 374). So lässt sich die Zuverlässigkeit dieser drei Regeln wie folgt hierarchisieren: ‚我们/$_{r}$ + 翻译/$_{v}$' [*wir* + *übersetzen*] > ‚*jemand* + menschliches Verhalten' > ‚Pronomen + Verb'.

Anhand der vorhandenen sprachlichen Regeln aller drei Typen soll exemplarisch der Satz <他今天翻译了两篇文章> (/tā jīntiān fānyì le liǎng piān wénzhāng/; *Er hat heute zwei Artikel übersetzt*) mit intelligenten Pinyin-Eingabemethoden eingegeben werden. Wie in Kap. 4.5.1 gezeigt, werden nach der Silben- sowie Wortsegmentation der eingetippten Pinyin-Kette die entsprechenden homophonetischen Wörter und angepassten sprachlichen Regeln im Kontext abgerufen und überprüft. Die Verwendung der sprachlichen Regeln bei der Auswahl der homophonetischen Wörter wird in der nachstehenden Abbildung skizziert.

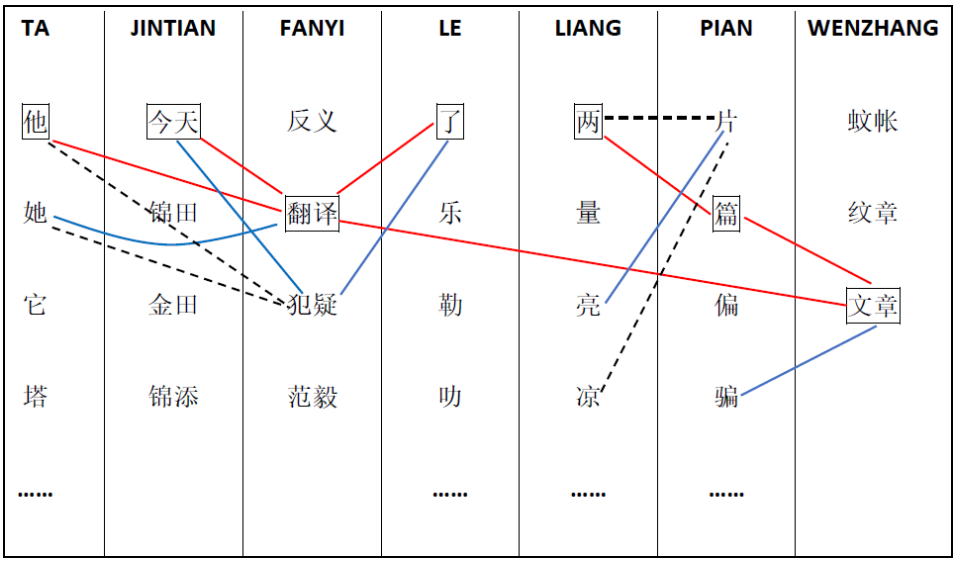


Abb. 4-32: Die Anwendung der sprachlichen Regeln zur Laut-Zeichen-Konversion von ‚TA/ JINTIAN/ FANYI/ LE/ LIANG/ PIAN/ WENZHANG'

Eine Variante von R1 (他/r + 翻译/v [*er* + *übersetzen*]) sowie R2 und R6 werden aus dem Beispielsatz gewonnen. Die für die Pinyin-Kette anwendbaren Regeln werden in dieser Abbildung durch Linien symbolisiert und in Tab. 4-16 erklärt. Korrekte Zeichenkandidaten werden umrahmt, die verwendbaren sprachlichen Regeln in derselben Tabelle angegeben, wobei die rote Linie die erste und die blaue die zweite Regel von den Pinyin-Wort-Regeln symbolisiert. Weitere Regeln werden mit schwarz-punktierten Linien dargestellt.

Regeln über Pinyin-Wort	**Regel Eins (Rot)**			**Regel Zwei (Blau)**		
	Regeltyp[260]	Regel in Zeichen & deu. Übersetzung	Häufigkeit[261]	Regeltyp	Regel in Zeichen & deu. Übersetzung	Häufigkeit
TA-FANYI[262]	S+K; abgeleitet von ‚jemand + 翻译 → Subjekt-Prädikat-Phrase'	他/r + 翻译/v [*er* + *übersetzen*]	70	abgeleitet von ‚jemand+翻译 → Subjekt-Prädikat-Phrase'	她/r + 翻译/v [*sie* + *übersetzen*]	13
JINTIAN-FANYI	G; abgeleitet von ‚Zeitnomen + Prädikatkopf-Verb → VP'	今天/nt + 翻译/v [*heute* + *übersetzen*]	0	abgeleitet von ‚Zeitnomen + Prädikatkopf-Verb → VP'	今天/nt + 犯疑/v [*heute* + *verzweifeln*]	0
FANYI-LE	G; abgeleitet von ‚Verb + 了 → Perfekt vom Verb'	翻译/v + 了/u [*übersetzen* + Partikel für Perfekt]	101	abgeleitet von ‚Verb + 了 → Perfekt vom Verb'	犯疑/v + 了/u [*verzweifeln* + Partikel für Perfekt]	2

[260] K steht für Regel mit Konstante (bestimmtes Wort), S für semantische Regel und G für grammatische Regel.
[261] Anzahl des Zusammenauftretens beider Wörter in einem Satz (im CN-Corpus Online mit 19.455.328 Zeichen).
[262] Sonstige Regeln, die von der Variable-Regel ‚Jemand/n/r/nh + 犯疑' (*jemand* + *verzweifeln*) abgeleitet sind; ihre Häufigkeit lautet: ‚他/r + 犯疑/v' 0-mal und ‚她/r + 犯疑/v' 1-mal.

Regeln über Pinyin-Wort	**Regel Eins (Rot)**			**Regel Zwei (Blau)**		
	Regeltyp[260]	Regel in Zeichen & deu. Übersetzung	Häufigkeit[261]	Regeltyp	Regel in Zeichen & deu. Übersetzung	Häufigkeit
FANYI-WENZHANG	K; (R2), gewonnen aus dem Korpus-Beispielsatz	翻译/$_{v}$ + 文章/$_{n}$ [*übersetzen* + *Artikel*]	6			
LIANG-PIAN[263]	G; abgeleitet von ‚M + ZEW → Attribut'	两/$_{m}$ + 篇/$_{q}$ [*zwei* + ZEW]	79	abgeleitet von ‚Adjektiv + Nomen → NP'	亮/$_{a}$ + 片/$_{n}$ [*hell* + *Stück*]	0
PIAN-WENZHANG	K; (R6), gewonnen aus dem Korpus-Beispielsatz	(m/r_det.) + 篇 + 文章 [ZEW + *Artikel*]	222	abgeleitet von ‚骗+Objekt-Nomen → VP'	骗/$_{v}$ + 文章/$_{n}$ [*betrügen* + *Artikel*]	0

Tab. 4-16: Typen und Häufigkeiten anwendbarer sprachlicher Regeln für die Pinyin-Kette ‚TA/ JINTIAN/ FANYI/ LE/ LIANG/ PIAN/ WENZHANG'

Tab. 4-16 zeigt, dass es verschiedene sprachliche Regeln zu zwei oder mehr Pinyin-Wörter geben kann. Aus diesem Grund ist es obligatorisch, die Regeln nach ihrer Häufigkeit und Zuverlässigkeit zu bewerten. In solchen Fällen werden Regeln mit höherer Wahrscheinlichkeit und Korrektheit zuerst berücksichtigt (vgl. Wang XL/Wang YL 1996: 57f).

Um die optimalen sprachlichen Regeln einiger Pinyin-Wörter auszuwählen, kann die so genannte Wahrscheinlichkeitsschlussmethode der minimalen Elemente verwendet werden. Ein Vektor zusammengefasster Bewertungen von Wahrscheinlichkeit sowie Zuverlässigkeit der sprachlichen Regeln und dem Fall der wenigsten segmentierten Pinyin-Wörter (siehe Kap. 4.4.2) wird dabei ausgerechnet (vgl. ibid.: 57, Xu ZM 2000: 53). Die Wahrscheinlichkeit einer sprachlichen Regel wird normalerweise von der Auftritts- oder Anwendungshäufigkeit einer Regel in einem Korpus errechnet. Ihre Verlässlichkeit kann des Weiteren anhand ihres Typs klassifiziert werden. Am Beispiel der sprachlichen Regeln von zwei bestimmten oder unbestimmten Wörtern, die sich auf R1 in Tab. 4-16 beziehen, gestalten sich die Klassen der Zuverlässigkeit wie folgt:

0-Niveau: 我们/$_{r}$ + 翻译/$_{v}$;
1-Niveau: *jemand*/$_{r/n/nh}$ + 翻译/$_{v}$;
2-Niveau: r/n/nh + 翻译/$_{v}$;
3-Niveau: *jemand*/$_{r/n/nh}$ + ‚menschliches Verhalten'/$_{v}$;
4-Niveau: r/n/nh + v.[264]

[263] Die sonstigen Regeln dieser zwei Pinyin-Wörter sind jeweils abgeleitet von denselben angegebenen Variable-Regeln, nämlich ‚两/$_{m}$ + 片/$_{q}$' und ‚凉/$_{a}$ + 片/$_{n}$' mit 0-maliger Häufigkeit.
[264] Zur Bedeutung der chinesischen Wörter siehe Tab. 4-16, für die verschiedenen POS-Tagging-Codes Tab. 4-9.

In Tab. 4-16 wurden die sprachlichen Regeln der Wörter (ergo die minimalen sprachlichen Elemente der segmentierten Pinyin-Kette) angegeben. Wahrscheinlichkeit und Typ solcher Regeln wurden mit berücksichtigt. Bei der automatischen Zeichenbestimmung werden mögliche sprachliche Regeln zusammen mit der syntaktischen Ebene überprüft. So können jene Zeichenkandidaten ausgelassen werden, die im Kontext ungültig sind. Bspw. kann das Pinyin-Wort ‚WENZHANG' nur einem Nomen (<文章> *Artikel* oder <蚊帐>, in etwa: *zeltförmiges Moskitonetz*) entsprechen, muss aus syntaktischer Sicht jedoch ein Objekt sein. Aus diesem Grund werden die zwei sprachlichen Regeln über <犯疑> (/fànyí/, *verzweifeln*), welches kein Objekt einleiten kann, nicht mehr weiter berücksichtigt. Im Gegenteil dazu kann der andere homophonetische Kandidat <翻译> (/fānyì/, *übersetzen*) zur Satzgeneration bestimmt werden und die Regeln über ihn werden der Satzausgabe entsprechend eingesetzt (siehe Regel Eins sowie Regel Zwei zu ‚TA-FANYI' in Tab. 4-16). Nach der Wahrscheinlichkeitsanalyse wird Regel Eins weiterhin vor Regel Zwei (70 vs. 13 Treffer) bevorzugt angewendet.

4.6 Zusammenfassung von Methodik und Effizienz der chinesischen Eingabemethoden

Wegen der Besonderheiten des Schriftsystems ist die tastaturbasierte Eingabe der chinesischen Schrift mit zahlreichen Inputcodierungsmöglichkeiten, einem hohen künstlichen Intelligenzniveau und einem relativ komplizierten Verarbeitungsprozess verbunden. Nach den in Kap. 4 dargestellten Informationen, den Forschungsergebnissen und dem Vergleich mit der Eingabe der alphabetischen Schriftsysteme (vgl. Kap. 2) lassen sich folgende drei Punkte über Methodik und Effizienz der chinesischen Eingabemethoden zusammenfassen:

1. Die Eingabemethodik hat sich im Kontrast zu alphabetischen Schriftsystemen bei der chinesischen Schrift aus Sicht der zu verarbeitenden schriftlichen Einheit, Tastaturanwendung und den Grundarbeitsphasen in verschiedene Richtungen entwickelt.

Funktional läuft das Schreiben schrittweise von kleineren bis zu größeren Einheiten ab; auf Satzebene also von Schriftzeichen/Buchstaben über das Wort bis zum Satz (vgl. hierzu Kap. 1.2.1, S. 12f). Aus diesem Grund fokussieren die meisten Eingabeverfahren und Textverarbeitungen die Codierung und die Darstellung der Zeichen. Für die Eingabe eines alphabetischen Schriftsystems ist der Buchstabe (als kleinste segmentale schriftliche Einheit) zugleich auch die Einheit der Tastenbelegung und die interne bzw. Austauschcodierung sowie die graphische Ausgabe. Aus diesen Gründen sind zur Erforschung solcher Eingabeverfahren die drei folgenden Problemfelder zu berücksichtigen: 1) die Buchstaben-Tasten-Repräsentation im Tasta-

turlayout, 2) die Zeichencodierung für interne Systeme sowie Datenaustausch und 3) die Entsprechung von Zeichen, Zeichencodes, Glyphen und Fonts (vgl. hierzu Kap. 2.6).

Im Vergleich dazu kann es im chinesischen Schriftsystem keine Vereinheitlichung von der kleinsten schriftlichen, der eingetippten und der verarbeiteten Einheit geben. Auch der Prozess des chinesischen Schreibens mit Stift und Papier ist mit dem Schreiben am PC kaum vergleichbar. Bei der Eingabe muss aber eine sprachliche Einheit, die ein Zeichen bis hin zu einem Satz umfassen kann, inputcodiert und eingetippt werden, damit zum Schluss die korrekte Zeichenformen ausgewählt werden kann. Der Schwerpunkt der Eingabe liegt hauptsächlich darin, eine der PC-Standardtastatur angemessene Inputcodierung zu entwerfen und die Inputcode-Zeichen-Konversion zu programmieren. Erkenntnisse der Grammatologie (zum Inputcodierungsentwurf), Grammatik, Semantik und Pragmatik (zur Ermöglichung Satzes als Eingabeeinheit) bilden die Grundlage der technischen Realisierungen.

2. Auf dem Markt gibt es für die Eingabe des Chinesischen eine vielfältige Anzahl an Inputcodierungen, Eingabeeinheiten und Softwares. Die Eingabeeffizienz ist zugleich von mehr Faktoren abhängig, als die eines alphabetischen Schriftsystems.

Zeichenform und Phonetik bieten zwei Grundperspektiven zum Design der Zeicheninputcodierung. Basierend auf einem oder beiden Attribut(en) wurden hunderte verschiedene Inputcodierungen erfunden (vgl. hierzu Kap. 4.1.1, S. 181). Gleichsam kann keine Inputcodierung unumstritten Verbreitung finden oder sogar als nationaler Standard festgelegt werden, da es nie eine ideale Inputcodierung der chinesischen Schrift geben kann. Ein wichtiger Grund hierfür liegt darin, dass der Inputcode ein Schriftzeichen nur unvollständig repräsentieren kann. Gemäß Abb. 3-5 (Kap. 3.4.1) sind Form, Aussprache und Sinninhalt unentbehrliche Elemente für ein Schriftzeichen, während eine Art von Inputcodierung meistens nur auf einer Dimension basiert. Eine zeichenformbasierte Inputcodierung wie Wubi hängt bspw. mit der akustischen Version sprachlicher Äußerungen kaum zusammen, so dass der Schreibgedankengang fortwährend gestört wird. Es kann teilweise auch zu Konkurrenzen mit dem Strukturoriginal kommen: Dieses verbindet Zeichenform und -sinninhalt eng miteinander, was wiederum den Erwerb dieser Inputcodierung irrational macht (siehe Kap. 3.3.4, S. 157f). Eine phonetische Zeichencodierung basiert zumeist auf der Vernachlässigung der Zeichenform und weist eine hohe Überschneidungsquote der Kandidaten auf, was die Effizienz stark vermindert.

Die Idee der satzstufigen Eingabe ist vor allem zur Verbesserung der phonetischen Inputcodierung entstanden. Die Anwendung der künstlichen Intelligenz ermöglicht ebenso eine relativ hohe Präzision der Laut-Zeichen-Konversion im Satz. Die Pinyin-Codierung des gesamten Satzes ist im Grunde die kombinierte Codierung nach Phonetik plus Sinninhalt. Unter al-

len Eingabemethoden ist diese Variante im Prinzip die zugangsfreundlichste und dem Schreibgedankengang am ehesten angepasst (vgl. Kap. 4.1.4, S. 202). Ein weiteres Problem der phonetischen Eingabemethoden ist die Häufung von Fehlern, die beim Eintippen, der Silben- sowie Wortsegmentation und der Laut-Zeichen-Konversion passieren.

Die Effizienz der Eingabe eines alphabetischen Schriftsystems ist sowohl von einem objektiven Grund (der Rationalität der Tastenbelegung) als auch von einem subjektiven Faktor (der individuellen Zugänglichkeit spezifischer Tastaturlayouts) abhängig. Wegen der komplizierteren Methodik steht die Effizienz der chinesischen Eingabe unter weit mehr objektiven Bedingungen.

Für die Effizienzevaluation müssen zudem unterschiedliche Kriterien für die zeichen- oder wortbasierte sowie satzstufige intelligente Eingabe erstellt werden. Im Allgemeinen ist die Bewertung einer Eingabemethode die Begutachtung der Inputcodierung, die durch die in Kap. 4.1.1 (S. 184f) skizzierten sechs Kriterien eingeschätzt werden kann. Die Statistiken sowie Algorithmen der folgenden vier Punkte müssen dabei mindestens berücksichtigt werden: das Inventar des Konversionslexikons (die Menge der eingebbaren Schriftzeichen), die Standardisierung der Inputcodierung, die durchschnittliche Codelänge und die Überschneidungsquote (die Häufigkeit der Tastenanschläge für die Kandidatenauswahl) (vgl. Zhang YH/Zhou 2004: 51). Die Evaluation einer intelligenten Pinyin-Eingabesoftware fokussiert das Künstlich-Intelligenz-Niveau, weshalb die Korrektheit der Laut-Zeichen-Konversion und der Silben- sowie Wortsegmentation dazu entscheidend ist. Laut eines Forschungsteams des Harbin Institute of Technology beträgt die höchste Präzision der intelligenten Laut-Zeichen-Konversion 63,3% (vgl. Tang et al. 2008: 51).

3. Zur Eingabe des Chinesischen sind viele computerlinguistische Anwendungen vonnöten, um die sprachlichen Erkenntnisse für die automatische Zeichenauswahl zu unterstützen.

Solche für intelligente Pinyin-Eingabemethoden unentbehrlichen computerlinguistischen Techniken beinhalten bspw. Wortsegmentation, POS-Tagging, Phrasenerkennung, Satzanalyse sowie -generation, welche ausnahmslos auf Basis von sprachlichen Wissensdatenbanken programmiert werden. Wie diese zusammenhängend in einer intelligenten Eingabesoftware funktionieren, wird in Abb. 4-33 zusammengefasst dargestellt. Die Software übernimmt dadurch in den meisten Fällen die Aufgabe des menschlichen Gehirns, die Zeichenform auszuwählen. Hat man sich in der Praxis an die relativ hohe Präzision der Laut-Zeichen-Konversion gewöhnt, verringert sich das Bewusstsein für manuelle Korrekturen, so dass Schreibfehler mit der intelligenten Pinyin-Eingabesoftware am häufigsten passieren (vgl. Tang et al. 2008: 51, Wang YM/Yang 2005).

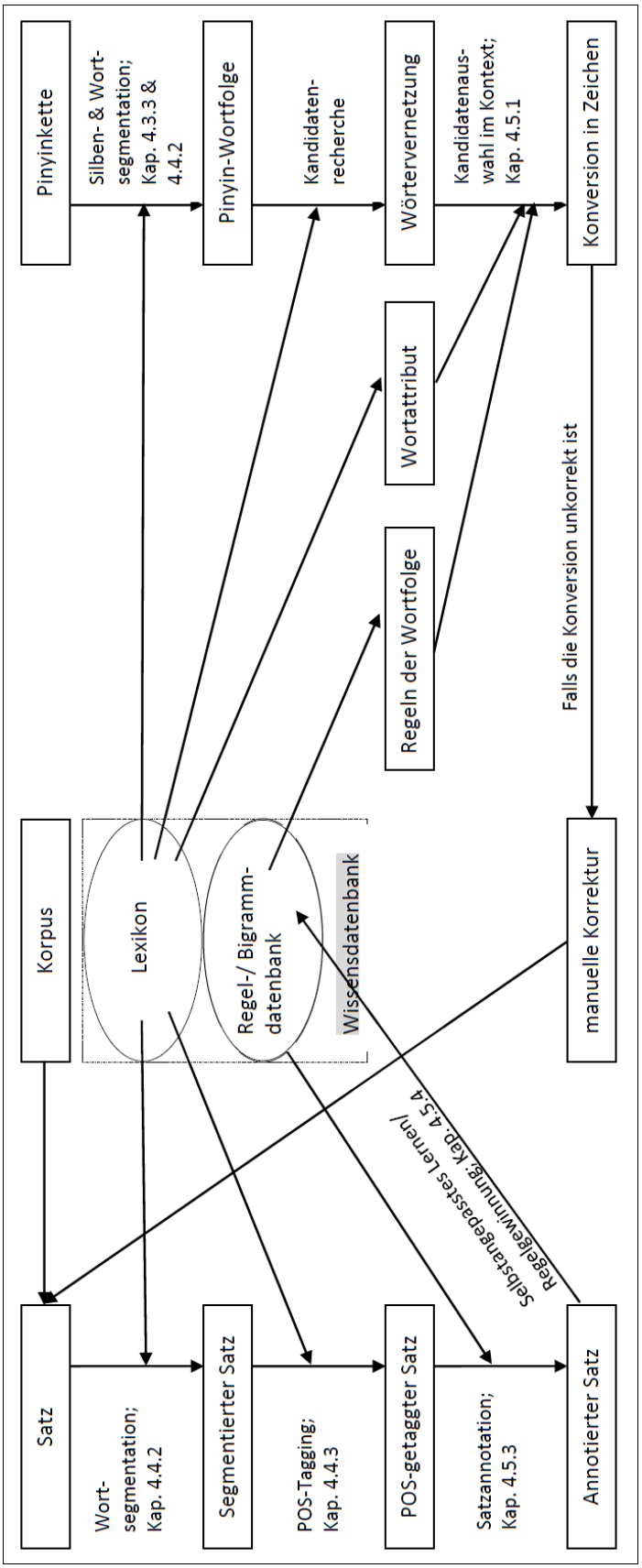


Abb. 4-33: Zusammenfassung von den computerlinguistischen Anwendungen bei intelligenten Pinyin-Eingabesoftwares

Künstliche Intelligenz ist deswegen ein zweischneidiges Schwert bei der Eingabe. Sie ermöglicht einerseits die satzstufige Laut-Zeichen-Konversion mit hoher Effizienz. Andererseits scheinen sich die sprachlichen sowie schriftlichen Kenntnisse von PC-Benutzern durch die maschinelle Intelligenz zu reduzieren, was eine negative kulturelle Entwicklung verursachen könnte. Entwicklungsperspektivisch stellt sich daher die Frage, ob die Intelligenz der Laut-Zeichen-Konversion weiter erhöht werden soll oder neue Eingabemöglichkeiten erfunden werden müssen, um die intelligente Pinyin-Eingabemethode zu ersetzen.

5 Projektplan für ein multilinguales Eingabesystem

Kap. 2 und 4 widmeten sich den Eingabeverfahren verschiedener Schriften sowie Schriftsysteme. Auf dieser Basis können Arbeitsprinzipien unterschiedlicher Eingabemöglichkeiten geschlussfolgert werden, auf deren Basis das Design einer neuen Eingabesoftware möglich wird. Aufgrund der Tatsache, dass sich bislang kein multilinguales Eingabesystem auf dem Markt durchsetzen konnte, erstelle ich die Konzeption für ein multilinguales Eingabesystem, das die Eingabe verschiedener Sprachen ohne Sprachumstellung gleichzeitig unterstützt. Nach Gumm und Sommer sind zur Softwareentwicklung folgende Arbeitsschritte nötig:

1. Problemanalyse und Anforderungsdefinition
2. Modellierung und fachlicher Entwurf
3. Software-technischer Entwurf
4. Programmierung und Modultest
5. Systemintegration und Systemtest
6. Installation, Betrieb und Weiterentwicklung

(Gumm/Sommer 2011: 831)

Im fünften Kapitel dieser Arbeit sollen abschließend die ersten drei Schritte des Projektplans einer Software umgesetzt, ergo die theoretische Vorarbeit geleistet, der Entwurf einer Softwarestruktur skizziert sowie das fachliche und technische Design konzeptionskriteriell antizipiert werden.

5.1 Anforderungsanalyse und Grundkonstruktion

Wie angegeben sind Problemanalyse und Anforderungsdefinition der erste Schritt des Entwurfs. In Kap. 5.1 werden zuerst die grobe Softwarestruktur und das -aussehen dargelegt. Die dafür benötigten Module werden in Kap. 5.2 dann detailliert nach technischen sowie linguistischen Aspekten analysiert und entworfen.

5.1.1 Anforderungsdefinition, Anwendungsfälle und Problemanalyse

Sowohl Individuen als auch Unternehmen und Organisationen zeichnen sich heutzutage durch Mehrsprachigkeit aus. Beim schriftlichen kommunikativen Austausch (und der Schreibtätigkeit generell) sind der weltweite Zeichencodierungsstandard (Unicode) sowie Textverarbeitungsprogramme, die über mehrere Sprachen verfügen (wie Microsoft Office Word), gebräuchlich. Zur Eingabe von mehreren Sprachen aber muss man meistens verschiedene Tastaturlayouts mechanisch oder visuell wechseln, bestimmte Eingabeverfahren für jede Sprache

installieren und situationsangemessen zwischen verschiedenen diesen umschalten. Wenn ein multilingualer Text geschrieben wird, sind sowohl die Arbeitseffizienz des Computers als auch das menschliche Gehirn stark belastet. Um diese Situation zu bessern, ist eine multilinguale Eingabesoftware sinnvoll.

Zwar gibt es bereits Erfindungen für multilinguale Eingabesysteme, wie z.B. das Patent von NAKASATO Shigemi (中里茂美) von Toshiba, welches die Eingabe in Japanisch, Englisch und Chinesisch unterstützt und dessen Arbeitsprinzipien für die meisten Schriftsysteme durchführbar sind (vgl. Nakasato 1998). Aus verschiedenen Gründen aber ist kein multilinguales Eingabesystem besonders beliebt und verbreitet. Nach Erforschungen der Eingabeverfahren verschiedener Schriftsysteme bin ich der Meinung, dass eine Software mit hoher Intelligenz zur Spracherkennung und Inputcode-Zeichen-Konversion eine gute Lösung für multilinguale Eingaben sein kann. Die Realisierung von Spracherkennung und Konversion auf Zeichen-, Wort- und Satzstufe ist der rote Faden für diesen Softwareentwurf. Die Wissensgrundlage der Softwareentwicklung basiert auf den Erkenntnissen und Forschungsergebnissen zu linguistischen Eigenschaften, Tastaturlayouts und Techniken zur Zeichencodierung, -inputcodierung sowie -ausgabe, die in Kap. 1 bis 4 vorgestellt und untersucht wurden.

Eine nach diesem Entwurf entwickelte Software könnte auf PCs privater User, die mindestens drei Sprachen beherrschen, sowie Computern internationaler Organisationen und Unternehmen, die von polyglotten Mitarbeitern benutzt werden, installiert werden. Ihr Design orientiert sich vor allem an der mit verschiedenen Schriften verbundenen Mehrsprachigkeit. Sie wird deswegen sowohl in multilingualen als auch multischriftlichen Fällen angewendet. Das zu verarbeitende Zeicheninventar stimmt erwartungsgemäß mit den 65.536 Zeichen aus der ersten Unicode-Ebene (Basic Multilingual Plane) überein.

Diese Software kann auf CD-ROM übertragen und auf Computern installiert werden. Zur Installation muss ein Computer die Voraussetzung mit sich bringen, Glyphen der meisten BMP-Zeichen zu enthalten. Da die Software immer mit einem komplexen Textverarbeitungsprogramm parallel interagieren muss, ist es auch notwendig, verschiedene Schriftrichtungen unterstützt werden und für mehrere Sprachen zuständig sein können.

Die größten Schwierigkeiten des Softwareentwurfs sind durch die Mehrsprachigkeit bedingt und können in den folgenden vier Punkten zusammengefasst werden:

1) Inputcodierung und das angesetzte Tastaturlayout.

Zur Anwendung einer multilingualen Eingabesoftware muss ein bestimmtes mechanisches Tastaturlayout benutzt werden, durch das schriftliche Ausdrücke in verschiedenen Sprachen

codiert werden können. Wie die multilingualen Symbole mit begrenzt belegbaren Tasten inputcodiert werden können, so dass die eingegebenen Sprachen sowie Informationen relativ effizient erkannt und ausgegeben werden können, ist das erste zu lösende Problem.

2) Techniken der Spracherkennung und Inputcode-Zeichen-Konversion.

Damit die Software multilinguale schriftliche Symbole anhand von Inputcodes anbieten kann, muss sie in der Lage sein, die eingegebene Sprache zu erkennen (oder die Wahrscheinlichkeit auf wenige Sprachen zu reduzieren) und den Inputcode in die betroffenen Kandidaten möglicher Sprachen zu konvertieren. Dazu müssen nicht nur Algorithmen mehrerer Module entwickelt werden, sondern sprachliche Ressourcen beigefügt sein.

3) Einschränkung des Speicherplatzes.

Wie oben dargestellt, müssen im Prinzip Wissensdatenbanken für die bearbeitbaren Sprachen zur Unterstützung der künstlich-intelligenten Spracherkennung und Konversion beigefügt werden. Es gibt in der heutigen Welt hunderte verschriftete Sprachen, die zum schriftlichen Austausch angewendet werden.[265] Aus diesem Grund muss die Sprachanzahl sowohl für PC-Benutzer als auch für eine multilinguale Eingabesoftware auf eine bestimmte Menge eingeschränkt sein. Nach meinem Plan beinhaltet das sprachliche Paket, das in der Software-CD-ROM gespeichert ist, voreingestellt nur Daten der zwanzig am häufigsten gesprochen Sprachen.[266] Der PC-Benutzer bestimmt bei der Installation, welche Sprachen eingegeben werden sollen. Die entsprechenden Datenbanken laden diese dann runter und wenden sie an. Falls die Datenbank einer gewünschten Eingabesprache im Paket nicht vorhanden ist, muss man sie zusätzlich herunterladen. Die erforderlichen Ressourcen für Zeichen-, Wort- und satzstufige Eingaben weichen von Sprache zu Sprache stark ab. Deswegen ist es notwendig, die einzugebenden Sprachen und ihre Verarbeitungsstufen in einem Computer individuell zu definieren, so dass nur die benötigten Ressourcen in die Softwareanwendung integriert werden.

4) Darstellung der Kandidaten in Wahlliste.

Derselbe Inputcode kann in verschiedenen Sprachen sehr viele Kandidaten betreffen. Wie solche Kandidaten besser angeordnet und in der Wahlliste angezeigt werden können, um die Schreibeffizienz zu erhöhen, ist ein weiterer wichtiger Punkt, den es in der Softwareentwicklung zu lösen gilt. Neben dem Problem der potentiell möglichen hohen Kandidatenüber-

[265] Als Richtschnur kann die freie Enzyklopädie Wikipedia betrachtet werden, die in 301 Sprachen vorliegt (vgl. *https://meta.wikimedia.org/wiki/List_of_Wikipedias* [Abruf: 2018-06-09]). Dabei muss beachtet werden, dass auch Welthilfssprachen wie Esperanto dazuzählen; die chinesischen Sprachen, die im Prinzip zu einem Schriftsystem zählen, werden einzeln betrachtet.

[266] Für die dazugehörigen Sprachen vgl. *https://www.ethnologue.com/statistics/size* [Abruf: 2018-06-09].

schneidung müssen die unterschiedlichen Schriftrichtungen berücksichtigt werden (siehe Kap. 1.3.6, S. 36f).

5.1.2 Lösungsvorschläge

Der Softwareentwurf beginnt zuerst mit Ideen zur Überwindung der vier durch die Mehrsprachigkeit bedingten Schwierigkeiten.

1) Das eingesetzte Tastaturlayout.

Zur multilingualen Eingabe muss zuerst ein Tastaturlayout ausgewählt werden, das weltweit bekannt und für die Eingabe aller Schriftsysteme möglich ist. Die 26 lateinischen Grundbuchstaben und internationalen Zahlzeichen werden deswegen bevorzugt. Das lateinische Alphabet ist einerseits die am meisten gebrauchte Schrift der Welt, andererseits besitzt fast jede andere Schrift wegen der Internationalisierung einen generell anerkannten Transliterations- oder Transkriptionsstandard (vgl. hierzu Kap. 1.3.4, S. 32). Mit einem US-amerikanischen Layout sind sprachliche Ausdrücke (ohne Sonderbuchstaben) in einem lateinalphabetischen Schriftsystem ohne Inputcodierung eingebbar. Um die Spracherkennung und Konversion wirksamer unterstützen zu können, werden drei neue Funktionstasten hinzugefügt oder belegt, die jeweils für *Markierung der graphischen Eigenständigkeit* (im Weiteren ‚CM-Taste'; Abk. für *character graphic mark*), *Konversion* (von *conversion*; ‚CV-Taste') und *Nicht-Konversion* (von *not-conversion*; ‚NC-Taste') stehen. Die Tasten ‚NC' oder ‚CV' werden nach einer Eingabeeinheit gedrückt, um zu entscheiden, ob die eingetippten Zeichen direkt ausgegeben oder ein Konversionsprogramm gestartet wird. Die ‚CM'-Taste wird nach dem Inputcode eines Syllabars oder morphologischen Schriftzeichens gedrückt und nur in alphasyllabischen, koreanischen, chinesischen und japanischen Schriftsystemen gebraucht. Das Tastaturdesign der geplanten Software wird in Abb. 5-4 und 5-5 abgebildet.

2) & 3) Techniken zur Spracherkennung und Inputcode-Zeichen-Konversion und das Sparen von Arbeitsspeicher.

Die zweite und dritte Schwierigkeit sind eng miteinander verwoben, so dass sie beim Softwareentwurf zusammen betrachtet werden müssen. Wie oben erwähnt wurde, läuft der Arbeitsprozess anders, je nachdem ob eine Eingabeeinheit mit ‚NC' und ‚CV' endet. Im ersten Fall sind Spracherkennung und Konversion unnötig und der Arbeitsprozess ist im Prinzip fast identisch mit auf Tasten-Zeichen-Repräsentation basierenden Eingabeverfahren. Wenn sie stattdessen mit ‚CV' endet, muss die Verarbeitung von künstlicher Intelligenz unterstützt werden, weshalb sprachliche Ressourcen wie die Transkriptionsliste und ein maschinenlesbares Lexikon einer Sprache usw. obligatorisch sind. Die Konversionsergebnisse sind die Kandida-

ten, denen der eingegebene Inputcode entspricht. Die Beispiele für die Eingabe von verschiedenen Sprachen mit dem geplanten multilingualen Eingabesystem werden in Tab. 5-1 (in Kap. 5.2.1) angegeben.

Da die Schreibfähigkeit und -häufigkeit in verschiedenen Sprachen individuell ausgeprägt sind, ist eine manuelle oder automatische Ranglistenumstellung mehrerer Sprachen sowie Kandidaten vonnöten. Das System berechnet einerseits die individuellen Schreibstatistiken für einen PC. Andererseits ist ein einzelner PC per Cloud Computing online mit Servern verbunden, so dass die individuellen Statistiken geteilt und generelle Schreibpräferenzen analysiert werden können. Auf der Basis von den automatisch, anhand der generellen und individuellen Statistiken umgestellten Ranglisten hat der Benutzer auch die Möglichkeit, die Liste für Sprachen und Zeichenkandidaten manuell umzuändern.

Für die Schreibeffizienzerhöhung und Arbeitsspeicherentlastung werden drei verschiedene Eingabeniveaus einer Sprache geplant, das simple, das mittlere und das fortgeschrittene Niveau: das simple Niveau für die Eingabe von einzelnen Buchstaben, Graphemen, Syllabare sowie Zeichen; das mittlere Niveau für die Eingabe eines Wortes; und das fortgeschrittene Niveau für Fälle, wenn ein Satz als größte Eingabeeinheit zu verarbeiten ist. Das Design der drei Niveaus ist einerseits von der Gebrauchshäufigkeit sowie Fertigkeit der benutzten Sprachen eines Individuums bedingt. In einer gut beherrschten und häufig geschriebenen Sprache ist die Verarbeitung in der Satzeinheit vonnöten. Hingegen ist in einer selten gesprochenen Sprache nur die Eingabe von einzelnen Symbolen erforderlich. Der Entwurf dreier Niveaus ist zugleich wegen der Arbeitsspeichereinschränkung vonnöten, denn für das fortgeschrittene Niveau einer Sprache sind Datenbanken mit viel größerem Speicherplatz erforderlich, als bei dem simplen Niveau. Zusammengefasst lauten die Beziehungen zwischen Eingabeniveaus, möglichen Kandidatenüberschneidungen und benötigten Daten wie folgt: Je höher das Eingabeniveau wird, desto weniger mögliche Kandidaten aus verschiedenen Sprachen gibt es im Schnitt. Auf höherem Niveau wird aber zugleich eine höhere künstliche Intelligenz angewendet. Die benötigten Datenmengen werden dementsprechend immer umfangreicher und komplexer. Beim Gebrauch dieses multilingualen Eingabesystems muss zielorientiert das gewünschte Niveau einer Sprache ausgewählt und die entsprechenden Ressourcen installiert werden. Die Gebrauchszwecke, die benötigten Ressourcen und der Arbeitsprozess der drei verschiedenen Eingabeniveaus werden in Kap. 5.2.2 und 5.2.3 ausführlicher vorgestellt.

4) Die Kandidatendarstellung in der Wahlliste.

Für die Kandidatenausgabe und -auswahl soll ein spezielles Fenster auf dem Bildschirm dargestellt und ein Touchpad angewendet werden, um die Auswahleffizienz zu erhöhen. Damit

die Kandidaten der verschiedenen Sprachen optisch getrennt angezeigt und schneller ausgewählt werden können, werden die Schriftsysteme je nach Schriftrichtung und graphischen Eigenschaften sowohl auf dem Bildschirm als auch auf dem Touchpad in vier Zonen aufgeteilt.

Die erste Zone steht für Äußerungen in einem europäischen Schriftsystem und international benötigten Sonderzeichen, die horizontal rechtsläufig geschrieben werden. Der zweiten Zone werden die Schriften des CJK-Kreises zugeordnet, während die dritte für die alphasyllabischen Schriften und die vertikal geschriebene mongolische Schrift zuständig ist. In der vierten Zone sind die Zeichen in arabisch-persischer Schrift. Entsprechend den Fingern einer Hand werden auf jeder Seite einer Zone nur vier Kandidaten angezeigt und das Scrollen zu der nächsten und letzten Seite ermöglicht. Ausführlicher wird auf den Zonenentwurf auf Bildschirm und Touchpad in Kap. 5.2.4 eingegangen.

5.1.3 Gesamte Softwarestruktur

Anhand der in Kap. 5.1.2 dargestellten Hauptideen soll diese multilinguale Eingabesoftware mindestens drei Arten von Modulen beinhalten: das Modul der Eingabe, die Module für die Inputcode-Zeichen-Konversion und das Modul der Kandidatenausgabe. Die grobe Softwarestruktur und den Verarbeitungsprozess mit der Software zeigt folgende Abbildung.

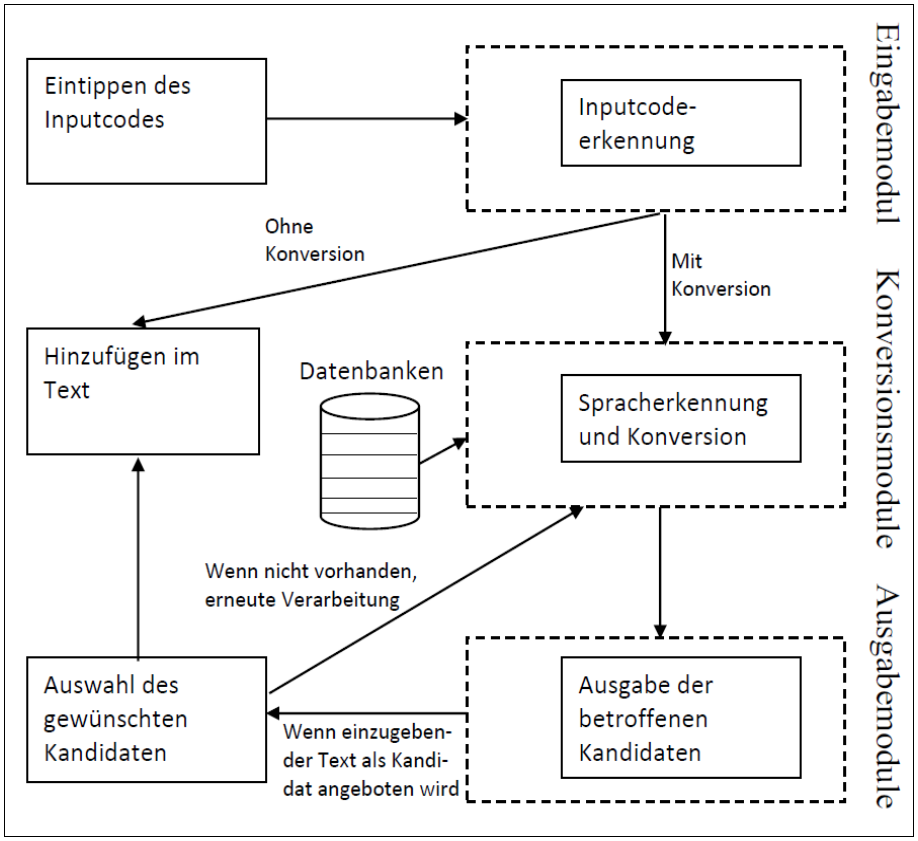


Abb. 5-1: Struktur der multilingualen Eingabesoftware und ihr Verarbeitungsprozess

Wie die Abbildung zeigt, beginnt die Verarbeitung mit der Inputcodeerkennung im Eingabemodul. Bedingt von der Taste ‚CV' oder ‚NC' läuft sie entweder gemäß des kürzeren Pfads direkt zum Ende oder nimmt den längeren Pfad, der die anderen Module durchlaufen muss.

Zu Konversionsmodulen gehören mehrere Teilmodule, die für bestimmte Eingabefälle zuständig sind. Ein Eingabefall ist von der Sprachenklassifikation (geteilt durch den Typ ihrer dargestellten Schrift) und dem Eingabeniveau bedingt. In einem solchen Modul sind (wie angerissen) zwei Hauptschritte unentbehrlich: Spracherkennung und Inputcode-Zeichen-Konversion. Die beiden Schritte müssen durch Zugriff auf eine oder mehrere entsprechende Datenbanken, in denen linguistische Wissensdaten bestimmter Sprachen gelagert sind, unterstützt werden. Die Einteilung der Eingabefälle, die Verarbeitungsprozesse in solchen Teilmodulen und die benötigten Datenbanken werden in Kap. 5.2.2 erläutert.

Nach der Konversion werden die möglichen betroffenen Kandidaten im Ausgabemodul in Form einer Wahlliste dargestellt, die in einer oder zwei der vier Teilzonen angezeigt werden kann. Die Ideen im Detail und die dabei verwendeten Techniken für das Modul werden in Kap. 5.2.4 skizziert.

Wenn der zu schreibende Inhalt als Kandidat angeboten wird, wird er ausgewählt und dem Text hinzugefügt, womit der Verarbeitungsprozess endet. (Verfahren und Techniken zur Kandidatenauswahl werden ebenfalls in Kap. 5.2.4 vorgestellt.) Falls er aus verschiedenen Gründen nicht ausgegeben wird, muss die Konversion erneut in kleineren Segmenten (als einzelnes Wort oder Zeichen) erfolgen.

5.1.4 Look and Feel der Software

Nach dem Plan der gesamten Softwarestruktur wird in diesem Kapitel das Aussehen der Software, das sog. ‚Look and Feel', berücksichtigt. Gemeint ist „das optische Erscheinungsbild sowie die Bedienungseigenschaften einer Benutzeroberfläche" (Greulich 2003: 545).

Das *Look and Feel* dieser Software hat nach meinem Plan vier Symbolleisten. Bei der Anzeigesprache empfiehlt sich wegen ihrer Verbreitung eine Voreinstellung auf Englisch, die im PC individuell umgestellt werden kann. Die Symbolleisten stehen jeweils für die Festlegung einer Sprache (*language codifying*), die Anordnung der verarbeitenden Sprachen (*languages arranging*), die Umschaltung zwischen Voll- und Halbbreite (*full- or halfwidth*) und Tools für weitere Leisten:

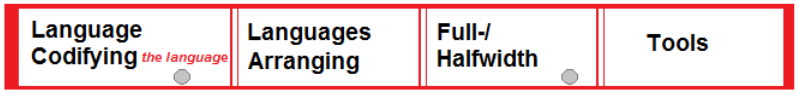


Abb. 5-2: Softwareerscheinungsbild *Look and Feel* des multilingualen Eingabesystems

Die Leiste ‚Language Codifying' bietet die Möglichkeit, eine bestimmte einzugebende Sprache festzulegen, damit die Eingabe zielorientiert effektiver ablaufen kann. Mit einem einmaligen rechten Mausklick auf das entsprechende Feld wird ein Menü für alle schreibbaren Sprachen des Computersystems angezeigt und die Festlegung einer bestimmten Sprache kann aktiviert werden. Der Status für die Aktivierung der Sprachfestlegung wird mit einem gefüllten Kreis symbolisiert und die festgelegte Sprache wird in roter kleiner Schrift angegeben. Im Gegenteil dazu steht der Status mit einem leeren Kreis für multilinguale Eingabe. Im Status der Sprachfestlegung kann auf die Anwendung der drei Funktionstasten verzichtet werden. Die Eingabe läuft dann im Prinzip nicht anders als ein transkriptionsbasiertes, einzelsprachorientiertes Eingabeverfahren. Wenn ein in einer einzelnen Sprache verfasster relativ langer Text zu schreiben ist, ist die Aktivierung der Sprachfestlegung empfehlenswert.

Mit ‚Languages Arranging' kann die Reihenfolge der eingebbaren Sprachen manuell angeordnet werden, so dass die Kandidaten in häufig geschriebenen Sprachen zuvorderst recherchiert, erzeugt und ausgegeben werden. Wie in Kap. 5.1.2 (S. 315) erwähnt, wird die Anordnung der Sprachen bei der Softwareanwendung sowohl automatisch (im Einzel-PC und per Cloud Computing) als auch manuell durchgeführt.

Für die Zeichenanzeige eines chinesischen Zeichens ist die doppelte Breite wie für die eines lateinischen Buchstabens erforderlich, falls sie in derselben Schriftgröße dargestellt sind. Aus diesem Grund sind im CJK-Kulturkreis die beiden Begriffe Voll- und Halbbreite (eng.: full- & halfwidth, chi. & jap.: 全角 & 半角) entstanden. Die beiden Termini beziehen sich dabei auf „the relative glyph size of characters" (Lunde 2009: 25). Auf den Punkt gebracht werden Sinogramme (Hanzi, Kanji und Hanja) und koreanische Syllabare immer in Vollbreite ausgegeben, während Buchstaben aus einem europäischen alphabetischen Schriftsystem immer in Halbbreite auszugeben sind. Diese Leiste der Eingabesoftware wird deswegen erst aktiviert und angezeigt, wenn Chinesisch, Japanisch oder Koreanisch als Eingabesprache des Eingabesystems ausgewählt wird. Wie in Kap. 1.3.1 (S. 20f) ausgeführt, sind manche internationalen Zeichen generell in allen Schriftsystemen benutzbar. In einem CJK-Text können sie – wie z.B. lateinische Majuskel, arabische Ziffer und Interpunktionszeichen – sowohl in Voll- als auch in Halbbreite dargestellt werden. Dasselbe gilt auch für Hiragana-, Katakana, Jamo- und Zhuyin-Zeichen (vgl. ibid.: 26).

Im Normalfall werden die Interpunktionszeichen und Sonderzeichen aus dem CJK-Schriftkreis vom System in Vollbreite voreingestellt und die lateinischen Buchstaben sowie Zahlzeichen in Halbbreite verarbeitet. Mit dieser Leiste hat man die Möglichkeit, die Breiteform solcher Zeichen nach Wunsch umzustellen. Wenn in dieser Leiste ein gefüllter Kreis

(‚Vollmond') steht, ist die schriftliche Anzeige in Vollbreite aktiviert. Hingegen markiert die Halbmondform den Status der Halbbreite. Wenn kein Symbol angezeigt wird, switcht das System automatisch zwischen Halb- und Vollbreite, je nachdem welche Schriftzeichen eingegeben werden.

Durch das Klicken der ‚Tools'-Leiste kann ein Menü mit weiteren Funktionen sowie Leisten geöffnet werden. Sie umfassen bspw. die Sprachlistenverwaltung des Systems (bspw. das Hinzufügen, Löschen und die Änderung des Eingabeniveaus der einzugebenden Sprachen, die voreingestellte Schriftart der Zeichen aus einem Schriftsystem usw.), das Aussehen der vier Teilzonen der Kandidatenwahlliste auf dem Bildschirm (bspw. ihre Position, Größe sowie Hintergrund oder Schriftfarbe) und die Umstellung der Anzeigesprache. Die Sprachliste am Beispiel von acht ausgewählten Sprachen könnte wie in Abb. 5-3 aussehen.

			low	middle	high	typeface	Region
中文	Chinese	ZH	○	○	●	Song-Ti	PR China
Deutsch	German	DE	○	○	●	Times New Roman	Germany
English	English	EN	○	●	○	Times New Roman	USA
日本語	Japanese	JA	○	●	○	Mincho	
рýсский	Russian	RU	○	●	○	Times New Roman	Russia
한국어	Korean	KO	●	○	○	Batang	ROK
हिन्दी	Hindi	HI	●	○	○	Mangal	
Tiếng Việt	Vietnamese	VI	●	○	○	Times New Roman	
العربية	Arabic	AR	●	○	○	Arabic Typesetting	Egypt

+/- languages

Abb. 5-3: Entwurf der Sprachliste des multilingualen Eingabesystems[267]

5.2 Zielorientierte Modulentwicklung, benötigte Ressourcen und Techniken

In Kap. 5.1.2 wurde die Lösungsvorschläge und in Kap. 5.1.3 die Softwarestruktur geplant. Kap. 5.2 geht ausführlicher auf den Entwurf jedes Moduls und die benötigten Elemente der Software ein. Die Reihenfolge der Untersuchungselemente stimmt dabei mit dem in Abb. 5-1 dargestellten drei Arbeitsschritten überein: 1) Eingabe (inkl. Inputcodierung und Tastaturlay-

[267] Jede Sprache wird jeweils in ihrer ursprünglichen Schrift, in Englisch und im ISO-Code angegeben. Mit ‚+/- languages' kann die Sprachliste verwaltet werden. Das Eingabeniveau einer Sprache wird mit einem gefülltem Kreis markiert (unter ‚low', ‚middle' und ‚high'). Die voreingestellte Schriftart und Region kann mit dem Pfeil rechtsunten umgestellt werden.

out), 2) Spracherkennung und Konversion (inkl. Klassifikation der Eingabeniveaus und -fälle, Eigenschaften der Sprachen und beigefügte sprachliche Ressourcen), 3) Ausgabe und Auswahl (inkl. Entwurf der benötigten Hardware, Methoden und Techniken für Kandidatenausgabe und -auswahl).

5.2.1 Design für die Inputcodierung verschiedener Sprachen und das Tastaturlayout

In Kap. 5.1.2 wurden die drei hinzugefügten Funktionstasten ‚CM', ‚CV' und ‚NC' sowie ihre Gebrauchssituationen eingeführt. Ihr Design hängt mit den verschiedenen Arten der Schriften zusammen, wie in Kap. 1.3.2 (vgl. hierzu S. 21-25) erläutert wurde. Verknappt geschildert werden die drei Tasten sowie die ‚Leer-' und ‚Shift-Taste' angewendet, um die eingegebene Sprache je nach den Merkmalen ihres Schriftsystems ausgefiltert zu bestimmen. Die erkennbaren Merkmale eines Schriftsystems bei der Eingabe sind die folgenden:

1. Ob das/der einzugebende Buchstabe / Wort / Satz im Umfang der 26 lateinischen Grundbuchstaben eingeschränkt ist. Nur wenn die einzugebende Sprache im lateinischen Alphabet geschrieben wird und kein Sonderbuchstabe in der Eingabeeinheit vorhanden ist, wird ‚NC' (*Nicht-Konversion*) benötigt. In anderen Fällen muss ‚CV' (*Konversion*) am Ende einer Eingabeeinheit gedrückt werden.
2. Ob es in einem Schriftsystem Unterscheidung zwischen Groß- und Kleinschreibung gibt. Wie auf S. 22 erwähnt, ist die Majuskel-Minuskel-Unterscheidung Eigenschaft der fünf vollalphabetischen Schriften, weshalb für die in diesen Alphabeten dargestellten Schriftsysteme ‚Shift' eine Rolle spielt: Falls in der Inputcodierung ‚Shift' eingetippt wird, muss diese eingegebene Sprache vollalphabetisch geschrieben sein.
3. Ob Buchstabe oder syllabisches/morphologisches Zeichen in dem Schriftsystem als graphische Einheit fungiert. Wenn ein Buchstabe als das kleinste funktionale und graphische Segmental im Text fungiert, ist kein ‚CM' (*Zeichenmarkierung*) gebräuchlich. Wenn ein zweidimensional zusammengesetztes Syllabar (in den indischen alphasyllabischen Schriften und im koreanischen Hangul), Silbenzeichen oder morphologisches Schriftzeichen als kleinste graphische Einheit auftritt, muss nach ihm ‚CM' eingetippt werden.
4. Ob ein Spatium zwischen den Wörtern auftritt. Wie in Kap. 1.3.1 (S. 13) skizziert, sind Spatien für alphabetische Schriftsysteme entscheidend, damit Lesen und Schreiben störungsfrei funktionieren. Hingegen ist es in chinesischen und japanischen Texten wegen der Anwendung der morphologischen Schriftzeichen nicht vonnöten. Wenn die Eingabeeinheit auf Wort- oder Satzebene ist und kein Spatium gebraucht wird, so muss es sich um Chinesisch/Japanisch handeln. Wenn eine Eingabeeinheit nur aus einem Wort besteht und hinter

dem Wort ein anderes Interpunktionszeichen als Spatium gesetzt werden soll, müssen das Interpunktionszeichen und das Spatium hinter ihm (falls nötig) zusammen zur Eingabeeinheit gehören, um Ambiguitätsfälle zu vermeiden.

Um den Gebrauch der vier Tasten zu erklären, nehme ich zwei Begriffe aus acht verschiedenen Sprachen zum Beispiel: *Computer* und *Tee*. Grund für die Auswahl der beiden Begriffe ist ihre identische etymologische Herkunft sowie ihre daher vergleichbare Phonetik.

Sprache	**Beispiel 1 (*Computer*)**[268]		**Beispiel 2 (*Tee*)**[269]	
	Eingabe	**Ausgabe**[270]	**Eingabe**	**Ausgabe**
Englisch	computer SP NC	computer•→	tea SP NC	tea•→
Deutsch	SH+c omputer SP NC	Computer•→	SH+T ee SP NC	Tee•→
Russisch[271]	komp'yuter SP CV	компьютер•→	chj SP CV	чай•→
Arabisch[272]	hasoub SP CV	←• حاسوب	shay SP CV	←• تي
Hindi[273]	kan CM pyoo CM ta CM ra CM SP CV	कंप्यूटर•→	chaay SP CV	चाय•→
Koreanisch[274]	keom CM pyu CM tel CM SP CV	컴퓨터•→	cha CM SP CV	차•→
Japanisch[275]	ko CM n CM p CM yu CM - CM ta CM - CM CV	コンピューター→	cha CM CV	茶→
Chinesisch[276]	dian CM nao CM CV	电脑→	cha CM CV	茶→

Tab. 5-1: Beispiel für Inputcodierung in acht verschiedenen Schriftsystemen

In Kap. 1.3.4 (S. 32) wurden *Transkription* sowie *Transliteration* definiert und ihre Anwendung für Eingabesoftwares verschiedener nichtlateinalphabetischer Schriftsysteme eingeführt. Bei der Erforschung der sonstigen Schriftsysteme in Kap. 2 und 3 wurden Transliterationen sowie Transkriptionen von Hindi, Koreanisch, Arabisch und Chinesisch tabellarisch angegeben. Wie erstellt, umfasst das Inventar der Transkriptionen die 26 lateinischen Grundbuchstaben in Majuskel sowie Minuskel, Binderstrich (z.B. in der japanischen Transkription für Langvokalmarkierung) und Apostroph (wie in der arabischen Transkription für [ʔ]). Da die für Inputcodierung eingesetzten Transkriptionen meistens nicht eins-zu-eins Schriftzeichen des

[268] Die Bezeichnung *Computer* ist ursprünglich im Amerikanisch-Englischen entstanden und wurde aus dem Lateinischen abgeleitet. Wegen der Internationalisierung wird diese Bezeichnung in den meisten Sprachen der Welt phonetisch entlehnt (Ausnahme sind das Arabische und Chinesische).

[269] Die Bezeichnung für *Tee* hat sich (zusammen mit dem Getränke) von China aus in anderen Ländern verbreitet, nämlich /te/aus dem Minnan-Dialekt (meistens in den europäischen Sprachen) und /cha/ auf Basis der nordchinesischen Aussprache (meistens in den Sprachen Asiens, des Nahen Ostens und Osteuropas) (vgl. *http://wals.info/chapter/138* [2017-05-08]).

[270] Der Punkt in der Mitte steht für Leerzeichen und der Pfeil für Tabulator.

[271] Die Inputcodierung des Russischen basiert an dieser Stelle auf ISO 9.

[272] Vgl. für die angewendete Arabisch-Transkription Tab. 2-21 (S. 116f).

[273] Vgl. für die angewendete Hindi-Transkription Tab. 2-13 & 2-14 (S. 92f).

[274] Vgl. für die angewendete Koreanisch-Transkription Tab. 2-17, 2-18 und 2-19 (S. 105ff).

[275] Die Inputcodierung des Japanischen basiert auf Romanji; *Computer* /konpyūtā/ wird mit sieben Zeichen geschrieben. Die Entsprechung lautet: コ/ko/, ン /n/, ピ /p/, ュ /yu/ (kleingeschrieben, um Silbenbildung mit vorderem Zeichen zu markieren); ー (markiert Langvokal; inputcodiert mit ‚-'; zweimaliger Auftritt) und タ/ta/.

[276] Die Inputcodierung des Chinesischen basiert auf Pinyin ohne Ton.

ursprünglichen Schriftsystems repräsentieren können, muss die Eingabesoftware per Worterkennung im Kontext die wahrscheinlichen Kandidaten anbieten. Manuelles Wahltreffen ist somit in den meisten Fällen obligatorisch.

In dem für diese multilinguale Eingabesoftware eingesetzten Tastaturlayout werden Apostroph und Bindestrich als Sondertasten ausgewählt und in den Abb. 5-4 und 5-5 blau gerahmt. Eine solche Sondertaste hat nach meinem Design spezielle Funktionen: Wenn sie in einem Modul der Nicht-Konversion (mit ‚NC' beendete Eingabeeinheit) eingetippt wird, wird das Zeichen direkt ausgegeben; sie wird hingegen als Teil des Inputcodes behandelt, wenn sie in einem Konversionsmodul (mit ‚CV' beendete Eingabeeinheit) eingetippt wird.

Gesetzt den Fall, dass eine gänzlich neue mechanische Tastatur für mein multilinguales Eingabesystem produziert würde, gälte es drei neue Tasten im Tastenblock hinzuzufügen. Da die Leertaste um ein vielfaches breiter als die anderen Tasten ist und bei der Anwendung des Eingabesystems häufig mit den drei neu designten Tasten zusammenarbeitet, kann eine Tastaturreformierung an der Stelle der Leertaste durchgeführt werden. Konkret gilt es die Breite der Leertaste zu reduzieren, um Platz für die drei neuen Tasten zu schaffen. Abb. 5-4 zeigt nachfolgend eine Skizze des alphanummerischen Tastenblocks.

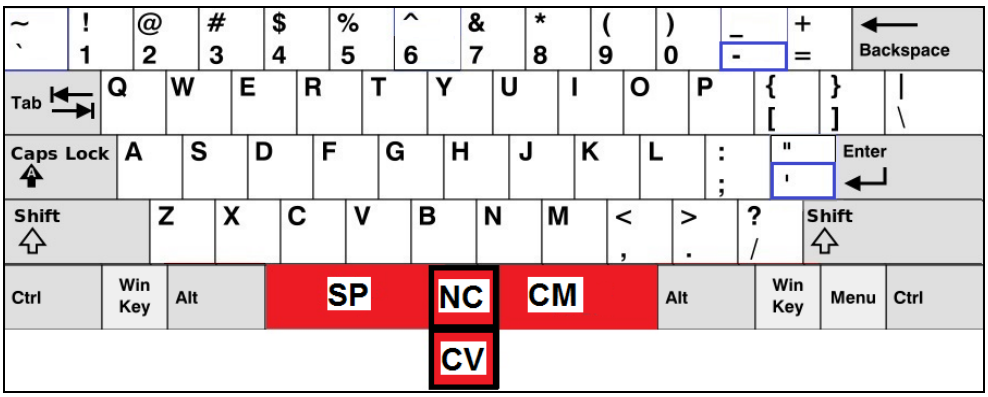


Abb. 5-4: Entwurf des alphanummerischen Tastenblocks mit neu hinzugefügten Tasten

Wie aus dem Design hervorgeht, wird ‚SP' mit dem linken Daumen kontrolliert. Gegenüber steht die ‚CM'-Taste, welche bei einem asiatischen Schriftsystem extrem häufig benutzt wird, bei europäischen Schriftsystemen jedoch unbenutzt bleibt. Zwischen den beiden befindet sich die ‚NC'-Taste, die sowohl mit dem linken als auch mit dem rechtem Daumen greifbar ist. Unterhalb wird die ‚CV'-Taste eigenständig in der neuen Tastenreihe belegt.

Da die Wahrscheinlichkeit für die Produktion einer solcher spezifischen Tastatur gering ist, habe ich ein alternatives Tastaturlayout entworfen, bei dem die drei vorhandenen Funktionstasten (die beiden Alt- und die rechte WinKey-Taste) jeweils zusätzlich die Leistungen von ‚NC', ‚CV' sowie ‚CM' übernehmen müssten:

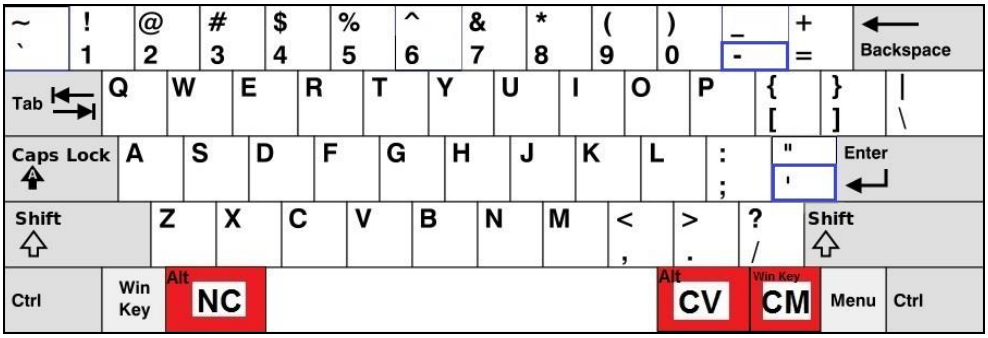


Abb. 5-5: Entwurf des alphanummerischen Tastenblocks ohne hinzugefügte Tasten

Der Hauptgrund für die Wahl der beiden ‚Alt'-Tasten liegt darin, dass sie kaum Ambiguitätskonkurrenzen bei der Anwendung mit ‚NC' sowie ‚CV' haben können. Im allgemeinen Fall muss eine ‚Alt'- immer mit einer (oder zwei) anderen Tasten gleichzeitig gedrückt werden, um ein alternatives Zeichen oder Befehle zu erzeugen. Im Gegenteil dazu muss ‚NC' oder ‚CV' nach einer Buchstabenfolge gedrückt werden, um eine Eingabeeinheit zu beenden und diese weiter vom System bearbeiten zu lassen. ‚Alt' verfügt nach meinem Entwurf deswegen über multiple Funktionen, neben ihrer ursprünglichen auch über die neu hinzugefügte Funktion von Nicht-Konversion sowie Konversion. Im Gegenteil dazu verliert die dritte ausgewählte Taste – die rechte ‚WinKey'-Taste – bei der Anwendung des multilingualen Eingabesystems ihre ursprüngliche Funktion und übernimmt komplett die ‚CM'-Funktion.

Neben der US-amerikanischen Tastatur kann auch eine andere lateinalphabetische europäische Tastatur Eingabegerät dieser Software sein. Wenn bspw. eine deutsche mechanische Tastatur zur Verfügung steht, muss unter ‚Tools' die Leiste für ‚Tastaturumstellung' geöffnet und die angewendete Tastatur gemäß des deutschen Layouts geändert werden. Die dem deutschen Schriftsystem zugehörigen Sonderbuchstaben sind in diesem Fall ohne Codierung eingebbar. Auch die mit Akut, Gravis und Zirkumflex gebildeten Sonderbuchstaben können durch einfache graphische Zerlegung transkribiert werden. Viele europäische Sprachen benötigen die Funktionstaste ‚CV' somit nicht.

5.2.2 Definition von Eingabeniveau und -fall sowie ihre Verarbeitungsprozesse

In Kap. 5.1.2 und 5.1.3 wurde die Lösungsvorschläge für Eingabeniveaus und -fälle skizziert, während im letzten Kapitel Beispiele für die Inputcodierung verschiedener Schriftsysteme gegeben wurden. In diesem Kapitel wird näher analysiert, wie Eingabeniveaus und -fälle nach der Art der Inputcodierung zu klassifizieren sind.

Das simple Niveau ist für die Eingabe von Sprachen, über die man Anfangskenntnisse besitzt und in der man nur einzelne Buchstaben, Grapheme sowie Zeichen schreiben möchte. Das mittlere Niveau ist für Benutzer geeignet, die den Grundwortschatz einer Sprache beherr-

schen. Das fortgeschrittene Niveau ist für die Eingabe von Sprachen angemessen, die man verhandlungssicher beherrscht und in denen man relativ häufig schreibt. Anders als bei dem allgemeinen Eingabeverfahren für alphabetische Schriftsysteme sind bei der Eingabe automatische Analysen und grammatische Korrekturen möglich.

Im Grunde genommen stehen das simple, mittlere und fortgeschrittene Niveau in einer Pyramidenstruktur: Je höher das Eingabeniveau einer Sprache ist, desto effektiver kann die Eingabe in dieser Sprache theoretisch sein. Gleichzeitig sind immer mehr Daten über das sprachliche Wissen sowie komplizierte Arbeitsschritte notwendig. Mögliche Überschneidungen verschiedener Sprachen sowie innerhalb von einer Sprache reduzieren sich mit der Niveauerhöhung mithilfe des erweiterten Kontextes. Wegen der Einschränkung der sprachlichen Datenbanken gibt es immer sprachliche Phänomene, die nicht bearbeitet werden können.

Anhand der eingesetzten Funktionstasten in einer Eingabeeinheit können die eingegebene Sprache sowie ihr Eingabeniveau nur in kleinerem Umfang bestimmt werden, was an dieser Stelle als Eingabefallbestimmung bezeichnet wird. Wie in Tab. 5-1 (S. 319) anhand von Beispielen erklärt wurde, sind der benötigte Schrifttyp und das Eingabeniveau (‚CM', ‚SP' und ‚CV'/‚NC') erkennbar, auf dessen Basis Eingabefälle kategorisiert werden. Im Folgenden werden solche Fälle in fünf Ober- sowie zehn Unterkategorien klassifiziert.

A) Der Fall der Nicht-Konversion. Wenn eine Eingabeeinheit mit ‚NC' endet, muss sie in einem lateinalphabetischen Schriftsystem formuliert werden und keinen Sonderbuchstaben beinhalten. Je nach ihrer Länge und den eingetippten Funktionstasten unterscheidet sich der Fall weiter in drei sprachliche Niveaus.

   A1) simples Niveau (kein ‚SP'/Interpunktion vorhanden);
   A2) mittleres Niveau (einmal ‚SP'/Interpunktion nach der Buchstabenfolge);
   A3) fortgeschrittenes Niveau (mehrmals ‚SP').

B) Der Fall der Konversion im simplen Niveau. Wenn eine Eingabeeinheit mit ‚CV' endet und in ihr kein ‚SP' (und ‚CM' maximal einmal) auftritt, ist sie im simplen Niveau einer Sprache. Dieser Fall kann nach ‚CM'-Einsatz in zwei Unterfälle kategorisiert werden.

   B1) In einem generellen alphabetischen Schriftsystem (kein ‚CM' vorhanden);
   B2) In einer indischen Schrift sowie Hangul oder im Chinesischen sowie Japanischen (endet mit ‚CM'): in der ersten Variante ist ein gruppiertes Syllabar die wesentliche schriftliche Einheit der Zeichen- sowie Glyphenausgabe; in der zweiten ist ein Schriftzeichen die kleinste segmentale Einheit.

C) Der Fall der Konversion in einem alphabetischen Schriftsystem im mittleren Niveau. Wenn eine Eingabeeinheit einmal mit ‚SP' nach einer Buchstabenfolge endet, muss sie einem

Wort aus einem alphabetischen Schriftsystem im mittleren Eingabeniveau entsprechen. Wie in B-Fall existieren auch bei ‚CM' zwei Unterfälle.

C1) In einem generellen alphabetischen Schriftsystem (kein ‚CM' vorhanden);
C2) In einer alphasyllabischen Schrift oder Hangul (‚CM' vorhanden).

D) Der Fall der Konversion in einem alphabetischen Schriftsystem im fortgeschrittenen Niveau. Wenn ‚SP' mehrmals in einer Eingabeeinheit eingetippt wird, muss sie im fortgeschrittenen Satzniveau verarbeitet werden, die in einem alphabetischen Schriftsystem formuliert wird. Erneut gilt es zwei Unterfälle zu unterscheiden.

D1) In einem generellen alphabetischen Schriftsystem (kein ‚CM' vorhanden);
D2) In einer alphasyllabischen Schrift oder Hangul (‚CM' vorhanden).

E) Im Fall der Konversion auf mittlerem oder fortgeschrittenem Niveau im Chinesischen oder Japanischen. Merkmale dafür sind das mehrmalige Auftreten von ‚CM' ohne ‚SP'. Die beiden Sprachen haben im Vergleich zu den meisten Schriftsystemen der Welt zwei entscheidende Unterschiede: Sie sind nicht alphabetisch und Leerzeichen zwischen den Wörtern sind ungebräuchlich. Ohne Leerzeichen sind Wort und Satz formell kaum zu unterscheiden, deswegen werden das mittlere und fortgeschrittene Eingabeniveau der beiden Sprachen im selben Eingabefall kategorisiert. Das simple Niveau wird dem Fall B2 zugeordnet, in dem ein einzelnes Schriftzeichen eingegeben wird. Wenn stattdessen eine Zeichenfolge eingegeben wird, müssen Spracherkennung und -analysen des sprachlichen Niveaus im nächsten Schritt mithilfe von Wörterbüchern durchgeführt werden.

Wenn sich solche Eingabefälle mit bestimmten Sprachen kombinieren lassen, hängen die folgenden Sprachen des simplen, mittleren und fortgeschrittenen Niveaus jeweils mit den folgenden Fällen zusammen:

Englisch: A1, A2 und A3;
Deutsch/Vietnamesisch: A1/B1, A2/C1 und A3/D1;
Russisch/Arabisch: B1, C1 und D1;
Hindi/Koreanisch: B2, C2 und D2;
Chinesisch/Japanisch: B2, E und E.

Nach der Bestimmung des Eingabefalls müssen die möglichen Eingabesprachen (eine oder mehrere) mit der Sprachliste des Eingabesystems im PC (siehe Abb. 5-3, S. 319) herausgefunden werden. Die sprachlichen Ressourcen der möglichen Sprachen werden dementsprechend abgerufen und das auf den Eingabefall bezogene Modul startet. Die konkreten Arbeitsschritte sind wie folgt.

A) Im A-Fall werden die eingetippten Informationen ohne Beteiligung des Konversionsmoduls ausgegeben, wie in Abb. 5-1 dargestellt. Aber die Korrektursysteme können (je nach den individuellen Bedürfnissen) mitinstalliert werden, um die Schreibfehler zu verringern, wenn die Eingabeeinheit im mittleren sowie fortgeschrittenen Niveau ist.

An dieser Stelle fungiert der folgende deutsche Satz mit falscher Konjugation als Beispiel: ‚*Er* SP *trinken* SP *Tee.* NC‘. Nach der Eingabefallerkennung wird auf die Wörterbücher aller möglichen Sprachen zugriffen, bzw. jenen Sprachen, die dem A-Fall zugehören und bei der Softwareoperation dem fortgeschrittenem Eingabeniveau zugerechnet werden. Nach der individuell eingestellten Reihenfolge der Sprachen im System wird die Eingabeeinheit mit den Wörterbüchern abgeglichen, um die Sprache und die eingegebenen Wörter zu erkennen. Nachdem diese Kette dem Deutschen zugeordnet wurde, werden die Wörter im grammatischen Kontext überprüft. Falls es Rechtschreib- oder Grammatikfehler gibt, bietet das System außer der original eingegebenen Version auch eine oder mehrere korrigierte Versionen an; für das Beispiel sowohl ‚*Er trinken Tee.*‘ als auch ‚*Er trinkt Tee.*‘. Der PC-Benutzer muss dann bestätigen, welche Variante dem Ziel entspricht.

B) Im B-Fall (simples Niveau) kann es (wie erwähnt) zu den meisten Überschneidungen unter den eingebbaren Sprachen kommen. Einerseits müssen alle eingebbaren Sprachen das simple Niveau erreichen, andererseits könnte ohne den Wort- oder Satzkontext ein Inputcode in jeder Sprache mehrere Kandidaten aufweisen. In diesem Fall müssen die Transkriptionslisten aller möglichen Sprachen (entweder im Fall von B1 oder B2) abgerufen und mit der Eingabeeinheit abgeglichen werden. Alle möglichen Kandidaten werden im nächsten Schritt in einer Wahlliste in einer bestimmten Reihenfolge dargestellt.

C) Die Wörterbücher der Sprachen, die in der Sprachliste das mittlere Niveau erreicht haben und im Fall C1 oder C2 behandelt werden müssen, werden abgerufen. Nach dem Abgleich wird die eingegebene Sprache erkannt und das Wort in der Wahlliste angeboten, wenn der Wortinputcode im Wörterbuch existiert.

Dieses Eingabeniveau ist von der Größe des Lexikons abhängig, d.h. die unregistrierten Wörter einer Sprache können nicht ausgegeben werden. Falls es in den Wahllisten keinen gewünschten Kandidaten gibt, so klickt man in der Wahllistezone dort doppelt, wo die gewünschte Sprache auftreten soll. Dann wird die einzugebende Sprache manuell bestimmt. Danach werden die Kandidaten nach dem Matching mit Transkriptionsliste und Konversionslexikon erneut abgerufen. Zuletzt wird das einzugebende Wort in einzelne Buchstaben / Syllabare / Zeichen zerlegt und schließlich ausgewählt.

D) Im fortgeschrittenen Niveau vom D-Fall kann die eingegebene Sprache in den meisten Fällen problemlos erkannt werden. Einerseits sind in der Regel nur wenige Sprachen, die im fortgeschrittenen Eingabeniveau festgelegt werden, einem bestimmten Eingabefall zugehörig. Andererseits tragen im Satzkontext Wörterbuch und grammatische Regeln zur Spracherkennung und Konversion bei. Nachdem eine Eingabeeinheit als D-Fall erkannt wurde, werden die Wörterbücher aller möglichen Sprachen (alle im fortgeschrittenen Eingabeniveau festgelegten Sprachen, außer Chinesisch und Japanisch) abgerufen. Wenn über die Hälfte der Wörter aus dem Inputcode im Wörterbuch einer Sprache recherchiert werden können, wird diese Eingabeeinheit als Formulierung derselben festgelegt. Danach wird die Wissensdatenbank mit den grammatischen Regeln der erkannten Sprache abgefragt. Anhand des Wörterbuchs und der grammatischen Regeln können sprachliche Analysen sowie die Konversion durchgeführt werden.

Wegen der imperfekten sprachlichen Fertigkeiten können häufig Fehler auftreten, weshalb die Korrektursysteme mit dem fortgeschrittenen Eingabeniveau kombiniert durchgeführt werden. Aufgrund der Begrenzung der erfassten Wörter sowie grammatischen Regeln usw. werden jedoch manche korrekt eingetippten Informationen als Fehlermeldung ausgegeben. Damit die Eingabe trotz Fehlermeldung störungslos abläuft, werden nach meinem Design die korrigierte, die originale und die lückenhafte Version gleichzeitig in der Wahlliste angezeigt. Diese Idee wird in Abb. 5-6 am Beispiel eines russischen Satzes visualisiert.

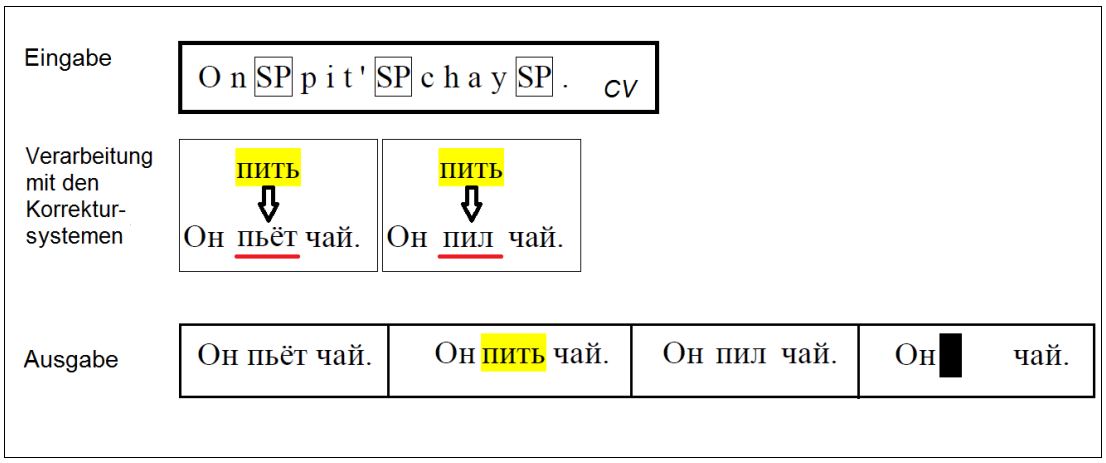


Abb. 5-6: Beispiel für den D-Eingabefall unter Beteiligung der Korrektursysteme

In der Abbildung wird der Inputcode ‚On SP pit' SP chay. CV' eingegeben. Durch die angewendeten Funktionstasten, die ausgewählten Sprachen des fortgeschrittenen Niveaus und das Wörterbuch wird diese Kette als Russisch erkannt. Nach dem Abgleich können die Wörter <он> (/on/, *er*) und <чай> (/chay/, *Tee*) problemlos konvertiert und ausgegeben werden. Die Transkription /pit'/ entspricht aber <пить> (der Infinitiv des Verbs *trinken*), das im Satz als

grammatisch unkorrekt gilt. Unter Beteiligung der Korrektursysteme wird es bevorzugt in die dritte Person Singular (Präsens sowie Präteritum) berichtigt: <пьёт> sowie <пил>. Die korrigierten Versionen werden dann als der erste sowie dritte Kandidat ausgegeben. Die mit einer Fehlmeldung markierte originale Version wird ebenso für den Fall angezeigt, dass sie mit dem Eingabeziel übereinstimmt. Zuletzt hat man die Möglichkeit, zunächst nur die beiden korrekten Wörter mit einer zu ergänzenden Lücke abzurufen (vierter Kandidat). Bei der Lücke kann der Inputcode, hier z.B. /pit'/, Buchstabe für Buchstabe konvertiert und manuell bestimmt werden, wie auch beim Prozess des simplen Niveaus, wenn der Inputcode korrekt ist. Falls er unkorrekt ist, kann der Wortinputcode erneut eingetippt werden.

E) Im E-Eingabefall sind sowohl polysyllabische Wörter als auch Sätze als Eingabeeinheit im Chinesischen oder Japanischen möglich. Nachdem ein Inputcode als E-Fall festgestellt wurde, müssen zuerst die Sprach- sowie Niveauerkennung durchgeführt werden, die auf der Basis der Zeichenaussprache und des Wörterbuchs der beiden Sprachen erfolgt. Wie in Kap. 3.4.2 dargelegt wurde, muss die Aussprache eines Zeichens im Chinesischen einer bestimmten Silbe entsprechen, die zu der Liste der ca. 415 Silbenvarianten gehören muss (vgl. Tab. 3-4, S. 169f). Im Gegenteil dazu kann die Aussprache eines Zeichens im japanischen Schriftsystem polysyllabisch (als Kanji), monosyllabisch (als Kanji oder Kana-Zeichen), ein Laut (wie das klein geschriebene < ㇷ゚ > /p/) oder ein Nulllaut (wie die Langvokalmarkierung <ー>) sein. Wenn die Eingabeeinheit zum fortgeschrittenen Niveau zählt, läuft die Wortsegmentation als Vorphase vor den Satzanalysen ab. Die konkreten Arbeitsschritte wurden in Kap. 4.5.1 zusammengefasst und in Abb. 4-28 (vgl. S. 286) versinnbildlicht. Der einzige Unterschied liegt in der Aussparung der Silbensegmentation, weil in diesem multilingualen Eingabesystem die Silbengrenze mit ‚CM' eindeutig markiert werden muss. Struktur und Arbeitsphasen der Konversionsmodule laufen schematisch wie folgt ab:

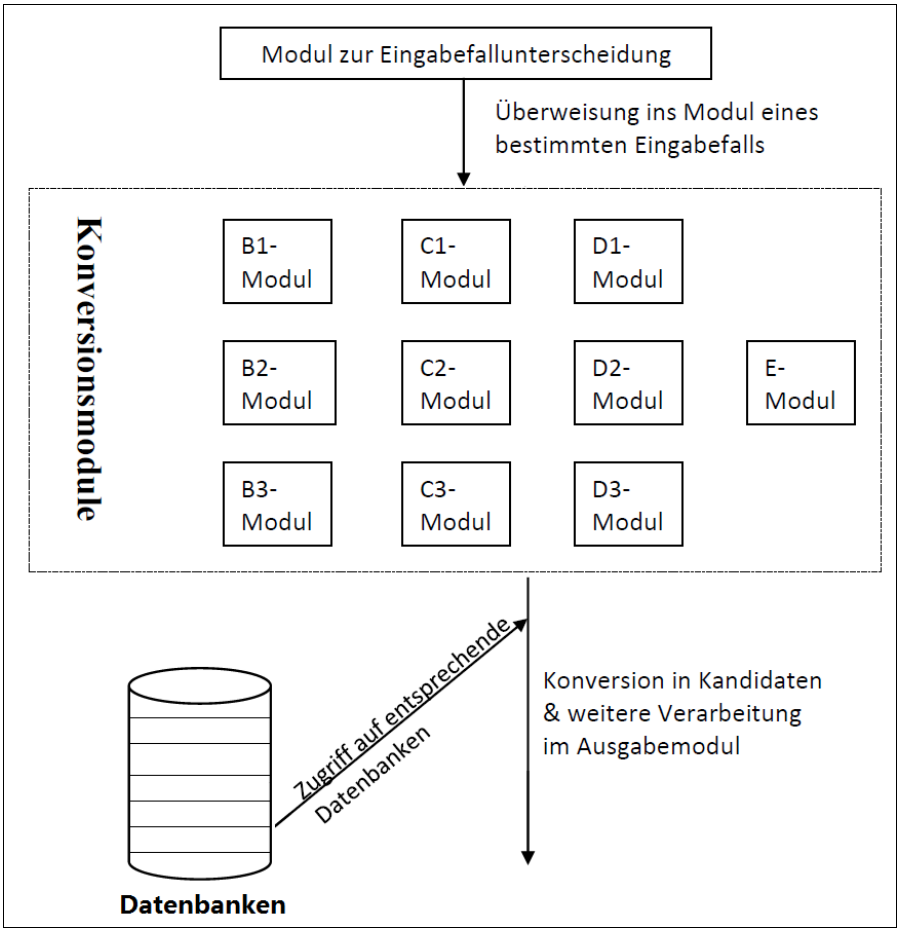


Abb. 5-7: Struktur und Verarbeitungsprozess der Konversionsmodule

Nachdem die möglichen Kandidaten abgerufen und erzeugt wurden, liegt der Schwerpunkt auf den Techniken zur Darstellung von Wahlliste sowie Kandidatenauswahl. Wie in Kap. 5.1.2 (S. 315f) erwähnt, wird die Zonenverteilung nach Schrifttyp sowie -richtung designet. Dieses Arbeitsprinzip gilt sowohl bei der Wahllistenanzeige im Bildschirm als auch beim Berührungsfeld zur Auswahl. Die Beziehungen zwischen der Zonenverteilung und dem Schriftsystem der gewünschten Sprache sind dabei wie folgt.

1. Erste Zone (rechtsläufiges Alphabet): Die auf dem lateinischen, kyrillischen sowie griechischen Alphabet basierten Schriftsysteme; sie befindet sich horizontal im unteren Bereich.
2. Zweite Zone (CJK-Schriften): Hier werden die drei ostasiatischen Schriftsysteme dem Chinesischen, Japanischen und Koreanischen zugeordnet, die sowohl horizontal als auch vertikal schreibbar sind und zum CJK-Schriftkreis zählen; sie befindet sich vertikal im rechten Bereich.
3. Dritte Zone (alphasyllabisch usw.): Zu der Zone gehören die alphasyllabischen und die anderen vertikal geschriebenen alphabetischen Schriftsysteme außerhalb des CJK-Kreises, wie Hindi, Tibetisch, Mongolisch usw.; sie befindet sich vertikal im linkem Bereich.

4. Vierte Zone (linksläufiges Alphabet): Sprachen im Konsonantenalphabet, die in linksläufiger Schriftrichtung geschrieben werden, gehören zu dieser Zone; sie befindet sich im oberen Bereich.

In der Phase der manuellen Kandidatenauswahl braucht der PC-Benutzer nur die Zone zu durchschauen, in der die gewünschten Schriftzeichen auftreten. Der gewünschte Kandidat kann per Berührungspunkt in der entsprechenden Zone im Berührungsfeld ausgewählt werden. Kap. 5.2.4 geht detaillierter auf die Techniken der Ausgabe und manuellen Auswahl ein.

5.2.3 Benötigte sprachliche Ressourcen

Die sprachlichen Ressourcen stellen (wie ausgeführt) die unentbehrlichen Datenbanken der multilingualen Eingabesoftware dar, um die Konversionsmodule zu unterstützen. Sie sind ebenso jene Elemente der Software, die am meisten Speicherplatz erfordern. Während oder nach der Softwareinstallation auf einen Computer muss das gewünschte Eingabeniveau einer zu schreibenden Sprache bestimmt werden, so dass die entsprechenden Ressourcen weiter installiert werden (siehe auch Kap. 5.1.2, S. 315).

Derlei Datenbanken einer Sprache können in drei Arten klassifiziert werden: Transkriptionsdatenbank, maschinenlesbares Wörterbuch und Wissensdatenbank mit sprachlichen Regeln. Auf simplem Niveau ist nur die Transkriptionsdatenbank gebräuchlich, beim mittleren Niveau ist ein Wörterbuch zusätzlich vonnöten und beim fortgeschrittenen Niveau sind alle drei Arten von Datenbanken obligatorisch.

In der Transkriptionsdatenbank wird die Transkription einer bestimmten Sprache in lateinischen Grundbuchstaben erfasst. Ihre Struktur und Größe sind von den grammatologischen Eigenschaften sowie dem Zeicheninventar eines Schriftsystems abhängig. Bei einem lateinalphabetischen Schriftsystem ist eine Liste der Sonderbuchstabenumschrift erforderlich. In der Transkriptionsdatenbank eines nichtlateinischen alphabetischen Schriftsystems muss die Umschrift von jedem Buchstaben sowie phonemtragenden Graphem erfasst werden. Die Transkription einer in alphasyllabischer oder sonderalphabetischer Schrift geschriebenen Sprache muss mindestens die allgemeingebräuchlichen Silbenblöcke umfassen.[277] In der Transkriptionsdatenbank des Chinesischen und Japanischen müssen tausende allgemeingebräuchliche Schriftzeichen und Wörter indexiert enthalten sein.

Die Eingabe des mittleren und fortgeschrittenen Niveaus muss via Wörterabgleich unterstützt werden, wozu das Lexikon (für beide Niveaus) und die Datenbank der Grammatiken (nur für das fortgeschrittene Niveau) einer Sprache als obligatorische Ressource zur Verfü-

[277] Sonderalphabete sind z.B. Hangul oder die mongolische Schrift.

gung stehen müssen. Da solche Ressourcen ebenso unentbehrlich für die Funktion der ‚Rechtschreib- und Grammatikprüfung' von Textprogrammen wie ‚Microsoft Office Word' sind, können sie via Filesharing für die Eingabe und Korrektursysteme einer bestimmten Sprache (außer Chinesisch) fungieren. Um die Effizienz des Filesharings und des Schreibens zu erhöhen, müssten kombinierte Anwendungen von Eingabe- und Korrekturprozessen ersonnen werden. Wenn die Eingabeeinheit Wort oder Satz ist, so wird sie mit den Wort- sowie Satzanalysen verarbeitet und in die korrekten Schriftzeichen konvertiert. Wort- sowie Grammatikfehler werden ebenfalls überprüft. Auch beim Schreiben eines lateinalphabetischen Schriftsystems, dessen Eingabeeinheit meistens mit ‚NC' endet, können Fehlermeldungen und korrigierte Vorschläge angezeigt werden. Der PC-Benutzer bestimmt weiter, ob die ursprüngliche oder die vorgeschlagene Version ausgegeben wird. Wie in Abb. 5-6 dargestellt, werden Ergebnisse der Spracherkennung und Korrekturvorschläge angeboten, wenn ein Fehler bei einer mit ‚CV' endenden Eingabeeinheit gefunden wird. Wenn der zum Schreibziel geeignete Satz wegen der Datenbankeinschränkung nicht automatisch erzeugt werden kann, ist eine Rückkehr zum mittleren oder simplen Niveau (eine Wort/Zeichen-für-Wort/Zeichen-Bestimmung) nötig. Dies entspricht dem Schritt der ‚erneuten Konversion', der in Abb. 5-1 visualisiert wird.

Das im Eingabesystem angewendete Lexikon einer Sprache enthält eine einfache und eine komplizierte Version. Auf dem mittleren Niveau ist das Wörterbuch hauptsächlich für die Sprach- sowie Worterkennung zuständig. Für dieses Ziel reicht eine Wörterliste, in der die gemeingebräuchlichen Wörter in originalen sowie flektierenden Wortformen eingetragen werden, ergo die einfache Version des Wörterbuchs. Im Gegensatz dazu sind Informationen über Wortattribute (wie Wortart, Grammatikalitäts- sowie Semantikbedingungen) für die Satzverarbeitung im fortgeschrittenen Niveau obligatorisch. Anders formuliert muss das Wörterbuch in diesem Fall zu einer komplizierten Version erweitert werden, so dass sprachliche Analysen zum Zweck der präziseren Konversion und der automatischen Korrektur durchgeführt werden können.

In Kap. 1.4.5 (S. 57) wurde erwähnt, dass wegen der Besonderheiten des Sprach- sowie Schriftaufbaus kein Korrektursystem des Chinesischen in Textverarbeitungsprogrammen besonders verbreitet ist. Dazu zählen (in Anlehnung an Kapitel 3 und 4) der morphologische Schrifttyp, der isolierende Sprachbau ohne Wortflexion, die Abstinenz von Leerzeichen für Wortmarkierungen, komplizierte Beziehungen zwischen Wortarten und Satzgliedern, die entscheidende Rolle der Semantik bei der Formulierung usw. Aufgrund der Tatsache, dass es kein ausgereiftes Korrektursystem des Chinesischen gibt, ist Filesharing in diesem Fall unmöglich. Der erforderliche Speicher des Eingabesystems weicht aufgrund der zu verarbeiten-

den Sprachen daher stark ab. Wenn Chinesisch als dem fortgeschrittenen Eingabeniveau zugehörig definiert wird, wird der Speicher stark belastet. Um die nötigen Ressourcen des multilingualen Eingabesystems konkret vorzustellen, sollen sie exemplarisch anhand von neun verschiedenen Sprachen sowie Schriftsystemen tabellarisch erklärt werden.

Sprache	**Simpel (Buchstabe, Graphem, Syllabar, Zeichen)**	**Mittel (Wort)**	**Fortgeschritten (Phrase, Satz)**	**Eingabefall & Anmerkungen**
Englisch		Lexikon (Filesharing mit Korrektursystemen)	Lexikon, Datenbank der grammatischen Regeln sowie Statistiken (Filesharing)	A1, A2 & A3
Deutsch	Umschriftliste (nur Sonderbuchstaben)	Umschriftliste, Lexikon (Filesharing)	Umschriftliste, Lexikon, Datenbank der grammatischen Regeln sowie Statistiken (Filesharing)	A1/B1, A2/C1 & A3/D1
Vietnamesisch	wie oben	wie oben	wie oben	wie oben
Russisch	Transkriptionsdatenbank (inkl. Buchstaben und Graphemen)	Transkriptionsdatenbank und Lexikon (inkl. Wortform in Inputcode; Filesharing)	Transkriptionsdatenbank, Lexikon und Datenbank der grammatischen Regeln sowie Statistiken (Filesharing)	B1, C1 & D1
Arabisch	Transkriptionsdatenbank (inkl. aller Buchstaben, diakritischen Zeichen und festen Ligaturen)	wie oben	wie oben	B1, C1 & D1; Verarbeitung in linksläufiger Textrichtung
Hindi	Transkriptionsdatenbank (inkl. Grundbuchstaben und abhängigen Vokale)	wie oben	wie oben	B2, C2 & D2
Koreanisch	Transkriptionsdatenbank in RRK (inkl. allen Jamo und syllabischen Blöcken)	wie oben	wie oben	B2, C2 & D2; zielorientierte Installation des Programms der Hangul-Hanja-Konversion
Japanisch	Transkriptionsdatenbank in Romanji (inkl. Kana-Zeichen und allgemeingebräuchlichen Kanji)	Lexikon (inkl. häufig festgelegter Phrasen; Filesharing)	Transkriptionsdatenbank, Lexikon, Datenbank der Grammatiken, Semantiken und Statistiken usw. (teils zum Filesharing)	E (‚CM‘ nach jedem Kana- und Kanji-Zeichen)
Chinesisch	Transkriptionsdatenbank in Pinyin (mindestens inkl. der allgemeingebräuchlichen Schriftzeichen)	Lexikon (inkl. festgelegten Phrasen, Affixen, Idiomen usw.)	Transkriptionsdatenbank, Lexikon, Bank der Grammatiken, Semantiken und Statistiken usw. (nicht zum Filesharing)	E; zusätzliche Funktionen: eine andere Eingabemöglichkeit für unbekannte Zeichen, selbstangepasstes Lernen usw.

Tab. 5-2: Die benötigten Ressourcen einer Sprache im multilingualen Eingabesystem

5.2.4 Techniken zur Kandidatenausgabe und -auswahl

Wie in Kap. 5.1.2 und 5.1.3 grob vorstrukturiert und designet wurde, ist eine Zonenverteilung zur Kandidatenausgabe und -auswahl nötig, damit die Effizienz trotz der hohen Kandidatenüberschneidungen erhöht werden kann. Um dies zu realisieren, muss einerseits das Ausgabemodul neu programmiert, andererseits spezielle Hardware entwickelt oder bereits bestehende reformiert werden. Dieses Kapitel fokussiert diese Probleme.

Da das multilinguale Eingabesystem und ein Textverarbeitungsprogramm gleichzeitig ablaufen müssen, treten die Fenster der beiden Softwares zusammen auf. Nach meinem Plan steht das Fenster des Textprogramms im Zentrum und das *Look and Feel* des Eingabesystems steht verkleinert unten rechts innerhalb des Textprogrammfensters. An vier Rändern werden die vier Zonen der Kandidatenliste angezeigt: der untere Rand für die erste Zone, der rechte für die zweite, der linke für die dritte und der obere für die vierte. In Abb. 5-8 bis 5-10 werden die Wahllisten am Beispiel eines bestimmten Inputcodes abgebildet.

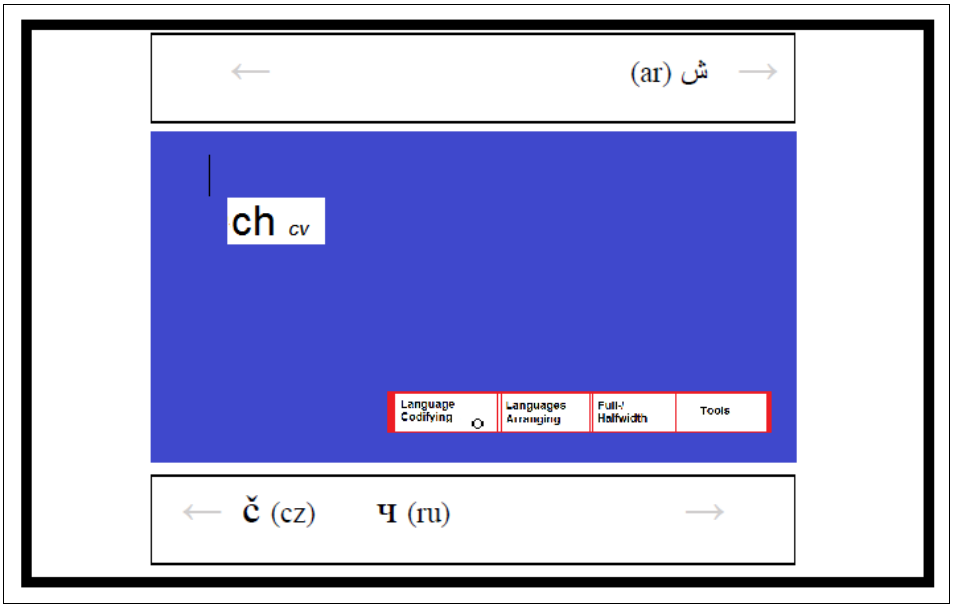


Abb. 5-8: B1-Eingabefall am Beispiel von ‚ch' mit Ausgabe in der ersten und vierten Zone[278]

[278] Der blaue Hintergrund steht für das Fenster des Textprogramms. Der weiße Kasten symbolisiert das Eingabefeld. Das Fenster mit rotem Rahmen und weißem Inhalt steht für das *Look and Feel* des Eingabesystems. An den vier Rändern des Bildschirms (normalerweise nur zwei davon besetzt) sind die Wahllisten. Jede Wahlseite bietet höchstens vier Kandidaten an und den Pfeil zum Scrollen. Die Farbe des Pfeils (hell oder schwarz) zeigt, ob es Kandidaten auf anderen Seiten gibt.

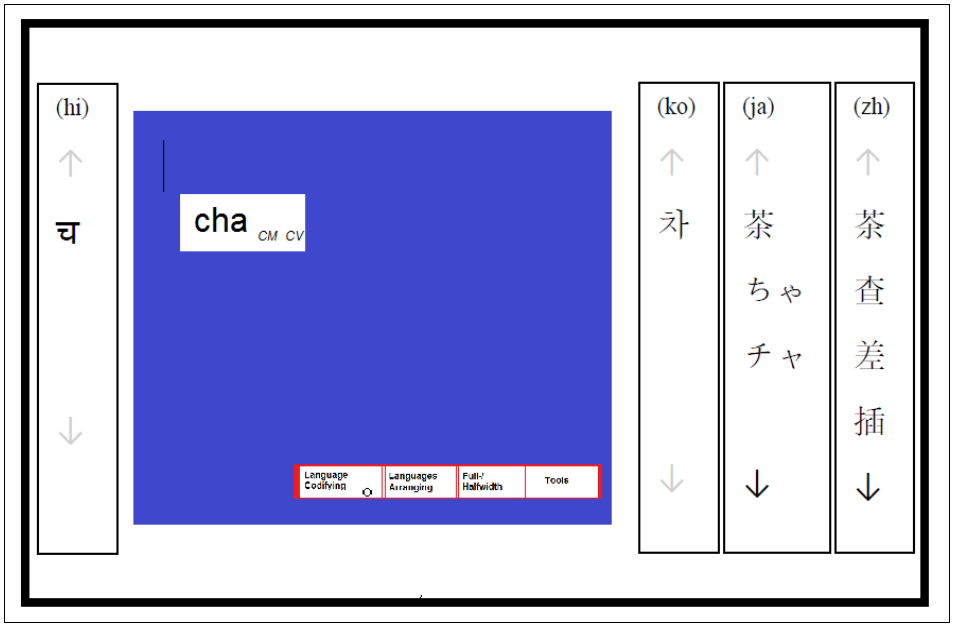


Abb. 5-9: B2-Eingabefall am Beispiel von ‚cha' mit Ausgabe in der zweiten und dritten Zone

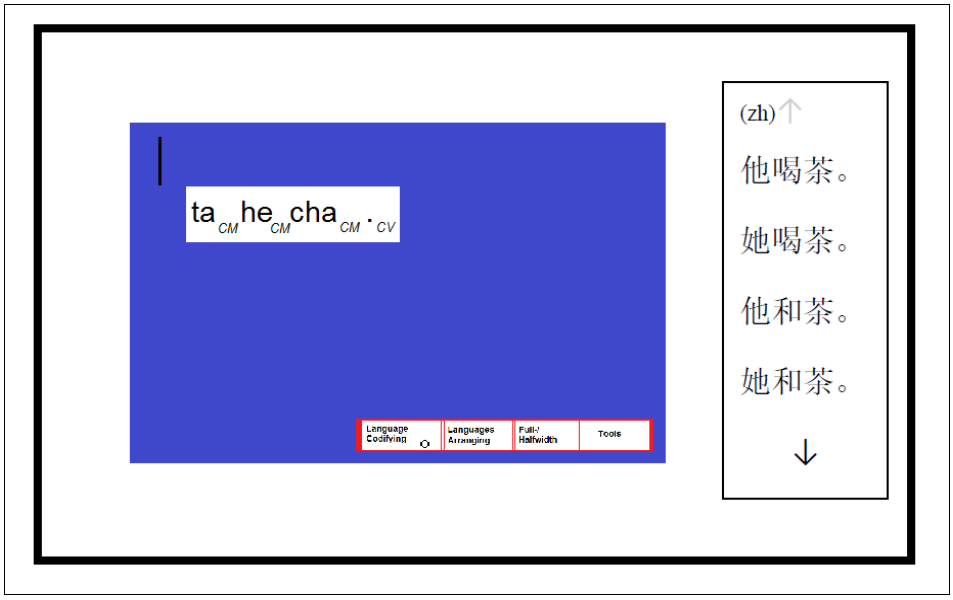


Abb. 5-10: E-Eingabefall am Beispiel von ‚ta he cha.' mit Ausgabe in zweiter Zone

Die Verarbeitungsphasen werden in Kap. 5.1.3 erklärt (vgl. S. 316f). Die Darstellungsart der Wahllisten in solchen Fällen muss ebenso designet werden. In Abb. 5-8 gibt es nur einige Kandidaten aus verschiedenen Sprachen, die jeweils in der ersten sowie vierten Zone dargestellt werden. Das Beispiel in Abb. 5-9 hat hingegen mehrere Kandidaten auch innerhalb eines bestimmten Schriftsystems, bedingt von den homophonetischen Eigenschaften der drei ostasiatischen Sprachen. Damit die Kandidatenausgabe trotzdem visuell auffällig ist und ihre Auswahl effektiv abläuft, werden die Kandidaten der zweiten Zone nach Sprachen geteilt. Bei der Auswahl muss zuerst die gewünschte Sprache bestimmt werden, was entweder durch die Cursortasten (Taste ←/→) oder durch das Zugreifen in der ersten Zone des Touchpads geschieht,

zu der keine Kandidaten zugehörig sind. Die Kandidaten der ersten und vierten Zone müssen horizontal angeordnet werden, während die der zweiten und vierten Zone vertikal dargestellt werden. Durch die Anwendung der vier Funktionstasten bedingt, können höchstens zwei der vier Zonen gleichzeitig besetzt sein, entweder oben und unten oder links und rechts (mit Ausnahme des Falls, dass klassisches Mongolisch die Eingabesprache ist). Die Größe einer Zone wird nach der Menge der betroffenen Sprachen sowie Kandidaten automatisch umgestellt. Auf einer Seite werden höchstens vier Kandidaten und die Scrollfläche auf der vorherigen sowie nächsten Seite gleichzeitig angezeigt. Dies ist durch das selbstdefinierte ‚Fünf-Finger-Berührungsprinzip' einer Hand bei der Kandidatenauswahl bedingt. Unter ‚Tools' des *Look and Feel* der Software können die angezeigten Sprachen und Positionen der Zonen individuell orientiert umgestellt werden.

Wenn klassisches Mongolisch die Eingabesprache ist, gibt es einige Ausnahmen anzumerken.[279] Wegen seiner vertikalen Schriftrichtung wird es für die dritte Zone geplant: Anders als die anderen Sprachen aus der zweiten Zone, deren Kandidaten horizontal verfasst und vertikal angeordnet werden (siehe Abb. 5-7 & 5-8), muss ein mongolischer Ausdruck vertikal geschrieben werden. Die Kandidaten lassen sich deswegen horizontal links-rechts anordnen, falls es mehr als eine Möglichkeit gibt. Aufgrund des Sondertasteneinsatzes sind Verwechslungen mit anderen Sprachen der zweiten und dritten Zone ausgeschlossen. So kann eine Richtungskonkurrenz vermieden werden. (Vgl. zu Grundeigenschaften des Mongolischen Kap. 1.3.1 und 1.3.6, S. 23 & S. 37.)

Den vier Zonen auf dem Bildschirm entsprechend wird das Touchpad in vier Berührungszonen unterteilt, wobei jede Zone sechs Berührungspunkte besitzt. Eine Zone wird mit horizontaler oder vertikaler Handhaltung sowie linker oder rechter Hand kontrolliert. Die fünf Finger einer Hand sind jeweils für die vier Kandidaten sowie das Scrollen nach oben/unten zuständig. Das ‚Fünf-Finger-Prinzip' entspricht den Wahllisten-Finger-Beziehungen. Konkreter erläutert muss für die erste und vierte Zone die linke sowie die rechte Hand horizontal aufgelegt werden, um schneller zuzugreifen. Für die zweite und dritte Zone hingegen ist eine vertikale Handhaltung vonnöten. Mit den Daumen kann die Seite nach unten gescrollt werden. Falls man den richtigen Kandidaten verpasst hat, kann mit dem kleinen Finger nach oben gescrollt werden. Der erste bis vierte Kandidat ist jeweils mit dem kleinen Finger, Ringfinger, Mittelfinger und Zeigefinger zu berühren. Abb. 5-11 visualisiert dieses Prinzip.

[279] Das in der mongolischen Schrift geschriebene Mongolisch in der Inneren Mongolei, VR China, wird klassisches Mongolisch genannt.

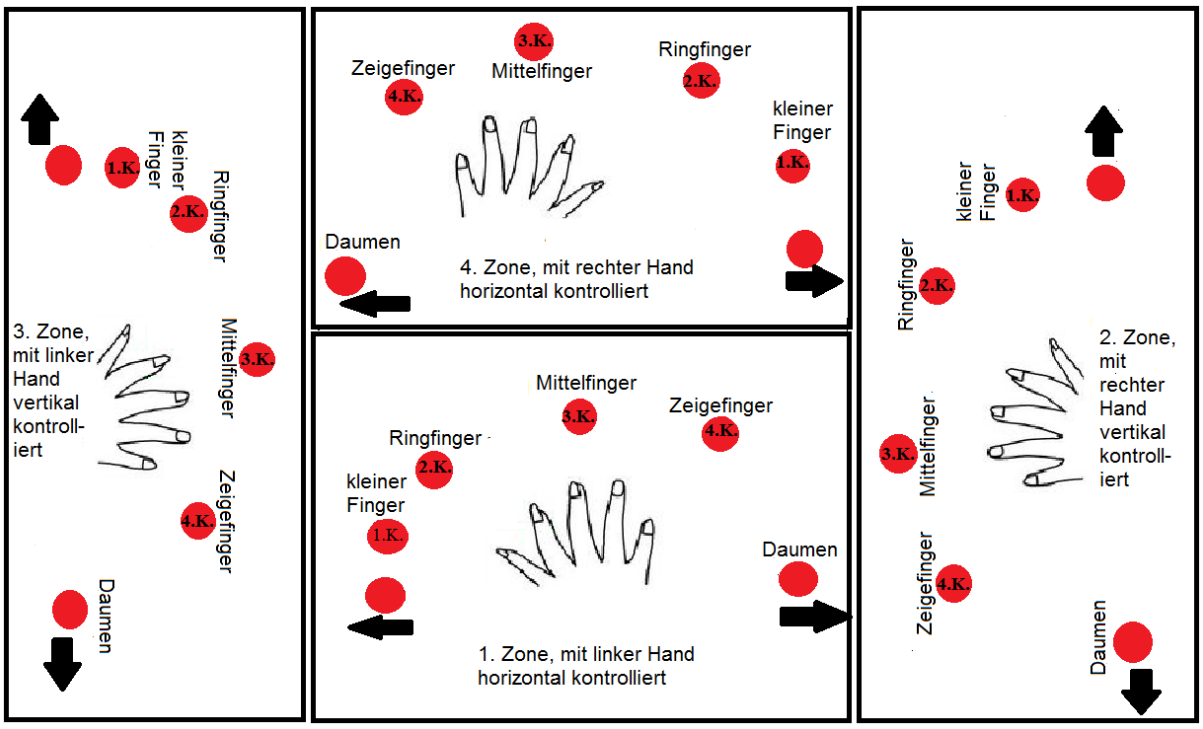


Abb. 5-11: Das Design des Touchpads für die Kandidatenauswahl

Um diesen Entwurf zu realisieren, müsste eine spezielle Hardware erdacht werden, die im Rahmen dieser Arbeit als ‚Wählgerät des multilingualen Eingabesystems' bezeichnet wird. Das Wählgerät ist im Prinzip ein spezielles Touchpad, das sich unterhalb von der Tastatur befindet oder eigenständig mit USB-Kabel an den PC angeschlossen ist.

Im Allgemeinen gelten Touchpads als Ersatz für die Maus, auf dem sich „der Mausanzeiger durch Fingerdruck auf einer berührungsempfindlichen Fläche bewegt" (Greulich 2003: 904). Nach Abb. 5-11 muss ein Wählgerät auf der Basis eines Touchpads so reformiert werden, dass eine Teilung in vier Zonen und eine Markierung von 24 Berührungspunkten (sechs Punkte pro Zone) erfolgt. Im Vergleich zu normalen Touchpads funktioniert das Wählgerät eher ähnlich wie eine Berührungsbildschirmtastatur. Durch den Druck eines Berührungspunkts wird entweder ein Kandidat ausgewählt oder die Wahllistenseite einer Zone gescrollt. D.h. jeder Berührungspunkt ist nur für einen bestimmten Kandidaten oder die Scrollfunktion zuständig, nach einmaligem Drücken der Kandidatenauswahl. Zur Kandidatenauswahl wird die Hand horizontal oder vertikal leicht gebogen und mit ausgestrecktem Finger gehalten. Nach Analysen verschiedener Handgrößen lässt sich die minimale Größe der Wählgerätfläche auf 125*180mm bestimmen. Damit die Punkte schnell und präzise gefunden werden können, können die Grenzen der Zone und die aktiven Berührungspunkte markiert werden, wie die folgende Abbildung zeigt:

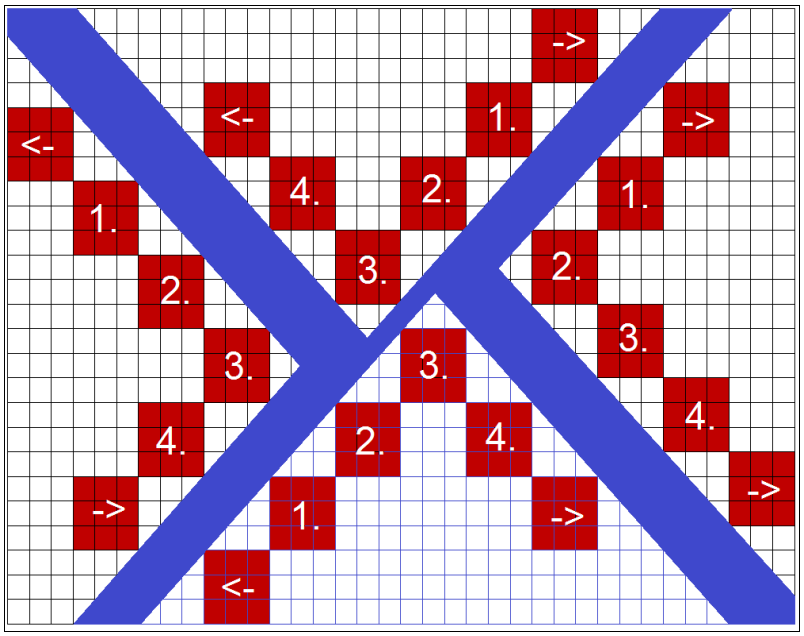


Abb. 5-12: Das Aussehen des benötigten speziellen Touchpads

Da das Wählgerät über das multilinguale Eingabesystem hinaus kaum anderen Zwecken dient, ist eine Produktion unrealistisch. Damit das Eingabesystem ohne Hardwareeinschränkung benutzt werden kann, habe ich eine Tastenersatzmöglichkeit geplant.

Die mit Nummern belegten Tasten aus dem alphanummerischen sowie nummerischen Block vertreten in diesem Fall die Berührungspunkte. Für die horizontal dargestellte erste sowie vierte Zone steht die nummerische Reihe des alphanummerischen Blocks zur Verfügung, die ebenso komplett horizontal angeordnet wird. Der Kandidaten-Tasten-Zusammenhang stimmt mit der Schriftrichtung überein. Der erste bis vierte Kandidat der ersten Zone wird daher mit den Tasten ‚1' bis ‚4' kontrolliert, dem hingegen werden die Kandidaten der vierten Zone von rechts nach links jeweils durch die Tasten ‚0' bis ‚7' ausgewählt. Der auf der Tastatur rechts zu findende nummerische Block ist besser für die zweite sowie dritte Zone geeignet. So entsprechen die vier Tasten der rechten Spalte von oben nach unten (‚9', ‚6', ‚3' und Del-Taste) den vier Kandidaten der zweiten Zone. Demgegenüber repräsentieren die vier Tasten der linken Spalte (‚7', ‚4', ‚1' und ‚0') die Kandidaten der dritten Zone. In der Nähe dieser Tasten werden zwei weitere Tasten festgelegt, die für das Scrollen nach oben/unten zuständig sind. Der Tastenersatzplan wird in Abb. 5-13 und Tab. 5-3 angegeben, wobei die Tasten für die erste, zweite, dritte und vierte Zone jeweils schwarz, rot, grün und orange markiert werden:

Abb. 5-13: Tastenersatz für das Wählgerät des multilingualen Eingabesystems

	oben	**1. Kan.**	**2. Kan.**	**3. Kan.**	**4. Kan.**	**unten**
1. Zone (schwarz)	#01; ‚`‘	#02; ‚1‘	#03; ‚2‘	#04; ‚3‘	#05; ‚4‘	#06; ‚5‘
2. Zone (rot)	#96; ‚8‘	#101; ‚9‘	#102; ‚6‘	#103; ‚3‘	#104; ‚Del‘	#98; ‚2‘
3. Zone (grün)	#83; ↑	#91; ‚7‘	#92; ‚4‘	#93; ‚1‘	#99; ‚0‘	#84; ↓
4. Zone (orange)	#12; ‚-‘	#11; ‚0‘	#10; ‚9‘	#09; ‚8‘	#08; ‚7‘	#07; ‚6‘

Tab. 5-3: Plan für den Tastenersatz zur Kandidatenauswahl

Mit der Leiste ‚Option‘ (die sich unter den ‚Tools‘ von *Look and Feel* der Software befindet) kann zwischen Wählgerät und Tastenersatz geswitcht werden, je nachdem, welche Hardware zur Verfügung steht. Die Tasten zur Kandidatenauswahl jeder Zone können auch individuell definiert werden.

5.2.5 Verarbeitung der Sonderzeichen

In Kap. 5.1.3, 5.2.2 und 5.2.4 wurde die Verarbeitung der schriftlichen Informationen in verschiedenen Schriftsystemen vorgestellt, welche die Eingabe sowie die systemgestützte Informationsverarbeitung in verschiedenen Eingabefällen und Kandidatenausgaben bzw. -auswahlen beinhaltet. Neben den von derlei Schriftzeichen übertragenen Informationen sind die Zahl-, Interpunktions- und sonstigen Sonderzeichen weitere wichtige Elemente der Schrift, deren Eingabe berücksichtigt werden muss.

Identisch wie in allgemeinen Fällen ist die Eingabe solcher Zeichen davon abhängig, ob sie auf dem US-amerikanischen Layout belegt werden (ihr Inputcode ist in diesem Fall ‚NC‘) oder nicht (‚CV‘). Um die Verarbeitung der sonstigen Zeichen zu analysieren, werden drei Fragen gestellt: A) Welche Zeichen können zusammen zu einer Eingabeeinheit gehören? B) Wie kann ein solches Zeichen inputcodiert werden, wenn es nicht auf der Tastatur belegt wird? C) In der Wahlliste von welcher Zone wird es ausgegeben und ausgewählt? Da diese Schriftzeichen unterschiedlichen Typen angehören, werden sie in vier Klassen unterteilt, wobei die Fragen für jede Klasse separat beantwortet werden:

1. Die erste Gruppe bezieht sich vor allem auf international verbreitete fachliche Schriftzeichen in mathematischen, naturwissenschaftlichen, geisteswissenschaftlichen, technischen

Bereichen usw., wie <‰> (per mille; Promille) oder <♂> (male; männlich). Solche Schriftzeichen sind regional unabhängig unter Fachleuten besonders häufig gebräuchlich.

Nach meinem Entwurf muss ein Schriftzeichen dieser Art immer als eigenständige Eingabeeinheit verarbeitet werden. Die Inputcodierung dazu kann entweder auf seiner international verbreiteten englischen Bezeichnung oder der graphischen Zerlegung basieren. Außer der Konversionstaste ‚CV' darf keine andere Funktionstaste beteiligt sein, egal in welchem Kontext oder welcher Sprache es vorkommt. Bspw. kann der Inputcode für <‰> sowohl ‚%0CV'[280] als auch ‚permilleCV'[281] sein. Ebenso unabhängig vom sprachlichen Kontext wird das Zeichen immer in der ersten Zone ausgegeben. Wenn es im Arabischen vorkommt, seine Schriftrichtung und manchmal seine Glyphen also umgekehrt dargestellt werden müssen, wird zuerst die allgemeine Form ausgewählt. Das System stellt danach Richtung und Glyphen automatisch um.

2. Die zweite Gruppe orientiert sich an solchen Interpunktionszeichen, die trotz derselben Bezeichnung und vergleichbaren Gebrauchssituationen in verschiedenen Schriftsystemen anders dargestellt werden. Typische Beispiele dafür sind Anführungszeichen (in Deutsch „…", in Englisch "…", in Französisch « … » und im [traditionellen] Chinesisch/Japanisch 「…」)[282], Punkt (in einem lateinbasierten Schriftsystem meistens ‚.' und in Chinesisch/Japanisch ‚。') und Ellipse (in einem alphabetischen Schriftsystem ‚…' [mit drei Punkten] und in Chinesisch/Japanisch ‚⋯⋯' [mit sechs Punkten]). Neben solchen Einzelfällen bei bestimmten Zeichen gibt es in bestimmten Schriften zudem allgemeine Regeln, etwa die voreingestellte Zeichendarstellung in Vollbreite im CJK-Schriftkreis und die horizontal widergespiegelte Zeichendarstellung in der arabischen Schrift. Die verschiedenen Glyphen desselben Sonderzeichens in verschiedenen Schriftsystemen können als Varianten angesehen werden.

Ein Interpunktionszeichen kann nach einem Wort/Satz sowohl als zusammengehörige Eingabeeinheit eingetippt werden, als auch eine eigenständige Eingabeeinheit darstellen. Voraussetzung hierfür ist jedoch, dass Buchstabenkette und Interpunktionszeichen beide mit CV oder NC enden. Wie in Abb. 5-10 dargestellt wurde, werden sie gemeinsam in mögliche Kandidaten konvertiert und ausgegeben. Falls Konkurrenzen zwischen CV oder NC auftreten, muss das Interpunktionszeichen eigenständig verarbeitet werden.

[280] Bei der Inputcodierung muss das Maximalprinzip gelten, die graphische Zerlegung in möglichst wenige Zeichen.
[281] Die englische Bezeichnung des Zeichens lautet *per mille*. Bei der Inputcodierung wird das Leerzeichen zwischen den beiden Wörtern weggelassen, damit das System die Bezeichnung besser als einheitlich erkennen und verarbeiten kann.
[282] Zwischen den öffnenden und schließenden Anführungszeichen steht das Auslassungszeichen für den geschlossenen Inhalt.

Bei der Verarbeitung eines solchen Interpunktionszeichens muss zuerst betrachtet werden, ob die gewünschte Glyphe auf einer Taste belegt ist. Wenn ja, so wird es ohne Konversion eingetippt. Wenn nicht, muss es mit der belegten variierten Zeichenform oder Bezeichnung inputcodiert und mit ‚CV' markiert werden. Nur Anführungszeichen stellen eine Ausnahme dar, da sie trotz graphischer Unterschiede (der öffnenden und schließenden Form) auf derselben Taste belegt werden. Wird ein Anführungszeichen eingetippt, so muss das System der eingegebenen Sprache entsprechend die Glyphe umstellen, egal ob die Eingabe mit ‚CV' oder ‚NC' endet.

Wenn das Interpunktionszeichen unbelegt ist, muss es durch graphische Zerlegung oder seine kurze Bezeichnung auf der zu schreibende Sprache inputcodiert werden. Bspw. kann das Auslassungszeichen durch die drei- oder sechsmalige Wiederholung des Punkts erkannt werden. Im deutschen Kontext kann es auch als ‚Auslassung' und im chinesischen als ‚slh' (Halbpinyin von ‚shengluehao') inputcodiert werden.

Damit die Eingabe solcher Zeichen problemlos funktioniert, muss die Glyphe nach der Spracherkennung vom System automatisch umgestellt werden. Bevorzugt wird dies nach der Sprache des Inputcodes bestimmt, wenn die Bezeichnung in der zu schreibenden Sprache dem Inputcode entspräche oder das Interpunktionszeichen mit einer sprachlichen Eingabeeinheit zusammen eingegeben würde. Falls die Spracherkennung innerhalb der Eingabeeinheit unmöglich ist, muss sie nach dem Kontext, vor allem dem Vorkontext, erkannt werden.

Die Ausgabezone solcher Interpunktionszeichen ist von dem Ergebnis der Spracherkennung bedingt. Sie werden demgemäß in der Wahlzone angegeben, zu der die erkannte Sprache zugehörig sein sollte. Falls die Sprache aus verschiedenen Gründen nicht erkannt werden kann, muss der PC-Benutzer in der Zone, in der der Kandidat auftreten soll, doppelt klicken. Das System bietet danach die korrekte Glyphe in der angeklickten Zone.

3. Die dritte Gruppe sind Währungszeichen, die trotz internationaler Anerkennung nur in bestimmten Regionen häufig benutzt werden, wie <¥> (für chinesisches *Yuan* [元 /yuán/] und japanisches *Yen* [円 /en/]), <₽> (für russisches *Rubel;* eng.: ruble; rus.: рубль /rubl′/) und <₩> (für süd- sowie nordkoreanisches *Won*; kor.: 원 /won/).

Die englische Bezeichnung für solche Zeichen ist meistens an die ursprüngliche Sprache angelehnt. Situationsbedingt aber gibt es bisweilen Unterschiede zwischen der lateinalphabetischen Transkription der ursprünglichen Sprache und der englischen Bezeichnung. Der Inputcode solcher Schriftzeichen muss auf der Transkription des Wortes in der ursprünglichen Sprache beruhen, wenn diese Sprache im multilingualen Eingabesystem eines PCs als eine

Eingabesprache gewählt wird. Wie bei der ersten Gruppe müssen die Schriftzeichen dieser Art ebenso eigenständig als Eingabeeinheit dargestellt werden, die stets mit ‚CV' endet. So lautet der Inputcode für die als Beispiel genannten drei Währungszeichen: ‚yuan'/‚en', ‚rubl'' und ‚won'. Unabhängig von der auftretenden Sprache werden sie zudem stets in der ersten Zone ausgegeben. Wenn diese Sprache keine Eingabesprache ist, kann das Währungszeichen durch den Ländercode der Währung plus Dollarzeichen inputcodiert werden. So kann *Yuan* in diesem Fall mit ‚CN$ CV' ausgedrückt werden.

4. Zur vierten Gruppe gehören die Zeichen, die regional und nur in bestimmten Schriftsystemen eingeschränkt verwendet werden. Die Zeichen dieser Gruppe können weiter in Zahl- (wie die arabisch-indischen, ostarabisch-indischen sowie Devanagari-Zahlzeichen) und Interpunktionszeichen unterteilt werden.

Wie in Kap. 1.3.6 (S. 38) angerissen, sind im indischen Schriftkreis, dem ostarabischen Gebiet und manchen weiteren Ländern Westasiens andere dezimale Ziffersysteme als die internationalen westarabischen Ziffern gebräuchlich, obwohl sie mit Indien dieselbe Quelle teilen und im Prinzip ‚Allographen' sind. Die Verarbeitung solcher Schriftsysteme muss ebenso ihr natives Ziffersystem umfassen.

Bei der Eingabe darf nur die Ziffer aus einem dezimalen System eine Eingabeeinheit darstellen, d.h. entweder ein eigenständiges Zahlzeichen oder eine Ziffernkette. Wie bei anderen Sonderzeichen wird die Taste ‚CV' nach dem Inputcode eingetippt. Nach meinem Entwurf gibt es zwei Methoden, solche nicht-westarabischen Ziffern zu inputcodieren, nämlich entweder nach der Phonetik in der ursprünglichen Sprache oder in Form der internationalen Ziffern. Die erste Variante erfordert einen relativ langen Code und gute Sprachkenntnisse der PC-Benutzer. Die Überschneidungsquote der möglichen Kandidaten ist dafür gering. Im Gegenteil dazu ist die zweite Variante zwar einfacher einzutippen, hat aber Kandidaten in mehreren verschiedenen Sprachen. Damit die zweite Inputcodierung einwandfrei funktionieren kann, muss der PC-Benutzer bei der Ausgabezone, in der die einzugebende Sprache auftauchen kann, doppelt klicken. Dadurch werden die Kandidaten zuerst ausgefiltert und nur Kandidaten aus wenigen Sprachen angeboten. Unabhängig von der Inputcodierungsart werden die gewünschten Kandidaten nur in der Wahlliste von der Zone ihrer Herkunftssprache angezeigt. Mit derselben Methode können auch römische Ziffern verarbeitet werden, die in bestimmten Situationen gebräuchlich sind. Die chinesischen Schriftzeichen für die zehn Ziffern können ebenso auf vergleichbare Art und Weise behandelt werden, aber ihre Eingabe muss zusätzlich mit ‚CM' markiert und nach der Lesart der chinesischen Sprache umgestellt werden, wie z.B.

‚2 CM 0 CM 1 CM 7 CM CV' für <二零一七> (/èrlíngyīqī/, *[das Jahr] 2017*). In Tab. 5-4 werden die Eingaben solcher Zeichen am Beispiel der Devanagari-Zahlzeichen erklärt.

Zahlzeichen	Hindi-Wort	Transliteration	Inputcode 1)	Inputcode 2)
०	शून्य	/śūny[a]/	shuunay	0
१	एक	/ek/	ek	1
२	दो	/do/	do	2
३	तीन	/tīn/	tiin	3
४	चार	/cār/	chaar	4
५	पाँच	/pə̃c/	paanch	5
६	छ	/cha/	chha	6
७	सात	/sāt/	saat	7
८	आठ	/āth/	aath	8
९	नौ	/nau/	nau	9

Tab. 5-4: Inputcode für die Devanagari-Zahlzeichen

Wegen der eindeutigen Unterscheidungen bei Schrifttyp und -richtung sind spezifische Interpunktionszeichen in Sprachen des CJK-, indischen sowie arabischen Schriftkreises recht häufig. Im Chinesischen sind dafür das Buchtitelzeichen (gebraucht für Buch-, Artikel-, Dateititel etc.; chi.: 书名号 /shūmínghào/; Form:《…》) und Aufzählungskomma (gebraucht für Aufzählungen einzelner Zeichen sowie Wörter; chi.: 顿号 /dùnhào/; Form: 、) typische Beispiele (vgl. dazu Unicode 12.0 Character: U+3000-U+303F). Spezielle Satzzeichen in Devanagari sind bspw. Danda, Doppeldanda, Avagraph und Abkürzungszeichen. Im Arabischen können Datumtrennzeichen und Tatweel genannt werden. Form, Unicode und Zeichengebrauchszweck werden in Tab. 2-16 (S. 97) und 2-23 (S. 120) angegeben.

Diese Interpunktionszeichen können im Prinzip wie die Zeichen aus der zweiten Gruppe verarbeitet werden, indem sie gemeinsam mit Wort/Satz oder eigenständig eingegeben und in der zugehörigen Zone ihres Schriftsystems als Kandidat angeboten werden. Der Hauptunterschied im Vergleich zu der Zeichenverarbeitung der zweiten Gruppe liegt an der Inputcodierung: Da solche Zeichen auf dem internationalen Tastaturlayout nicht belegt werden und nur in bestimmten Schriftsystemen bekannt und gebräuchlich sind, können sie nach meinem Entwurf nur nach ihrer Bezeichnung auf der ursprünglichen Sprache inputcodiert werden. Da die vollständige Bezeichnung manchmal aus relativ langen Buchstabenketten besteht, wird die kurze Form der Bezeichnung bevorzugt. Im Chinesischen z.B. ist Halbpinyin (die ersten Buchstaben jeder Zeichenumschrift) einsetzbar. Die chinesische Eingabe des Buchtitelzeichens mit und ohne sprachlichen Inhalt läuft wie folgt ab:

Mit Inhalt: SMH hong CM lou CM meng CM SMH CV → 《红楼梦》[283]
Ohne Inhalt: smh SP CV → 《SP》

Wie bei der Zeichenverarbeitung der zweiten Gruppe spielt die Spracherkennung eine ebenso entscheidende Rolle bei spezifischen Interpunktionszeichen. Ergebnisse der Spracherkennung, Ausgabe- sowie Auswahlzone sind vergleichbar mit der Situation der Zeichen der zweiten Gruppe. Für Zeichen, die wie Klammern oder Anführungszeichen paarweise auftreten, ist nach meinem Entwurf sowohl eine gemeinsame als auch einzelne Eingabe möglich. Wie das obige Beispiel zeigt, muss ein Zeichen zweimal codiert werden, falls der Text in derselben Eingabeeinheit eingegeben wird, nämlich sowohl für den öffnenden als auch den schließenden Teil. Im Gegenteil kann es einheitlich inputcodiert und verarbeitet werden, wenn kein Text zwischen den beiden Teilen miteingetippt wird.

In Tab. 5-5 werden acht spezielle Interpunktionszeichen im chinesischen Schriftsystem, die entweder der vierten oder der zweiten Gruppe angehören, exemplarisch erklärt. Vergleichbar mit der Eingabe im multilingualen Eingabesystem wird die Zeichen-Tasten-Repräsentation solcher Zeichen im chinesischen Tastaturlayout angegeben.

Glyphe[284]	**chi. Ausdruck**	**Eng. & Deu. Ausdruck**	**Unicode**	**Gebrauch & Anmerkungen**	**Eingabe mit chi. Tastatur[285]**	**Eingabe mit multilingualem System**
、	顿号 /dùnhào/	ideographic comma & Aufzählungszeichen	3001	gesetzt bei Separation der koordinierten Wörter oder nach Ordinalzahl	Taste ‚/' [#55]	dh CV
【…】	方头括号 /fāngtóu kuòhào/	lenticular bracket & Linsenklammer	3010 & 3011,	für die Markierung zu erklärender Wörter; Angabe der Nachrichtenagentur am Textanfang	einzelne Eingabe; L: Taste ‚[' [#27]; R: Taste ‚]' [#28].	Eingabe im Paar: ftkh CV; einzelne Eingabe: […] CV.
·	间隔号 /jiān'gé hào/	interpunct & Mittelpunkt	00B7	Bei der Trennung von Namen, Titeln usw.[286]	Taste <`>; [#01]	jgh CV
。	句号 /jùhào/	full stop & Punkt	3002	Satzendmarkierung in CJK-Schriften	Taste ‚.' [#54]	. CV

[283] Um Interpunktionszeichen von den Inputcodes des Textinhalts zu unterscheiden, wird die Umschrift des Buchtitelzeichens an dieser Stelle in Großschreibung empfohlen;《红楼梦》/hónglóumèng/ („*Der Traum der Roten Kammer*", einer der sog. vier großen klassischen Romane der chinesischen Literatur).

[284] Zwischen dem öffnenden und schließenden Teil des im Paar auftretenden Zeichens werden drei Punkte für Inhaltsauslassungen dargestellt.

[285] Der Tastenname basiert auf dem US-amerikanischen Tastaturlayout, in der eckigen Klammer wird der Scan-Code parallel angegeben. Wenn es mehrere Formen sowie Variantenglyphen zu einem Zeichen gibt, wird die alternative Glyphe in dieser Spalte weggelassen. Bei der Erklärung der generellen Form werden die Abkürzungen für Oben (O), Unten (U), Links (L) und Rechts (R) gebraucht.

[286] Bei Namen außerhalb des CJKV-Kulturkreises, genauer vollständigen Namen mit umgekehrter Reihenfolge von Familien- und Vorname im Vergleich zu CJKV-Namen, wird dieses Zeichen benötigt: 汉斯·穆勒 (/hànsī mùlè/; *Hans Müller*).

Glyphe[284]	chi. Ausdruck	Eng. & Deu. Ausdruck	Unicode	Gebrauch & Anmerkungen	Eingabe mit chi. Tastatur[285]	Eingabe mit multilingualem System
——	破折号 /pòzhé hào/	dash & Gedankenstrich	2-mal 2014	für Anmerkungen sowie erweiterte Erklärungen mancher Elemente oder Umwandlungen des Klangs sowie der Bedeutung	SH + Taste ‚-' [#12]	pzh CV
……	省略号 /shěngluè hào/	ellipsis & Auslassungspunkt	2-mal 2026	für Auslassung des Inhalts sowie Fortsetzung der Semantik	SH + Taste ‚6' [#07]	slh CV oder CV
《…》 〈…〉	(双/单) 书名号 /(shuāng/ dān) shūmíng hào/	(double/ single) guillemet & (doppeltes/ einfaches) Buchtitel-zeichen	300A/300B & 3008/3009	für die Angabe von Titeln verschiedener Werke, wie Büchern, Filme usw.; die Doppelform wird primär gebraucht	einzelne Eingabe; LO: SH + Taste <,> [#53]; RO: SH+ Taste <.> [#54]; LU & RU: keine Belegung	Eingabe im Paar: Oben: ssmh CV; Unten: dsmh CV
'…' "…" Alt. : 『…』 「…」	(单/双) 引号 /(dān/ shuāng) yǐnhào/	(single/ double) quotation mark & (einfaches/ doppeltes) Anführungszeichen	FF02 FF07; Alt.: 300E/300F & 300C/300D	für direkte Zitate, spezielle Markierungen (z.B. der Distanz) usw.; die Doppelform wird primär gebraucht	einzelne Eingabe; BO: Taste <'> [#41]; BU: SH+ Taste <'> [#41]	einzelne Eingabe: O: ' CV; U: " CV; oder Eingabe im Paar: O: dyh CV; U: syh CV

Tab. 5-5: Die Eingabe der chinesischen Interpunktionszeichen mit einer Pinyin-Eingabemethode sowie mit dem multilingualen Eingabesystem

5.3 Weitere Probleme der multilingualen Software

In Kap. 5.1 und 5.2 wurden die ersten drei Schritte für den Softwareentwurf – Problemanalysen und Anforderungsdefinition, Modellierung und fachlicher sowie softwaretechnischer Entwurf – skizziert. Vor der Programmierung müssen einige weitere Probleme vertieft analysiert werden. 1) Anhand welcher linguistischen Erkenntnisse und auf welche Art und Weise kann die Spracherkennung effektiver wirken? 2) Ist die potentielle Effizienz dieser Software hoch genug, um erforschenswert und anwendungsbereit zu sein? 3) Welche weiteren Funktionen können dieser Software hinzugefügt werden, um Effizienz und Zugangsfreundlichkeit zu erhöhen? In Kapitel 5.3.1 bis 5.3.3 werden diese Fragen analysiert.

5.3.1 Linguistische Theorien über Spracherkennung und Konversion

Linguistische Theorien spielen eine entscheidende Rolle beim Eingabesoftwareentwurf eines Schriftsystems und den Verarbeitungsprozessen. Wie in Kap. 2 und 4 gezeigt wurde, sind linguistische Erkenntnisse bzgl. Schrifttyp und Zeicheninventar die obligatorische Wissensgrundlage. Für eine intelligente multilinguale Eingabesoftware müssen die linguistischen Informationen mehrerer Sprachen strukturell erfasst werden, damit Spracherkennung und intelligente Eingabe einwandfrei ablaufen können (vgl. auch Kap. 5.1.3, S. 316f). In diesem Kapitel wird analysiert, wie die Linguistik auf den Verarbeitungsprozess einwirkt und wie hoch Spracherkennungsfähigkeit und Kandidatenüberschneidungsquote in verschiedenen Sprachen sein können.

Nach dem in Kap. 5.2.2 vorgestellten Verarbeitungsprozess beginnt die Spracherkennung nach der Eingabefallbestimmung. Dabei ist die Spracherkennung die Unterscheidung zwischen jenen Sprachen, die zum selben Eingabefall gehören können. Wegen der zahlreichen Varianten im Verarbeitungsprozess, die in jedem Computer wegen den individuell bestimmten eingebbaren Sprachen und des festgelegten Eingabeniveaus unterschiedlich sind, muss diese Erforschung auf einer konkreten Variante basieren. An dieser Stelle wird die in Abb. 5-3 (S. 319) geschilderte Variante als Forschungsobjekt genommen. Darauf konkret bezogen sind die Zusammenhänge zwischen Sprachen und ihrem Eingabefall wie folgt:

A1 (simples Niveau ohne Konversion): Englisch, Deutsch (tb[287]), Vietnamesisch (tb);
A2 (mittleres Niveau ohne Konversion): Englisch, Deutsch (tb);
A3 (fortgeschrittenes Niveau ohne Konversion): Deutsch (tb);
B1 (simples Niveau mit Konversion von Buchstaben als Grundeinheit): Deutsch (TT), Vietnamesisch (TT), Russisch, Arabisch;
B2 (simples Niveau mit Konversion von Syllabar / Silbenzeichen / Logogramm als Grundeinheit): Chinesisch, Japanisch, Koreanisch, Hindi;
C1 (mittleres Niveau mit Konversion von Buchstaben als Grundeinheit): Deutsch (TT), Russisch;
C2 (mittleres Niveau mit Konversion von Syllabar als Grundeinheit): keine;
D1 (fortgeschrittenes Niveau mit Konversion von Buchstaben als Grundeinheit): Deutsch (TT);
D2 (fortgeschrittenes Niveau mit Konversion von Syllabar als Grundeinheit): keine;
E (Sonderfall des mittleren/fortgeschrittenen Niveaus mit Konversion): Chinesisch, Japanisch.

[287] tb – ‚teils betroffen', markiert, dass nur ein begrenzter Teil dieser Sprache zum Eingabefall gehören kann, im Deutschen etwa Buchstaben, Wortteile, Wörter und Sätze ohne Sonderbuchstaben. Dementsprechend bedeutet TT ‚Teilsweise-Transkription'.

Im A-Fall wird die eingetippte Buchstabenkette ohne Konversion ausgegeben, daher ist keine Spracherkennung obligatorisch. Zusammengefasst geht es bei der Spracherkennung anhand der angegebenen Sprachliste vor allem darum, zwischen 1) Deutsch, Vietnamesisch, Russisch und Arabisch im B1-Fall, 2) Chinesisch, Japanisch, Koreanisch und Hindi im B2-Fall, 3) Deutsch und Russisch im C1-Fall und 4) Chinesisch und Japanisch im E-Fall zu unterscheiden. Je höher das Eingabeniveau ist, desto weniger Sprachen sind möglich und desto mehr Merkmale gibt es für die Spracherkennung. Im simplen Niveau B1 und B2, bei dem Silbe / Morphem / Zeichen die längsten Einheiten darstellen, können nur die linguistischen Phänomene innerhalb des Wortes betrachtet werden, wie bspw. Silbenstruktur, Zeichen- sowie Grapheminventar, Affixe der grammatischen Flexion und zur Wortbildung. Auf mittlerem Niveau ist es im Normalfall möglich, die verarbeitete Sprache und das eingegebene Wort durch den Wörterbuchabgleich zu erkennen. Im fortgeschrittenen Niveau gibt es wegen der am wenigsten zu definierenden Sprachen und der Wissensdaten über Syntax und Wortschatz kaum Ambiguitäten bei sich überschneidenden Sprachen und Kandidaten.

Im B1-Fall, bei dem es die meisten Kandidaten verschiedener Sprachen gibt, ist bspw. eine kontrastive Liste der möglichen Sprachen, in der Informationen über Transkription, Phonetik, Morphologie usw. eingetragen werden, obligatorisch. Durch Matching der Eingabeeinheit mit derlei eingeschriebenem linguistischem Wissen können die Kandidaten schrittweise eliminiert werden, wie die nachstehende Tabelle zeigt:

Art	Deutsch[288]	Vietnamesisch[289]	Russisch[290]	Arabisch[291]
Silbenaufbau[292]	(K)(K)(K) VK(K)	(K)(W)V^{Ton}K	(K)(K)(K)(K) V(K)(K)(K)(K)	generell: KV, KV:, KVK; relativ selten: KV:K, KVKK [K und V: werden schriftlich repräsentiert]
Grundbuchstaben[293]	A/a	A/a	*A/a* = A/a	*a* = 1) = ´ [Fatha, optional] 2) = ا/ٱ/أ [nach Kontext]
	B/b	B/b	*B/b* = Б/б	*b* = ب
	C/c	C/c	*C/c* = Ц/ц	*c* = s = س [wie s2]
	D/d	D/d	*D/d* = Д/д	*d* = 1) = <d>= د 2) = <ḍ> = ض
	E/e	E/e	*E/e* = E/e	
	F/f		*F/f* = Ф/ф	*f* = ف
	G/g	G/g	*G/g* = Г/г	*g* = ğ = ج [wie j]

[288] Vgl. Eisenberg 1989: 61f, Altmann 2010: 150, Plath 2014: 9-12 & Clément 2000: 40ff & 133ff.
[289] Vgl. Coulmas 1996a: 543, Boscher 1989: 10 & Edmonson 2006: 1149-1153.
[290] Transkription nach ISO-9; für weitere Informationen vgl. Daum/Schenk 2002: XIII - XVI, Hippsley 1998, Ostapenko 2005 & Ward 2006.
[291] Vgl. Procházka 2006 & Watson 2006.
[292] Formeln für Silbenaufbau werden mit Unterstrich markiert.
[293] Um die ursprünglichen Zeichen von Inputcodes eines Schriftsystems zu unterscheiden, werden Inputcodes kursiv dargestellt; Zwischencodes (meistens Sonderbuchstaben der Transliteration) werden in spitzer Klammer angegeben.

Art	Deutsch[288]	Vietnamesisch[289]	Russisch[290]	Arabisch[291]
	H/h	H/h		*h* = 1) = <h> = ه 2) = <ḥ> = ح
	I/i	I/i	*I/i* = И/и	*i* = 1) = ِ [Kasra, optional] 2) = ا/ئ/ [nach Kontext]
	J/j		*J/j* = Й/й	*j* = <ǧ> = ج
	K/k	K/k	*K/k* = К/к	*k* = ك
	L/l	L/l	*L/l* = Л/л	*l* = ل
	M/m	M/m	*M/m* = М/м	*m* = م
	N/n	N/n	*N/n* = Н/н	*n* = ن
	O/o	O/o	*O/o* = О/о	*o* = أ [nach Kontext]
	P/p	P/p	*P/p* = П/п	*p* = b = ب [wie b]
	Q/q	Q/q		*q* = ق
	R/r	R/r	*R/r* = Р/р	*r* = ر
	S/s	S/s	*S/s* = С/с	*s* = 1) = <ṣ> = ص 2) = <s> = س
	T/t	T/t	*T/t* = Т/т	*t* = 1) = ṭ = ط 2) = ṯ= ت
	U/u	U/u	*U/u* = У/у	*u* = 1) = ُ [Damma, optio-nal] 2) = أ/ع [nach Kontext]
	V/v	V/v	*V/v* = В/в	*v* = f = ف [wie f]
	W/w			*w* = و
	X/x	X/x		
	Y/y	Y/y	*Y/y* = Ы/ы	*y* = ي
	Z/z		*Z/z* = З/з	*z* = 1) = z = ز 2) = ẓ = ظ
andere vokalische Buchstaben/Zeichen	*Ae/ae* = Ä/ä	Aw/aw= Ă/ă	*Yo/yo* = <ë>= Ё/ё	*a-* = ā= أ
	Oe/oe = Ö/ö	Aa/aa= Â/â	*Eh/eh* = <ė> = Э/э	*i-* = ī = إ
	Ue/ue = Ü/ü	Ee/ee= Ê/ê	Ju/ju / Yu/yu = Ю/ю	*u-* = ū= أ
		Oo/oo= Ô/ô	Ja/ja / Ya/ya = Я/я	' = <’> = أ/ا/ء/ئ/ق
		Ow/ow= Ơ/ơ		*an* = ً [Fathatān für Akkusativendung auf Alif, Hamza oder Tamabuta]
		W/w= Ư/ư		*in* = ٍ [Masratān für Genitivendung]
				un = ٌ [Dammatān für Nomivativendung]
				(vokallos) = ْ [Sukūn]
andere konsonantische Buchstaben	*ss* = ß (nur teils identisch)	Dd/dd= Đ/đ	*Kh/kh* = <ch> = Х/х	*th* = <ṯ> = ث
			Ch/ch = <č> = Ч/ч	*sh/ch* = š = ش
			Sh/sh = <š> = Ш/ш	*kh* = <ḫ> = خ
			Shch/shch = <šč> = Щ/щ	*dh* = <ḏ>= ذ
			zh = <ž> = Ж/ж	*gh* = ġ = غ

Art	Deutsch[288]	Vietnamesisch[289]	Russisch[290]	Arabisch[291]
sonstige Sonderschriftzeichen	*`A/`a* = À/à [gleich wie für È/è, Ì/ì, Ò/ò, Ù/ù]	V$^{2.\text{Ton}}$: V + 'f' – *af* = à [in gleicher Weise gibt es die Codes *ef*, *if*, *of*, *uf*, *yf*, *awf*, *aaf*, *eef*, *oof*, *owf* und *ef*.	′ = <″> = Ъ/ъ	ء [Hamza, Umschrift nach Kontext; über/unter Buchstaben oder allein stehend]
	´A˝a = Á/á [gleich wie für É/é, Í/í, Ó/ó, Ú/ú, Ý/ý]	V$^{3.\text{ Ton}}$: V + 's' – *as* = á [in gleicher Weise gibt es Codes: *es*, *is*, *os*, *us*, *ys*, *aws*, *aas*, *ees*, *oos*, *ows*, *ws*].	' = <'> = Ь/ь	ة [Ta Marbuta, Umschrift nach Kontext, Affix für feminines Genus]
	^A/ ^a = Â/â [gleich wie Ê/ê, Î/î, Ô/ô, Û/û]	V$^{4.\text{Ton}}$: V + 'j' – *aj* = ạ [in gleicher Weise gibt es Codes: *ej*, *ij*, *oj*, *uj*, *yj*, *awj*, *aaj*, *eej*, *ooj*, *owj*, *wj*].		
		V$^{5.\text{ Ton}}$: V + 'r' – ar = å [in gleicher Weise gibt es Codes: *er*, *ir*, *or*, *ur*, *yr*, *awr*, *aar*, *eer*, *oor*, *owr*, *wr*].		
		V$^{6.\text{ Ton}}$: V + 'x' – ax = ã [in gleicher Weise gibt es Codes: *ex*, *ix*, *ox*, *ux*, *yx*, *awx*, *aax*, *eex*, *oox*, *owx*, *wx*].		
Dia- & Trigraphe	ah, ch, ck, eh, ie, ng, nn, oh, ph, qu, ts, tz, uh, üh (*ueh*); sch, tsch	ch, gh, gi, kh, nh, ng, ph, qu, th, tr	[fast immer Eins-zu-eins-Entsprechung zwischen Phonem und Graphem]	[fast immer Eins-zu-eins-Entsprechung zwischen Phonem und Graphem]
Diphthonge	ai, au, ay, äu (*aeu*), ei, eu, ey, ui	ia, iê (*iee*), ua, uô (*uoo*), ưa (*wa*), ươ (*wow*)	[kaum]	[kaum]
Silbenränder aus zwei Buchstaben	gs, ks, pf, sp, st, tr	[kaum]	*gm*, *kn*, *kt*, *mn*, *pn*, *pt*, *tk*, *tm*, *vm*, *vn*, *zhm* (жм) etc.[294]	[kaum]
Silbenränder aus drei oder mehreren Buchstaben	chs, dsch, rst, sch, schr, tsch	ngh	fkl, ftr, stl, str, skhv (схв), vbr, vgl, vsk, vzv, vzb, zdr, zbr, zbl, zgl, zgn; fskr, fspl, fspr, fstr, fskhl (фсхл), vzbr, vzdr, bzgl, vzgr etc.	[kaum]

[294] Sie werden wegen der Platzeinschränkung nur in Form der Transkription angegeben. In Ambiguitätsfällen wird die originale Form in kyrillischem Alphabet in der Klammer angedeutet. Dasselbe gilt auch für alle folgenden Zeilen.

Art	Deutsch[288]	Vietnamesisch[289]	Russisch[290]	Arabisch[291]
Wortbildungs-suffixe	-heit, -keit, -ung, -in; -lich, -isch, -ig, -bar etc.		*-es*, *-l'shch'ikh* (льшчьих), *-nik*, *-tel'*, *-'ist*, *-'ik*, *-shch'ikh* (шчьих) etc.	[kaum]
Präfix & Zirkumfix zur Wortbildung	ab-, ent-, un-, wider-, zu-; be-…-t, ver-…ern etc.		*v-*, *vy-*, *voeh-* (воэ), *do-*, *za-*, *iza-*, *na-*, *ot-*, *pod-*, *pri-*, *pro-*, *so-*, *u-*, *raeh…sja* (раэ...ся), *so…sja* (со...ся) etc.	[kaum]
Flexionsaffix aus einem/zwei Buchstaben	-e, -es, -en, -er, -s, -te, -et, -t, -st, etc.		*-a*, *-am*, *-at*, *-akh* (ах), *-aja* (ая), *-e*, *-ev*, *-ee*, *-ej*, *-em*, *-et*, *-jo* (ё), *-yom* (ём), *-ie*, *-ii*, *-ij*, *-im*, *-it*, *-ikh* (их), *-iju* (ию), *-ija* (ия), *-j*, *-mi*, *-o*, *-ov*, *-oe*, *-oj*, *-om*, *-si*, *-s'*, *-t'*, *-ti*, *-u*, *-ut*, *-uju* (ую), *-ch'* (чь), *-ye*, *-yj*, *-ym*, *-ykh* (ых), *-'*, *-'ju* (ью), *-ju* (ю), *-juju* (юю), *-jut* (ют), *-ja* (я), *-jam* (ям), *-jat* (ят), *jakh* (ях), *-jaja* (яя) etc.	*mu-*, *ta-*, *ya-*, *yu-*, *'i-* etc.
Flexionsaffix aus mehreren Buchstaben	ge-...-t, -end, -ste, -sten etc.		*-ajshij* (айший), *-amm*, *-ego*, *-ejshij* (ейший), *-emu*, *-ete*, *-esh'* (ешь), *-iej*, *-imm*, *-ite*, *-ish'* (ишь), *-ogo*, *-omu*, *-ymi*, *-jamm* (jamm) etc.	*mun-*, *muta-*, *'in-* etc.

Tab. 5-6: Vergleich der vier möglichen Sprachen im B1-Eingabefall

Wie in Kap. 5.2.2 (S. 324) vorgestellt wurde, ist die Eingabe im B1-Fall mit Buchstaben, Silben oder Wortteilen (wie Affixen der grammatischen Flexion oder der Wortbildung) möglich. Die Anwendung solch einer Liste zur Kandidateneliminierung kann nach meinem Entwurf in sechs Schritten durchgeführt werden:

1. Erkennung durch ‚Shift'-Taste.

Wie im Kontext des zweiten erkennbaren Merkmals des Schriftsystems in Kap. 5.2.1 erwähnt wurde, muss der Eingabesprache ein vollalphabetisches Schriftsystem zugrunde liegen, wenn innerhalb des Inputcodes ‚Shift' benutzt wird. Sprachen, in deren Schriftsystem es keine Majuskel-Minuskel-Unterscheidung gibt, können in diesem Fall disambiguiert werden.

2. Inputcodeerkennung im einzelnen Buchstaben.

In diesem Prozess wird jeder Buchstabe des Inputcodes von links nach rechts einzeln betrachtet, um die Artikulationsklasse (Konsonant oder Vokal) zu bestimmen und diese mit dem Zeicheninventar des Schriftsystems sowie der Transkription einer Sprache abzugleichen. Sprachen, deren Schriftsystem oder Transkription keinen Buchstaben beinhalten, welcher in dem zu verarbeitenden Inputcode auftritt, können disambiguiert werden. Existiert bspw. in der Eingabeeinheit der Buchstabe ‚w', kann Russisch aus der möglichen Sprachliste entfernt werden, da es in diesem Transkriptionssystem ungebräuchlich ist. Vietnamesisch hingegen wird behalten, da dort ‚w' für das Umschreiben von Sonderbuchstaben gebraucht wird.

3. Unterscheidung zwischen teil- und volltranskribierten Sprachen.

Wie dargestellt, können die möglichen Sprachen im B1-Eingabefall in drei verschiedene Gruppen unterschieden werden: die im lateinischen Alphabet geschriebenen Sprachen mit Sonderbuchstaben (teilweise transkribiert), die nichtlateinischen vollalphabetischen sowie die konsonantenalphabetischen Schriftsysteme (beide volltranskribiert). Nur wenn es im Inputcode ein Bi- oder Trigramm gibt, das der Umschriftliste eines teilweise transkribierten Schriftsystems entspricht, wird diese Sprache weiter berücksichtigt. Wenn nicht, wird diese Sprache ausgeschlossen, denn die Funktionstaste für die Konversion wird erst gebraucht, sobald mindestens ein lateinischer Sonderbuchstabe geschrieben wird. Durch diesen Prozess kann der Großteil der lateinalphabetischen Schriftsysteme von der möglichen Liste gestrichen werden. Im umgedrehten Fall kann die Wahrscheinlichkeit einer betroffenen Sprache entsprechend verstärkt werden.

4. Recherche der für einzelne Buchstaben stehenden Bigramme und Trigramme.

In vielen Fällen wird ein einzelner Buchstaben des ursprünglichen Schriftsystems in einer Buchstabenfolge (meistens Bi- oder Trigramm) transkribiert, wie z.B. ‚ü = *ue*' im deutschen und ‚Щ = *Shch*' im russischen Schriftsystem. Ziel dieses Arbeitsschritts ist, die für einzelne Buchstaben stehenden Bi- und Trigramme der Transkription/Inputcodierung zu erkennen.

An dieser Stelle wird nach meinem Entwurf vor allem die Methode der Merkmalbegründung angewendet. Für Deutsch bspw. fungieren ‚e' hinter den Vokalen a/o/u und die Verdopplung von ‚s' als Merkmale. Mit demselben Verfahren wird die Merkmalliste jedes möglichen Schriftsystems verfasst und mit der Eingabeeinheit abgeglichen. Die Merkmalliste jeder möglichen Sprache (außer denen in vorherigen Schritten ausgeschlossenen) lässt sich mit dem Inputcode vergleichen. Die Verarbeitungsreihenfolge sollte mit der Reihenfolge des Sprachge-

brauchs im multilingualen Eingabesystem übereinstimmend sein. Die erkannten Bi- sowie Trigramme werden, angeordnet nach der Sprachreihenfolge, für begrenzte Zeit registriert.

5. Graphen- und Silbenerkennung.

In diesem Schritt wird zuerst überprüft, ob es in der eingetippten Buchstabenkette einen festgelegten Graph eines Schriftsystems (eine Buchstabenfolge für eine Lautung oder einen Diphthong) gibt, der durch die Informationen in der Zeile ‚Bi- & Trigraphe' und ‚Diphthonge' untersucht wird. Die als ein Graph erkannte Buchstabenfolge wird dann in ihrer Gesamtheit betrachtet. Danach beginnt der Silbenerkennungsprozess, wobei der Inputcode dahingehend analysiert wird, ob er eine Silbe (mit mindestens einem Vokal) sein kann. Wenn ja, so wird er als eine Silbe erkannt und mit der Silbenstruktur der möglichen Sprachen abgeglichen. Sprachen, deren Silbenstruktur mit der eingetippten Silbe konkurriert, werden eliminiert. Wenn der Inputcode aus mehr als einer Silbe besteht, wird er in Silben segmentiert und die segmentierten Silben weiter durch Abgleiche verarbeitet. Wenn es innerhalb des Inputcodes keine Silbe geben kann, bleibt er in diesem Arbeitsschritt unverarbeitet. Da ein Kurzvokal in einem konsonantenalphabetischen Schriftsystem nicht schriftlich dargestellt werden muss, muss er ebenso nicht inputcodiert werden. Daher ist die Silbenerkennung für Sprachen wie Arabisch ungültig. Anders formuliert können solche Sprachen in diesem Schritt nicht eliminiert werden, egal wie die Ergebnisse der Silbenerkennung ausfallen.

6. Recherche der möglichen Affixe und Silbenränder.

In diesem Arbeitsschritt wird überprüft, ob es im Inputcode typische Affixe (für Wortbildung sowie Flexion) und Silbenränder einer Sprache (außer den in den letzten vier Schritten schon beseitigten Sprachen) gibt. Wenn ja, so wird diese Sprache wahrscheinlicher. Der betroffene Kandidat aus dieser Sprache wird in der Wahlliste ebenso an vorderster Stelle angezeigt.

Besteht der Inputcode mindestens aus drei Buchstaben, ist eine grobe Spracherkennung prinzipiell möglich. Die für diese Liste benötigten Informationen können zuerst als zusätzliche Daten der Transkriptionsliste einer Sprache gespeichert werden. Nachdem die benötigten Ressourcen der einzugebenden Sprachen runtergeladen wurden (siehe Kap. 5.2.3, S. 330ff), kann so eine kontrastive Liste automatisch vom System erzeugt und angewendet werden.

Mit dem damit ähnlichen Verfahren aus Tab. 5-6 kann ebenfalls eine kontrastive Liste für Chinesisch, Japanisch, Koreanisch und Hindi im B2-Fall erstellt werden. Im Gegenteil dazu bilden jedoch die Möglichkeiten für Zeichenmarkierung, generelle Silbenstruktur, Varianten für konsonantische, vokalische Lautungen und gemeingebräuchliche Silben und grammatische Merkmale (flektierende Affixe im Hindi und agglutierende Merkmale im Japani-

schen/Koreanischen) die Schwerpunkte. Da die zum indischen und CJK-Schriftkreis zugehörigen Sprachen (alle möglichen Sprachen des B2-Niveaus) im Vergleich zur Sprachengruppe B1 weniger sind und zu verschiedenen Sprachfamilien gehören können, könnte die Spracherkennung bei dieser Gruppe effektiver wirken. Aber gleichzeitig verfügen die drei ostasiatischen Sprachsysteme über ein relativ umfangreiches Homophoninventar. Dies bringt mit sich, dass es für eine Sprache viele verschiedene Kandidaten geben kann.

Auf die nötigen linguistischen Erkenntnisse und groben Verarbeitungsschritte im mittleren sowie fortgeschrittenen Eingabeniveau (C1-, C2-, D1-, D2- und E-Fall) geht Kap. 5.2.2 (S. 324f) ein. Mit der Erhöhung des Eingabeniveaus und erweiterten linguistischen Erkenntnissen wird die Spracherkennung immer präziser und effektiver. Falls manche Wörter und grammatische Fälle wegen der Begrenzung der Wissensdatenbank (Lexikon und linguistische Wissensdatenbank) nicht verarbeitet werden können, werden sie mit den Arbeitsphasen des niedrigeren Eingabeniveaus weiter behandelt.

5.3.2 Analysen über die Effizienz der Software

Effizienz ist einer der wichtigste Faktoren für den Entwicklungs- und Anwendungswert eines Eingabeverfahrens. In diesem Kapitel wird die Effizienz dieser multilingualen Eingabesoftware für verschiedene Sprachen im Vergleich mit Eingabeverfahren analysiert, die sich an einzelnen Sprachen orientieren (im weiteren Verlauf ‚EzS-Eingabeverfahren' genannt).

Wie in Kap. 5.1.4 geplant, kann der Status zwischen Aktivierung und Deaktivierung des ‚Language Codifying' (Sprachfestlegung) von PC-Benutzern situationsangemessen umgestellt werden. Die Leistung in den beiden Status für eine bestimmte Sprache weicht dabei erheblich ab. Da die Grundarbeitsprinzipien eines EzS-Eingabeverfahrens bei verschiedenen Typen der Schriftsysteme unterschiedlich sind, werden Effizienzanalysen in vier Gruppen (jeweils für lateinalphabetische, nichtlatein-alphabetische, alphabetisch-syllabisch-hybride und nichtalphabetische Schriftsysteme) durchgeführt und in acht Fälle gegliedert.

1. Erste Gruppe: Eingabe eines lateinalphabetischen Schriftsystems.

Da das US-amerikanische Tastaturlayout (mit Umänderungen) als Hardware für diese Eingabesoftware bestimmt wird, ist die Eingabe in dieser Sprachgruppe kaum unterschiedlich zu EzS-Eingabeverfahren. Die Unterschiede liegen nur im mehrmaligen Tastenanschlag für Sonderbuchstaben (gilt für beide Status), der Einsetzung der Funktionstaste ‚NC' oder ‚CV' und der durch die Mehrsprachigkeit bedingten, möglicherweise längeren Verarbeitungszeit für Spracherkennung, Konversion und Kandidatenauswahl (nur bei dem Status der Deaktivierung der Sprachfestlegung).

2. Zweite Gruppe: Eingabe eines nichtlateinischen aber alphabetischen Schriftsystems.

Im Vergleich zu der ersten Gruppe erfordern die Schriftsysteme der zweiten Gruppe vor allem das zusätzliche Umcodieren der schriftlichen Symbole, was dem jeweiligen PC-Nutzer zufällt und das Schreibtempo in vielen Fällen verlangsamt. Dasselbe Problem gilt auch für die zwei anderen Gruppen. Um mit dieser Eingabesoftware andere Sprachen außerhalb des lateinischen Schriftkreises eingeben zu können, muss ein PC-Benutzer die standardisierte Transliteration oder Transkription solcher Sprachen beherrschen.

Wenn die Eingabe im Status der Sprachfestlegungsaktivierung stattfindet, ist ihre Effizienz identisch mit auf der Transkription basierenden EzS-Eingabeverfahren. Mit entsprechender Einübung kann damit im Prinzip ebenfalls ein hohes Schreibtempo erreicht werden. Ist sie deaktiviert, so ist die Effizienz eingeschränkt durch mehrmaligen Tastenanschlag (bei ‚CV' als Satzende) und Softwareverarbeitungszeit – gleich wie bei Sprachen der ersten Gruppe.

3. Dritte Gruppe: Eingabe eines alphabetisch-syllabisch-hybriden Schriftsystems.

Wie bei den Sprachen der zweiten Gruppe wird die Effizienz bedingt durch Transkription, Umgewöhnung des Tastaturlayouts (bei beiden Status), mehrmaligen Tastenanschlag (‚CM' und ‚CV') und eine relativ langen Verarbeitungszeit verzögert (nur im Status der Deaktivierung der Sprachfestlegung). Im EzS-Tastaturlayout dieser Gruppe werden Konsonanten und Vokale meist auf linke und rechte Hälfte verteilt, was das Auswendiglernen der Tastenbelegung vereinfacht und die Tippgeschwindigkeit erhöht. Das US-amerikanische Layout hat deswegen im Vergleich dazu unvermeidbare Nachteile. Ein ideales Schreibtempo ist schwer zu erreichen.

4. Vierte Gruppe: Eingabe von Chinesisch und Japanisch.

Für die Eingabe der Sprachen in dieser Gruppe haben die intelligenten EzS-Eingabemethoden am meisten Gemeinsamkeiten mit dieser multilingualen Eingabesoftware. Inputcodierung, Konversion, Kandidatenangebot und -auswahl auf Zeichen-, Wort- oder Satzstufe sind unentbehrliche Elemente. Mit dieser Software können Chinesisch und Japanisch mit wenigen Vorkenntnissen und Schwierigkeiten eingegeben werden.

Neben dem mehrmaligen Tastenanschlag für die Tasten ‚CM' und ‚CV' (im deaktivierten Status) kann die Schreibeffizienz der Software auch wegen der mangelnden Qualität der sprachlichen Ressourcen rückschrittlicher als die allgemeinen intelligenten Eingabemethoden sein. Durch Begrenzungen des Speichers und multilingualen Erforderlichkeiten können die Datenbanken einzelner Sprachen unmöglich so umfangreich sein, wie die der EzS-Methoden.

Aus diesem Grund könnte es deutlich mehr sprachliche Phänomene geben, die mit der multilingualen Eingabesoftware nicht verarbeitet werden können.

Betrachtet man die vier Gruppen zusammengefasst wird ersichtlich, dass die Effizienz dieser Software fast immer etwas niedriger ausfällt, als die der EzS-Verfahren. Das Schreibtempo ist meist trotzdem akzeptabel. Eine multilinguale Eingabesoftware ist ein Gerät mit multiplen Funktionen, das für die Realisierung des ‚All-Set'- oder ‚Alles in einem'-Prinzips auf Spezifikationen bestimmter Funktionen verzichten muss.

Wegen der eingeschränkten Effizienz ist diese Software nicht für fachkundiges Schreiben langer Texte geeignet. Die empfohlenen Anwendungssituationen sind bspw. die Eingabe einer relativ selten gesprochenen Sprache, das Schreiben kurzer und solcher Texte mit häufigen Sprachwechseln (so wie diese Arbeit). Auch für öffentliche PCs, deren Benutzer aus verschiedenen Nationen kommen, ist sie geeignet. Bei der Systemsteuerung eines PCs sollte diese multilinguale Eingabemethode als zweite Alternative hinter einem EzS-Eingabeverfahren definiert werden, wenn diese Sprache eine häufige geschriebene Sprache ist.

5.3.3 Verbesserungsvorschläge

Die Grundkonstruktion der Eingabesoftware kann in vielen Details verbessert werden, um sie zugangsfreundlicher und praktischer zu machen. Nach meiner Vorstellung könnten u.a. folgende Module und Funktionen zusätzlich designet werden.

1. Das Hinzufügen des selbstangepassten Moduls.

Wie viele intelligente chinesische Eingabemethoden, kann diese geplante Eingabesoftware ebenfalls ein selbstangepasstes Modul enthalten. Seine Beziehungen zu den anderen Modulen und die Art und Weise des systemgestützten Selbstlernens sind vergleichbar mit denen des Chinesischen, die in Kap. 4.5.1 (S. 285-290) analysiert wurden. Dieses Modul dient einerseits dazu, die Datenbanken einer bestimmten Sprache zu ergänzen. Andererseits können gemäß des Softwaregebrauchs mit einem PC Wörter und sprachliche Ausdrücke idiolektal umgeordnet werden, so dass sich die Software individuell am Nutzer orientiert.

2. Zusammenfügen der regionalorientierten Zelldatenbanken zu den standardsprachlichen Datenbanken.

Eine Sprache, besonders eine, die in mehreren Ländern und von einer großen Bevölkerung gesprochen wird, hat häufig mehrere Varianten. Bei Wortschatz, Aussprache, Grammatik usw. kommt es dementsprechend zu Abweichungen. Im Englischen gilt es bspw. britisches, ameri-

kanisches, afrikanisches und australisches Englisch usw. zu differenzieren. Die chinesischen Sprachen haben (neben den verschiedenen Dialekten und Sprachen in der VR China) zusätzlich Varianten in Taiwan, Singapur, Malaysia und nicht zuletzt unter den Menschengruppen chinesischer Herkunft in anderen Ländern.

Neben sprachlichen Zweigen innerhalb der muttersprachlichen Sprechergemeinschaft bildet eine Sprache zudem spezielle Eigenschaften aus, wenn sie von Fremdsprachlern ausgeübt wird, die eine gemeinsame Ausgangssprache oder verschiedene zu derselben Sprachgattung gehörigen Sprachen teilen. Es kommt zu Einflüssen, die sich etwa dadurch äußern, dass Ausdrücke etc. geprägt werden, die den Muttersprachlern fremd vorkommen. Je weiter Ziel- und Ausgangssprache typologisch/sprachfamiliär voneinander entfernt sind und sich etwa in Schrifttyp und Grammatik unterscheiden, desto eher kommt es zu diesem Phänomen. Bspw. wird Sprachlernanfängern in China häufig Chinglisch (Englisch mit chinesischen Einflüssen) und in Europa und Amerika häufig ‚westernisiertes Chinesisch' (Chinesisch mit Einflüssen aus indogermanischen Sprachen) erworben. So gibt es etwa in einem von Chinesen verfassten Text häufig englische Ausdrücke von chinesischer Grammatik und im von Westlern gesprochenen Chinesisch häufig einen übertriebenen Partikelgebrauch.

Drei Funktionen können dazu beitragen, die lokalisierten Varianten einer Sprache mit dieser Software besser zu verarbeiten: das selbstangepasste Modul, ergänzende Zelldatenbanken einer variierten Sprache und die Verwaltung des individuellen Wörterbuchs nach Sprachgebrauch.

Die erste Funktion wurde im letzten Punkt erwähnt. Die Realisierung der zweiten Funktion setzt eine Unterscheidung der allgemein verbreiteten Sprache von den variierten regionalen Versionen voraus. Die am meisten verbreitete Variante einer Sprache samt gemeingebräuchlichem Wortschatz wird voreingestellt in der CD-ROM der Software gespeichert und angewendet, falls diese Sprache als Eingabesprache definiert wird. Wenn der PC-Benutzer eine andere Variante bevorzugt, können Zelldatenbanken, in denen alternative regionale Wörter und sprachliche Regeln eingeschrieben sind, zusätzlich im Internet runtergeladen und ergänzend zu den generellen sprachlichen Daten verwendet werden.

Mit der dritten Funktion können idiolektale Wörter (sowohl regionale als auch Code-Switching-Wörter) eingespeichert werden. Dies gilt sowohl für einzelne Wörter als auch eine große Wörtergruppe. Dies ist z.B. nötig, wenn ein deutscher Sinologe ein deutschsprachiges Buch über chinesische Geschichte schreibt, in dem viele historische Fachbegriffe sowohl in chinesischer Schrift als auch transkribiert in Pinyin vorkommen müssen. Um eine höhere Effizienz zu erzielen, hat er die Möglichkeit, das fachhistorische Zellwörterbuch des Chinesi-

schen komplett dem individuellen Wörterbuch hinzuzufügen und mit deutschen Datenbanken zu verbinden. Das individuelle Wörterbuch kann jeder Zeit nach den Schreibwünschen des PC-Benutzers definiert und verwaltet werden.

3. Funktion der Online-Handschrifteingabe.

Eine transkriptionsbasierte Inputcodierung hat häufig Probleme bei der Eingabe. Für ein nichtlateinalphabetisches Schriftsystem gibt es meist mehrere Transkriptionsstandards, die den PC-Benutzern ungewöhnlich und verwirrend erscheinen können. In der chinesischen Schrift stößt man wegen der Zeichenvielfalt immer wieder auf unbekannte Schriftzeichen. Damit man trotz mangelnder Transkriptions- und Zeichenkenntnisse die gewünschte Sprache eingeben kann, kann eine Online-Handschrifteingabemöglichkeit integriert werden. Wie bei einer intelligenten chinesischen Eingabemethode kann diese Funktion (unter ‚Tools' des *Look and Feel*) gestartet werden. Ein Handschriftfenster wird sodann geöffnet, in dem man per Maus oder Eingabestift die einzugebenden schriftlichen Symbole zeichnen kann. Durch Schrift- sowie Textanalysen und Abgleich mit der Glyphendatenbank wird die zugehörige Schrift und das gezeichnete Symbol erkannt und dem Text hinzugefügt. Die per Handschrift eingebbaren Einheiten umfassen einzelne Zeichen (wie Buchstabe [außer den 26 lateinischen], Syllabar sowie Sinogramm), festgelegte, aus mehreren Buchstaben zusammengesetzte Grapheme (eines in Voll- sowie Konsonantenalphabet geschrieben Schriftsystems) sowie syllabische Einheiten (in einem alphasyllabischen oder dem koreanischen Schriftsystem) und Radikale der vereinheitlichten CJK-Logogramme.

Die Glyphendatenbank muss im Prinzip alle allgemeingebräuchlichen Schriftzeichen und festgelegten Grapheme aus denen im System definierten Sprachen umfassen. Damit die Handschrifterkennung effizienter abläuft, ist die Bestimmung der zugehörigen Schrift bzw. Zeichensorte obligatorisch. Anhand meines Entwurf werden dazu fünf Faktoren berücksichtigt: äußerliche Gestalt, Schreibweise, proportionale Konstellation, Statistiken der Handschrifteingabe in dem Cloud Computing und Vorkontext.

Zeichen aus dem CJK-Schriftkreis entsprechen äußerlich meist einem Quadrat, da Breite und Höhe ungefähr identisch sind. Hingegen gleicht ein Buchstabe aus einer alphabetischen oder alphasyllabischen Schrift eher einem Rechteckt, mit Abweichungen bei Länge und Breite. Typographisch wird ein CJK-Zeichen in der Regel in Vollbreite dargestellt, während ein Nicht-CJK-Zeichen in Halbbreite repräsentiert wird.

Mit Schreibweise ist die Art des Zeichenschreibens gemeint, die sich in verschiedenen Schriftarten erheblich unterscheidet. Die graphischen Elemente eines Voll- sowie Konsonan-

tenalphabets bestehen normalerweise nur aus Kurven, Graden und Punkten, weshalb ein Buchstabe mit wenigen Zügen geschrieben werden kann. Der Hauptunterschied zwischen Voll- und Konsonantenalphabeten liegt in der Richtung: Der Zug für europäische Buchstaben läuft generell von links von rechts, bei arabischen umgekehrt. Im Gegenteil zu den alphabetischen werden für das Schreiben der chinesischen Schrift vielfältigere und kompliziertere Striche eingesetzt. Durchschnittlich wird ein Zeichen überdies (in der Regelschriftart) mit mehr als zehn Strichen geschrieben und die Strichfolge ist generell zuerst von oben nach unten sowie von links nach rechts (siehe Kap. 3.3.3, S. 156). Die Schreibweisekomplexität eines alphasyllabischen Syllabars lässt sich zwischen logographischen und alphabetischen Schriften verorten. In der Devanagari bspw. wird ein Syllabar mit drei bis sechs Strichen geschrieben. Dabei wird zuerst der graphisch-unterscheidende Strich unterhalb des Horizontals (wie der erste Zug ै für के /ke/), dann das Vertikal (क), das abhängige Vokalzeichen (े) und zuletzt das Horizontal der graphischen Grundlinie (क) geschrieben (vgl. Friedrich 2006: 25f).

Auf Basis beider Faktoren kann im Prinzip der zugehörige Schriftkreis des handgeschriebenen Zeichens bestimmt werden (voll-, konsonantenalphabetischer, alphasyllabischer oder CJK-Schriftkreis). Weiterhin kann die verwendete Schrift anhand der im System definierten eingebbaren Sprachen und weiterer Schriftfaktoren bestimmt werden. Besonders innerhalb des CJK-Schriftkreises, in dem die Zeichenmenge wie ausgeführt sehr hoch ist, ist eine nähere Bestimmung obligatorisch.

Unter der Voraussetzung, dass alle drei Sprachen als Eingabesprachen des multilingualen Eingabesystem definiert werden, kann das handgeschriebene Zeichen aus dem CJK-Schriftkreis vor dem exakten Glyphenmatching zuerst in eine von fünf Kategorien bestimmt werden: das simple Sinogramm, das komplexe Sinogramm, Hiragana, Katakana und Hangul-Syllabar. Anhand der geometrischen Konstellation und Strichmerkmalen sowie -beziehungen kann die Kategorisierung durchgeführt werden.

Unter den drei Arten des einfachen Aufbaus können die japanischen Silbenzeichen zuerst durch die Diakritika < ゛> oder < ゜> erkannt werden, falls sie dem Grundsyllabar beigefügt sind. Hiragana und Katakana können weiter durch Strichmerkmale unterschieden werden. Erstgenannte bezieht sich auf kurven- sowie rundförmige Formen. Die letztere indes weist fast identische Strichmerkmale zur Regelschriftart der chinesischen Schrift auf. Da es graphische Überschneidungen von Katakana- und simplen chinesischen Schriftzeichen geben kann, nachdem ein Zeichen als CJK-Zeichen des einfachen Aufbaus mit Regelschriftstrichen erkannt wurde, wird es zuerst mit der relativ kleinen Katakana-Glyphendatenbank (mit 48 Grundzeichen) abgeglichen. Danach wird es weiter mit den ca. 280 simplen chinesischen

Schriftzeichen[295] und Radikalen gematcht. Alle Zeichenglyphen, die graphische Ähnlichkeiten zum handgezeichneten Symbol aufweisen, werden dann in einer Wahlliste angeboten.

Unter den CJK-Zeichen des komplizierten Aufbaus – ergo den komplexe Sinogrammen und den Hangul-Syllabaren – sind die geometrische Konstellation und Schreibweise divergent. In Kap. 2.4 wurde ausgeführt, dass ein koreanisches Syllabar entweder zwei- oder dreiteilig (aus An-, In- und optional Auslaut) besteht und seine Struktur insgesamt sechs Varianten hat (siehe Abb. 2-11, S. 105). Währenddessen gibt es drei Grund- und mehrere mehrstufige sowie Nebenarten der proportionalen Konstellation der komplexen Sinogramme (siehe Kap. 3.3.5, Abb. 3-4, S. 161). Die ein bis vier Schreibzüge des Jamo verlaufen hauptsächlich horizontal, vertikal und kreisförmig und unterscheiden sich somit von den Strichen der chinesischen Schrift eindeutig. Wegen des relativ kleinen Inventars von 24 einfachen Jamo kann ein koreanisches Syllabar ebenso mit wenigen Schwierigkeiten bestimmt werden.

Die Statistiken aus der Rechnerwolke entscheiden auf Basis der Schrifterkennung, welche Kandidaten aus welchen Sprachen wahrscheinlicher sind: Anhand der geteilten Eingabestatistiken von Benutzern verschiedener Länder (bzw. mit unterschiedlichen Muttersprachen) werden Statistiken geschlussfolgert, welche Sprachen generell und von einer bestimmten muttersprachlichen Gruppe am häufigsten per Handschrift eingegeben werden. Bei der Analyse des Vorkontexts wird dabei die Sprache desselben erkannt, woraufhin diese als wahrscheinlichste Eingabesprache handgeschriebener Zeichen festgelegt wird.

4. Funktion der automatischen Inputcodierungskorrektur.

Wegen der Nachteile der phonetischen Inputcodierung, die im letzten Punkt erwähnt wurden, kann es beim Eintippen des Inputcodes leicht zu Fehlern kommen. Um gewisse typische Fehler automatisch zu erkennen und zu korrigieren, wird eine Inputcode-Korrekturliste für diese Funktion designet. Sie wird entweder durch rechtes Mausklicken geöffnet, wenn ein fehlerhafter Inputcode eingegeben wird, oder unter ‚Tools'. Diese Liste kann dann manuell bearbeitet werden, damit ein typischer Fehler zukünftig in die entsprechende korrekte Form umgeändert werden kann. Eine exemplarische Liste (für das Chinesische) könnte wie folgt aussehen:

[295] Vgl. *http://xh.5156edu.com/page/z2714m8730j18605.html* [2017-12-04].

Chinese		
[Example] Character	the wrong form	the right form
词	zi	ci
西	si	xi
呀	ja	ya
字	si	zi

Abb. 5-14: Design der Inputcode-Korrekturliste[296]

Wenn die falsche Inputcode-Form innerhalb eines Sprachsystems nicht für anderen Zeichen stehen kann, wird sie immer in die korrekte Form umgewandelt und zu den Kandidaten konvertiert. Wenn die falsche Form zufällig der richtige Inputcode für andere Kandidaten ist, wird sie sowohl in originaler als auch in korrigierter Form weiter verarbeitet. Der PC-Benutzer bestimmt die zum Schreibziel geeignete Variante, nachdem die betroffenen Kandidaten angeboten werden.

Zur Findung typischer Inputcodefehler einer bestimmten Sprache kann auch das Cloud Computing dienen. Anhand der Benutzerstatistiken wird analysiert, welche Fehler gehäuft auftreten. Dies hängt von Muttersprache, Dialekt und Sprachgewohnheit der Benutzer ab. Wenn beim Eintippen ein Fehler dieser Art unterläuft, kann das Computersystem mithilfe der Rechnerwolke den Fehler automatisch entdecken und korrigieren.

5. Maschinelle Übersetzung im Wort als Eingabemöglichkeit.

Nicht nur in der Fremdsprache kommt es unweigerlich zu Wortschatzlücken. Damit unbekannte Wörter mit dieser Software dennoch eingegeben werden können, wäre eine Eingabemöglichkeit durch eine andere Sprache vorstellbar. Um dies zu realisieren, müssten Worteinträge verschiedener Sprachen, die gleiche oder ähnliche Bedeutungen tragen, Querverweise bilden, z.B. durch Markierungen via Indikationsnummer.

Hat ein Chinesischschreiber bspw. das Zeichen für *Kaffee* vergessen, kann er durch *coffee* das chinesische Wort ausgeben lassen. Dafür muss nach der Eingabe des englischen Wortes

[296] Diese Liste kann bspw. für einen Chinesischsprecher aus Deutschland definiert werden. Die definierten Fehler werden durch die verschiedenen Rechtschreibungsregeln von Deutsch und dem chinesischem Pinyin verursacht. In der ersten Spalte wird der/das einzugebende Buchstabe / Syllabar / Schriftzeichen erfasst (für Chinesisch wird wegen der großen Menge an Homophonen nur ein Beispielzeichen für eine Silbe benötigt); unter ‚the wrong/right form' werden jeweils die falsche (die unerkannte Transkription) und die richtige Inputcodeform (die erkannte) angegeben. In den Kästchen können weitere typische Fehler definiert werden.

im Eingabefenster mit der rechten Maustaste getippt und danach die Funktion ‚Translation' und die Zielsprache mit der linken Maustaste ausgewählt werden. Das chinesische (Lehn-)Wort <咖啡> /kāfēi/ kann dann durch Kandidatenwahltreffen abgerufen werden.

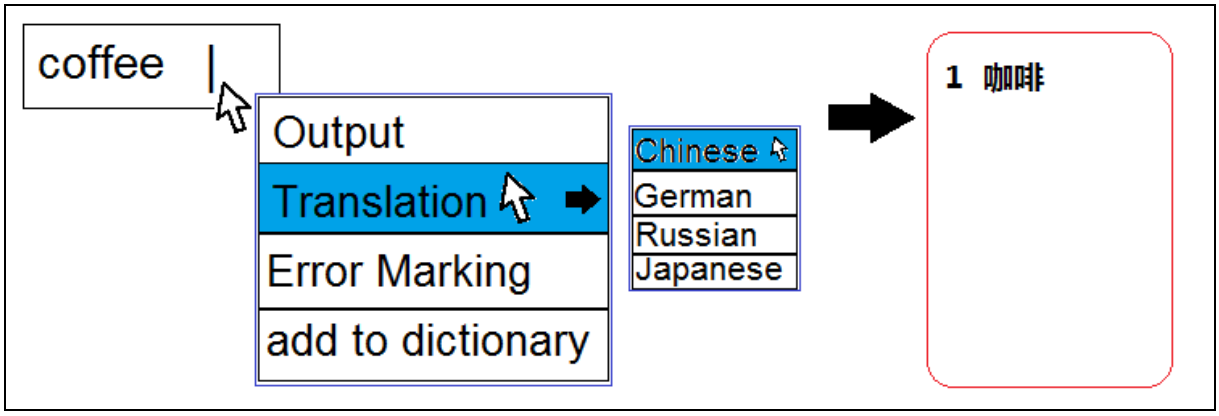


Abb. 5-15: Beispiel für die Eingabe-Möglichkeit der maschinellen Übersetzung[297]

[297] Wegen der Bedeutungsüberschneidungen zwischen verschiedenen Sprachen werden im Normalfall mehrere Kandidaten in der Kandidatenwahlliste angeboten.

6 Fazit, Schlussfolgerungen und Ausblick

In den ersten vier Kapiteln wurden die von Schriftsystemen bedingten Eingabetechniken zunächst überblickshaft (in Kap. 1), später detailliert beleuchtet: Zunächst die alphabetischen Schriften (Kap. 2), danach das komplexere morphologische Schriftsystem des Chinesischen (Kap. 4). Auf Basis der dargestellten Theorien wird in Kap. 5 schließlich eine multilinguale Eingabesoftware designt. Innerhalb der besagten Kapitel wurden Lösungssätze für die vier in der Einleitung angegebenen Leitfragen gesucht. Sie lassen sich wie folgt zusammenfassen:

- Leitfrage 1: Wie lassen sich Terminologien über Schriftsysteme nebst unterschiedlicher Schreibkulturen/-technologien einheitlich definieren und klassifizieren?

Wie in Kap. 1.3 vorgestellt, kann ein Schriftsystem (wie z.B. das deutsche) anhand des Typs (alphabetisch), der Schriftgattung (vollalphabetisch), der verwendeten Schrift (lateinisches Alphabet), dem Schriftkreis (der lateinische) und der Schreibrichtung (horizontal-rechtsläufig) kategorisiert werden. Die Schreibtechnologien weichen für jedes Schriftsystem dementsprechend ab, wie in Kap. 1.4 aus vier Perspektiven erklärt wurde: den vier Arten der Zeichencodierung (interne, Austausch-, Ausgabe- und Inputcodierung von verschiedenen Schriftzeichenarten), der Grundlage der Eingabetechniken (Zeichen-Tasten-Repräsentation oder Anwendung einer Inputcodierung), dem Zeichen-Glyphen-Zusammenhang und den computerlinguistischen Anwendungen zur Textverarbeitung.

- Leitfrage 2: Wie funktionieren die Arbeitsprinzipien der alphabetischen Eingabeverfahren?

Trotz der Zeichen-Tasten-Repräsentation als Grundlage weichen die Eingabetechniken wegen der verschiedenen alphabetischen Gattungen erheblich ab. In den Kapiteln 2.1 bis 2.5 werden Eingabefunktionalitäten von fünf Schriftsystemen an Beispielen schrittweise analysiert und in einem Zwischenfazit (Kap. 2.6) zusammengefasst. Entscheidend für Methodik und Effizienz der alphabetischen Eingabeverfahren sind zwei Faktoren: 1) die Bestimmung des Zeicheninventars für Codierung sowie Tastenbelegung und 2) eine für relativ effizientes Schreiben nutzbare effiziente Zeichen-Tasten-Repräsentation. Vier signifikante technische Schwierigkeiten fallen auf: 1) die Behandlung der abhängigen Zeichen in alphabetischen, alphasyllabischen und konsonantenalphabetischen Schriftsystemen; 2) die Glyphenausgabe einer alphasyllabischen Schrift; 3) die Buchstaben-Silben-Konversion des Koreanischen; 4) die linksläufige sowie bidirektionale Textverarbeitung der arabischen Schrift.

- Leitfrage 3: Fragen zur Grammatologie der chinesischen Schrift, die von ihr bedingten verschiedenen Inputcodierungen sowie Eingabemethoden und die Funktionalität der satzstufigen intelligenten Inputcode-Zeichen-Konversion.

Zur Beantwortung der dritten Leitfrage wurden in Kap. 3 linguistische Erkenntnisse über die chinesische Schrift eingeführt, während in Kap. 4 die technische Eingaberealisierung des Chinesischen skizziert wurde. Der morphologische Schrifttyp und das immense Zeicheninventar sind Ursachen für die Notwendigkeit von Inputcodierungen und Eingabesoftware. Grammatologisch bedingt – gemeint sind die drei Zeichengrundattribute – kann eine Zeicheninputcodierung nach Zeichenform- oder Ausspracheaspekten entworfen werden. Der erste Aspekt wird linguistisch in Kap. 3.3 erläutert und technisch in Kap. 4.1.2 am Beispiel der Wubi-Zixing-Eingabemethode vorgestellt. Der zweite Aspekt wird in Kap. 3.4 sprachwissenschaftlich fundiert; Informationen zur Pinyin-Eingabemethoden werden zudem in Kap. 4.1.3 und 4.1.4 zusammengefasst. Die meisten heute benutzten Pinyin-Eingabemethoden basieren auf den Techniken der satzstufigen intelligenten Laut-Zeichen-Konversion, die von Wissensdatenbanken (inkl. Wörterdatenbank und linguistischen Datenbank) unterstützt werden müssen.

- Leitfrage 4: Entwurf des multilingualen Eingabesystems.

Aufbauend auf den Lösungsansätzen der zweiten und dritten Leitfrage können die Eingabeverfahren in zwei Gruppen gegliedert werden: die von der Tastenbelegung und die von der Inputcode-Zeichen-Konversion abhängigen Eingabetechniken. Hauptidee des selbsterstellten multilingualen Eingabesystems ist, die beiden Eingabemöglichkeiten als verschiedene Module derselben Software zu integrieren. Eine Spracherkennung anhand schriftlicher und sprachlicher Eigenschaften eines Schriftsystems ist dabei erforderlich.

Auf Basis von Kap. 1 bis 5 können Lösungen für die Hauptfrage geschlussfolgert werden, die genauso nach Unterschieden und Gemeinsamkeiten tastaturbasierter Eingabeverfahren verschiedener Schriftsysteme fragt, wie nach der möglichst effizienten Nutzung der computergestützten Textverarbeitung und dem Zusammenhang zur Computerlinguistik. Ob das Zeicheninventar eines Schriftsystems auf dem Tastaturlayout belegbar oder ob es der funktionalen Tastatur angepasst ist, entscheidet über die angewendete Eingabemöglichkeit, die entweder auf der Zeichen-Tasten-Repräsentation oder der Inputcode-Zeichen-Konversion basiert. Die beiden Eingabemöglichkeiten funktionieren zwar in entgegensetzten Richtungen, haben aber Gemeinsamkeiten beim allgemeinen Tastaturaufbau (mit 47 oder 48 Zeichen-repräsentierenden Tasten) und der Codierungsumwandlung (von [Inputcode zu] Scan-Code, über internen Code, bis hin zu Ausgabecode). Die Effizienz der ersten Variante ist hauptsächlich

von der Tastenbelegung bedingt, während die letzte von Konversionstechniken und -qualität abhängig ist. Eingabeverfahren der ersten Variante hängen kaum direkt mit computerlinguistischen Anwendungen zusammen, werden aber indirekt von den Korrektursystemen verbessert. Im Gegensatz dazu ist intelligente Konversion, die für Eingabeverfahren der zweiten Varianten meist vonnöten ist, ohne computerlinguistische Erkenntnisse unrealisierbar.

Wie in Kap. 2.6 dargelegt, ist die Effizienzerhöhung des alphabetischen Eingabeverfahrens zwar durch Rationalisierung des Tastaturlayouts möglich, aber wegen der Schreibgewohnheit schwer durchzuführen, besonders bei einem lateinalphabetischen Schriftsystem. Rekurrierend auf das Zwischenfazit (Kap. 4.6) kann die Schreibeffizienz eines nicht-alphabetischen Schriftsystems zwar durch die Erhöhung der künstlichen Intelligenz verbessert werden, die Schriftkultur gleichzeitig aber negativ beeinflussen. Die Anwendung der hohen künstlichen Intelligenz in der Textverarbeitung kann sich rezessiv auf die schriftsprachlichen Fähigkeiten auswirken, sowohl bei einem alphabetischen Schriftsystem durch Korrektursysteme als auch bei einem nicht-alphabetischen durch automatische Laut-Zeichen-Konversion. Es ist deshalb ungewiss, in welche Richtung sich die verschiedenen tastaturbasierten Eingabeverfahren weiterentwickeln. Dem hingegen zeigen Eingabetechniken von Smartphones und Tablets eine rasante Entwicklung mit immer mehr Multilingualität sowie Vielfältigkeit und höherer künstlicher Intelligenz sowie Schreibeffizienz. Für die allgemeine Eingabe per visueller Tastatur (mit T9 oder vollständigem Layout) können Funktionen wie Sprach- sowie Worterkennung, automatische Korrektur, Prognose des kommenden Textes usw. eingesetzt werden. Neue Eingabemöglichkeiten durch Textscannen, Online-Handschriften, Sprechen, Übersetzungen usw. werden ebenfalls ermöglicht.

Ich bin persönlich der Meinung, dass in Zukunft verschiedene zielorientierte Eingabeverfahren sowie Textverarbeitungsmodule einer bestimmten Sprache existieren werden. Von den vielen denkbaren Anwendungskontexten sollen nachfolgend fünf exemplarisch durchgespielt werden. Dabei werden zugleich situationsangemessene Ideen bzgl. des Korrektursystemmoduls der deutschen Textverarbeitung (Beispiel für die alphabetischen Schriftsysteme) und der Eingabemethoden des Chinesischen vorgestellt.

1) In der Büroarbeit, bei der hohe Präzision und große Effizienz der Verarbeitung und Abschreibung der Texte erforderlich sind.

Bei der deutschen Textverarbeitung sind in dieser Situation eine präzise Überprüfung eines großen Wortschatzes sowie grammatische Phänomene und die Verringerung von Fehlermeldungen die wichtigsten Ziele. Sie können bspw. mit der Anwendung von Cloud Computing

erreicht werden, bei der Berufstätige und Facharbeiter derselben Branche miteinander Daten (inkl. Wörtern, syntaktischen sowie pragmatischen Informationen, Schreibvorlagen einer bestimmten Textsorte usw.) teilen können. In der chinesischsprachigen Büroarbeit muss die Laut-Zeichen-Konversion einer Pinyin-Eingabesoftware auf Ebene von Präzision und Effizienz gleichermaßen berücksichtigt erhöht werden. In Kap. 4.1.4 (S. 205) und 4.2.2 (S. 214) wurde die Anwendung der Zellwörterdatenbanken und des Cloud Computings in die moderne Pinyin-Eingabe angerissen. Mit Blick auf die Zukunft können die beiden Techniken kombiniert eingesetzt werden, so dass Statistiken jedes Büroberufstätigen unter der Voraussetzung des Datenschutzes geteilt, analysiert und zur Konversion verwendet werden können. Da Schreibfehler mit einer intelligenten Pinyin-Eingabesoftware (inkl. der Tippfehler der Inputcodierung und Konversionsfehler) relativ häufig passieren, dabei jedoch nur reduziert, nicht entfernt werden können, müssen zusätzliche Funktionen für die Büroeingabesoftware entwickelt werden. Ein Testschreibfeld ist aus diesem Grund vorzuschlagen: Textteile mit dem Umfang von weniger als 120 Zeichen werden dort zunächst eingegeben und können erst nach der manuellen Korrektur und Bestätigung in ein Textverarbeitungsprogramm hinzugefügt werden. Die korrigierten Textstellen können auch im selbstangepassten Lernmodul verarbeitet werden, um neue sprachliche Regeln und typische Benutzerfehler zu akquirieren.

2) Professionelles berufliches Schreiben von Schriftstellern, Journalisten und literarischen Übersetzern, deren Schreibkunst und Wortschatz häufig die von Computern zu leistende Sprachverarbeitungsfähigkeit überschreiten.

In diesem Fall zielt ein Korrektursystem oder eine Eingabesoftware hauptsächlich darauf ab, die Schreibstilkunst möglichst wenig zu stören. Dazu können in der deutschen Textverarbeitung einerseits die von Literaturarbeitern geteilten Daten in Cloud Computing und Korporaanalysen einer bestimmten Sprache verwendet werden. Andererseits muss das Korrekturmodul möglichst individuell orientiert sein. Ein verstärktes Selbstlernmodul, das idiolektale Wörter und Formulierungsstile eines bestimmten PC-Besitzers akquirieren kann, ist deswegen vonnöten. Ganz ähnlich lassen sich so Verbesserungsvorschläge für eine äquivalente chinesische Schreibsoftware zusammenfassen: Kompatibilität mehrerer Inputcodierungen (außer Pinyin- mindestens eine zeichenformbasierte Inputcodierung); umfangreichere fachliche Zellwörterdatenbanken; verstärkte Anwendung von Regelakquisition anhand klassischer sowie moderner literarischer Korpora; ein dezidiert selbstangepasstes Lernmodul.

3) Schreibübungen von mutter-/sekundärsprachlichen Schülern zur positiven Entwicklung ihrer Sprachkompetenz.

Zur Unterstützung von minderjährigen Muttersprachlern bei der Textverfassung müssen Korrektursysteme und intelligente Eingabesoftwares positiv auf deren Sprachgewohnheiten einwirken. Dafür ist es nötig, die Abhängigkeit von der künstlichen Intelligenz und Störungen durch Fehlermeldungen nach Möglichkeit zu vermeiden.[298] Um die deutschsprachige Textverarbeitung für Schüler geeigneter zu machen, kann das Korrektursystem mit einer Online-Wörterbuch-Webseite verlinkt werden. Wenn ein Schreibfehler gefunden wird, zeigt das System anstelle von Korrekturvorschlägen Informationen zum betroffenen Worteintrag an. Auch die im Text auftretenden Wörter, Phrasen etc., bei denen der Schreiber unsicher ist, können mithilfe des Korrektursystems sofort per Klick auf die rechte Maustaste nachgeschlagen werden. Weiterhin kann das Korrektursystem typische Fehler eines PC-Benutzers bei Buchstabierung (wie z.B. Fehler bei Phonem-Graphem-Repräsentationen), Morphologie (z.B. Flexion) und Syntax (z.B. Dependenzgrammatik und Phraseologie) analysieren, diese kategorisieren und entsprechende sprachliche Daten abrufen. Bei einer intelligenten Pinyin-Eingabesoftware ist bei minderjährigen Chinesen einerseits eine Toninputcodierung vonnöten. Andererseits muss das Bewusstsein über Zeichenformen trotz der automatischen Satzgeneration berücksichtigt werden. Wie in Kap. 4.1.3 erwähnt, zählt der Ton zu einem der drei wesentlichen phonetischen Attribute, so dass ein betonter Pinyin-Code viel weniger Zeichenkandidaten betrifft. Damit die satzstufige Laut-Zeichen-Konversion möglichst wenig zur ‚sprachlichen Faulheit' der Schüler beiträgt, können statt eines kompletten Satzes Wörter und Zeichen einzeln angeboten werden. Das System zeigt anhand von Kontextanalysen die wahrscheinlichsten Kandidaten ganz vorne in der Wahlliste, so dass die Eingabe trotz der Satzzerlegung in akzeptabler Geschwindigkeit ablaufen kann. Außer der phonetischen muss in der Eingabesoftware mindestens eine Zeichenform-, Form-Aussprache-Kombination- oder Aussprache-Form-Kombination-basierte Inputcodierung parallel kompatibel sein. Die Handschrift-Eingabealternative (entweder durch OCR oder Online-Handschreiben) ist ebenso empfehlenswert.

[298] Zu letztgenannten zählen solche Wortformen oder Satzteile, die vom System als Fehler erkannt werden, obwohl sie richtig sind. Es kann in diesem Sprachstadium noch nicht vorausgesetzt werden, dass Schüler die Inkorrektheit der angebotenen Wort- und Satzteile erkennen.

4) Freizeitschreiben in sozialen Netzwerken und anderen Kontexten, in denen ‚schnelles Schreiben' bevorzugt wird.

Die zum Unterhaltungsziel geeignete Textverarbeitung, wie z.B. der Chat in einem sozialen Netzwerk, erfordert ein besonders hohes Schreibtempo. Die umgangssprachliche Eingabemethode ist deswegen zu bevorzugen. Es ist zu erwarten, dass diese Eingabemöglichkeit dank verbesserter Datenaustauschtechniken, schnellerem Internet, größerem Arbeitsspeicher und der Verbreitung von Smartphones tendenziell effektiver und qualitativ hochwertiger wird. Technische Lücken der Sprach-Text-Konversion müssen durch manuelle Korrekturen ergänzt werden.

5) Schreiben in einer Fremdsprache – Didaktisch motivierte Fehlerkorrektur.

Zur Entwicklung von Textverarbeitungstechniken, die sich Fremdsprachlern anpassen, müssen sowohl die im ersten (Büroarbeit) als auch im dritten Punkt (minderjährige Muttersprachler) erwähnten Probleme berücksichtigt werden. Es handelt sich gewissermaßen um eine Kombination beider Fälle, weshalb die deutschsprachigen Korrektursysteme und intelligenten chinesischen Eingabesoftwares angewendet werden können: Wenn ein Fremdsprachler in Deutsch schreibt, hat er die Möglichkeit, zwischen präzisen automatischen Korrekturvorschlägen und dem Abruf der verlinkten Wortinformationen zu entscheiden. Ein fremdsprachlicher Benutzer mit einer chinesischen Pinyin-Eingabesoftware kann ebenso zwischen dem Status des Satzangebots von hoher Präzision und der zerlegten Auswahl von einzelnen Wort-/Zeichenkandidaten umschalten. Wegen verschiedenen Vorkenntnissen über die chinesische Schrift werden verschiedene Inputcodierungen für die Fremdsprachler außer- und innerhalb des CJK-Schriftkreises empfohlen. Es ist für japanische und koreanische Chinesischlerner wichtig, diese möglichst wenig durch ihre muttersprachliche Zeichenaussprache zu verwirren, da sie bereits vor Lernbeginn ca. 2.000 Schriftzeichen beherrschen. Empfehlenswert ist deshalb eine phonetische Inputcodierung. Bei anderen Fremdsprachlern ist hingegen entscheidend, das Zeichenbewusstsein auszubilden und zu verstärken. Eine relativ häufige Online-Handschrift sowie zeichenformbasierte oder Zeichenform-Aussprache-Inputcodierungen sind deswegen sinnvoll.

Durch die allgemeine Verbreitung von künstlicher Intelligenz werden elektronische Medien wie Computer und Smartphones in der Zukunft immer häufiger menschliche Aufgaben der Verfassung, Bearbeitung, Übersetzung usw. von Texten übernehmen. Die voll- oder semiautomatisch von Computern verarbeiteten Texte reichen immer näher an menschliche Sprachformulierungen heran. Maschinelle ‚Schreibassistenten' prägen die natürlichen Sprachstilfor-

mulierungen und Schreibgewohnheiten stark mit. Anstelle menschlicher Kreationen können Texte so mit immer weniger geistiger Arbeit ‚industriell' produziert werden, was wiederum die nächste Textverarbeitungsreformierung mit sich bringen könnte. In diesem Szenario ließe sich eine Blütezeit der effizienten Büroarbeit genauso imaginieren, wie eine , Epoche des Literaturuntergangs'.[299] Es müsste vor diesem Hintergrund Berücksichtigung finden, wie Eingabetechniken unter Schonung der Schriftkultur entwickelt werden können.

Wegen der Umfangsbegrenzung können viele weitere Faktoren des Forschungsbereichs nicht berücksichtigt werden. Dazu zählen bspw. alternative Tastaturlayouts des lateinischen Alphabests (wie z.B. die Davorak-Belegung), die in Smartphones angewendeten multilingualen intelligenten Eingabemethoden, verschiedene japanische Eingabemethoden, der von Eingabemethoden verursachte Sprachwandel usw.

In der Einleitung habe ich die Forschungsabsicht dieser Arbeit mit einem Tropfen Wasser innerhalb eines aus hundert Flüssen entstandenen Meeres verglichen, der in seiner Größe zwar eingeschränkt, aber inhaltlich informationsreich ist. Der Verweis auf weitergehende Forschungen ist somit kein Allgemeinplatz – sondern im Zeitalter der Digitalisierung zwangsläufig.

[299] Wie nah derlei Szenarien bereits an der Realität sind, wird unter anderem anhand eines ‚neuen' Harry-Potter-Kapitels ersichtlich, welches von einer künstlichen Intelligenz geschrieben und 2017 von den Medien aufgegriffen wurde (Vgl. z.B. *https://www.businessinsider.de/there-is-a-new-chapter-in-harry-potters-story-and-it-was-written-by-artificial-intelligence-2017-12?r=US&IR=T* [Abruf: 2018-06-14]). Obgleich die literarische Qualität als unterirdisch rezipiert wurde, zeigt das Beispiel die Potentiale voll- oder semiautomatischer Textproduktionssoftwares, wie sie im Fließtext beschrieben werden.

7 Anhang

7.1 Parallelindex der Fachbegriffe in Deutsch, Englisch und Chinesisch

Nr.	Deutsch	Englisch	Chinesisch	Kapitel [Seite]	bezog. Begriffe[300]
1	Abhängiger Vokal	dependent vowel	依附元音	2.3.1 [87ff], 2.5.1 [118f].	2, 3, 6, 19, 41.
2	Abjad / Konsonantenschrift / partielles Alphabet	abjad / consonant script	辅音音素文字	1.3.2 [22], 1.3.5 [34f], 2.5 [115-128].	6, 9.
3	Abugida / Alphasyllabar / alphasyllabische Schrift	abugida / alphasyllabary	元音附标文字	1.3.2 [22f], 1.3.5 [33ff], 2.3 [87-101].	1, 19, 41.
4	Alphabet / alphabetische Schrift	alphabet	字母文字	1.3.2 [22f], 1.3.5 [33ff], 2 [59-131].	20, 56, 62, 125.
5	ASCII	ASCII (American Standard Code for Information Interchange)	ASCII 代码	1.4.2 [47f].	31, 101, 102.
6	arabisches Alphabet	Arabic alphabet	阿拉伯字母	1.3.5 [34f], 2.5 [115-128].	2, 9.
7	BMP/ Plane 0	Basic Multilingual Plane (BMP)	基本多文种平面/ 第零平面	1.4.2 [49ff]	101, 102
8	Basiszeichen	base character	基本字符	1.4.1 [44f], 1.4.2 [47].	20, 31, 50.
9	bidirektionale Schrift / BIDI-Schrift	bidirectional writing / BIDI-writing	双向文字	1.3.6 [36ff], 2.5 [115-128].	2, 6.
10	chinesische Schrift	Chinese scripts	汉字	1.3.5 [33ff], 3 [133-176].	14, 34, 35, 47, 109.
11	chinesisches Schriftsystem	Chinese writing system	汉语文字系统 / 中文系统	3.1-3.4 [133-170].	10, 28, 40, 58, 61.
12	Chinesische Schriftzeichen / Sinogramm	Chinese character / Sinogram	汉字	1.3.5 [33ff], 3 [133-176].	10, 14, 28, 34, 35, 47, 51, 79, 80, 86.
13	chinesische Zeichencodierung	Chinese character encoding	汉字编码	1.4.2 [48ff], 3.2.2 [143ff], 4.1.1 [181ff].	15, 16, 17, 18, 39, 40.
14	CJK vereinheitlichte Ideographen	CJK Unified Ideographs	中日韩统一表意文字	1.4.2 [49f], 3.2.3 [145f].	12, 34, 35, 47, 101, 102.

[300] Entspricht der Nummerierung der Termini in dieser Tabelle.

Nr.	Deutsch	Englisch	Chinesisch	Kapitel [Seite]	bezog. Begriffe[300]
15	Codierung nach der Aussprache-Form-Kombination	coding by reading-structure-combining	音形编码	4.1.1 [183]	13, 68, 124.
16	Codierung nach der Form-Aussprache-Kombination	coding by structure-reading-combining	形音编码	4.1.1 [184]	13, 68, 124
17	Codierung nach der Phonetik	coding by reading	字音编码	3.4.3 [168ff], 4.1.1 [183], 4.1.3 [195-202].	13, 57, 68, 69.
18	Codierung nach der Zeichenform	coding by structure	字形编码	3.3.4 [157-161], 4.1.1 [181ff], 4.1.2 [187-195].	13, 117, 118, 124.
19	Devanagari	Devanāgarī	天城文	2.3 [87-101]	3, 41.
20	Diakritika / diakritisches Zeichen	diacritic / diacritical mark	附加符号	1.4.1 [44f], 2.1.3 [71-74], 2.2.2 [79ff], 2.6 [129f].	8, 46, 56.
21	Eingabegerät	input device	输入设备	1.2.2-1.2.3 [15-18], 1.4.3 [51ff].	59, 75, 92, 94.
22	Eingabemethode	input method	输入法	1.1 [8f], 1.4.3 [53f], 1.4.5 [57f], 2.4.3 [108-111], 4 [177-309].	23, 24, 57, 93, 99.
23	Eingabesoftware	input software	输入法软件	4 [177-309]	22, 57, 93, 99.
24	Eingabeverfahren	input procedure	录入法	1.1 [7-10], 1.4.3 [51-54], 2.6 [128-131], 4.6 [305-309].	22, 92, 125.
25	Font	font	字体/字库	1.4.1 [45], 1.4.4 [54ff].	27, 119.
26	funktionales Tastaturlayout	functional layout	功能布局	1.4.3 [53]	59, 110, 125.
27	Glyphe	glyph	字形	1.4.1 [42-46], 2.3.3 [94f].	25, 119.
28	Grammatologie	grammatology	字法学	3 [133-176]	12
29	Grammwurzel	gram-root	字根	3.3.1 [147], 3.3.4 [157], 4.1.2 [187-195].	51, 53.
30	Graphem	grapheme	字位	1.3.5 [33ff], 2.1.1 [62].	77, 79.
31	graphisches Zeichen	graphic character	可视字符	1.4.2 [47]	5, 8, 101, 102.
32	Halbpinyin	half pinyin	简拼	4.1.3 [197f], 4.3.2 [230f].	60, 68, 111.
33	Hangul	hangul	谚文	1.3.5 [35], 2.4 [101-114].	22, 43, 44.
34	Hanja / sinokoreanisches Schriftzeichen	hanja	朝鲜汉字	3.5 [171-176].	12, 14, 33.
35	Hanzi	hanzi	（中国）汉字	3 [133-176]	12, 14, 98, 108.

Nr.	Deutsch	Englisch	Chinesisch	Kapitel [Seite]	bezog. Begriffe[300]
36	Hiragana	hiragana	平假名	3.5.1 [171f]	20, 45, 46.
37	Homophone	homophone	同音字	3.4.1 [163f], 4.1.3 [196-199], 4.3.1 [227f], 4.5.1 [287ff].	57, 68.
38	horizontales Schreiben	horizontal writing	横排书写	1.3.6 [36-41]	2, 3, 9, 10, 19, 56, 76, 109.
39	Informationsverarbeitung des Chinesischen	Chinese information processing	中文信息处理	1.4.5 [57f], 4.1.1 [178-186], 4.3-4.6 [223-309].	40, 42, 93, 105, 106.
40	Informationsverarbeitung der chinesischen Schrift	Chinese character information processing	汉字信息处理	1.4.5 [57f], 4.1.1 [178-186], 4.3 [223-242].	13, 15, 16, 17, 18, 42, 107, 120.
41	inhärentes Vokal	inherent vowel	固有元音	2.3.3 [91f]	1, 3, 19.
42	Inputcodierung / Eingabeschema	input code	输入码	1.4.1 [42], 1.4.3 [51-54], 4.1 [177-207], 4.6 [305-309], 5.2.1 [319-322].	15, 16, 17, 18, 40, 99.
43	intelligente Eingabesoftware	intelligent input software	智能输入法	4.1.4 [202-207], 4.2-4.6 [208-309], 5.2 [319-344].	57, 62, 74, 105, 112, 116.
44	Jamo	jamo	谚文字母	2.4 [101-114]	33
45	japanisches Schriftsystem	Japanese writing system	日语文字系统/日文系统	3.5 [171-176]	22, 36, 47, 48, 57.
46	Kana	kana	假名	3.5.1 [171f]	36, 48.
47	Kanji / sinojapanisches Schriftzeichen	kanji	和氏汉字	3.5 [171-176]	12, 45.
48	Katakana	katakana	片假名	3.5.1 [171f]	20, 45.
49	Klassenhaupt	indexing component	部首	3.2.1 [140], 3.4 [157-161].	12, 51, 53, 85.
50	kombinierendes Zeichen	combining mark	组合字符	1.4.1 [44f], 1.4.2 [47], 2.3 [87-101], 2.5 [114-128].	1, 8, 20, 41.
51	Komponente (der chinesischen Schriftzeichen)	(Chinese character) component	（汉字）部件	3.3 [146-161]	80, 86, 124.
52	komplexes Wort	compound word	复合词	4.4 [242-284]	84, 113, 115.
53	komplexes Zeichen	compound character	合体字	3.3 [146-161]	51, 85.
54	Konversion	conversion	变换	4.2 [208-223], 5.1.3 [316f], 5.2 [319-344].	55, 57, 63.

Nr.	Deutsch	Englisch	Chinesisch	Kapitel [Seite]	bezog. Begriffe[300]
55	Konversionslexikon	conversion dictionarty	转换字典	4.1.1 [178f], 5.2.3 [330ff].	22, 54, 57.
56	lateinisches Alphabet	Latin alphabet	拉丁文字	1.3.4 [29-33], 2.1-2.2 [59-86].	4, 125.
57	Laut-Zeichen-Konversion	syllable-to-word conversion	音字转换	4.1.3 [196], 4.2 [207-223], 4.5.1 [285-290].	22, 54.
58	logographische Schrift / Logographie	logography	表语文字	1.3.2 [21f], 1.3.5 [33ff], 3.1.2 [135ff].	10, 61, 62.
59	mechanisches Tastaturlayout	mechanical layout	机械布局	1.4.3 [53]	26, 110.
60	Mischpinyin	mix pinyin	混拼	4.1.3 [197f], 4.3.2 [230f].	32, 68, 111.
61	morphologische Schrift	morphology / morphemic writing	语素文字	3.1.2 [135ff]	10, 58.
62	nicht-alphabetisches Schriftsystem	non-alphabet	非拼音文字系统	1.3.1 [20f], 4.6 [305-309].	4, 10, 36, 46, 48, 91.
63	Nicht-Konversion	non-conversion	无变换	2.6 [128-131], 5.1.3 [316f].	54, 57.
64	Parsing	parsing	语法解析	4.4.4-4.4.5 [271-284], 4.5.3 [295-300].	43, 74
65	POS-Tagging/ Wortart-Tagging	part-of-speech tagging / POS tagging	词性标注	4.4.3 [261-270]	43, 74.
66	phonetische Eingabemethode	input method by reading	语音输入法	4.1.1 [183], 4.1.3-4.1.4 [195-207].	57, 68.
67	Phonographie / phonographische Schrift	phonography	表音文字	1.3.1-1.3.5 [19-35]	2, 3, 4, 91.
68	Pinyin	pinyin	拼音	3.4.3 [168ff], 4.1.3 [195-202], 4.3 [223-242].	11, 17, 69, 70.
69	Pinyin-Eingabemethode	pinyin input method	拼音输入法	4.1.3-4.1.4 [195-207], 4.2-4.6 [208-309].	57, 68.
70	Putonghua / modernes Standardchinesisch	putonghua	普通话	3.1.1 [134], 3.4.2 [166f].	11, 68
71	QWERTY-Tastaturbelegung	QWERTY keyboard layout	QWERTY 键盘布局	1.1 [7f], 1.2.2 [16], 1.4.3 [51ff], 2.1-2.2 [59-86].	92, 94, 125.
72	Satzgeneration	sentence generation	句子生成	4.5.4 [302-305]	73, 105.
73	satzstufige Eingabemethode	sentence-level input method	语句级输入法	4.1.3-4.2.3 [195-223]	22, 43, 74, 105.
74	satzstufige Laut-Zeichen-Konversion	sentence-level syllable-to-word conversion	语句级音字转换	4.1.3-4.2.3 [195-223]	43, 57, 73, 105.
75	Scan-Code	scancode	扫描码	1.1 [9f], 1.4.3 [51f].	21, 71, 94, 125.

Nr.	Deutsch	Englisch	Chinesisch	Kapitel [Seite]	bezog. Begriffe[300]
76	Schriftrichtung	writing directionality	文字方向	1.3.6 [36-41], 2.5.2-2.5.3 [122-128].	38, 109.
77	Schriftsystem	writing system	文字系统	1.3 [19-41]	24, 78, 79.
78	Schrifttyp	script type	文字类别	1.3.1-1.3.2 [19-25]	4, 58, 91.
79	Schriftzeichen	character	文字符号	1.3 [19-41], 1.4.1 [41-46].	77, 78.
80	sechs Schriften / sechs Klassen der Schriftzeichen	the six writings	六书	3.3.1-3.3.2 [147-155]	12, 90, 98, 108, 124.
81	Segmentationseinheit	segmentation unit	切分单位	4.4.2 [251-261]	52, 84.
82	selbstangepasstes Lernen	self-adaptation learning	自适应学习	4.5 [284-305]	43, 105.
83	Silbensegmentation	syllable segmentation	音节切分	4.3.2-4.3.3 [228-237]	68, 69, 107.
84	simples Wort	simple word	单纯词	4.4.1 [242-245]	52, 81.
85	simples Zeichen	simple character	单体字	3.3 [146-161]	53, 86.
86	Strich	stoke	笔画	3.3.3 [155f]	12, 51, 76, 88, 89.
87	Strichanzahl	stroke count	笔数	3.3.3 [155f]	12, 86.
88	Strichmerkmal	stroke feature	笔形	3.3.3 [155f]	12, 86.
89	Strichordnung	stroke order	笔顺	3.3.3 [155f]	12, 86.
90	Strukturoriginal	structure origin	结构理据	3.3.4 [157-161]	12, 49, 51, 53, 80.
91	Syllabar / Silbenschrift / syllabische Schrift	syllabary	音节文字	1.3.2 [22-25], 3.5.1 [171f].	3, 33, 36, 46, 48, 78.
92	tastaturbasierte Eingabe	input via keyboard	键盘输入	1.1 [7-10], 1.4 [41-58], 4.1 [177-207].	21, 42, 125.
93	tastaturbasierte Eingabemethode der chinesischen Zeichencodierung	Chinese character coding keyboard input method	汉字编码键盘输入法	4.1 [178-207]	15, 16, 17, 18, 69, 117.
94	Tastaturlayout/-belegung	tastatur layout	键盘布局	1.1 [7f], 1.4.3 [51f], 2 [59-131], 5.2.1 [320-323].	21, 125.
95	Textverarbeitung	word processing	文字处理	1.2 [10-18], 1.4.5 [56ff], 2.6 [128-131], 4.6 [305-309].	22, 24, 39, 40.
96	Ton	tone	音调	2.2 [77-86], 3.4 [162-170].	11, 70.
97	Tonem	toneme	调位	3.4.2 [166f]	11, 70.
98	traditionelles chinesisches Schriftzeichen	traditional Chinese characters	繁体字	3.1.1 [134]	12, 35, 108.

Nr.	Deutsch	Englisch	Chinesisch	Kapitel [Seite]	bezog. Begriffe[300]
99	Transkription	transcription	音标	1.3.4 [32], 2.2.4 [85f], 3.4.3 [167-170], 5.2.1 [321f].	6, 11, 33, 42, 45, 56, 68, 70, 100.
100	Transliteration	transliteration	转写	1.3.4 [32]	6, 56, 99.
101	Unicode	Unicode (the Unicode Standard)	万国码	1.4.1-1.4.2 [41-52]	5, 7, 14, 102.
102	Unicode-Block	Unicode block	万国码字符集	1.4.2 [49ff]	5, 7, 14, 101.
103	unoriginaler Abbau	unoriginal disassembly	无理拆分	3.3.4 [157-161]	90, 124.
104	unregistriertes Wort	unregistrered word	未登录词	4.4.4 [271-280]	52, 106, 116.
105	Verarbeitung in Satz	sentence processing	句处理	4.1.1 [186], 4.1.3 [200f], 4.5 [284-305].	43, 73, 74, 106, 107.
106	Verarbeitung in Wort	word processing	词处理	4.1.1 [185f], 4.1.3 [200f], 4.4 [242-284].	55, 65, 105, 106, 115, 116.
107	Verarbeitung in Zeichen	character processing	字处理	4.1.1 [185], 4.1.3 [199], 4.3 [223-242].	55, 83, 105, 106.
108	vereinfachte chinesische Schriftzeichen	simplified Chinese characters	简体字	3.1.1 [134]	12, 35, 98.
109	vertikale Schreibung	vertical writing	纵排书写	1.3.6 [36-41]	10, 76.
110	visuelles Tastaturlayout	visual layout	可视化布局	1.4.3 [53]	26, 59.
111	Vollpinyin	full pinyin	全拼	4.1.3 [197f], 4.3.2-4.3.3 [228-237].	32, 60, 68.
112	Wissensdatenbank	knowledge database	知识库	4.2.3 [217-223], 5.2.3 [330ff].	43, 74.
113	Wortbildung	word formation	组词	4.4.1 [246-251], 4.4.4 [271-280].	52
114	Wortfrequenz	word frequency	词频	4.3.4 [237-242]	52, 81, 84.
115	Wortsegmentation	word segmentation	词语切分	4.4.2 [251-261], 4.5.1 [285f].	52, 81, 84.
116	Wörterdatenbank	word database	词库	4.2.3 [217-223]	52, 81, 84, 123.
117	Wubi-Zixing-Eingabemethode	Wubi-zixing input method	五笔字型输入法	4.1.2 [186-195]	18, 118.
118	Wubi-Zixing-Schema	Wubi-zixing encoding	五笔字型编码	4.1.2 [186-195]	18, 117.
119	Zeichen	character	字符	1.4 [41-58]	3, 8, 31, 50.

Nr.	Deutsch	Englisch	Chinesisch	Kapitel [Seite]	bezog. Begriffe[300]
120	Zeichencodierung	character encoding	字符编码	1.4.1-1.4.3 [41-54], 2.6 [129f], 3.2.2-3.2.3 [142-146], 4.1.1 [181-184].	5, 13, 42, 101.
121	Zeichenfrequenz	character frequency	字频	2.1.2 [67f], 2.1.4 [74f], 4.3.4 [237-242].	12, 30, 114.
122	Zeicheninventar	character quantity	字量	1.4.2 [47ff], 2.1.1 [60ff], 2.2.2 [79ff], 2.3.3 [91-94], 2.4.2 [103-107], 2.5.1 [114-120], 3.2.2 [142-145].	77, 101.
123	Zeichenlexikon	character dictionary	字典	3.2.1 [140ff]	12, 49, 51, 68, 80, 90.
124	Zeichenstruktur	character structure	字体结构	3.3.4-3.3.5 [157-161]	51, 80, 90.
125	Zeichen-Tasten-Repräsentation	character-key-representation	键位字符对应	1.4.3 [51f], 2.6 [128-131].	24, 94.

7.2 Verzeichnis der Abbildungen

Abb. 1-1: ISO 9995-3: 2010, die internationale Norm für Tastaturbelegung (Pentzlin 2010: 1)..... 8
Abb. 1-2: Der vereinheitlichte Scan-Code dieser Dissertation (von dem alphanumerischen Block der PC-Tastatur mit 102 Tasten) und die Tastenkontrolle nach Zehnfingersystem 10
Abb. 1-3: Der Arbeitsprozess des Computers bei Textverarbeitung 18
Abb. 1-4: Klassifikation der Schriften und Schriftsysteme 25
Abb. 1-5: Vergleich der Grapheme in verschiedenen Schriftsystemen 34
Abb. 1-6: Die Ambiguitäten des Chinesischen wegen verschiedenen Schriftrichtungen 39
Abb. 1-7: Varianten für die Datumsangabe in verschiedenen Regionen 41
Abb. 2-1: DIN 32743, Teil 8, Endgeräte für die Textkommunikation; Nationaler Teletex-Schriftzeichenvorrat (DIN-Taschenbuch 210: 91) 61
Abb. 2-2: Die deutsche Tastaturbelegung nach DIN 2137-2 (DIN-Taschenbuch 210: 21) 65
Abb. 2-3: Das belgische Tastaturlayout 73
Abb. 2-4: Das schweizerische Tastaturlayout 73
Abb. 2-5: Das Tastaturlayout des vietnamesischen Nationalstandards – TCVN 6064:1995 (Đỗ 2005: 14) 83
Abb. 2-6 [links]: Die Inputcodierung der Eingabemethode Tiếng Việt VNI 85
Abb. 2-7 [rechts]: Beispiel für die ‚Eingabemethode per automatischer Worterkennung' 85
Abb. 2-8: Schriftliche Darstellung des Hindi-Wort /pūrti/ (Unicode 12.0 Chapters: Kap. 2.2: 17)..... 90
Abb. 2-9: Graphische Silbenstruktur für Glyphen-Zusammensetzungen 99
Abb. 2-10: Das Hindi-Inscript-Tastaturlayout 100
Abb. 2-11: Varianten zur proportionalen Konstellation der Komponenten in Syllabaren 105
Abb. 2-12: Schriftform des Wortes *Hanja* in Hangul 108
Abb. 2-13: Das koreanische Standardtastaturlayout 110
Abb. 2-14: Die dem Silbenzeichen <한> /han/ entsprechende, gebräuchliche Hanja 111
Abb. 2-15: Die neun möglichen Kombinationsarten des Jamo <ㅈ> /j/ 113
Abb. 2-16: Das arabische Tastaturlayout Sakhr/MSX 121
Abb. 2-17: Das arabische Tastaturlayout IBM-PC 121
Abb. 2-18: Schriftform vom arabischen Beispielwort /ktāb/ 123
Abb. 2-19: Der arabische Beispieltext von zwei Schriftrichtungen 124
Abb. 2-20: Die Ausgabe des arabischen bidirektionalen Beispieltexts mithilfe von direktionalen Formatzeichen 128
Abb. 3-1: Abschnitt von CJK Unified Ideographs mit Variantenglyphen (Unicode 12.0 Character: U+4E0D-U+4E10) 146
Abb. 3-2: Segmentierung des Beispielzeichens <德> /dé/ in Komponenten 159
Abb. 3-3: Segmentierung des Beispielzeichens <國> /guó/ in Komponenten 159
Abb. 3-4: Die Varianten der proportionalen Konstellation der chinesischen Schrift 161

Abb. 3-5: Die Dreieckbeziehungen zwischen Zeichenform, -aussprache und -sinninhalt eines chinesischen Schriftzeichens 163
Abb. 4-1: Der allgemeine Arbeitsprozess der chinesischen Eingabemethoden (Wu 1999: 7 [Übersetzung der Verfasserin]) 178
Abb. 4-2: Das Wubi-Tastaturlayout, Version-86 189
Abb. 4-3: Tastaturbelegung für Doppelpinyin 198
Abb. 4-4: Eingabe in einzelnen Zeichen am Beispiel <是> (/shì/, *sein*) 199
Abb. 4-5 [links]: Eingabe in Wort am Beispiel von <事实> (/shìshí/, *Tatsache*) 201
Abb. 4-6 [rechts]: Eingabe im Satz am Beispiel <这是事实> (/zhè shì shìshí/, *Das ist Tatsache*) 201
Abb. 4-7: Korrektur des Tippfehlers von ‚Y' zu ‚Z' bei Satzeingabe 203
Abb. 4-8: Einstellung des verschwommenen Pinyin der Sogou-Pinyin-Eingabesoftware 204
Abb. 4-9: Eingabe per Online-Handschreiben mit Sogou-Pinyin-Eingabesoftware 205
Abb. 4-10: Wortprognose 206
Abb. 4-11: Satzprognose 206
Abb. 4-12 [links]: Struktur der auf sprachlichem Verstehen basierenden Eingabesoftware (Chen YF/Zhu 2002b: 13 [Übersetzung der Verfasserin]) 216
Abb. 4-13 [rechts]: Struktur der auf Sprachmodell-Matching basierenden Eingabesoftware (ibid.: 16) 216
Abb. 4-14: Struktur der auf Zusammenhängen im Kontext basierenden Eingabesoftware (ibid.: 17) 217
Abb. 4-15: Die vier verschiedenen Verarbeitungsniveaus der chinesischen Korpora 222
Abb. 4-16: Die Eingabe eines Worts mit ER-Laut, in dem /er/ als eigenständige Silbe behandelt werden muss 225
Abb. 4-17: Die automatische Korrektur zur aktuellen Standard-Aussprache eines Heteronyms mit variierten Aussprachen 226
Abb. 4-18: Der Verarbeitungsprozess von ‚XIAN' zu Schriftzeichen 230
Abb. 4-19: Beispiel für die Ambiguität der Segmentation einer Pinyin-Kette in Halbform 231
Abb. 4-20: Definition der Knoten bei Silbensegmentation (nach Liu ZY/Wu/Liu 2008: 36 [Übersetzung der Verfasserin]) 234
Abb. 4-21: Die erste (1), die zehnte (2) und die 87ste Seite (3) der Wahlliste bei der Eingabe der Pinyin-Kette ‚XIANGANG' der Sogou-Pinyin-Eingabesoftware 236
Abb. 4-22: Die erste Seite der Wahlliste bei der Eingabe der Pinyin-Kette ‚XIANGANG' von Microsoft-Pinyin-IME 236
Abb. 4-23: Die Eingabe von ‚YEDL' für das Idiom <掩耳盗铃> (der 2. Kandidat) in Halbform 237
Abb. 4-24: Die erste Seite der Wahlliste von ‚SHISHI' bei zwei verschiedenen Eingaben 242
Abb. 4-25: Die Vernetzung der Wortsegmentation von der Zeichenkette ‚中国人民生活' (vgl. Wang XL 2005: 36) 258
Abb. 4-26: Die möglichen Wörter von der Pinyin-Kette ‚TA'SHI'DE'GUO'REN' und die Pfade der möglichen Segmentation 260
Abb. 4-27: Wahlliste von ‚TA'SHI'DE'GUO'REN' mit Sogou-Eingabesoftware 270
Abb. 4-28: Strukturschaubild einer verbreiteten tastaturbasierten satzstufigen intelligenten chinesischen Eingabesoftware (nach: Xu ZM et al. 2000: 54 [Übersetzung der Verfasserin]) .. 286

Abb. 4-29: Schaubild der maximalen Zeichenkandidatenmöglichkeiten der segmentierten Pinyin-Kette TA/SHI/DEGUO/REN 287
Abb. 4-30: Die allgemeine syntaktische Struktur bei Sätzen mit sechs Satzgliedern 293
Abb. 4-31: Die Dependenzgrammatik vom Beispielsatz , (勤劳的) 我们 || [很快地] 翻译 <完了> | (这篇) 文章‘ 296
Abb. 4-32: Die Anwendung der sprachlichen Regeln zur Laut-Zeichen-Konversion von ‚TA/ JINTIAN/ FANYI/ LE/ LIANG/ PIAN/ WENZHANG‘ 303
Abb. 4-33: Zusammenfassung von den computerlinguistischen Anwendungen bei intelligenten Pinyin-Eingabesoftwares 308
Abb. 5-1: Struktur der multilingualen Eingabesoftware und ihr Verarbeitungsprozess 316
Abb. 5-2: Softwareerscheinungsbild *Look and Feel* des multilingualen Eingabesystems 317
Abb. 5-3: Entwurf der Sprachliste des multilingualen Eingabesystems 319
Abb. 5-4: Entwurf des alphanummerischen Tastenblocks mit neu hinzugefügten Tasten 322
Abb. 5-5: Entwurf des alphanummerischen Tastenblocks ohne hinzugefügte Tasten 323
Abb. 5-6: Beispiel für den D-Eingabefall unter Beteiligung der Korrektursysteme 327
Abb. 5-7: Struktur und Verarbeitungsprozess der Konversionsmodule 329
Abb. 5-8: B1-Eingabefall am Beispiel von ‚ch‘ mit Ausgabe in der ersten und vierten Zone 333
Abb. 5-9: B2-Eingabefall am Beispiel von ‚cha‘ mit Ausgabe in der zweiten und dritten Zone 334
Abb. 5-10: E-Eingabefall am Beispiel von ‚ta he cha.‘ mit Ausgabe in zweiter Zone 334
Abb. 5-11: Das Design des Touchpads für die Kandidatenauswahl 336
Abb. 5-12: Das Aussehen des benötigten speziellen Touchpads 337
Abb. 5-13: Tastenersatz für das Wählgerät des multilingualen Eingabesystems 338
Abb. 5-14: Design der Inputcode-Korrekturliste 359
Abb. 5-15: Beispiel für die Eingabe-Möglichkeit der maschinellen Übersetzung 360

7.3 Verzeichnis der Tabellen

Tab. 1-1: Die Beziehungen zwischen Zeichen, Glyphen und Font mit Beispielzeichen aus verschiedenen Schriften 46
Tab. 1-2: Zeichencodierung verschiedener Schriften im Unicode 51
Tab. 1-3: Die Belegung des US-amerikanischen Standardtastaturlayouts 52
Tab. 2-1: Der Buchstabenhäufigkeits-Tastenbelegungs-Zusammenhang im Deutschen 68
Tab. 2-2: Die 25 häufigsten Bigramme im Deutschen 69
Tab. 2-3: Die Häufigkeit der Sonderzeichen in der deutschen Schriftsprache 71
Tab. 2-4: Der Vergleich zwischen der Verarbeitung der Sonderbuchstaben in deutschen Texten 73
Tab. 2-5: Die Zeichencodierung mancher mit diakritischen Zeichen abgeleiteten Sonderbuchstaben, die in der lateinischen Schrift vorhanden sind 74
Tab. 2-6: Die Häufigkeit und die Tastenbelegung der 26 Grundbuchstaben des Englischen, Deutschen, Französischen und Spanischen 75
Tab. 2-7: Die Evaluation der QWERTY-Belegung im Buchstaben-Tasten-Zusammenhang für das englische, deutsche, französische und spanische Schriftsystem 76
Tab. 2-8: Die 35 häufigsten Bigramme der vier Sprachen 77
Tab. 2-9: Der vokalische Buchstaben-Phonem-Zusammenhang des vietnamesischen Alphabets 79
Tab. 2-10: Die sechs Töne der vietnamesischen Sprache 80
Tab. 2-11: Die Vokal-Varianten mit Tonzeichen 81
Tab. 2-12: Eingabeprozess des Hindi-Worts /pūrti/ *Erfüllung* 90
Tab. 2-13: Vokalische Buchstaben der Devanagari 92
Tab. 2-14: Konsonantische Buchstaben der Devanagari 93
Tab. 2-15: Die inhärenten sekundären Zeichen der Devanagari 97
Tab. 2-16: Die Zahlzeichen und nativen Interpunktionszeichen der Devanagari 97
Tab. 2-17: Die konsonantischen Buchstaben des Hangul 106
Tab. 2-18: Vokale und Vokalkombinationen als Inlaute des Hangul 107
Tab. 2-19: Die Konsonantencluster als Auslaute des Hangul 107
Tab. 2-20: Eingabeprozess und Zeichencodes von *Hanja* 108
Tab. 2-21: Die 28 Buchstaben des arabischen Schriftsystems 117
Tab. 2-22: Die diakritischen Hilfszeichen des arabischen Schriftsystems 119
Tab. 2-23: Zahl-, mathematische und Interpunktionszeichen des arabischen Schriftsystems 120
Tab. 2-24: Eingabeprozess von dem arabischen Beispielwort /ktāb/ 123
Tab. 2-25: Der Eingabeprozess für den arabischen bidirektionalen Beispieltext 125
Tab. 3-1: Die Evolution der chinesischen Schrift mit zehn piktographischen Beispielzeichen 139
Tab. 3-2: Beispielzeichen für sechs Schriften und ihre Konstruktionsprinzipien 155
Tab. 3-3: Beispiele für gemeinsame Zeichen in Chinesisch, Japanisch und Koreanisch (Unicode 12.0 Chapters: Kap. 18.1: 708) 165
Tab. 3-4: Initiale sowie Finale des Standardchinesischen und der Vergleich mit den drei Transkriptionssystemen 170

Tab. 4-1: Die Komponenten der Wubi-Inputcodierung, Version-86 191
Tab. 4-2: Die 45 häufigsten Silben im Standardchinesischen (nach Li M: Kap. 2.3) 228
Tab. 4-3: Möglichkeiten zur Silbensegmentation und entsprechende Zeichenketten 235
Tab. 4-4: Die Rangliste der Zeichenhäufigkeit und Gesamthäufigkeit (nach Peng 1994: 44 [Appendix-1], nach: Chen ZW/Jin 1988: o.S.) 239
Tab. 4-5: Die Überdeckungsquote (ÜDQ) von Wörtern auf verschiedenen Häufigkeitsniveaus (Chen YF/Zhu 2002: 10) 240
Tab. 4-6: Die zehn häufigsten Homophonen von /shishi/ und ihre Frequenz 241
Tab. 4-7: Die Wortsegmentation mit der maximalen Matching-Methode (FMM und OMM) bei Pinyin-Ketten ‚TA'SHI'DE'GUO'REN' 259
Tab. 4-8: Alle möglichen Segmentationsvarianten von ‚TA'SHI'DE'GUO'REN' 261
Tab. 4-9: POS-Tag-Standard der Wortart und deren Erklärung im Sprachgebrauch 265
Tab. 4-10: Das POS-Tagging der Pinyin-Kette ‚TA'SHI'DE'GUO'REN' 269
Tab. 4-11: Die Wortbildungsgrammatik und die Semantik der chinesischen Kompositionswörter ... 275
Tab. 4-12: Die Phrasenarten im Chinesischen 283
Tab. 4-13: Die Wortart-Satzglied-Beziehungen im Chinesischen (nach: Lu 2008: 112; ergänzt von der Verfasserin durch Zhao 1992: 73-103) 292
Tab. 4-14: Struktur und Erklärung jedes Satzglieds mit Beispiel 294
Tab. 4-15: Die Prädikatvarianten im Chinesischen und ihre Grundsatzformeln 299
Tab. 4-16: Typen und Häufigkeiten anwendbarer sprachlicher Regeln für die Pinyin-Kette ‚TA/ JINTIAN/ FANYI/ LE/ LIANG/ PIAN/ WENZHANG' 304
Tab. 5-1: Beispiel für Inputcodierung in acht verschiedenen Schriftsystemen 321
Tab. 5-2: Die benötigten Ressourcen einer Sprache im multilingualen Eingabesystem 332
Tab. 5-3: Plan für den Tastenersatz zur Kandidatenauswahl 338
Tab. 5-4: Inputcode für die Devanagari-Zahlzeichen 342
Tab. 5-5: Die Eingabe der chinesischen Interpunktionszeichen mit einer Pinyin-Eingabemethode sowie mit dem multilingualen Eingabesystem 344
Tab. 5-6: Vergleich der vier möglichen Sprachen im B1-Eingabefall 349

7.4 Literaturliste

I Normen und Werkzeugbücher[301]

Agency for Cultur Affairs Goverment Japan [文化庁] (2010): Die Liste der häufigen Kanji [常用漢字表] (**„Jouyou Kanji Hyou“ 2010**). Kündigung des japanischen Konzils am 30.11, das 22. Jahr Heisei (2010 n. Chr.) [平成 22 年 11 月 30 日内閣告示]. Online verfügbar unter *http://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/kijun/naikaku/pdf/joyokanjihyo_20101130.pdf* [Abruf: 2018-06-27].

Arab Organization for Standardization and Metrology (ASMO): **ISO 8859-6** [auch: ASMO 708]. Online verfügbar unter: *http://www.langbox.com/codeset/iso8859-6.pdf* [Abruf: 2016-11-01].

Boscher, Winfried (1989): Wörterbuch Vietnamesisch-Deutsch. Leipzig: Verlag Enzyklopädie, 5. Aufl.

Chinese General Administration of Quality Supervision [国家技术监督局] (1990): **GB 12200.1-90**. Chinese information processing – Vocabulary, Part 01: Fundamental Terms [汉语信息处理词汇 01 部分: 基本术语].

Chinese General Administration of Quality Supervision [国家技术监督局] (1994): **GB/T 12200.2-94**. Chinese information processing – Vocabulary, Part 02: Chinese and Chinese Character [汉语信息处理词汇 02 部分: 汉语和汉字].

Chinese General Administration of Quality Supervision [国家技术监督局] (1996): **GB/T 16159-1996**. Basic rules for Hanyu Pinyin Orthography [汉语拼音正词法基本原则]. Beijing: China Standard Press. Online verfügbar unter: *http://www.zyschool.com.cn/research/kgsj/201712/987.html* [Abruf: 2018-07-27].

Chinese Language Government [国家语言文字委员会]/ Ministery of Eduacation of the PRC [国家教育部] (1985): „**Putonghua Yidu Ci Shenyin Biao“ 1985** (Die Liste der Wörter mit alternativer Aussprache im Putonghua) [普通话异读词审音表].

Chu, BongFoo [朱邦復] (1990): Cangjie-Eingabeschema - die fünfte Generation [第五代仓颉输入法] (**Cangjie-5**). Online verfügbar unter: *http://www.cbflabs.com/book/ocj5/ocj5/index.html* [Erstellung: 2002-01-27, Abruf: 2015-02-08].

CNS [全字庫] (Chinese [Taipei] National Standard, 2017): Chinese Code Introduction [中文碼介紹]. Online verfügbar unter: *http://www.cns11643.gov.tw/AIDB/encodings_en.do* [Abruf: 2018-06-27], 2017©.

Computational Linguistical Laboratory of the Institute for Applied Linguistice, Ministery of Education of the PRC [教育部语言文字应用研究所计算语言学研究室] (©2018): **CN-Corpus Online** [语料库在线]. Online verfügbar unter: *http://www.cncorpus.org/*.

[301] Die Nomenbezeichnungen, die als Literaturangabe im Text und in den Fußnoten angegeben werden, werden in dem Literaturverzeichnis fett und mit Unterstrich aufgehoben.

Ministery of Education of the PRC [中华人民共和国教育部] (2013): Table of General Standard Chinese Character [通用规范汉字表] (**„Tongyong Guifan Hanzi Biao“ 2013**). Beijing: Secretary of State Council General Office.

Daum, Edmund/ Schenk Werner (2002): Schulwörterbuch Russisch. Berlin · München · Wien · Zürich · New York: Langenscheidt, 10. Aufl.

DIN Deutsches Institut für Normung e.V. (1993): **DIN-Taschenbuch 210**. Zeichenvorräte und Codierung für den Text- und Datenaustausch. Normen (Informationstechnik 10). Berlin/Wien/Zürich: Beuth Verlag GmbH, 2. Auflage.

DIN Deutsches Institut für Normung e.V. (2015): **DIN-Taschenbuch 343**. Bibliotheks- und Dokumentationswesen: Bibliotheks- und Archivbau, Zitierregeln, Umschriften, Digitale Langzeitarchivierung, Codierungen. Normen (Informationstechnik 10). Zürich: Beuth Verlag GmbH.

HUMANUM [人文網]: Chinese Character Database – With Word-Formations, Phonologically Disambiguated According to the Cantonese Dialect [粵語蕃音配詞字庫] (**Yueyu Fanyin Peici Ziku**). Online verfügbar unter *http://humanum.arts.cuhk.edu.hk/Lexis/lexi-can/* [Abruf: 2018-06-27].

Indian Standard: Indian Script Code for Information Interchange – **ISCII. IS 13194: 1991**. Online verfügbar: *https://law.resource.org/pub/in/bis/S04/is.13194.1991.pdf* [Abruf: 2015-07-19].

Jisho Organization: Jisho Japanese-English dictionary. **Jisho**. Online verfügbar: *http://jisho.org/* [Abruf: 2018-01-23].

Klemm, Erika (1997): Wörterbuch Hindi-Deutsch. Leipzig-Berlin-München-Wien-Zürich-New York: Langenscheidt Verlag Enzyklopädie, 4. Aufl.

Korean Graphic Chracter Set for Information Interchange: **KS X 1001:1992**. In: Lunde, Ken (2009): CJKV Information Processing, AppL.

State Commitee for the Reform of the Chinese Written Language [国家文字改革委员会] (1958): Scheme for the Chinese Phonetic Alphabet [汉语拼音方案] (**„Hanyu Pinyin Fang'an“ 1958**). Online verfügbar: *http://www.moe.gov.cn/ewebeditor/uploadfile/2015/03/02/20150302165814246.pdf* [Abruf: 2018-06-27].

Standardization Administration of the PRC [中国国家标准化管理委员会] (1980): The Guobiao Standard of Character Code for Information Interchange – Basic Set [信息交换用信息编码字符集——基本集]. **GB 2312-80**. In: Lunde, Ken (2009): CJKV Information Processing, AppE.

Standardization Administration of the PRC [中国国家标准化管理委员会]: **GB/T 13715-92**. Guobiao Standard of Contemporary Chinese Language Word Segmentation Specification for Information Processing [信息处理用现代汉语分词规范]. Publikation: 1992-10-04.

Standardization Administration of the PRC [中国国家标准化管理委员会]: **GB/T 20532-2006**. Guobiao Standard of POS Tag of Contemporary Chinese for CIP [信息处理用现代汉语词类标记规范]. Publikation: 2006-09-18.

State Language Affairs Commission [国家语言文字工作委员会]: **GF 3001-1997**. Chinese Character Component Standard of GB 13000.1 Character Set for Information Processing [信息处理用 GB 13000.1 字符集汉字部件规范]. Online verfügbar unter: *http://old.moe.gov.cn/ewebeditor/uploadfile/2015/01/12/20150112165337190.pdf* [Abruf: 2018-06-27].

State Language Affairs Commission [国家语言文字工作委员会] (2008): Lexicology of Common Words in Contemporary Chinese (the Draft) [现代汉语常用词表（草案）] (**„Xiandai Hanyu Changyong Ci Biao“ 2008**). Beijing: The Commercial Press.

Taiwan (National) Academy for Educational Research [中華民國教育部] (2017): Dictionary of Chinese Character Variants [異體字字典] (**Yitizi-Lexikon**), die 6. Aufl. Online verfügbar unter: dict.variants.moe.edu.tw [Abruf: 2019/03/01].

The Science and Technology Information Center (Abk.: STIC) [行政院國家科學委員會]: **CNS 11643-1992 Plane 2**. National Chinese Character Interchange Code. In: Lunde, Ken (2009): CJKV Information Processing AppG.

The Unicode Consortium (2019): **ArabicShapping.txt**, Online verfügbar: *https://www.unicode.org/Public/UCD/latest/ucd/ArabicShaping.txt*. [Aktualisiert 2019-02-21].

The Unicode Consortium (2019): Unicode 12.0. Character Code Charts (**Unicode 12.0 Character**). Online verfügbar: *http://www.unicode.org/charts/* [Aktualisiert: 2019-02-25].

The Unicode Consortium (2019): Unicode Standard 12.0 Chapters (**Unicode 12.0 Chapters**). Online verfügbar: *https://www.unicode.org/versions/Unicode12.0.0/* [Aktualisiert: 2019-03-25].

The Unicode Consortium (2019): Glossary of Unicode Terms (**Unicode Glossary**). Online verfügbar: *http://www.unicode.org/glossary/* [Aktualisiert 2019-03-11].

Wangma Group [王码集团] (repräsentiert von Wang Yongmin): The Homepage of Wangma [王码] Online verfügbar: *http://www.wangma.com.cn/* [Abruf: 2019-03-04].

Xinhua-Lexikon Verlag [新华辞书社]: Xinhua-Zeichenlexikon – Online-Version [在线《新华字典》] (**Xinhua-Lexikon Online**). Online verfügbar: *http://xh.5156edu.com/* [Abruf: 2015-02-08].

Xu, Shen [许慎] (100 n. Chr.): Shuowen-Jiezi – Online-Version [在线《说文解字》]. Online verfügbar: *http://ctext.org/shuo-wen-jie-zi/zh* [Abruf: 2015-02-08].

Zhang, YuShu [张玉书]/ Chen, TingJing [陈廷敬] etl. (1716): Kangxi-Zeichenlexikon Online-Version [在线《康熙字典》]. Online verfügbar: *http://xh.5156edu.com/kxbs.html* [Abruf: 2015-02-08].

II. Literatur in Deutsch und Englisch

Altmann, Hans/ Ziegenhain, Ute (2010): Prüfungswissen Phonetik, Phonologie und Graphemik. Göttingen: Vandenhoeck und Ruprecht.

Ameling, Walter/ Kreft, Lothar (1996): Technische Kodierung. In: Günter/Ludwig (1996): Schrift und Schriftlichkeit, HSK 10.2. Berlin & New York: Walter de Gruyter, Art. 148, S.1629-1638.

Atsuji, Tetsuji (1994): Der Kulturkreis der chinesischen Schriftzeichen (hànzì). In: Günter/Ludwig (1994): Schrift und Schriftlichkeit, HSK 10.1. Berlin & New York: Walter de Gruyter, Art. 32, S. 436-450.

Baier, Peter E. (1996): Maschinenschreiben und forensische Urheberidentifizierung. In: Günter/Ludwig (1996): Schrift und Schriftlichkeit, HSK 10.2. Berlin & New York: Walter de Gruyter, Art. 89, S.1056-1068.

Barcodemann: PC Keyboard Scan Codes. PC Keyboars, 102 Keys. Online verfügbar: *http://www.barcodeman.com/altek/mule/scandoc.php*, [Abruf: 2013-09-06]

Bauer, Friedrich L. (1997): Entzifferte Geheimnisse. Methoden und Maximen der Kryptologie, 2. Aufl. Berlin / Heidelberg / New York: Springer.

Bauer, Thomas (1996): Das arabische Schriftsystem. In: Günter/Ludwig (1996): Schrift und Schriftlichkeit, HSK 10.2. Berlin & New York: Walter de Gruyter. Art. 123, S.1433-1437.

Bergerhausen, Johannes/ Poarangan, Siri (2011): Decodeunicode: Die Schriftzeichen der Welt. Mainz: Hermann Schmidt.

Best, Karl-Heinz (2006): Quantitative Linguistik. Eine Annährung, 3. Auf Göttinger: Peust & Gutschmidt Verlag.

Bodmer, Frederick (1997): Die Sprachen der Welt. Linzenzausgabe. Köln: Parkland.

Brekle, Herbert E. (1994a): Die Buchstabenformen westlicher Alphabetschriften in ihrer historischen Entwicklung. In: Günter/Ludwig (1994): Schrift und Schriftlichkeit, HSK 10.1. Berlin & New York: Walter de Gruyter. Art. 12, S. 171-204.

Brekle, Herbert E. (1994b): Typographie. In: Günter/Ludwig (1994): Schrift und Schriftlichkeit, HSK 10.1. Berlin & New York: Walter de Gruyter. Art. 13, S. 204-228.

Busch, Albert/ Stenschke, Oliver (2008): Germanistische Linguistik, 2. Aufl.. Tübingen: Narr.

Bußmann, Hadumod (2002): Lexikon der Sprachwissenschaft. Stuttgart: Alfred Kröner Verlag.

Carstensen, K.-U., etl. (2010): Computerlinguistik und Sprachtechnologie: Eine Einführung, 3. Aufl Heidelberg: Spektrum Akademischer Verlag.

Chen, ShenYuan/ Zhao, Hai/ Wang, Rui (2015): Neutral Network Language Model for Chinese Pinyin Input Method Engine. In: Processings of 29th Pacific Asia Conference on Language, Information and Computation. Online verfügbar: *http://www.aclweb.org/anthology/Y15-1052* [Abruf: 2018-01-07].

Chen, Zheng/ Lee, Kai-Fu (2000): A New Statistical Approach to Chinese Pinyin Input. In: Proceedings of the 38th Annual Meeting of the Association for Computational. Online verfügbar: *http://aclweb.org/anthology/P00-1031* [Abruf: 2015-10-29].

Clément, Denièle (2000): Linguistisches Grundwissen. Eine Einführung für zukünftige Deutschlehrer, 2. Aufl. Wiesbaden: Westdeutscher Verlag.

Coulmas, Florian (1994): Theorie der Schriftgeschichte. In: Günter/Ludwig (1994): Schrift und Schriftlichkeit, HSK 10.1. Berlin & New York: Walter de Gruyter. Art. 15, S. 256-264.

Coulmas, Florian (1996a): The Blackwell Encyclopedia of Writing Systems. Cambridge: Blackwell Publishers Inc.

Coulmas, Florian (1996b): Typology of Writing Systems. In: Günther/Ludwig (1996): Schrift und Schriftlichkeit, HSK 10.2, Berlin & New York: Walter de Gruyte. Art 118, S. 1380-1387.

Daley, Bill A. (2016): Computer Are Your Future 2007, Complete, 9 th Edition. Facts101 Textbook Key Facts. Cram101 Textbook Reviews.

Davis, Mark/ Lanin, Aharon/ Glass, Andrew (2019): Unicode Bidirectional Algorithm. In: Unicode Organization: Unicode Standard Annex #9, Version Unicode 12.0.0. Online verfügbar: *http://www.unicode.org/reports/tr9/* [Aktualisiert: 2019-02-04].

Deshwal, Priyendra/ Deb Kalyanmoy (2003): Design of an Optimal Hindi Keyboard for Convenient and Efficient Use. KanGAL Report Number 2003004. Online verfügbar unter: *http://www.iitk.ac.in/kangal/papers/k2003004.pdf* [Abruf: 2015-05-28].

Diab, Mona/ Hacioglu, Kadri/ Jurafsky, Daniel (2007): Automatic Processing of Modern Standard. Arabic Text. In: Soudi et al. (2007): Arabic Computational Morphology. Dordrecht: Springer, S. 159-179.

Đỗ Bá Phước (2005): Computing in Vietnamese: Progress & Challenges. Online verfügbar unter: *http://imug.org/presentations/VietComputing.pdf* [Abruf: 2015-09-17].

Dürscheid, Christa (2006): Einführung in die Schriftlinguistik, 3. Aufl. Göttingen: Vandenhoeck & Ruprecht.

Edmondson, J (2006): Vietnamese. In: Brown, Keith (Hrsg.): Concise Encyclopedia of Languages of the World. Amsterdam: Elservier, S.1149-1153.

Ehlich, Konrad E. (1994): Funktion und Struktur schriftlicher Kommunikation. In: Günter/Ludwig (1994): Schrift und Schriftlichkeit, HSK 10.1. Berlin & New York: Walter de Gruyter, Art. 2, S. 18-40.

Eilts, John (1995): Commentary (of Wien 1999: "Nine Problems Concerning Arabic"). In: Byrum, John D./ Madison Olivia (1995): Multi-script, Multilingual, Multi-character Issues for the Online Environment. Istanbul: IFLA, S. 39-41.

Eisenberg, Peter/ Günter, Hartmut [Hgg.] (1989): Schriftsystem und Orthographie. Tübingen: Max Niemeyer Verlag.

Eisenberg, Peter (1996a): Das deutsche Schriftsystem. In: Günter/Ludwig (1996): Schrift und Schriftlichkeit, HSK 10.2. Berlin & New York: Walter de Gruyter, Art. 127, S.1451-1456.

Eisenberg, Peter (1996b): Sprachsystem und Schriftsystem. In: Günter/Ludwig (1996): Schrift und Schriftlichkeit, HSK 10.2. Berlin & New York: Walter de Gruyter, Art. 117, S.1368-1380.

Etemad, Elika J. (2005): Robust Vertical Text Layout. Using the Unicode Bidi algorithm to handle complexities in typesetting multi-script vertical text. In: 27th Internationlization and Unicode Conference. Berlin, Germany, April 2005. Online verfügbar: *http://unicode.org/notes/tn22/RobustVerticalLayout.pdf* [Abruf: 2015-08-22].

Evert, Stefan/ Frötschl, Bernhard/ Lindstrot, Wolf: Statistische Grundlagen. In: Carstensen et al. (2010): Computerlinguistik und Sprachtechnologie. Heidelberg: Spektrum, S. 114-158.

Fliedner, Gerhard (2010): Korrektursysteme. In: Carstensen et al. (2010): Computerlinguistik und Sprachtechnologie. Heidelberg: Spektrum, S. 555-565.

Friedrich, Elvira (2006): Einführung in die indischen Schriften – Devanāgarī, 2. Aufl. Hamburg: Buske.

Gallmann, Peter (1996): Interpunktion. In: Günter/Ludwig (1996): Schrift und Schriftlichkeit, HSK 10.2. Berlin-New York: Walter de Gruyter, Art. 128, S.1456-1467.

Gelb, I. J. (1963): A study of writing. Revised Edition. Chicago & London: The University of Chicago Press.

Gibbon, Dafydd (2010): Lexika für multi-Modale Systeme. In: Carstensen et al. (2010): Computerlinguistik und Sprachtechnologie. Heidelberg: Spektrum Akademischer Verlag, 3. Auflage. S. 515-523.

Greulich, Walter [Red.] (2003): Der Brockhaus. Computer und Informationstechnologie. Fachlexikon für Hardware, Software, Multimedia, Internet, Telekommunikation. Leipzig-Mannheim: F.A. Brockhaus.

Günter, Hartmut/ Ludwig, Otto (1994): Vorwort [in Deutsch]. In: Günter/Ludwig (1994): Schrift und Schriftlichkeit, HSK 10.1. Berlin & New York: Walter de Gruyter. S. V-XXII.

Gumm, Heinz-Peter/ Sommer, Manfred (2013): Einführung in die Informatik. München: Oldenbourg Verlag, 10. Auflage.

Haarmann, Harald (1994): Entstehung und Verbreitung von Alphabetschriften. In: Günter/Ludwig (1994): Schrift und Schriftlichkeit, HSK 10.1. Berlin & New York: Walter de Gruyter, Art. 25, S. 329-347.

Habash, Nizar Y. (2010): Introduction to Arabic Natural Language Processing. Morgan & Claypool Publishers.

Heilmann, T. A. (2012) : Textverarbeitung. Eine Geschichte des Computers als Schreibmaschine. Bielefeld: Transcript Verlag.

Hippsley, Andrew R (1998): Indexed Stems and Russian Word Formation: A Network Morphology Account of Russian Personal Nouns. University of Kentucky: UKnowledge, Linguistics Faculty Publications. Online verfügbar: *http://uknowledge.uky.edu/lin_facpub/43*. [Abruf: 2017-6-23].

Hagenbruch, André: Flache Satzverarbeitung. In: Carstensen (2010): Computerlinguistik und Sprachtechnologie. Heidelberg: Spektrum Akademischer Verlag, S. 264-279.

Huang, Jin-Xia/ Bae, Sun-Mee/ Choi, Key-Sun (2004): A Statistical Model for Hangul-Hanja Conversion in Terminology Domain. In: Proceedings of the Third SIGHAN Workshop on Chinese Language Processing. Online verfügbar: *http://aclweb.org/anthology/W04-1111* [Abruf: 2015-10-29].

Hundt, Eckart/ Maderlechner (1994): Elektronische Lese- und Schreibtechnologien. In: Günter/Ludwig (1994): Schrift und Schriftlichkeit, HSK 10.1. Berlin & New York: Walter de Gruyter, Art. 9, S. 130-146.

James, E (1995): Commentary (of John 1995: „The Unicode Standard. An Overview with Emphasis on Bidirectionality"). In: Byrum, John D./ Madison Olivia (1995): Multi-script, Multilingual, Multi-character Issues for the Online Environment. Istanbul: IFLA, S. 112-123.

John, M. (1995): The Unicode Standard. An Overview with Emphasis on Bidirectionality. In: Byrum, John D./ Madison Olivia (1995): Multi-script, Multilingual, Multi-character Issues for the Online Environment. Istanbul: IFLA, S. 95-111.

Kappenberg, Bernd (2012): Zeichen setzen für Europa. Der Gebrauch europäischer lateinischer Sonderzeichen in der deutschen Öffentlichkeit. Stuttgart: Ibidem-Verlag.

Kessel, Katja/ Reimann, Sandra (2010): Basiswissen Deutsche Gegenwartssprache, 3. Aufl. Tübingen & Basel: A. Francke Verlag.

King, Ross (1996): Korean Writing. In: Daniels, Peter T. (1996): The world's writing systems. New York [u.a.]: Oxford University Press, S. 218-227.

Klabunde, Ralf (2010): Automatentheorie und Formale Sprache. In: Carstensen et al. (2010): Computerlinguistik und Sprachtechnologie. Heidelberg: Spektrum, S. 66-93.

Koji, Ishii/ Lunde, Ken (2019): Unicode Vertical Text Layout. Unicode Technical Report #50, Version: Unicode 12.0.0. Online verfügbar: *http://www.unicode.org/reports/tr50/* [Aktualisiert: 2019-02-04].

Krebernik, Manfred/ Nissen, Hans J. (1994): Die summerisch-akkadische Keilschrift. In: Günter/Ludwig (1994): Schrift und Schriftlichkeit, HSK 10.1. Berlin & New York: Walter de Gruyter, Art. 18, S. 274-289.

Kunzmann, Robert Walter (1979): Hundert Jahre Schreibmaschinen im Büro – Geschichte des maschinellen Schreibens. Rinteln: Merkur Verlag.

Langer, Hagen (2010): Syntax und Parsing. In: Carstensen et al. (2010): Computerlinguistik und Sprachtechnologie. Heidelberg: Spektrum, S. 280-329.

Lenders, Winfried/ Willée, Gerd (1986): Linguistische Datenverarbeitung. Ein Lehrbuch. Opladen: Westdeutscher Verlag.

Li, Jie (1996): Das chinesische Schriftsystem. In: Günter/Ludwig (1996): Schrift und Schriftlichkeit, HSK 10.2. Berlin & New York: Walter de Gruyter, Art. 120, S.1404-1412.

Lobin, Henning (2010): Computerlinguistik und Texttechnologie. Paderborn: W. Fink.

Lobin, Henning (2014): Engelbarts Traum. Wie der Computer uns Lesen und Schreiben abnimmt. Frankfurt & New York: Campus Verlag.

Ludwig, Otto (1983): Einige Vorschläge zur Begrifflichkeit und Terminologie von Untersuchungen im Bereich Schriftlichkeit. In: Günter, Klaus-Burkhard/ Günter Hartmut (Hg.) (1983): Schrift,

Schreiben, Schriftlichkeit. Arbeiten zur Struktur, Funktion und Entwicklung schriftlicher Sprache. Tübingen: Max Niemeyer Verlag, S. 1-15.

Ludwig, Otto (1994): Geschichte des Schreibens. In: Günter/Ludwig (1994): Schrift und Schriftlichkeit, HSK 10.1. Berlin & New York: Walter de Gruyter. Art. 4, S. 48-65.

Ludwig, Otto (2005): Geschichte des Schreibens. Band 1. Von der Antike bis zum Buchdruck. Berlin & New York: Walter de Gruyter.

Lunde, Ken (2009): CJKV Information Processing. Beijing • Cambridge • Farnham • Köln • Sebastopol • Taipei • Tokyo: O'Reilly Media, 2. Edition.

Majidi, Mohammad-Reza (2006): Einführung in die arabisch-persische Schrift. Hamburg: Buske, 3. Aufl.

Mazal, Otto (1994): Traditionelle Schreibmaterialien und -techniken. In: Günter/Ludwig (1994): „Schrift und Schriftlichkeit", HSK 10.1. Berlin & New York: Walter de Gruyter, Art. 18, S. 122-130.

Meier, Helmut (1978): Deutsche Sprachstatistik, I/II. Hildesheim & New York: Georg Olms Verlag.

Molitor-Lübbert, Sylvie (1996): Schreiben als mentaler und sprachlicher Prozess. In: Günter/Ludwig (1996): Schrift und Schriftlichkeit, HSK 10.2. Berlin & New York: Walter de Gruyter. Art. 85, S.1005-1027.

Möbius, Bernd/ Haiber, Udo (2010): Verarbeitung gesprochener Sprache. In: Carstensen et al. (2010): Computerlinguistik und Sprachtechnologie. Eine Einführung. Heidelberg: Spektrum Akademischer Verlag, S. 214-235.

Mrayati M./Alam Y. Meer/at-Tayyan M.H. (2003): Arabic Origins of Cryptology, Vol. 2., Riyadh: KFCRIS & KACST.

Müller, Didier (2002): Analyse des fréquences en espagnol. Online verfügbar unter: *http://www.apprendre-en-ligne.net/crypto/stat/espagnol.html* [Aktualisiert: 2002-08-07, Abruf: 2014-06-12].

Müller, Didier (2003): Analyse des fréquences en français (avec les accents). Online verfügbar unter: *http://www.apprendre-en-ligne.net/crypto/stat/francais2.html* [Aktualisiert: 2003-02-12, Abruf: 2017-11-15].

Müller-Yokota, Wolfram (1994): Die chinesische Schrift. In: Günter/Ludwig (1994): Schrift und Schriftlichkeit, HSK 10,1. Berlin & New York: Walter de Gruyter, Art. 26, S. 347-382.

Müller-Yokota, Wolfram (1994): Weiterentwicklungen der chinesischen Schrift. In: Günter/Ludwig (1994): Schrift und Schriftlichkeit, HSK 10,1. Berlin & New York: Walter de Gruyter, Art. 27, S. 382-405.

Nguyễn, Đình-Hoà (1990): Vietnamese. In: Comrie, Bernard (1990): The major languages of East and South-East Asia (1990). London: Routledge, S. 49-68.

Ong, Walter J. (1987): Oralität und Literalität. Die Technologisierung des Wortes. Originalausgabe in Englisch: „Orality and Literacy. The Technologizing of the Word", 1982. Opladen: Westdeutscher Verlag.

Ostapenko, Olesya (2005): The Optimal L2 Russian Syllable Onset. LSO Working Papers in Linguistics 5: Proceedings of WIGL 2005, 140-151. Online verfügbar: *https://www.csun.edu/~bashforth/301_PDF/301_P_P/RussianSyllableOnsets.pdf* [Abruf 2017-06-22].

Pentzlin, Karl [23.10.2010]: Information about the Revision of ISO/IEC 9995-3. Online verfügbar: *http://www.pentzlin.com/info2-9995-3-V3.pdf* [Abruf: 2013-12-05].

Plath, Verena (2014): Deutsche Wortbildung. Studienbibliographie Sprachwissenschaft 44. Tübingen: Julius Groos Verlag Brigitte Narr GmbH.

Pommerening, Klaus (2007): Kryptologie – Zeichenhäufigkeit in Deutsch. JGU Mainz. *http://www.staff.uni-mainz.de/pommeren/Kryptologie/Klassisch/1_Monoalph/deutsch.html* [letzte Änderung: 2007-12-09].

Pommerening, Klaus (2014): Kryptologie – Bigramm-häufigkeiten. JGU Mainz. *https://www.staff.uni-mainz.de/pommeren/Kryptologie/Klassisch/6_Transpos/Bigramme.html* [Aktualisiert: 2014-05-14]

Pospeschill, Markus (1996): Schreiben mit dem Computer. In: Günter/Ludwig (1996): Schrift und Schriftlichkeit, HSK 10.2. Berlin & New York: Walter de Gruyter, Art. 90, S.1068-1074.

Preverelli, Peter (2015): The History of of Modern Chinese Grammar Studies. Berlin & Heidelberg: Springer Verlag.

Procházka, S (2006): Arabic. In: Brown, Keith (Hrsg.): Concise Encyclopedia of Languages of the World. Amsterdam: Elservier, S. 42-50.

Radvan, Florian (2013): Digitales Schreiben im Deutschunterricht. In: Lobin, Henning et al. (Hg.) (2013): Lesen, Schreiben, Erzählen. Kommunikative Kulturtechniken im digitalen Zeitalter. Frankfurt & New York: Campus Verlag.

Robinson, Andrew (2013): Bilder, Zeichen, Alphabete. Die Geschichte der Schrift. Englische Originalausgabe "Writing and Script. A very short Introduction" von Josef Billen 2009. Darmstadt: Lambert Schneider Verlag.

Scheffler, Christian: Kalligraphie. In: Günter/Ludwig (1994): Schrift und Schriftlichkeit, HSK 10.1. Berlin & New York: Walter de Gruyter, Art. 14, S. 228-255.

Schenkel, Wolfgang (1994): Die ägyptische Hieroglyphenschrift und ihre Weiterentwicklungen. In: Günter/Ludwig (1994): Schrift und Schriftlichkeit, HSK 10.1. Berlin & New York: Walter de Gruyter, Art. 19, S. 289-297.

Sen, Subhadra (1996): The Devanagari Writing System. In: Günter/Ludwig (1996): Schrift und Schriftlichkeit, HSK 10.2. Berlin & New York: Walter de Gruyter, Art. 122, S.1428-1432.

Schenkel, Wolfgang (1994): Die ägyptische Hieroglyphen und ihre Weiterentwicklungen. In: Günter/Ludwig (1994): Schrift und Schriftlichkeit, HSK 10.1. Berlin & New York: Walter de Gruyter, Art. 19, S. 289-297.

Smith, Janet S. (1996): Japanese Writing. In: Daniels, Peter T. (1996): The world's writing systems. New York [u.a.]: Oxford University Press, S. 209-217.

Sohn, Homin (2001): The Korean Language. Cambridge: Cambridge University Press.

Sproat, Richard (2010): Language, Technology and Society. Oxford: Oxford University Press.

Stalph, Jürgen (1996): Das japanische Schriftsystem. In: Günter/Ludwig (1996): Schrift und Schriftlichkeit, HSK 10.2. Berlin & New York: Walter de Gruyter, Art. 121, S.1413-1428.

Stubbs, Michael (1996): The Englisch Writing System. In: Günter/Ludwig (1996): Schrift und Schriftlichkeit, HSK 10.2. Berlin & New York: Walter de Gruyter, Art. 125, S.1441-1445.

Tsai, Jia-Lin (2005): Using Word-Pair Identifier to Improve Chinese Input System. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing. Online verfügbar: *http://aclweb.org/anthology/I05-3* [Abruf: 2015-10-29].

Tsai, Jia-Lin (2006): Using Word Support Model to improve Chinese Input System. In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. Online verfügbar: *http://aclweb.org/anthology/P06-2108* [Abruf: 2015-10-29].

Umamaheswaran, V.S. (2002): UTF EBCDIC. In: Unicode Technical Report #16 – Version 8. Online verfügbar unter: *http://www.unicode.org/reports/tr16/* [Aktualisiert: 2002-04-16].

Ward, D (2006): Russian. In: Brown, Keith (Hrsg.): Concise Encyclopedia of Languages of the World. Amsterdam: Elservier, S. 905-908.

Watson J C E (2006): Arabic as an Introflecting Languages. Brown, Keith (Hrsg.): Concise Encyclopedia of Languages of the World. Amsterdam: Elservier, S. 50-53.

Wien, Charlotte (1995): Nine Problems Concerning Arabic. In: Byrum, John D./ Madison Olivia (1995): Multi-script, Multilingual, Multi-character Issues for the Online Environment. Istanbul: IFLA, S. 25-38.

Willbertz, Veronika (1994): Die arabische Schrift. In: Günter/Ludwig (1994): Schrift und Schriftlichkeit, HSK 10.1. Berlin & New York: Walter de Gruyter, Art. 22, S. 312-317.

Whistler, Ken (2019): Unicode Normalization Forms. In: Unicode Organization: Unicode Standard Annex #15, Version Unicode 12.0.0. Online verfügbar: *http://unicode.org/reports/tr15/* [Aktualisiert: 2019-02-04].

Whistler, Ken/ Davis, Mark/ Freytag, Asmus (2008): Unicode Technical Report #17. Unicode Character Encoding Model, Revision 7. Online verfügbar: *http://unicode.org/reports/tr17/* [Aktualisiert: 2008-11-11].

Zhang, MeiShan/ Zhang, Yue/ Che, WanXiang/ Liu, Ting (2014): Character-Level Chinese Dependency Parsing. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Online verfügbar: *http://aclweb.org/anthology/P14-1125* [Abruf: 2015-11-04].

Zikmund, Hans (1996): Transliteration. In: Günter/Ludwig (1996): Schrift und Schriftlichkeit, HSK 10.2. Berlin & New York: Walter de Gruyter, Art. 143, S.1591-1604.

III: Literatur in Chinesisch[302]

Bai, ShuanHu [白栓虎] (1992): The Study and Realization of Statistics Based Approach to Tagging Chinese Corpus [基于统计的汉语语料库词性自动标注的研究与实现]. In: Huang, Changning [黄昌宁] (1996): Wissenschaftliche Arbeiten über Informationsverarbeitung [语言信息处理专论]. Beijing: Tsinghua University Press, S. 37-72.

Cai, YongFei [蔡勇飞] (2005): A Comparativ Study of Written Chinese and English [汉英文字比较研究]. Hangzhou: Zhejiang University Press.

Chen, AiWen [陈爱文] (1986): Die Theorien und Praxis der Zeichenkodierung der chinesischen Schrift [汉字编码的理论与实践]. Shanghai: Xuelin Press.

Chen, DunHua [陳敦化] (1984) (TW): Die Theorien des vertikalen und horizontalen Schreibens der chinesischen Schrift [中國文字縱橫論]. In: Chinese Culture University Hwa Kang Journal of Engineering [華岡工程學報], Vol.2/ Juli 1984, S. 29-44.

Chen, QiGuang [陈其光] (1993): Die Schriften des chinesischen Schriftkreises [汉字系文字]. In: Xu, ShouChun [许寿椿] (1993): Wissenschaftliche Arbeiten über Vergleich der Schriften [文字比较研究散论]. Beijing: Minzu University of China Press.

Chen, XiaoHe [陈小荷] (1999): Automatische Analyse der chinesischen modernen Sprachen [现代汉语自动分析]. Beijing: Beijing Language and Culture University Press (BLCUP).

Chen, XiaoHe [陈小荷]/ Cui, YongHua [崔永华] (1996): Die Hypothese über die Begründung der großen chinesischen Treebanken [关于建立大规模汉语数据库的设想]. In: Luo, ZhenSheng [罗振声] (1996): Die Erforschung an der chinesischen Sprache und Schrift im Zeitalter des Computers. Beijing: Tsinghua University Press, S. 303-314.

Chen, YiFan [陈一凡]/ Zhu, Liang [朱亮] (2002a): Die Statistiken über die allgemeine Frequenz der Zeichen sowie Wörter im Chinesisch am Anfang des 20. Jh. [二十世纪初汉语字、词流通频度统计]. In: Chinese Information Processing Society of China, Committee of Chinese Character Coding, 8th Symposium [中国中文信息学会汉字编码专业委员会第八届年会], S. 8-11.

Chen, YiFan [陈一凡]/ Zhu, Liang [朱亮] (2002b): Die Überblick von den intelligenten Verarbeitungssoftwares der tastaturbasierten Eingabe der chinesischen Schrift [汉字键盘输入智能处理软件综述]. In: Chinese Information Processing Society of China, Committee of Chinese Character Coding, 8th Symposium [中国中文信息学会汉字编码专业委员会第八届年会], S. 12-18.

[302] Die Personal- sowie Ortsnamen werden nach der Pinyin-Rechtschreibung umgeschrieben. Der erste Buchstabe der zweiten Silbe des Vornamen wird auch groß geschrieben. Bei der Literaturangabe im Text wird der erste Buchstabe jeder Vornamensilbe auch angegeben, falls der Familienname des Autors sehr häufig ist. Der Titel der Literaturen wird nach der Übersetzung der Verfasserin in Deutsch angegeben (außer den Materialien mit originalem englischem Titel). Die meisten Verlagen sowie wissenschaftliche Zeitschriften haben festgelegten englische Bezeichnung und (nur) in Englisch angegeben. Die nicht aus Festlandchina gestammten Literaturen werden hinter der Jahrangabe mit TW (aus Taiwan, originales in Langzeichen), HK (aus Hongkong, originales in Langzeichen) oder SG (aus Singapur, originales in Kurzzeichen) markiert.

Chen, Yue [陈越] (1955): Die Entwicklung der Kulturtechnik und die Probleme bei schriftlicher Revolution [文化技术的发展和中国文字改革的问题]. Shanghai: Eastern Bookstore Publishing House.

Cheng, ChunYou [成春有] (2002): Die Forschung an der On-Lesung der Kanji [日语汉字音读研究]. Hefei: University of Science and Technology of China Press.

Chia, Shih-Yar [谢世涯] (1989) (SG): Die Erforschung der vereinfachten Schriftzeichen in Singapur, China und Japan [新中日简体字研究]. Beijing: Language & Culture Press.

Ding, DaBin [丁大斌]/ Huang ChangNing [黄昌宁] (2009): Chinese Homophones: An Information Processing Perspective [从信息处理角度看汉语同音词]. In: Applied Linguistics [语言文字应用], No. 4/ Nov. 2009, S. 132-142.

Dou, ShuoHua [窦硕华] (2008): The Evolution of Chinese Characters in Japanese — Comparison of Simplified Characters Popularly in China and Japan [日本汉字演进轨迹——兼及中日现行简化字比较]. In: Journal of Changchun University of Science and Technology: Band of Social Science [长春理工大学学报：社会科学版], Vol.21/ No.2/ Mar. 2008, S. 112-114.

Fei, JinChang [费锦昌] (1996): Wie kann sich der Zeichenabbau in Komponenten im Bereich der Informatik und Linguistik übereinstimmen [计算机界和语文界在汉字部件切分上如何求同]. In: Luo, ZhenSheng [罗振声], 1996: [计算机时代的汉语与汉字研究]. Beijing: Tsinghua Universität Press, S. 443-448.

Feng, ZhiWei [冯志伟] (1999): Die Allgemeintheorien der Anwendungslinguistik [应用语言学综论]. Guangzhou: Guangdong Education Publishing House.

Feng, ZhiWei [冯志伟] (2001a): The computer processing of Chinese characters and Chinese language [汉字和汉语的计算机处理]. In: Contemporary Linguistics [当代语言学], Vol.3/ No.1/ 2001, S. 1-21.

Feng, ZhiWei [冯志伟] (2001b): Die Erforschung der Computerlinguistik [计算语言学探索]. Harbin: Heilongjiang Education Verlag.

Feng, Zhiwei [冯志伟] (2001c): Some grammatical factors to determine the segmentation elements [确定切词单位的某些语法因素]. In: Terminology Standardization & Information Technology [术语标准化与信息技术], No.2/2001.

Gao, GengSheng [高更生]/ Wang, HongQi [王红旗] (1996): Die Forschung der chinesisch-didaktischen Grammatik [汉语教学语法研究]. Beijing: Language & Culture Press.

Gao, GengSheng [高更生] (2000): Die Erforschung der chinesischen Schrift [汉字研究]. Ji'nan: Shandong Eduacation Press.

Gao, Jing [高晶] (1995): Publikationssystem des Computers [计算机出版系统]. Beijing: Graphic Communications Press (GCP).

Gao, WenHan [高文汉] (1990): Die Theorien der japanischen Wörter [日语词汇论]. Changchun: Jilin Education Press.

He, Sheng [贺胜]/ Lu, YaJun [卢亚军] (2007): Research of Tibetan Input Method Based on Nationally and Internationally Compliable Tibetan Coding (Basic Set) [基于藏文编码（基本集）国家暨国际标准的藏文输入法研究]. In: Library & Information [图书与情报], Vol.6/ 2007, S. 45-49.

Hou, Min [侯敏] (1999): Computerlinguistik und automatische Analyse der chinesischen Sprache [计算机语言学与汉语自动分析]. Beijing: Communication University of China Press.

Huang, ChangNing [黄昌宁] etl. (1996): Wissenschaftliche Arbeiten über Informationsverarbeitung [汉语信息处理专论]. Beijing: Tsinghua University Press.

Huang, Jinwen [黄金文]/ Jin, Hua [金华]/ Wang, Fan [王凡] et al. (2010): Research on Keyboard Layout for Chinese Pinyin IME [关于中文拼音输入法键盘字母布局的研究]. In: Journal of Chinese Information Processing, Vol. 24/ No. 6/ Nov. 2010.

Lan, BinHan [兰宾汉] (2002): Die Theorien und Praxis der grammatischen Analysen der chinesischen Sprache [汉语语法分析的理论与实践]. Beijing: China Social Science Press.

Li, BaoJia [李葆嘉] (1991): Der korrelative Zusammenhang zwischen dem Typ der Sprache und der Schrift [论语言类型与文字类型的制约关系]. In: Yuan, XiaoYuan [袁晓园] (1991): Wissenschaftliche Arbeiten bei Symposium für chinesische Schrift und Sprache [汉字汉语研讨会论文集], Vol. 2. Changchun: Jilin Education Press. S. 54-68.

Li, DeChun [李得春]/ Jin, JiShi [金基石] (1997): Die Kultur der chinesischen Schrift und die sinokoreanischen Schriftzeichen [汉字文化与朝鲜汉字]. In: Dongjiang Journal [东疆学刊], Vol.14/ No.3/ Juni 1997, S. 44-51.

Li, GuangBin [李光斌] (1993): Überblick von dem arabischen Schriftkreis [阿拉伯文字圈概况]. In: Xu, ShouChun [许寿椿] (1993): Wissenschaftliche Arbeiten über Vergleich der Schriften [文字比较研究散论]. Beijing: Minzu University of China Press, S. 10-18.

Li, Kai [李开] (2002): Die Linguistik der chinesischen Sprache und die Theorien der Chinesisch-als-Fremdsprache-Didaktik [汉语语言学和对外汉语教学理论]. Beijing: China Social Science Press.

Li, Mu [李牧] (TW): Das Messen und die Forschung an dem System Design der Sinographie [漢字系統工程的計量研究]. Kap. 2.3: Die statistischen Analysen und Diskussion über Homophonen [同音字统计及讨论]. *http://chinese.exponode.com/2_3.htm* [Abruf: 2015-01-16].

Li, WanFu [李万福] (2005): A analzsis of the inside structure of characters systems [论文字系统]. In: Journal of Chongqing Colleg of Education [重庆学报], Vol.18/ No.5/ Sep. 2005, S. 20-22.

Liu, Qian [刘迁]/ Jia, HuiBo [贾惠波] (2006): Die Forschung und künftige Perspektiven der Wortsegmentationstechniken bei der chinesischen Informationsverarbeitung [中文信息处理中自动分词技术的研究与展望]. In: Computer Engineering and Applications [计算机工程与应用], Vol.42/ No.3/ 2006, S. 175-177.

Liu, ZhengYi [刘政怡] (2007): Research on Chinese Sentential Intelligence Input Method [中文整句智能输入方法研究]. Supervisor: Prof. Wu, Jianguo [吴建国]. Dissertation submitted to Anhui University – For the Ph.D Degree, Hefei, VR China.

Liu, ZhengYi [刘政怡]/ Wu, JianGuo [吴建国]/ Liu, HuiTing [刘慧婷] (2008): Research on Syllable Segmentation Method [音节切分歧义方法研究]. In: „Computer Technology and Development“ [计算机技术与发展], Vol.18/ No.8/ 2008, S. 35-38.

Liu, ZhengYi [刘政怡]/ Wu, JianGuo [吴建国]/ Li, Wei [李炜] (2008): State Space Model Based on Sentence Input Method [基于整句输入法的状态空间模型]. In: „Computer Engineering and Application“ [计算机工程与应用], Vol.44/ No.30/ 2008, S. 153-156.

Lu, Chuan [鲁川], Wang, YuJiu [王玉菊] (2008): Sinogrammbasierte informatische Grammatik der chinesischen Sprache [汉字信息语法学]. Ji'nan: Shangdong Eduacation Press.

Lu, ZhongFa [陆忠发] (2009): Die neue Richtung der chinesischen Schriftlinguistik [汉字学的新方向]. Hangzhou: Zhejiang University Press.

Luo, WeiHua [骆卫华] / Luo, ZhenSheng [罗振声] / Gong, XiaoJin [宫小瑾] (2004): Study of Techniques of Automatic Proofreading for Chinese Texts [中文文本自动校对技术的研究]. In: Journal of Computer Research and Development, Vol. 41/ No.1/ Jan. 2004, S. 244-249.

Luo, ZhenSheng [罗振声] (1996): Die Erforschung an der chinesischen Sprache und Schrift im Zeitalter des Computers [计算机时代的汉语与汉字研究]. Beijing: Tsinghua University Press.

Lü, ShuXiang [吕叔湘] (1979): Die Probleme bei der Analyse der chinesischen Grammatik [汉语语法分析问题]. Beijing: The Commercial Press.

Nakasato, Shigumi [中里茂美] (1998): Multiple Language Input-system [多语言输入系统]. Type: Patent-Anmeldung [von ‚Patentministerium der Volksrepublik China‘]. Anmeldungs-Nr.: 98109769.3. Publikations-Nr.: CN1203398A, -Datum: 30.12.1998. Antragsteller: Toshiba Corporation. Auch veröffentlicht unter: CN1118771C & US6182099.

National Languages Committee [國語推行委員會] (TW): Die allgemeine Geschichte des Entwurf der standardisierten Glyphen [標準字體研訂簡史]. Online verfügbar: *http://www.edu.tw/files/site_content/M0001/std/no1.htm?open* [Abruf: 2015-02-05].

Pan, JiXing [潘吉星] (2010): Die chinesische Papierherstellungstechniken [中国的造纸术]. Beijing: China International Radio Press.

Peng, ShouQuan [彭寿全] (1994): Die Informationsverarbeitung in der chinesischen Schrift [汉字信息处理]. Chengdu: University of Electronic Science and Technology of China Press (UESTCP).

Qi, HuYang [齐沪扬] (2009): Skizzen der Anwendungslinguistik [应用语言学纲要]. Shanghai: Fudan University Press.

Qian, WeiGang [钱为钢] (2002): Die Anweisung zum Kurs des Anwendungschinesisch [应用汉语教程学习指导]. Shanghai: Shanghai Literature and Art Publishing Group.

Shen, Kuo [沈括] (ca. 1086-1093): Mengxi Bitan [梦溪笔谈] (wört.: *Essays der Pinselunterhaltungen am Traumbach*). Online verfügbar: *https://so.gushiwen.org/guwen/book_14.aspx* [Abruf: 2019-03-13].

Shen, XiLun [沈锡伦] (2011): Die aktuelle Situation der sinojapanischen Schrift [日本汉字现状]. Wochenpresse für Sprach- sowie Literaturwissenschaft [语言文学周报], No.004 [Publikation: 2011-11-09], S. 1f.

Sheng, XinHua [盛新华] (2006): Die Anwendung der Sprache und Schrift [语言文字应用]. Beijing: Unity Press.

„Shuowen-Jiezi“ Bianwei Hui [《说文解字》编委会] (2012): Shuowen-Jiezi [说文解字]. Beijing: The Chinese Overseas Publishing House.

Sun, JunXi [孙钧锡] (1988): Generelle Theorien über die chinesische Schrift [汉字通论]. Shijiazhuang: Hebei Education Press.

Tang, Ping [汤平] (2014): Der Vergleich von ein paar beliebtesten Pinyin-Eingabemethoden [几款主流拼音输入法的比较]. In: Electronic Technology & Software Engineering [电子技术与软件工程], No.1/ 2004, S. 94.

Tang, Jian [唐建] (1992): Die großen ärchaologischen theoretischen Bedeutungen von den Jiahu-Symbolen [贾湖遗址新时期时代甲骨契刻符号的重大考古理论意义]. In: Fudan Journal (Social Science Edition), No.03/1992, S. 94-107.

Wang, GuangZheng [王广正]/ Wang, XiFeng [王喜凤] (2008): A Method of POS Tagging Based on Priority of Rules [一种基于规则优先的词性标注方法]. In: Journal of Anhui University of Technology (Natural Science) [安徽工业大学学报（自然科学版）], Vol.25/ No.4/ Okt. 2008, S. 426-429.

Wangma-Unternehmen: Das Standard-Wubi-Zixing-Lehrmaterial des schnellen Lernens (Version 86) [标准五笔字型（86版）速成教材]. Online verfügbar unter: *http://www.wangma.net.cn/UploadFiles/otherfile/406a8cb1f76440628292f56ee6ddc2b2.pdf*. [Abruf: 2016-04-15].

Wang, XiaoLong [王晓龙] (2005): Die maschinelle Sprachverarbeitung mit Computer [计算机自然语言处理]. Beijing: Tsinghua University Press.

Wang, XiaoLong [王晓龙]/ Wang, KaiZhu [王开铸]/ Sun, XiWen [孙希文]/ Wang, YingWei [王英伟] (1993): Machine Learning in Pinyin-character Translation [音字转换中的机器学习研究]. In: Chinese J. Computers [计算机学报]. Vol.16/ No.5/ Mai 1993, S. 370-377.

Wang, XiaoLong [王晓龙]/ Wang, YouLong [王幼龙] (1996): Chinese Input by Sentence [语句级汉语输入技术]. In: Journal of Chinese Information Processing [中文信息学报], Vol.10/ No.4/ 1996, S. 50-59.

Wang, XiaoYan [王晓燕]/ Mao ShuFen [毛树芬] (2000): Schnelles Erwerben von Wubi-Zixing [五笔字型一册通]. Xi'an: Xidian University Press.

Wang, YongMin [王永民] (2005): The Three Principles of Computer Chinese Character Keyboard Design [计算机汉字键盘设计“三原理”]. Beijing: Wangma Group of China [中国王码集团].

Wang, YongMin [王永民] (2008): Research and Application in Chinese Input Technology for Numerical Keyboard [数字键汉字编码技术的研究和应用]. In: Chinese Journal of Computers [计算机学报], Vol.31/ No.6/ June 2008, S. 1046-1055.

Wang, YongMin [王永民]/ Yang, TaoYuan [杨桃源] (2005): Mahnung vor der von Pinyin-Eingabemethoden verursachten Reduzierung der Verwendungsfähigkeit der chinesischen Schrift [警觉拼音输入法对运用汉字能力的销蚀]. In: Guangming Daily [11.10.2005]. Online verfügbar: *http://edu.china.com.cn/2013-08/09/content_29675445.htm* [Abruf: 2018-06-27].

Wu, LiangZhan [吴良占] (1999): Die Eingabemethoden der chinesischen Schrift in der modernen Zeit [现代汉字输入法]. Hangzhou: Zhejiang Science and Technology Publishing House.

Xiao, GuoZheng [肖国政] (1988): Die Erklärung der Zweifelpunkten der modernen chinesischen Grammatiken [现代汉语语法释疑]. Wuhan: Central China Normal University Press.

Xu, ShouChun [许寿椿] (1991): On the Order of Character Set [字符集的序性]. In: Journal of Chinese Information Processing, Vol.5/ No1./ 1991, S. 28-35.

Xu, ShouChun [许寿椿] (1993): Wissenschaftliche Arbeiten über Vergleich der Schriften – neue Entdeckungen in dem Zeitalter des Computers [文字比较研究散论——计算机时代的新观察]. Beijing: Minzu University of China Press.

Xu, TongQiang [徐通锵] (2008): „Zeichenbasierender Aspekt“ und Sprachforschung [“字本位”和语言研究]. In: Lu, Chuan [鲁川], Wang, Yujiu [王玉菊] (2008): Sinogrammbasierte informatische Grammatik der chinesischen Sprache [汉字信息语法学]. Ji'nan: Shangdong Eduacation Press. Das Pralog des Buchs, S. 1-21.

Xu, YangChun [徐阳春] (2008): Die moderne chinesische Sprache [现代汉语]. Beijing: Higher Education Press.

Xu, ZhiMing [徐志明]/ Wang, XiaoLong [王晓龙]/ Jiang, ShouXu [姜守旭] (2000): A Sentence-Level Chinese Character Input [一种语句级汉字输入技术的研究]. In: Chinese High Technology Letters [高技术通讯], Vol.10/ No.01/ 2000, S. 51-55.

Yao, YaPing [姚亚平] (1997): Computerlinguistik in China [中国计算语言学]. Nanchang: Jiangxi Science and Technology Press.

Yin, GuangFu [殷光复] (1998): Die häufig gebrauchten Computersysteme in der chinesischen Schrift und die Eingabemethoden der chinesischen Schrift [常用计算机汉字系统及汉字输入法]. Beijing: China Water & Power Press.

Yu, JinFeng [余锦凤] (2002): Grundkurs der chinesischen Informationsverarbeitung [中文信息处理基本教程]. Beijing: Beijing University Press.

Yuan, XiaoYuan [袁晓园] (1991): Wissenschaftliche Arbeiten bei Symposium für chinesische Schrift und Sprache. [汉字汉语研讨会论文集]. Changchun: Jilin Education Press.

Yuan, Zhe [袁哲] (2008): Überblick von den Pinyin-Eingabemethoden und die Erneuerung derer Funktionalitäten [浅谈拼音输入法及其功能创新]. In: The Science Education Article Collects [科教文汇], Dec. 2008, S. 279f.

Yuan, Zhe [袁哲] (2010): Die Anwendung der künstlichen Intelligenz bei der Pinyin-Eingabesoftware [人工智能在拼音输入法中的应用]. In: Software Guide [软件导刊], Vol.9/ No.6/ June 2010, S. 10-12.

Zhai, LiTing [翟荔婷] (2012): Überblickende Vorstellung der Wortsegmentationsmethoden der chinesischen Texte [浅谈中文文本分词方法]. In: Manager Journal [经济管理者], No.18/ 2002, S. 258-259.

Zhang, Sen [章森]/ Zong, ChengQing [宗成庆]/ Chen, ZhaoXiong [陈肇雄]/ Huang, HeYan [黄河燕] (1997): Intelligent Approach: From PINYIN to Chinese Characters [语句拼音—汉字转换的智能处理机制分析]. In: Journal of Chinese Information Processing [中文信息学报], No.2/Vol.12/ 1998, S. 37-43.

Zhang, ShuDong [张树栋]/ Pang, DuoYi [庞多益]/ Zheng, RuSi [郑如斯] etl. (2004): The Concise General History of Chinese Printing [中华印刷通史]. Guilin: Guangxi Normal University Press (GROUP), 1. Auflage.

Zhang, Yuhua [张玉华]/ Zhou, Kelan [周克兰] (2004): The Design of Input Method Auto Evaluating System based on Rule Lib [基于规则库的汉字输入法自动评测系统的设计]. In: Journal of Chinese Information Processing, Vol.18/ No.4/ 2004, S. 50-54.

Zhang, YuJin [张玉金] (1991): Die Kultur von den Determinativphonetika [形声字文化]. In: Yuan, Xiaoyuan (1991): Die Sammlung der wissenschaftlichen Arbeiten beim Symposium für die chinesische Schrift und Sprache [汉字汉语学术研讨会论文集]. Changchun: Jilin Education Press [吉林教育出版社], S. 95-103.

Zhang, YuJin [张玉金] (2000): Die moderne chinesische Schriftlinguistik [当代中国文字学]. Guangzhou: Guangdong Education Publishing House.

Zhang, ZhouCai [张轴材] etl. (1991): Der aktuelle Stand und die Entwicklung der Technologie der chinesischen Textverarbeitung [中文信息处理技术的现状与进展]. Herausgabe von: The international Association for general Chinese Character Code (ACCC) [通用中文代码国际联合会].

Zhao, BoZhang [赵珀璋] (1987): Die Informationsverarbeitung der chinesischen Sprache mit Computer [计算机中文信息处理] – Unterband. Beijing: China Astronautic Publishing House.

Zhao, YongXin [赵永新] (1992): Essentials of Chinese Grammar [汉语语法概要]. Beijing: Beijing Language and Culture University Press (BLCUP).

Zhou, Gang [周钢]/ Zhu, YinNiu [朱荫牛] (2002): Die Generierungssystem der intelligenten Eingabe-Platte der chinesischen Schrift [智能汉字输入平台生成系统]. In: Chinese Information Processing Society of China, Committee of Chinese Character Coding, 8th Symposium [中国中文信息学会汉字编码专业委员会第八届年会], S. 41-44.

Zhou, Qiang [周强]/ Duan, HuiMing [段惠明] (1999): Die Wortsegmentation und POS Tagging bei der Verarbeitung der Korpora im modernen Chinesischen [现代汉语语料库加工中的切词与词性自动标注处理]. Herausgabe von: Institute of Computational Linguistics (ICL), Beijing University [北京大学计算语言学研究所], Beijing.

Zhou, Qiang [周强]/ Yu, ShiWen [俞士汶] (1993): Eine von Wortsegmentation und POS-Tagging zusammengestellten mehrstufigen Verarbeitungsmethode an chinesischen Korpora [一种切词和词性标注相融合的汉语语料库多级加工方法]. In: Symposium of the 2th China National Conference on Computational Linguistics/ CCL 1993 [第二届全国计算语言学会议].

Zhou, YouGuang [周有光] (1954): Die Geschichte der Buchstaben [字母的故事]. Shanghai: Shanghai Education Publishing House.

Zhou, YouGuang [周有光] (1980): Die Untersuchung bei der Aussprache der phonetikhinweisenden Komponente der Sinogramme [汉字声旁读音便查]. Changchun: Jilin People's Publishing House, 1. Auflage.

Zhou, YouGuang [周有光] (1996): Die Verbesserung der chinesischen Eingabetechniken mithilfe von den innerlichen sprachlichen Regeln [利用汉语的内在规律，改进中文的输入技术]. In: Luo, ZhenSheng [罗振声] (1996): Die Erforschung an der chinesischen Sprache und Schrift im Zeitalter des Computers [计算机时代的汉语与汉字研究]. Beijing: Tsinghua University Press, S. 389-394.

Zhu, WenXiong [朱文雄] (2000): Die Theorien der Didaktik der chinesischen Grammatik [汉语语法教学论纲]. Nanning: Guangxi People's Publishing House.

Zhu, XiaoXu [朱晓旭] (2002): Der Entwurf der Segmentationsmethoden von Wortketten im Pädagogik-System der Eingabe der chinesischen Schrift [汉字输入教学系统中词组切分方法的设计]. In: Chinese Information Processing Society of China, Committee of Chinese Character Coding, 8th Symposium [中国中文信息学会汉字编码专业委员会第八届年会], S. 19-22.

Zou, XiaoLi [邹晓丽] (2004): Generelle Theorien über die chinesische Schrift [汉字通论]. Shenyang: Shenyang Press.