LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

METHODS SHOWCASE ARTICLE

(Generalized Linear) Mixed-Effects Modeling: A Learner Corpus Example

Stefan Th. Gries

Department of Linguistics, University of California, Santa Barbara & Justus Liebig University Giessen, UC Santa Barbara, Santa Barbara, California, USA

Abstract: This methods showcase article provides a detailed overview of a mixed-effects modeling analysis of corpus data on the use of *that* in object and subject complementation by native speakers of English compared to its use by German and Spanish learners of English.

> "We emphasize that we do not claim that our illustrations are the only way to carry out these analyses, but the strategy outlined above has yielded satisfactory results." (Bates et al., 2018, p. 5)

Keywords Mixed-effects modeling; generalized linear modeling; regression; that-complementation

Introduction

General Introduction

Over the last 10 or so years, mixed-effects regression modeling has taken linguistics by storm. Although many linguistic subdisciplines have been using regression-based approaches for a long time—and I am including the kinds of linear models that are still often referred to by traditional names such as ANOVA or ANCOVA—since at least 2008, mixed-effects modeling (MEM) has seen a meteoric rise in probably most sub-fields in linguistics. This is because 2008 saw the publication of both:

Correspondence concerning this article should be addressed to Stefan Th. Gries, Department of Linguistics, University of California, Santa Barbara & Justus Liebig University Giessen, UC Santa Barbara, Santa Barbara, California, USA. E-mail: stgries@linguistics.ucsb.edu

The handling editor for this article was Emma Marsden.

DOI: 10.1111/lang.12448

- Baayen (2008), Chapter 7 of which was one of the first at least somewhat introductory discussions of MEM in linguistics, and
- a special issue of the *Journal of Memory and Language* in 2008 on "Emerging Data Analysis" (Volume 59, Issue 4), whose contributors included again Baayen, Davidson, and Bates but also Quené and van den Bergh, Barr, and others who showcased the power of this kind of approach.

The most important advantage of MEM is the way it modifies or enriches generalized linear models. Generalized linear models involve the assumption that the data points (or observations or—in the usual long, case-by-variable format used by statistical software—the rows of the spreadsheet) are statistically independent of each other, an assumption that ideally (a) informed the design of the data collection process or the data analysis and (b) was checked after the fact as part of model diagnostics (e.g., by checking, visually or otherwise, the structure of the residuals of linear models especially).

In what way might data points not be statistically independent of each other? Several possibilities are common. First, data points might be related because they share characteristics other than those encoded in predictors, as when subjects in experiments produce multiple/repeated measures of the response variable. That way, all measurements of a certain subject or on a certain stimulus might be affected by, for example, the subject's general idiosyncratic aptitude or motivation or the characteristics of the stimulus. In addition, subjects might differ not just in an overall tendency for certain (ranges of) values of the response variable but also in their reaction to changes of the predictor(s). Crucially, multiple repeated-measurements structures of this type can coexist in one experimental design, as when every subject contributes multiple measurements and when there are multiple measurements for each stimulus/item. This is often referred to as a *crossed random-effects structure*. Diagnostically, such repeated-measurements structures could show up in residual plots exhibiting structure (resulting from clumping of measurements from one speaker) or from notable results in influence measures. In addition to such overall tendencies and tendencies to react to predictors, subjects' data points might also be temporally related as when the multiple data points of a subject in an experiment are subject to learning, habituation, practice, or fatigue effects or when speakers in observational data exhibit priming and/or resonance effects. This means that even just knowing the previous value(s) of a response variable (perhaps even without knowing the values of any predictors) already helps predict the current value of the response (a situation referred to as *autocorrelation*).

Second, data points might also be related in a more specific version of sharing characteristics, namely, because they share multiple taxonomically or hierarchically organized characteristics (what Meteyard & Davies, 2020 call "multi-stage sampling"), giving rise to what are so-called *nested random effects*. A frequently-used example to explain this involves educational research. For instance, when, high school students are tested, all students within the same classroom share the teacher (whose teaching style might have a certain impact on all of his students), but they also share the school (whose parent-teacher association might work in a way that is systematically different from that of other schools), but they also share the school district (whose funding situation might affect all its schools differently from the way funding affects schools of other districts). This, too, might diagnostically show up in residual structure and/or influence measures.

Although the simplest kind of MEM is statistically equivalent to a *t* test for dependent samples, MEM can consider multiple crossed and nested random effects simultaneously (e.g., repeated measurements for subjects and items) in way that such *t* tests or different kinds of workarounds proposed for repeated-measures ANOVAs—quasi-F and then $F_1/F_2/\mathrm{min}F$' analyses widely used after Clark's (1973) seminal paper—cannot. For instance, an experimental set-up where each speaker and each stimulus contribute multiple data points is dealt with by including crossed random effects in a way far superior to averaging over subjects and/or items (see Baayen, 2008, Section 7.2; Baguley, 2012, pp. 732−733; Meteyard & Davies, 2020; and especially Brauer & Curtin, 2018, for discussion of the disadvantages of the traditional ANOVA).

To illustrate how MEM proceeds conceptually, I present a simple linear regression model in Equations 1 and 2 with a numeric dependent/response variable Y (e.g., a reaction time or a duration) and a numeric independent/predictor variable X (e.g., length or frequency of a word), where Equation 1 shows the typical mathematical notation and Equation 2 shows how this model might be written in R.

$$y = a + bx \tag{1}$$

$$\mathrm{lm}\,(\mathrm{Y} \sim 1 + \mathrm{X}, \ldots) \tag{2}$$

In this notation, the a in Equation 1 represents the intercept and, for all practical intents and purposes, the 1 in Equation 2 does the same. Whereas the b in Equation 1 represents the slope of the numeric predictor X, which the linear model in Equation 2 would estimate/return, and, from this intercept and slope, one can compute a regression line that visualizes the relation between Y and X. A small data set with 30 observations of a response Y and a predictor X from Speakers 1, 2, and 3 is plotted in Figure 1 such that the speaker names are

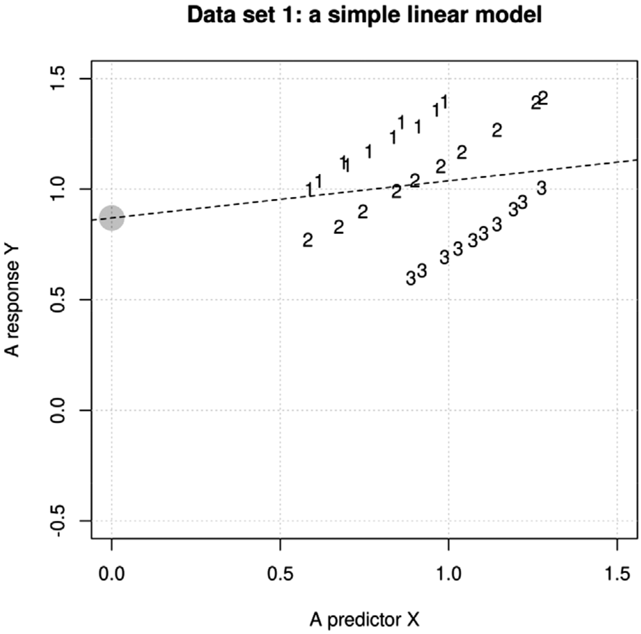

Figure 1 Scatterplot of Data Set 1 with a fixed-effects regression line (dashed).

used as point characters together with the regression line visualizing the trend for all three speakers combined (with an intercept *a*/1 of 0.87 represented by the grey point) and a slope of X of 0.168, which means that, for every one-unit increase of the numeric predictor X, the numeric response variable Y increases by 0.168 units.

It is obvious that the fit of the regression line for all three speakers together is weak (multiple $R^2 = .022$), but it is apparent that the speakers individually exhibit what looks like a perfect correlation between y and x and that the slopes of x for each speaker seem to be very similar. Now if all observations were independent of each other, using one intercept for all of them could theoretically make sense, but, because the 30 data points are contributed by only three speakers, then a mixed-effects model can take that into account. The simplest way it can do so statistically is by letting all speakers have their own intercept while still letting them share a single slope of X; more precisely, one would say there is also a shared intercept, but all speakers get their own adjustment to it.[1]

Figure 2 shows the result of such a mixed-effects model, where the solid lines are the speaker-specific regression lines with different intercepts (represented by the three grey points on the left when $X = 0$) but the same

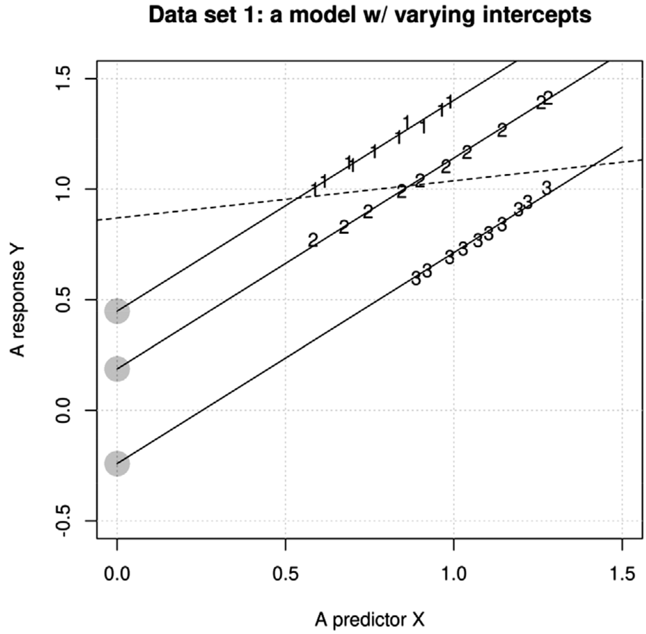

Figure 2 Scatterplot of Data Set 1 with fixed and mixed-effects regression lines (dashed and solid respectively).

slope (obvious from the regression lines being perfectly parallel) of 0.95. If Figure 2 exemplified reaction times in a lexical decision task (*Y*, which might be *z*-standardized/scaled) as a function of word lengths (*X*, which were transformed from the original values), then it would mean that all three speakers' reaction times increased in equal measures to word length (the slope is 0.954, meaning it is approximately six times as high as in the linear model in Figure 1), but the three speakers differed in their baseline reaction speed. The R^2 of .998 of this MEM indicates how much better and appropriate this model is for these data.[2]

A second kind of MEM can be motivated by the data shown in the left panel of Figure 3, where again the overall linear model (with an intercept *a*/1 of 0.94 and a slope of *X* of 0.41) results in a weak fit ($R^2 = .218$). However, the right panel shows what happens when, this time, the relatedness of the 10 data points of each speaker is accommodated not by each of the speakers getting their own intercept adjustment but by their own slope adjustment. If Figure 3 was a similar reaction time paradigm as above, this would mean in effect that all speakers are equally fast in general (as a kind of baseline, all speakers'

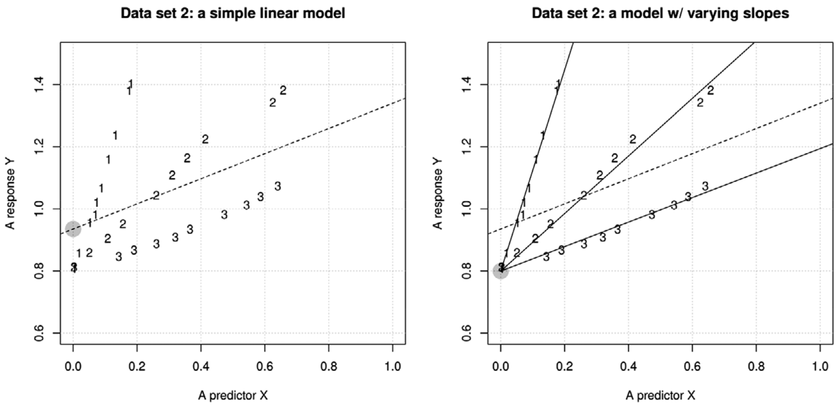

Figure 3 Scatterplot of Data Set 2 with fixed and mixed-effects regression lines.

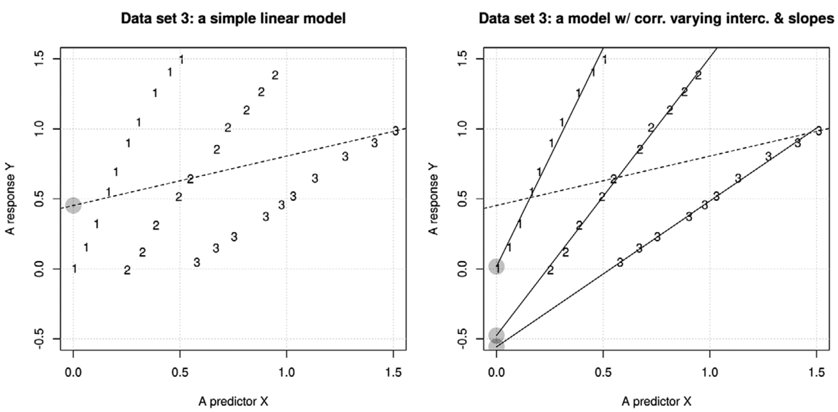

Figure 4 Scatterplot of Data Set 3 with fixed and mixed-effects regression lines.

intercept is 0.8) but become slower differently in response to the words getting longer. Speaker 1's slope is predicted to be 1.52 (the shared slope) plus that speaker's adjustment of 1.72 for an overall 3.24 whereas, for instance, Speaker 3's slope is predicted to be 1.52 (the shared slope) plus that speaker's adjustment of −1.13 for an overall 0.39. This model, which allows speakers' slopes to vary quite widely from the overall slope of 1.52, which would be used to predict results for unseen speakers, results in an R^2 of .999.

Finally, a MEM can also accommodate the relatedness of each speaker's data points by giving individual speakers their own intercept as well as their own slope. The left panel of Figure 4 shows a model (with an intercept *a*/1 of

0.45 and a slope of X of 0.35), which results in a weak fit ($R^2 = .099$), whereas the right panel shows the MEM with (correlated) speaker-specific intercepts and slopes resulting in an R^2 of .999. Figure 4 would therefore mean something like all speakers have different baseline reaction speeds but also react differently in response to the stimulus words getting longer.[3]

(For a more detailed discussion of these examples, see Gries to appear b, Section 6.1, where I have the space to also discuss the consequences of using the better-suited MEMs in terms of regression/residual diagnostics.) Even in this small example, some of the advantages of MEM should be apparent: Trivially, one can avoid violating the independence-of-data-points assumption of fixed-effects modeling. However, more important, because the MEM deals with speaker- or stimulus-specific idiosyncrasies separately, so to speak, it can help avoid incorrect decisions regarding the nature and the significance of effects. For instance, the fixed-effects regression slope in Figure 1 is not significantly different from 0, but the overall mixed-effects regression slope in Figure 2 is. Conversely, the fixed-effects slope in the left panel of Figure 3 is significantly different from 0, but the overall mixed-effects regression slope in Figure 2 is not. This of course also suggests that the fixed-effects results of a mixed-effects model (which one would use to predict new cases from previous, unseen speakers or for previously unseen items/stimuli) will be more robust and likely generalize better.

Data and Hypotheses

The data whose analysis I discuss here are a subset of the corpus data studied (differently) in Wulff, Gries, and Lester (2018). The data and code are available in the online supplementary material (Appendix S1; see https://osf.io/xmkpt/?view_only=0d2fb75051e8449c87dd30eded9683ab). The supplementary material also allows readers to see exactly how the analysis was conducted. The binary response variable for the present paper was the use or omission of an optional *that*-complementizer in subject and object complementation (see Example 1 and Example 2 respectively) by native speakers (NS) of English and German and Spanish nonnative speakers (NNS)/learners, of English; this variable was called COMPLEMENTIZER and had the levels *absent* (Examples 1a and 2a) and *present* (Examples 1b and 2b):

Example 1

(a) The problem was **Ø** the Vorlons refused to help in the war.
(b) The problem was **that** the Vorlons refused to help in the war.

Example 2

(a) I thought **Ø** the Vorlons could have helped earlier.
(b) I thought **that** the Vorlons could have helped earlier.

For this methods showcase article, I considered the following variables as having been found to be correlated with the choice to (not) realize *that* in such cases. I will explain them using the example sentence in Example 3:

Example 3

Seriously, I really hope very much **that** the Vorlons will help us against the Shadows.

- REGISTER: was the example *spoken* or *written*?
- TYPE: the complementation type: *object* or *subject*—here *object*;
- L_MATRBE4S: the length of any material before the matrix clause (in characters)—here 9 (*Seriously*);
- L_MATRSUBJ: the length of the matrix clause subject (in characters)—here 1 (*I*);
- L_MATRS2V: the length of any material between the subject and the verb of the matrix clause—here 6 (*really*);
- L_MATRV2CC: the length of any material between the verb of the matrix clause and the complement clause—here 9 (*very much*);
- L_COMP: the length of the complement clause—here 44 (*the Vorlons will help us against the Shadows*);
- L_COMPSUBJ: the length of the complement clause subject—here 11 (*the Vorlons*);
- L_COMPREST: the length of the rest of complement clause—here 32 (*will help us against the Shadows*).

However, the main new predictor variables of interest in this particular analysis were the following:

- DPCW, which represents $\Delta P_{\text{construction2word}}$, a unidirectional association statistic which measures the degree to which *that* omission prefers the specific matrix clause verb;
- DPWC, which represents $\Delta P_{\text{word2construction}}$, a unidirectional association statistic, which measures the degree to which the matrix clause verb prefers *that* omission; given the temporal order of the two events, this is the one with the stronger expectation of an effect;
- SURPRISAL: the degree to which the last word of the matrix clause (*much* in Example 3) makes the first word of the complement clause (the first

the in Example 3) surprising based on the (usual) negative binary log of the weighted conditional probability *p* (*the*|*much*) from the British National Corpus.

For each of these three predictors, Hypothesis 1 was that it would have an effect on whether or not speakers use a complementizer. In addition, for each of these predictors, Hypothesis 2 made the prediction that the NS would behave differently from the learners because of the additional processing effort that comes with producing in a language that is not one's first language (L1).

Hypotheses 1 and 2, the effects of DPCW and DPWC, were motivated by the fact that research has shown that lexical items have sometimes very strong constructional preferences (Gries & Stefanowitsch, 2004; Stefanowitsch & Gries, 2003). The preference has been used in priming but now also in complementation research (see, e.g., Jaeger, 2010, who operationalized them as surprisal). These two verb-specificity hypotheses were operationalized with the above-mentioned ΔP statistics (see Ellis, 2007, and Gries, 2013) from a distinctive collexeme analysis based on the present data set.[4]

Hypothesis 3, the effect of SURPRISAL, was motivated by the growing body of work documenting predictability/surprisal effects in NS sentence processing (see, e.g., Jaeger & Snider, 2008; Lester, 2018; Linzen & Jaeger, 2014) and the interest in determining whether this would also play a role in NNS.[5]

In this learner corpus context and for the fixed effects (henceforth, *fixef*), I was also interested in the variable L1, which had the levels *English*, *German*, and *Spanish*. L1 is required to interact with the three main predictors of interest to see whether any of the effects of these predictors differ between the levels of L1.

In terms of potential random effects (henceforth, *ranef*), I will consider FILE (i.e., the speaker who produced the example) and MATCHLEMMA (i.e., the matrix clause verb lemma, i.e., *hope* in Example 3). These variables are exactly the sources of repeated-measures variation (i.e., most speakers contribute more than one data point, most matrix clause verbs occur more than once in the data) that, in an ANOVA context, would correspond to by-subjects/by-items calculations.

The data to be analyzed in this showcase article had the overall basic distribution shown in Table 1. The examples are from the British Component of the International Corpus of English (Nelson, Wallis, & Aarts, 2002) for the NS and from the International Corpus of Learner English (Granger, Dagneaux, Meunier, & Paquot, 2009) and the Louvain International Database of Spoken English Interlanguage (Gilquin, De Cock, & Granger, 2010) for the NNS.

Table 1 Overview of the distribution of the data set

	COMPLEMENTIZER		
First Language	*absent*	*present*	Total
English	2,493	1,176	3,669
German	727	642	1,369
Spanish	549	587	1,136
Total	3,769	2,405	6,174

In the Methods section, I discuss all aspects of the modeling process that I can include here. In the Methods for Interpreting Mixed-Effects Models section, I discuss the presentation of results of the modeling process before presenting the Discussion and Concluding Remarks section. It recommends this article be read together with the online supplementary material mentioned above.

[T]here is no single correct way to implement an LMM, and … the choices they [researchers] make during analysis will comprise one path, however justified, amongst multiple alternatives. (Meteyard & Davies, 2020, pp. 1−2)

Methods

Let me first give an overview of a general structure of steps. The discussion below will shed more light on each of the steps, and the online supplementary material exemplifies everything in much detail (see also Meteyard & Davies's, 2020, best-practice list and Gries, in press-b, Sections 6.2−6.5 for a much more detailed discussion).

Often the question arises as to what to include in the final write-up. The answer to that question is actually easy: Everything from Table 2 with the exceptions of the descriptive visualization of Step 1 and the things one does not do (e.g., if one does not do model amalgamation, it does not need to be discussed). This is how the methods section becomes comprehensive enough (a) to provide the full context for the results and their interpretation and (b) to ensure replicability (for more information on reporting, etc., see Norris, Plonsky, Ross, & Schoonend, 2015, and Meteyard & Davies, 2020, especially Table 7 and their bullet list for best practice). Ideally, readers will, of course, have access to the full analysis as is the case in the online supplementary material to this methods showcase article held at https://osf.io/xmkpt/?view_only=0d2fb75051e8449c87dd30eded9683ab.

Table 2 General overview of a mixed-effects model process: What needs to be done/considered

Step 1. Exploration/preparation	
descriptive stats and visualization	every variable on its own
	every fixef & ranef with response
	every combination of fixefs
	every combination of ranefs with relevant predictors
correcting data	e.g., correcting wrongly entered data
discarding/trimming data	e.g., discarding outliers (before/after first model),rare levels of fixefs/ranefs
merging/conflating data	e.g., rare levels of fixefs/ranefs, highly correlated variables
form-changing transformations	e.g., log, inverse, sqrt, (i)logit, power (e.g., Box-Cox), …
form-preserving transformations	e.g., centering, *z*-standardizing
setting contrasts for factors	e.g., treatment contrasts, ordinal factors, planned contrasts
Step 2. Initial model formulation	
fixef structure	What hypotheses are being tested?
	What needs to be controlled for?
ranef structure	What is the maximal ranef structure that seems necessary?
Step 3. Model/variable selection	
single hypothesis test or model/variable selection or model amalgamation?	
if model/variable selection, …	ranefs, then fixefs (Zuur et al., 2009)
	direction of model selection? forward? backward? hybrid?
	criterion of model selection? *p*? information criteria?
(If model amalgamation, what is the range of reasonable models?)	
dealing with problems	convergence, collinearity, overdispersion, …

(*Continued*)

Table 2 (Continued)

Step 4. Model diagnostics	
(depending on this, one might have to go one to three steps back up again)	
residuals, collinearity, overdispersion, influence measures, distribution of ranef adjustments …	
Step 5. Model validation	
How well does the model generalize? cross-validation (nontrivial for mixed-effects models)	
Step 6. Model interpretation	
overall model statistics	significance test, R^2s
	accuracy (compared to the baseline[s]), precision, recall, *C*-score
fixefs	coefficients, confidence intervals, and significance tests
	effects plots of predictions
ranefs	dotcharts of adjustments (optional, but useful)
	ranef-specific slopes (optional, but useful)

Note. fixef = fixed effects; ranef = random effects.

Exploration and Preparation

As Table 2 indicate, the first analytical step is one that must in fact precede all kinds of regression analyses, not just MEMs: a thorough exploration of one's data. For (combinations of) categorical variables, this exploration should minimally involve (cross-)tabulation; for (combinations of) numeric variables, this should minimally involve numeric summaries and differently binned histograms, ecdf plots;[6] and/or scatterplots (while paying attention to nonlinear trends as well); and for combinations of categorical and numeric variables, this should minimally involve boxplots, spineplots, or similar graphic displays of the data. The purpose is always to find and maybe address distributional peculiarities such as (extreme) skew, potential outliers, gaps in distributions, missing data, or infrequent or nonexistent factor (variable) levels or their combinations (which might rule out the study of interactions or the use of certain random slopes) and to pay attention to variables in need of transformations or factorization/binning or to the need for interactions or required or useful contrast settings.

As the supplementary material shows, a variety of decisions were made that are worth mentioning. For instance, the tabulation of MatchLemma indicated the extreme Zipfian distribution of the lemmas.[7] Thus, to avoid data sparsity and, consequently, likely convergence issues later, the data set was reduced to all those levels of MatchLemma with a frequency of ≥10 (which is already low; see the supplementary material for an alternative way to proceed). The tabulation of File indicated a similar, but less extreme, Zipfian distribution and, for the same reasons as above, the data set was reduced to all those levels of File with a frequency of ≥4, leading to a final sample size of 5,187 data points.

The variable L1 was recoded with orthogonal a priori contrasts to make the regression output more straightforwardly interpretable: The first contrast pitted NS against NNS, the second German against Spanish learners, and the contrasts were scaled such that the coefficients in the summary regression table reflected differences in logits directly.

As one can see in the supplementary material, many of the numeric controls and predictors exhibited very long right tails, which were dealt with in different ways:

- Some numeric variables were factorized on the basis of the results from separate conditional inference trees in which they were the only predictor of Complementizer, leading to new predictors L_MatrBe4S.fac and L_MatrS2V.fac (with two levels) and L_MatrSubj.ord (an ordinal variable with four levels). Three important comments should be made here: (a) Factorization is not obviously unproblematic given the information loss it incurs. However, the extreme skew of some of these variables was considered more damaging to subsequent modeling than factorization. (b) Factorization could also have been done on the basis of the variables' univariate distribution. (c) It should be noted that L_MatrSubj.ord is treated here as an ordered factor/an ordinal variable. Most studies that use binning or inherently ordinal predictors (such as an animacy hierarchy) that I see do not do that, which also results in information loss and has no advantage of which I know.

- Some numeric variables were Box-Cox transformed leading to the new predictors L_Comp.bcn, L_CompSubj.bcn, and L_CompRest.bcn.
- The numeric variable L_MatrV2CC had to be discarded because of its near-constancy: 99.77% of all its values were 0.

For the two association predictors DPCW and DPWC, I first reversed their polarity to make them more compatible with what the regression would try to predict (presence, not absence, of the complementizer, i.e., I multiplied them by -1, and then I checked whether they were highly correlated.). However, although their linear correlation was very high ($r > .9$), this was due to two classes of verbs that made an otherwise weaker correlation seem very strong; visualization, by contrast, showed that clearly. Transformations did not change the picture much and monofactorial spineplots looked promising so these predictors were left untransformed. Also, Surprisal was first winsorized such that all values of $\leq$3.3 were set to the mean of those Surprisal values and then Box-Cox transformed as well into Surprisal.wind.bcn. This affected less than 1.5% of the data that otherwise were extremely likely to affect subsequent regression modeling negatively (especially given the otherwise fairly robust correlation of Surprisal with Complementizer).

A final check for pairwise correlations between all numeric variables confirmed that, as expected, especially the variables L_Comp.bcn and L_CompRest.bcn were highly correlated. Many studies faced with something like this pick just one of these to enter into their model (often the one with the highest correlation with the dependent variable. Instead, I computed a principal component analysis on the two correlated variables and included as a predictor the scores of the first principal component, which was called L_ComplClPC and retained approximately 98.4% of the information of the original two correlated variables.

The Initial Model

As Table 2 indicates, the next step is to formulate an initial regression model that embodies all one's expectations about the data (for critical predictors, control variables, and random effects). One needs to consider questions such as: What kind of modeling perspective does one adopt (frequentist vs. Bayesian)? What is the fixed-effects structure (FES) of the predictor/control-response relations to be included in a/the model (and does one need interactions and/or curvature)? What random-effects structure (RES) does the model structure require (maximally)? These questions of course raise the issue of what to consider fixed effects and what to consider random effects. As always, not everyone agrees on how to proceed here (see Gelman & Hill, 2006, pp. 245−246, for a discussion), but it seems that most authors consider a variable to be a random effect:

- if the levels of the variable in the sample do not exhaust the levels that the variable would have in the population. For example, if one does a judgment experiment on contemporary American English with some stimulus sentences, then the 40 speakers who participate are not the whole population, and the 20 stimulus sentences that the speakers rate do not exhaust all possible sentences that could be presented to the subjects;
- if the real interest is not so much in what exactly these particular speakers and stimuli do but in generalizing from them to the population and just controlling for speakers' idiosyncrasies and for those of the stimuli.

Gelman and Hill (2006, p. 246) claimed that "[t]hese two recommendations (and others) can be unhelpful." That may be so, but these two criteria certainly seem to underlie most of the work that I see in linguistics and other fields (see, e.g., Brauer & Curtin 2018, p. 392).

According to the discussion of *that* complementation above, the first/maximal model involved a FES with:

- the three main predictors of interest: DPCW, DPWC, and SURPRISAL.WIND.BCN;
- the variable L1 and its interaction with each of these predictors of interest to see whether the effects of the main predictors of interest differed between L1s;
- the general categorical controls REGISTER and TYPE;
- the matrix clause controls L_MATRBE4S, L_MATRSUBJ, and L_MATRS2V;
- the complement clause controls L_COMPSUBJ and L_COMPLCLPC.

The RES of the first/maximal model followed Barr, Levy, Scheepers, and Tily (2013) as much as possible, but with the expectation that it would need to be reduced considerably. Specifically, Barr et al. (2013) argued in favor of fitting a maximal RES, that is, varying intercepts and slopes for all predictors of interest, here DPCW, DPWC, and SURPRISAL.WIND.BCN. However, although this is a justifiable starting point in theory, in practice it often leads to extremely complex RESs, which in turn often lead to model convergence issues. Matuschek, Kliegl, Vasishth, Baayen, and Bates (2017) and Bates, Kliegl, Vasishth, and Baayen (2018) have argued that it is acceptable or even better to be more parsimonious in one's RES. For the present data, I first confirmed the relations of the three main predictors with the two sources of random-effects variation, FILE and MATCHLEMMA. Because DPCW and DPWC were predictable from MATCHLEMMA, slope adjustments for these two predictors for MATCHLEMMA were not useful. There were usually not many different values of DPCW, DPWC, and SURPRISAL.WIND.BCN for the levels of FILE, so this situation might become quite tricky later, but this meant the maximal RES for now consisted of:

- intercept and slope adjustments for DPCW, DPWC, and SURPRISAL.WIND.BCN for the levels of FILE; these were allowed to be correlated for the simple reason that it seems to be most people's default and it is the more complex model whose RES might then be simplified;
- intercept adjustments as well as slope adjustments for SURPRISAL.WIND.BCN for MATCHLEMMA; these were allowed to be correlated, too.[8]

This initial model was fit, summarized, and then subjected to a model selection process.

Model/Variable Selection

Model/variable selection is one of the thorniest, most controversial issues in MEM—in particular, model selection based on significance testing. Some authors, with compelling reasons, make it very clear that they are against it (Harrell, 2015, pp. 67–69; Heinze, Wallisch, & Dunkler, 2018; Thompson, 1995, 2001). Yet textbooks illustrate model selection based on significance testing (e.g., Crawley, 2013, pp. 390ff; Säfken, Rügamer, Kneib, & Greven, 2018; Zuur, Ieno, Walker, Saveliev, & Smith, 2009, Section 5.7), and there are packages/functions with associated reviewed publications for model selection in general and for MEM in particular such as `FWDselect::selection` by Sestelo, Villanueva, Meira-Machado, and Roca-Pardiñas (2016), `lmerTest::step` by Kuznetsova, Brockhoff, and Christensen (2017), `cAIC4::step` by Säfken et al. (2018), or `buildmer::buildlme` by Voeten (2020). In addition, one finds quotes in the relevant literature that indicate that in related areas also using predictive modeling, for example, "in machine learning, variable (or feature) selection seems to be the standard" (Heinze et al., 2018, p. 432).[9]

In this study, I followed the two-step strategy outlined by Zuur et al. (2009, Chapter 5). I first determined the best RES and then, with that RES, the best fixed-effects structure (FES).[10] In terms of variable selection, I only considered the targeted/relevant predictors for deletion/removal, but not the control variables. As for the selection criterion of figuring out the right FES structure, I used likelihood ratio tests (LRTs) with a significance threshold set to .1 (because, as far as I know, this is the first exploration of this specific kind of unidirectional association and surprisal effects in learner corpus research; for the RES, I stuck with the traditional threshold of .05; see supplementary material). At the same time, I also monitored Akaike information criterion (AICc) values (whose use for finding the best RES is not uncontroversial). During model selection, I also did some basic, but important, model diagnostics, namely, checking for overdispersion and multicollinearity.[11] After model selection, I did some more model diagnostics/model criticism (now also involving residuals, influence measures, etc.). Like most authors in linguistics, I used `lme4::glmer`, but there are of course alternatives (e.g., `nlme`, `lmerTest`, `glmmPQL`, `MCMCglmm`, or the Bayesian modeling packages `blme` and especially `brms`).

Finding the Best Random-Effects Structure

As so often happens, Model `m.01` came with a convergence warning, meaning that the optimizer algorithm did not succeed in finding a robust solution, and a PCA of the ranef covariance matrix suggested that I might need only one or two ranef/variance components for File and maybe only one for MatchLemma (because of high intercorrelations of the random effects of each variable, also indicated in the summary output). Although I will turn to the FES below, it was encouraging to see that there were already some significant results (including for the predictors of interest, but also for the NS vs. NNS contrast of L1), no outsized standard errors (often a sign of multicollinearity problems), and a preliminary round of diagnostics revealed no outrageous collinearity problems (but see the supplementary material for a discussion of the collinearity results) or overdispersion issues.

What to do with the convergence warning? Multiple options are available and the online supplementary material lists the most frequently used ones (with code and additional references). Among the lowest-hanging fruit are refitting the model but using as starting parameters for its estimation process the results of Model `m.01` and/or even refitting the model with a higher number of iterations in the hope that more modeling time will make the algorithm converge. I could also immediately simplify the RES of Model `m.01` by deleting one of the ranef components with the smallest variance/standard deviation.

Yet another option is to not be too concerned about convergence warnings, at least not paranoically. Bolker, the main maintainer of `lme4`, once gave the following advice regarding convergence warnings:

- The bottom line is that your models are no more nor less trustworthy than they used to be [after an R update]. If you have tried them with a variety of optimizers, and if the results make sense, it's *probably* the case that they're OK, and just a little unstable. (Bolker, 2014, para. 3)
- Quick thought: "nonconvergence" doesn't necessarily mean the fit is actually bad (false positives blah blah), and in most (all?) cases you actually get a working fitted model (i.e., you could get the C-index). (Bolker, 2019, para. 1)

Although I am sure that Bolker would not want to give carte blanche to ignore all convergence warnings, it should be noted that (a) over the evolution of `lme4`, the threshold values for the convergence warnings were adjusted a few times, (b) there is/was a discussion of even dropping them altogether, (c) the `lme4` reference manual states that "warnings will occur even for apparently well-behaved fits with large data sets" (p. 18), and (d) after a few hundred MEMs one gets a bit of a feeling for what values in the convergence warning are associated with dangerous volatility/unpredictability. If the volatility of the fixefs is low and the results "make sense" (see Bolker quote above), then I am not too reluctant to accept the results of a model if it comes with a not too outrageous convergence warning.

There are, unfortunately, no easy answers and each first decision made at this point can of course affect all that follow. In this case, I chose to first drop the random slopes for SURPRISAL.WIND.BCN for FILE (Model `m.02`) and then the random slopes for DPWC for FILE (Model `m.03`) because the random slopes accounted for the least variability and were both supported by LRT; their removals had no consistent effect on the convergence warnings. Therefore, I removed the intercept adjustments for MATCHLEMMA (Model `m.04`) which was again supported by an LRT (but see the discussion in the supplementary material); the attempt to then delete intercept adjustments for DPCW (Model `m.05`) failed (significant LRT). Next, I removed the correlation parameter for FILE

(Model `m.06`), which was accepted, as was removing slope adjustments for SURPRISAL.WIND.BCN for MATCHLEMMA (Model `m.07`), but removing intercept adjustments for FILE was not (Model `m.08`). In an attempt to improve Model `m.07`, it was refitted into Model `m.09` with different start parameters and more iterations, which reduced convergence warnings considerably.

Finding the Best Fixed-Effects Structure

The exploration of the FES started with computing LRTs for deletion for all predictors in the model, which returned the interaction term of L1:SURPRISAL.WIND.BCN as the droppable predictor with the highest non-significant *p* value (according to the principle of marginality, see Fox & Weisberg, 2019, Sections 4.6.2 and 5.3.4). However, another important alternative was to consider whether any of the main predictors did not have a linear but a curved effect. This is often not tested, it seems, although it is probably fair to assume that many variables, especially cognitive ones such as learning, forgetting, or priming, exhibit curved effects. Here, because (a) a linear effect of SURPRISAL has been attested in other studies and (b) a curved effect is plausible (such that there might be an upper limit beyond which SURPRISAL is not more effective than before, just like with frequency effects), I explored curvature of SURPRISAL.WIND.BCN with a second-degree polynomial. However, both an LRT and AICc indicated that the added curvature of SURPRISAL.WIND.BCN in Model `m.10` did not significantly improve the model (and convergence became worse), which is why I then removed the interaction L1:SURPRISAL.WIND.BCN for Model `m.11`. Trying to improve this model by refitting it with new start values and more iterations/modeling time (Model `m.12`) did not help, and no terms still in the model—predictors and controls—could be deleted from Model `m.11`, which was therefore considered the final model and henceforth (optimistically) called Model `m.final`.

Model Diagnostics

Before I discuss the results in more detail, some model diagnostics are necessary.[12] First, I checked how volatile the current fixefs and ranefs were to the choice of optimizer. There was fairly little variability there; only one of the optimizers returned results that were notably different from the others and for only a limited number of effects. There was some more variability in how the fixefs changed over the course of the model selection process (especially with regard to SURPRISAL.WIND.BCN). Second, an exploration of the residuals with the DHARMa package (Hartig, 2020) did not raise many red flags (in terms of uniformity or dispersion, at least; there were some outliers). Third, the final tests for multicollinearity did not give great cause for concern (all variance inflation factors ≤ 10), and no overdispersion was found.

With regard to the ranefs, the adjustments to the intercepts for FILE were fairly normally distributed (compared against 200 random normal distributions with the same means and standard deviations), whereas the adjustments to the slopes of DPWC for FILE exhibited a bit of a discrepancy, but seemingly nothing major. The residuals for the different levels of FILE did exhibit some outliers (which one could explore post hoc, if so desired), but no problems in terms of normality and dispersion. In sum, these diagnostic results were not picture-perfect but also did not seem to reveal massive problems.

Finally, I computed a version of influence measures to determine how much influence each level of FILE had on the final fixed-effects coefficients, but the fixef results were stable no matter which speaker/level of FILE I removed. The supplementary material contains additional code and results that one might consider.

Model Validation

Model validation is an important issue that is essentially concerned with the question of how well the model that one is considering would do if it was given different data. Specifically in cases like the present one, the model returned predicted probabilities for cases on which it was trained, a situation which I usually refer to as *classification*. However, one is usually interested in the performance of a model on cases on which it is not trained, a situation which I usually refer to as *prediction*. To assess the degree to which the model generalizes well to other data, one can use cross-validation. However, I think it is fair to say that cross-validation for MEMs is far from being an obvious standard (certainly not in linguistics), an assessment I feel is supported by the complete absence of the topic even in the main overview articles and in many of the standard references of the top people in the field who currently inform MEM applications in linguistics.[13]

Methods for Interpreting Mixed-Effect Models

Overall Model Results/Quality

The next series of steps was concerned with inspecting the model's quality and numeric results as well as visualizing its results.

Model Significance, R^2*, and Classification Statistics*

In order to obtain an overall significance test for Model `m.final`, I did an LRT between Model `m.final` and the so-called null model (i.e., a model with the same RES but no fixed effects: Model `m.nofixef`); that comparison indicated that Model `m.final` was indeed significant, $LR(18) = 951.68, p < 10^{-15}$. The difference in AIC scores between Model `m.final` and model `m.nofixef` was very high (915.502). We should also report R^2 values. For Model `m.final`, I computed (as per Nakagawa, Johnson, & Schielzeth, 2017) an R^2_{marginal} of .38 (the R^2 that summarizes only the explanatory power of the FES) and an $R^2_{\text{conditional}}$ of .51 (the R^2 that summarizes the combined explanatory power of the FES and the RES). It was good to see that the FES did considerably more work than the RES.

Finally, it is useful to provide an indication of the classification accuracy of Model `m.final.` For each case in the data, I computed the predicted probability of COMPLEMENTIZER being *present*, and if that probability was ≥ .5, Model `m.final` predicted *present*, otherwise, it predicted *absent*. These classifications were then cross-tabulated with the actual choices and returned an accuracy of 84.21%, which, according to exact binomial tests, is significantly better than the no-information baseline and the random-guessing baseline (see the supplementary material for definitions). Also and more important than the classification accuracy, I computed the *C*-score, a widely used measure in the interval [.5, 1.0]; in linguistics, .8 is usually seen as a good score, and the score of Model `m.final` was .91. The supplementary material also provides precision and recall scores for both outcomes of the response variable as well as code to bootstrap confidence intervals for the predicted probabilities and another possible follow-up analytical step.

Fixed Effects

Numeric Results

For numeric results, it is customary to report the summary table of a MEM, that is, the estimates/coefficients, their standard errors, test statistics (usually *t* or *z*), and *p* values, often presented in a way that comes straight from R. My own preferred format, which I have not always been able to use myself, is a bit different because the reader can often not easily understand the estimates and/or the intercept because not all authors state what the contrasts or reference levels for their categorical predictors were. I therefore often prefer a table like Table 3, (a) where the table title reveals which level of the response variable is predicted, all of which help understanding the estimates, (b) where explicit labeling of the coefficients is provided in the subscripts, and (c) which provides confidence intervals of estimates.[14]

Table 3 Coefficients and their significance tests for Model `m.final` (predicted level: COMPLEMENTIZER: *present*)

Variable	*b*	*95 % CI*	*SE*	*z*	*p*
Intercept	−3.125	[−3.220, −2.528]	0.477	−6.557	<.0001
$L1_{\text{NS vs. NNS}}$	−0.908	[−1.045, −0.687]	0.163	−5.568	<.0001
$L1_{\text{German vs. Spanish}}$	−0.432	[−0.784, 0.026]	0.253	−1.710	.0873
$DPCW_{0\rightarrow1}$	−2.942	[−3.064, −2.354]	0.689	−4.273	<.0001
$DPWC_{0\rightarrow1}$	4.804	[4.525, 5.212]	0.514	9.339	<.0001
$\text{SURPRISAL.WIND.BCN}_{0\rightarrow1}$	2.066	[1.308,2.131]	0.526	3.929	<.0001
$\text{REGISTER}_{\text{spoken}\rightarrow\text{written}}$	0.985	[0.811, 1.213]	0.144	6.831	<.0001
$\text{TYPE}_{\text{object}\rightarrow\text{subject}}$	0.485	[0.156, 0.823]	0.199	2.429	.0151
$\text{L\_MATRBE4S.FAC}_{0\rightarrow1-202}$	0.532	[0.361, 0.703]	0.123	4.311	<.0001
$\text{L\_MATRSUBJ.ORD}_{L}$	0.548	[0.304, 0.689]	0.133	4.115	<.0001
$\text{L\_MATRSUBJ.ORD}_{Q}$	0.027	[−0.110, 0.172]	0.110	0.244	.8072
$\text{L\_MATRSUBJ.ORD}_{C}$	0.201	[0.100, 0.484]	0.112	1.802	.0716
$\text{L\_MATRS2V.FAC}_{0\rightarrow2-49}$	1.269	[0.795, 1.489]	0.213	5.951	<.0001
$\text{L\_COMPLCLPC}_{0\rightarrow1}$	0.161	[0.093, 0.231]	0.032	5.086	<.0001
$\text{L\_COMPSUBJ.BCN}_{0\rightarrow1}$	0.779	[0.695, 0.833]	0.083	9.360	<.0001
$L1_{\text{NS vs. NNS}}:DPCW_{0\rightarrow1}$	2.524	[1.487, 2.914]	1.172	2.153	.0313
$L1_{\text{German vs. Spanish}}:DPCW_{0\rightarrow1}$	−1.131	[−2.728, 1.185]	1.918	−0.589	.5556
$L1_{\text{NS vs. NNS}}:DPWC_{0\rightarrow1}$	−1.338	[−1.973, −0.557]	0.827	−1.618	.1056
$L1_{\text{German vs. Spanish}}:DPWC_{0\rightarrow1}$	2.142	[0.865, 3.199]	1.425	1.503	.1328

Given that the model predicted COMPLEMENTIZER: *present*,

- positive *b* values indicate that the (combination of the) listed level of a categorical/ordinal predictor and/or a one-unit increase for a numeric predictor reflect increased (log) odds/predicted probabilities of COMPLEMENTIZER being *present*; whereas negative *b* values, obviously, reflect decreasing probabilities of COMPLEMENTIZER being *present*;
- for the predictor for which we created planned contrasts (L1), the coefficient indicates the difference between the conditions that we encoded in our contrasts.

For instance, the coefficient *b* of TYPE indicates that *subject*, as opposed to *object*, complementation increases the occurrence of *that*. In addition to the above, I also like to provide the LRT results for deletion of each predictor that, according to the principle of marginality, is droppable, that is, what in R would be returned as the result of `drop1`. This is for two reasons: First, because the summary table does not contain significance results for any predictor with more than one degree of freedom, that is, all categorical predictors and maybe the interactions in which they are participating. For example, from the above one cannot infer the one *p* value for the interaction L1:DPWC. Second, because the output of `drop1` contains only the droppable predictors, it focuses one's attention nicely, I find, on the variables to interpret: If the model contains a significant interaction A:B, then `drop1` will only return the *p* value of that interaction A:B and not also those of the main effects A and B, which should typically not be taken at face value because the whole point of the interaction A:B is that the effect of A/B is not constant across B/A respectively. Thus, for Model `m.final`, I would also want to report the results shown in Table 4.

However, the complexity of model outputs such as Table 3—especially once nonlinear effects or higher-order interactions are involved and/or when authors do not provide reference or predicted levels or explicit contrasts—means that I personally always want to see visualizations of the relevant predictors.

Visualization With Effects Plots and Interpretation

To me, the by far best representation of fixed-effects results for interpretation involves effects plots (see Fox, 2003; Fox & Weisberg, 2019), which plot predicted values (of the response variable in linear MEMs, or of logits or, after the ilogit transformation, of predicted probabilities of the second level of the response variable). I prefer these over what is often provided, namely, observed means or slopes, because (a) if one fits a model and wants to summarize

Table 4 Likelihood ratio test (LRT) results for all droppable predictors in Model `m.final`

Variable	No. of parameters	AIC	LRT	*p*
L_CompSubj.bcn	1	4566.7	88.714	<.0001
Register	1	4525.8	47.878	<.0001
L_MatrS2V.fac	1	4515.6	37.615	<.0001
L_ComplClPC	1	4504.3	26.341	<.0001
L_MatrSubj.ord	3	4500.7	26.749	<.0001
L_MatrBe4S.fac	1	4496.6	18.664	<.0001
Surprisal.wind.bcn	1	4494.9	16.974	<.0001
Type	1	4484.1	6.184	.0129
L1:DPWC	2	4482.0	6.013	.0495
L1:DPCW	2	4481.3	5.382	.0678

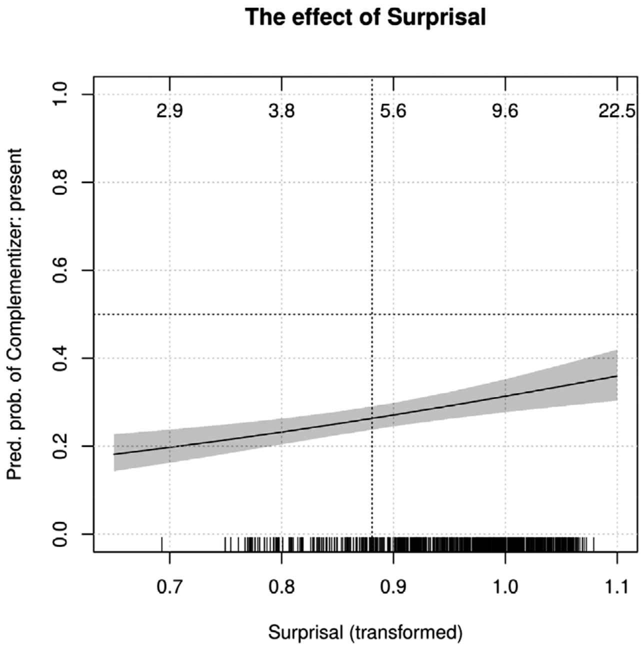

Figure 5 Plot of the effect of SURPRISAL.WIND in Model `m.final`.

that model, then one should visualize the output of the model and not an observed effect (also known as what one could visualize without even having a model) and, maybe even more important, because (b) the plots implemented in the `effects` package represent the effect of a predictor—one variable or an interaction—while holding all other predictors at typical values. Holding predictors at typical values means—by default at least—the mean of numeric predictors and, very nicely, proportional distributions of the levels of all categorical predictors (rather than just the most frequent level). This means that, especially for unbalanced and/or observational data, other effects are controlled for when the currently relevant effect is represented/discussed, which makes for a more accurate representation of an effect than mere observed means/slopes.[15]

I begin with the first main effect of interest, that is, SURPRISAL.WIND.BCN (Hypothesis 3), which is shown in Figure 5. The predictor is on the *x*-axis (attested values are indicated with rugs, its observed mean indicated with the dotted vertical line), the predicted probability of *that* being produced from Model `m.final` is on the *y*-axis (with the prediction cut-off point indicated

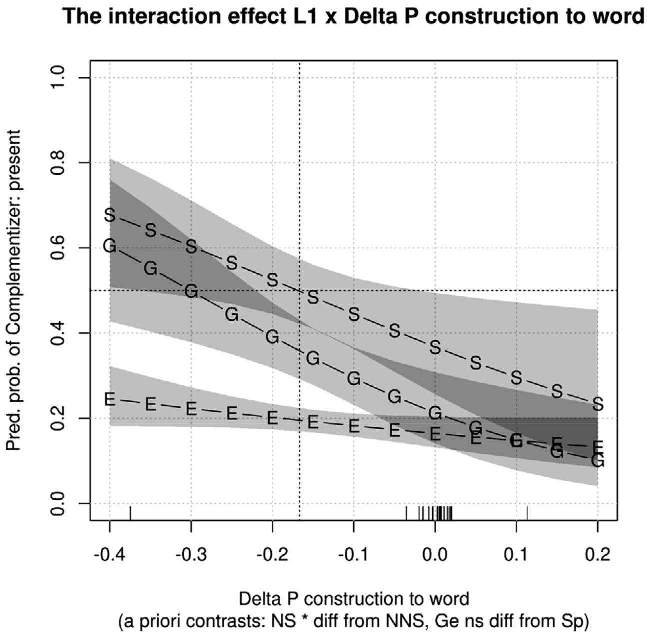

Figure 6 Plot of the effect of the interaction L1:DPCW in Model `m.final`. (The letters *E*, *G*, and *S* represent the L1s of the speakers.)

by the horizontal dotted line at $y = 0.5$), and the plot showing the regression line of the predictor with its 95% confidence band. This plot is based only on the FES because that is the result that one would apply to new speakers/files.

The result is relatively straightforward and compatible with previous work. Speakers are more likely to produce *that*, the more the first word of the complement clause is surprising given the last word of the matrix clause. Thus, with regard to the hypothesis, the result is mixed; surprisal has the expected effect, but there is no difference between native and NNS (because SURPRISAL.WIND does not interact with L1).

The next two effects are the two interactions of L1 with the ΔP values, that is, Hypotheses 1 (for DPCW) and 2 (for DPWC). In both Figure 6 and Figure 7, the ΔP predictors are on the x-axes, predicted probabilities are on the y-axes, and each regression line with its 95% confidence band represents one L1 background. As for the interaction L1:DPCW in Figure 6, the result is unexpected. For NS, there is a weak and only just about significantly different from 0 ($p = .038$) downward trend between DPCW and complementizer

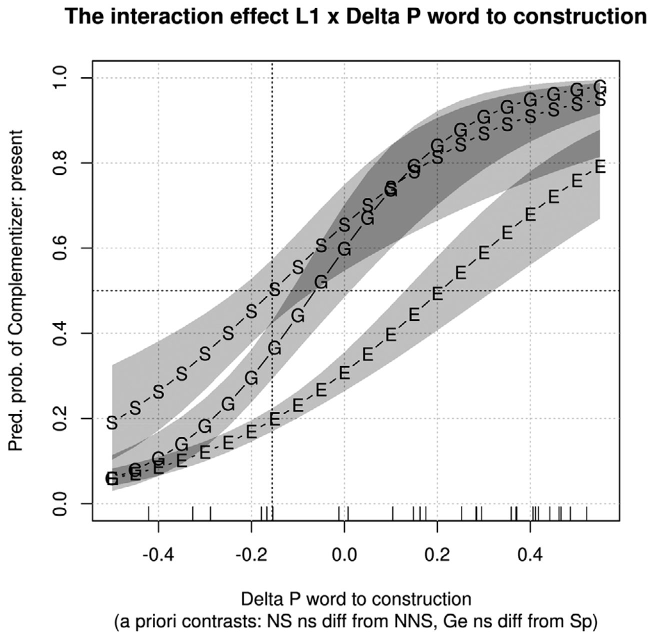

Figure 7 Plot of the effect of the interaction L1:DPWC in Model `m.final`.

realization, but the learners differ significantly from the NS ($p = .0313$): For the learners, there is an unexpected negative correlation such that if the construction COMPLEMENTIZER: *present* (i.e., use of *that*) "likes" the verb more (i.e., going from lower to higher *x*-axis values), the learners are predicted to use *that* less(!), with German and Spanish learners not differing from each other ($p = .5556$). In the supplementary material, I discuss possible reasons for, and ways to follow up on, this surprising effect.

The corresponding plot for the interaction L1:DPWC (Figure 7), on the other hand, exhibits the hypothesized effect direction. All three speaker groups exhibit an effect such that when DPWC increases, that is, when the verb is more and more attracted to COMPLEMENTIZER: *present*, then they are predicted to use *that* more, and this effect does not differ much either between NS and NNS or between the German and the Spanish learners.

As for the controls, which were not the focus here, they all behave as previous literature would lead one to expect. The more linguistic material (as measured in characters) that is found in the various slots for which we coded, that is, the higher the amount of processing load that we might associate with planning and producing the sentence, the more likely *that* is used, possibly because, given the speaker's planning time (while producing a complementizer that is easy to retrieve and articulate), it helps mark the syntactic structure of the sentence for the hearer (see the supplementary material for the corresponding graphs—it is obvious that, had there been any interesting deviations from the expectations, those could have been discussed.).

Random Effects

Many studies do not explore or use the ranef results in any way, which is a bit of shame. Not only can this be relevant for assessing model quality, the RES can also be instructive in its own right; patterns in the ranefs can be insightful. Miglio, Gries, Harris, Wheeler, and Santana-Paixão (2013), for instance, found in post hoc RES exploration that sizes and directions of varying intercepts for speakers were correlated with their dialects. It can therefore be useful to pay attention to the RES—either to explore it in a post hoc/bottom-up way or to correlate ranefs with other variables or just to visualize them to understand the spread that comes with the fixed-effects results.

In what follows, I discuss how much intercept and slope adjustments affect behavior of one predictor—DPWC—by computing and visualizing individual speakers' regression lines for that interaction of the predictor with L1 for some of the files that contributed the most data points (see Gries & Adelman, 2014, p. 49, for an early application; Murakami, 2016, in a language learning context; Meteyard & Davies, 2020, or Verbeke, Molenberghs, Fieuws, & Iddi, 2018, p. 18). To that end, I created a data frame whose rows contained all possible combinations of predictors (with frequencies approximating their frequency in the data as a crude (!) approximation of the above-mentioned logic of effects plots) for those speakers and their L1s and computed the predictions of Model `m.final` for these rows, which were then averaged over all other predictors (which result in the weighting-by-frequency/proportion) and plotted into a coordinate system like that of Figure 7. The result is shown in Figure 8 (the supplementary material uses color to distinguish L1 backgrounds).

This is of course just a small example but should nonetheless serve to highlight the potential of this exploration. The variability of this interaction is considerable and might, in a study focusing more on individual variation (see Gries & Wulff, in press), give rise to follow-up analyses. For example, one could use bottom-up exploratory statistics to see whether groups of speakers can be established or try to correlate intercept/slope adjustments with metadata regarding the speakers (much like one might do with influence measures).

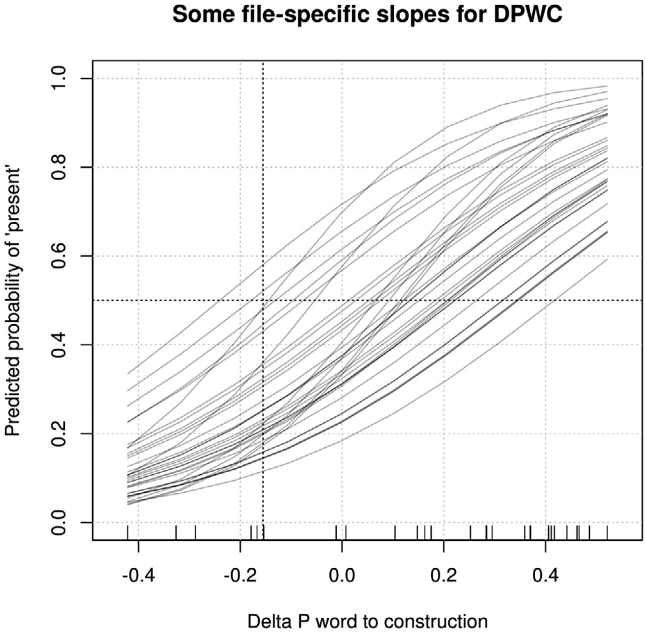

Figure 8 Plot of the effect of the interaction L1:DPWC in Model `m.final`.

Discussion and Concluding Remarks

Interim Summary

In sum, the main findings are that:

- the overall model was significant with a decent degree of explanatory power and classification accuracy and a good *C*-score; in that model, all control variables behaved as expected;
- one of the main new predictors—the one testing Hypothesis 1, DPCW—interacted significantly at the adopted significance level with L1 but behaved unexpectedly (especially for the two learner groups) indicating a mismatch between the preference of the construction and what speakers actually do;
- another new predictor—the one testing Hypothesis 2, DPWC—interacted significantly at the adopted significance level with L1 and did so in the hypothesized direction. There was a match between the constructional preference of the main-clause verb and what speakers actually do;
- the final main new predictor—the one testing Hypothesis 3, SURPRISAL.WIND.BCN—was significant at the adopted significance level and behaved as hypothesized—but for all three L1 groups alike.

However, the model comes with two bits of concern as well. Convergence issues and the fact that one may not want to fully trust the main effect of SURPRISAL.WIND.BCN because its significance was based on a tricky trimming-the-RES decision (the step from Model `m.03` to Model `m.04`).

Where to Go From Here

MEMs are here to stay, even if practices and standards are still evolving. New suggestions, techniques, and packages emerge at a high speed and in ways that make it really quite difficult to keep track of things. On the other hand, this rapid pace of development, testing, and implementation has raised the bar for what is possible and for what is getting published in ways that revolutionize the field (see Speelman, Heylen, & Geeraerts, 2018, for a recent collection of MEM papers in linguistics).

However, there are also limitations. One is that, compared to other kinds of predictive modeling, regression approaches, in general, do not necessarily have the highest degree of predictive power but excel at helping users of these approaches to interpret the results (whereas other, more black boxy, approaches such as support vector machines or random forests often predict better but can be harder to interpret [see, e.g., Breiman, 2001; Kuhn & Johnson, 2013, p. 50]). A more important limitation is how the assumptions, or data requirements, of MEM in particular can make it very difficult to conduct the analysis in which one is theoretically most interested, and that kind of limitation is exactly why data exploration and preparation accounts for such an exhaustingly long part of analyses and, here, the supplementary material. In particular with unbalanced data (e.g., in observational studies where one might also not be able to simply increase the sample size), there may be too low and/or uneven cell counts for predictors, controls, and their interactions and/or for random effects, and one may have to strike a delicate balance between creativity and resourcefulness, between what one wanted to do and what can be realistically done with one's data. For instance, in this analysis, the variable L_MATRV2CC was supposed to be entered as a control variable and was annotated accordingly, but then it had to be discarded given its extremely low variability.

In more extreme cases, alternative methods may have to be used. In linguistics, it seems as if tree-based approaches and particularly random forests are an alternative that many people find appealing (see, e.g., Hundt, 2018; Rezaee & Golparvar, 2017; Tagliamonte & Baayen, 2012, for applications in linguistics and Strobl, Malley, & Tutz, 2009, for an excellent overview). I think random forests can be a good alternative in particular for data whose distributions make MEM unlikely to succeed. However, as a prediction method, the goals of random forests are not necessarily the same as those of regression, and their variable importance values do not translate straightforwardly into what significance values or effect sizes are in regressions (see Efron, 2020). Also, I sometimes have the impression that trees/forests seem attractive for their perceived simplicity because, for instance, widely used implementations (e.g., `randomForest`, Liaw & Wiener, 2002; `party`, Hothorn, Hornik, & Zeileis, 2006; or `partykit`, Hothorn & Zeileis, 2015) seem to do away with everything many users hate about regression—coefficients, standard errors, and diagnostics. However, just because one needs only a one-liner to get variable importance values from a random forest does not mean there are not many similarly complex decisions to be made (e.g., tweaking hyperparameters, what algorithm to use for computing importance scores, how to compute partial dependencies, and even the difference between detecting and capturing interactions; see Efron, 2020; Gries 2020, for detailed discussion and exemplification). Thus, there are alternatives to MEM, but they are no magic wands. They also come with different goals/characteristics, and they definitely come with their own challenges that are only slowly being discussed in linguistics.

In terms of the continuing development of MEM practices, I think that the two most important current trends about which researchers should try to remain informed because of the potential impact on the field are (a) the increasing push toward including curvature in one's models with generalized additive mixed models, an extremely powerful but then also extremely complicated technique (see the above references, but also Baayen, van Rij, de Cat, & Wood, 2018, or Baayen, Vasishth, Kliegl, & Bates, 2016) and (b) the frequentist-versus-Bayesian discussion (Norouzian, de Miranda, & Plonsky, 2018) reinforced by the availability of powerful packages, especially such as `brms` (Bürkner, 2017, 2018; see especially or McElreath, 2020; Vasishth, Nicenboim, Beckman, Li, & Kong, 2018). That being said, probably just about any aspect of MEM is still undergoing lively discussion and active development:

- convergence warnings (when are they serious and not a false positive, when to flag them in `lmer's` output, and how to deal with them);
- model/variable selection and the role different information criteria play in it;
- validation of models with complex RES; and finally,
- power analysis for MEMs with complex RES (e.g., see Brysbaert & Stevens, 2018, as well as Meteyard & Davies, 2020, for recent discussions and the R packages of Donohue, 2020; Martin, 2020; Reich, Myers, Obeng, Milstone, & Perl, 2012; and Scheipl, Greven, & Kuechenhoff, 2008).

It is probably fair to say, though, that the increased degree of technicality and sophistication will not make access to all the fast-paced developments easy (I myself often have that feeling), but the degree to which MEMs and new developments can empower our statistically based findings is certainly worth every bit of effort we can invest.

Final revised version accepted 14 January 2021

Notes

1 Such a MEM is not the same as replacing the model lm(Y∼1+X) of Figure 1 with a model like lm(Y∼1+X*SPEAKER). Although the results can be similar, they are conceptually different in two important ways. First, the MEM does something that a fixed-effects-only model does not do: It employs something called *shrinkage*, which amounts to reducing estimates for levels of random effects like, here, SPEAKER, in a way that reflects their variance or, as Baayen (2008, p. 277) puts it, "considering the behavior of any given subject in the light of what it knows about the behavior of all the other subjects." For instance, if there are only few observations for a certain speaker, then the coefficient of a linear model for such a speaker would not be filtered or adjusted in any way due to the relative paucity of data, but in a MEM such a speaker's potentially extreme adjustments get shrunk towards the overall intercept or slope, which makes the results more robust (see Bell, Fairbrother, & Jones, 2019, Section 3, for a brief but excellent discussion and see Gelman & Hill, 2006, Sections 12.1–12.5, for a discussion of the way of partial pooling of MEM differs from the complete-pooling of a single fixed-effects model on all data or the no pooling of separate fixed-effects models for each level of a random effect). Second, this also means that an analyst needs to decide when to enter a variable into a model as a fixed effect or as a random effect, a question I will discuss in the section The Initial Model of the Methods section.

2 R^2 values are tricky beasts. Their meaning and computation are fairly uncontroversial in the case of linear models where an unadjusted R^2 is the proportion of variability of the response variable that is explained by all predictors. Although there are R^2 values for generalized linear models—with Nagelkerke's R^2 values being the default, as far as I can tell—these do not have the same function/interpretation and are therefore often called pseudo-R^2s (see discussion in Harrell, 2015, Section 10.8). For mixed-effects models, the situation is even more complex, given that multiple sources of variability can be involved in computing predictions. The R^2 value just reported here is the one that I see most often and that is implemented in at least two R packages, namely, a so-called $R^2_{\text{conditional}}$ of Nakagawa, Johnson, and Schielzeth (2017), which will also be used for the main analysis reported below and defined there.

3 For a more detailed discussion of these examples, see Gries (in press-b; Section 6.1) where I have the space to also discuss the consequences of using the better-suited MEMs in terms of regression/residual diagnostics.

4 The values had to be computed from all the data (as opposed to just from the NS data) to have association measures even for those verbs that were not used by NS. Given the frequency distribution of the L1s, the NS data influenced the association measures most strongly.

5 These are not, of course, all variables that one could consider. For example, I did not include any (self-) priming effects (Gries, 2016, 2019; Jaeger & Snider, 2008) and no time/counter variables to operationalize something like fatigue, habituation, and so on (see Doğruöz & Gries, 2012; Gries, 2019; Scheepers, 2003, again, or, much more sophisticated, Baayen, Vasishth, Kliegl, & Bates, 2016). Also, I did not have any proficiency scores for the learners who produced the essays, and I did not approximate them here with some text-based measures (as in Gries & Wulff, in press).

6 The term ecdf plots refers to *empirical cumulative distribution function* plots, which are extremely useful plots for seeing at one glance many central aspects of the distribution of a numeric variable by plotting for every observed value o of a numeric variable v the percentage of data points in v that are $\leq o$. The advantage of these plots over the better-known ones such as histograms and boxplots is that ecdf plots do not bin any data and, thus, provide a more fine-grained resolution of v (see Gries, in press-b, Section 3.1.3.3).

7 A distribution is Zipfian if a large number of types account for very few tokens, and a few types account for the majority of tokens. The standard example is one of word frequencies; often half of all word types in a corpus occur only once, but the 10 most frequent words types account for a third of all tokens.

8 The precision in the description of the RES should be noted. Many manuscripts using MEMs that I have reviewed say something like "SPEAKER was included as a random effect," which is much too imprecise. Were there intercept adjustments/random intercepts for SPEAKER, slope adjustments/random slopes but no random intercepts for SPEAKER, both kinds of adjustments, and, if there were both, were those correlated?

9 Part of the friction surrounding this kind of discussion might be attributable to how statistics and, for instance, machine learning differ in their emphases on attribution/interpretation and prediction, respectively (see Breiman, 2001, and the comments on the paper and Efron, 2020, for discussions that I think anyone modeling data in any way should read and that I wish I had been familiar with much earlier).

10 In the case of linear MEM, selecting the RES needs to be done with restricted maximum likelihood estimation (the default setting of the function lmer), whereas selecting the best FES needs to be done with maximum likelihood estimation

(which can be set with an argument to the lmer function called REML = FALSE; see Zuur, Ieno, Walker, Saveliev, & Smith, 2009, Section 5.6).

11 I still often see manuscripts where authors attempt to check for multicollinearity by checking pairwise correlations between predictor variables. Frankly, this is mostly useless. High pairwise correlations are a sufficient condition for multicollinearity, not a necessary one, because pairwise correlations are by definition unable to identify more complex collinear structures arising from multiple columns in the model matrix.

12 Model diagnostics above and beyond those related to convergence problems are often only discussed rather briefly even in the otherwise best existing overview articles (e.g., Bolker et al. 2009; Brauer & Curtin, 2018; Meteyard & Davies, 2020), but they can be extremely important for identifying problems in a model but also for determining how to address such problems (see Gries, in press-b, Section 6.3, for a worked example where diagnostics revealed the need for curvature for three control variables).

13 In addition, one needs to bear in mind that, if one were to do a 10-fold cross-validation, a potential model selection process would also have to be done 10 times, which, because I personally would never do a fully automatic model selection process, would increase demands on time and computational power considerably. If one considers only a single model, cross-validation will of course be more feasible computationally.

14 In cases involving polynomials or splines (for curvature) or ordinal predictors (like L_MatrSubj.ord here), of course not all estimates will be straightforwardly interpretable, but even then, all other coefficients would be more meaningful.

15 In the rare case of a perfectly balanced design with completely orthogonal predictors, etc., the predicted results will reduce to the observed, so plotting predicted results will generate what I think is virtually always the right kind of plot.

Open Research Badges

This article has earned Open Data and Open Materials badges for making publicly available the digitally-shareable data and the components of the research methods needed to reproduce the reported procedure and results. All data and materials that the authors have used and have the right to share are available at https://osf.io/fxujc/?view_only=eda3922a71844f49becd7c6220a0dadd and https://www.iris-database.org/. All proprietary materials have been precisely identified in the manuscript.

References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

Baayen, R. H., Davidson, D.J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2016). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234. https://doi.org/10.1016/j.jml.2016.11.006

Baayen, R. H., van Rij, J., de Cat, C., & Wood, S. (2018). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by generalized additive mixed models. In D. Speelman, K. Heylen, & D. Geeraerts (Eds.), 2018. *Mixed-effects regression models in linguistics* (pp. 48–69). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-69830-4_4

Baguley, T. (2012). *Serious stats. Houndmills*, Basingstoke, UK: Palgrave Macmillan.

Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*, 457–474. https://doi.org/10.1016/j.jml.2007.09.002

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2018). Parsimonouis mixed models. Manuscript submitted for publication. Retrieved from https://arxiv.org/abs/1506.04967

Bell, A., Fairbrother, M., & Jones, K. (2019). Fixed and random effects models: Making an informed choice. *Quality and Quantity*, *53*, 1051–1074. https://doi.org/10.1007/s11135-018-0802-x

Bolker, B. M. (2014, October 21). Convergence warnings in lme4 [Electronic mailing list message]. Retrieved from https://stat.ethz.ch/pipermail/r-sig-mixed-models/2014q4/022813.html

Bolker, B. M. (2019, May 29). Optimism introduced by non-converging models in bootstrap validation of GLMM (lme4) [Electronic mailing list message]. Retrieved from https://stat.ethz.ch/pipermail/r-sig-mixed-models/2019q2/027864.html

Bolker, B. M., Brooks, M. E., Clark, C. J., Grange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389–411. https://doi.org/10.1037/met0000159

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1). 9, 1–20. https://doi.org/10.5334/joc.10

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

Crawley, M. J. (2013). *The R book* (2nd ed.). Chichester, UK: John Wiley.

Doğruöz, A. S., & Gries, St. Th. (2012). Spread of on-going changes in an immigrant language: Turkish in the Netherlands. *Review of Cognitive Linguistics*, *10*(2), 401–426. https://doi.org/10.1075/rcl.10.2.07sez

Donohue, M. C. (2020). longpower: Power and sample size calculations for linear mixed models (R package version 1.0-21) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/longpower

Efron, B. (2020). Prediction, estimation, and attribution. *Journal of the American Statistical Association*, *115*(530), 636–655. https://doi.org/10.1080/01621459.2020.1762613

Ellis, N. C. (2007). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*(1), 1–27. https://doi.org/10.1093/applin/ami038

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software*, *8*(15), 1–27. https://doi.org/10.18637/jss.v008.i15

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Thousand Oaks, CA: Sage.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Gilquin, G., De Cock, S., & Granger, S. (Eds.). (2010). *Louvain international database of spoken English interlanguage [Handbook and CD-ROM]*. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (Eds.) (2009). *International corpus of learner English (Version 2)* [Handbook and CD-ROM]. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Gries, St. Th. (2013). 50-something years of work on collocations: What is or should be next … *International Journal of Corpus Linguistics*, *18*(1), 137–165. https://doi.org/10.1075/ijcl.18.1.09gri

Gries, St. Th. (2016). Variationist analysis: Variability due to random effects and autocorrelation. In P. Baker & J. A. Egbert (Eds.), *Triangulating methodological approaches in corpus linguistic research* (pp. 108–123). New York, NY: Routledge.

Gries, St. Th. (2019). Priming of syntactic alternations by learners of English: An analysis of sentence-completion and collostructional results. In J. A. Egbert & P. Baker (Eds.), *Using corpus methods to triangulate linguistic analysis* (pp. 219–238). New York, NY: Routledge.

Gries, St. Th. (2020). On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory*, *16*(3), 617–647. [R report]

Gries, St. Th. (in press-b).*Statistics for linguistics with R* (3rd ed.). Boston, MA: De Gruyter Mouton.

Gries, St. Th., & Adelman, A. S. (2014). Subject realization in Japanese conversation by native and non-native speakers: Exemplifying a new paradigm for learner corpus research. In J. Romero-Trillo, (Ed.), *Yearbook of corpus linguistics and pragmatics 2014: New empirical and theoretical paradigms* (pp. 35–54). Cham, Switzerland: Springer.

Gries, St. Th., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on ‘alternations’. *International Journal of Corpus Linguistics*, *9*(1), 97–129. https://doi.org/10.1075/ijcl.9.1.06gri

Gries, St. Th., & Wulff, S. (in press). Adverbial clause ordering in learner production data. *Applied Psycholinguistics*, https://doi.org/10.1017/S014271642000048X

Harrell, F. Jr. (2015). *Regression modeling strategies* (2nd ed.). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-19425-7

Hartig, F. (2020). DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models. (R package version 0.3.1) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/DHARMa/index.html

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection - A review and recommendations for the practicing statistician. *Biometrical Journal*, *60*, 431–449. https://doi.org/10.1002/bimj.201700067

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. https://doi.org/10.1198/106186006x133933

Hothorn, T., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partitioning in R. *Journal of Machine Learning Research*, *16*, 3905–3909.

Hundt, M. (2018). It is time that this (should) be studied across a broader range of Englishes: A global trip around mandative subjunctives. In S. C. Deshors (Ed.), *Modeling World Englishes: Assessing the interplay of emancipation and globalization of ESL varieties* (pp. 217–244). Amsterdam, The Netherlands: John Benjamins.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62. https://doi.org/10.1016/2Fj.cogpsych.2010.02.002

Jaeger, T. F., & Snider, N. E. (2008). Implicit learning and syntactic persistence: Surprisal and cumulativity. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the Cognitive Science Society* (pp. 1061–1066). Austin, TX: Cognitive Science Society.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Cham, Switzerland: Springer. https://doi.org/10.1007/978-1-4614-6849-3

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Lester, N. (2018). *The syntactic bits of nouns: How prior syntactic distributions affect comprehension* (Unpublished doctoral dissertation). Santa Barbara, CA: University of California. Retrieved from https://escholarship.org/uc/item/25r9w1t1

Linzen, T., & Jaeger, T. F. (2014). Investigating the role of entropy in sentence processing. In V. Bemberg & T. O'Donnell (Eds.), *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics* (pp. 10–18). Baltimore, MD: Association for computational Linguistics.

Martin, J. (2020). pamm: Power analysis for random effects in mixed models (R package version 1.121) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/pamm

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, R. H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). Boca Raton, FL: CRC Press.

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*(104092), 1–22. https://doi.org/10.1016/j.jml.2020.104092

Miglio, V. G., Gries, St. Th., Harris, M. J., Wheeler, E. M., & Santana-Paixão, R. (2013). Spanish lo(s)-le(s) clitic alternations in psych verbs: A multifactorial corpus-based analysis. In J. Cabrelli Amaro, G. Lord, A. de Prada Pérez, & J. E. Aaron (Eds.), *Selected proceedings of the 15th Hispanic Linguistics Symposium* (pp. 268–278). Somerville, MA: Cascadilla Press.

Murakami, A. (2016). Modeling systematicity and individuality in nonlinear second language development: The case of English grammatical morphemes. *Language Learning*, *66*, 834–871. https://doi.org/10.1111/lang.12166

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*(134), 1–11. https://doi.org/10.1098/rsif.2017.0213

Nelson, G., Wallis, S., & Aarts, B. (2002). *Exploring natural language: Working with the British component of the International Corpus of English*. Amsterdam, The Netherlands: John Benjamins.

Norouzian, R., de Miranda, M. A., & Plonsky, L. (2018). The Bayesian revolution in second language research: An applied approach. *Language Learning*, *68*, 1032–1075. https://doi.org/10.1111/lang.12310

Norris, J. M., Plonsky, L., Ross, S. J., & Schoonend, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning*, *65*, 470–476. https://doi.org/10.1111/lang.12104

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*, 413–425. https://doi.org/10.1016/j.jml.2008.02.002

Rezaee, A. A., & Golparvar, S. E. (2017). Conditional inference tree modelling of competing motivators of the positioning of concessive clauses: The case of a non-native corpus. *Journal of Quantitative Linguistics*, *24*(2–3), 89–106. https://doi.org/10.1080/09296174.2016.1265799

Reich, N. G., Myers, J. A., Obeng, D., Milstone, A. M., & Perl, T. M. (2012). Empirical power and sample size calculations for cluster-randomized and cluster-randomized crossover studies. *PLoS One*, *7*, e35564. https://doi.org/10.1371/journal.pone.0035564

Säfken, B., Rügamer, D., Kneib, T., & Greven, S. (2018). Conditional model selection in mixed-effects models with cAIC4. Manuscript submitted for publication. Retrieved from https://arxiv.org/abs/1803.05664

Scheepers, C. (2003). Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition*, *89*(3), 179–205. https://doi.org/10.1016/S0010-0277(03)00119-7

Scheipl, F., Greven, S., & Kuechenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis*, *52*, 3283–3299. https://doi.org/10.1016/j.csda.2007.10.022

Sestelo, M., Villanueva, N. M., Meira-Machado, L., & Roca-Pardiñas, J. (2016). FWDselect: An R package for variable selection in regression models. *The R Journal*, *8*(1), 132–148. https://doi.org/10.32614/RJ-2016-009

Speelman, D., Heylen, K., & Geeraerts, D. (Eds.). (2018). *Mixed-effects regression models in linguistics*. Cham, Switzerland: Springer.

Stefanowitsch, A., & Gries, St. Th. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, *8*(2), 209–243. https://doi.org/10.1075/ijcl.8.2.03ste

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/2Fa0016973

Tagliamonte, S. A., & Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language*

Variation and Change, *24*(2), 135–178. https://doi.org/10.1017/S0954394512000129

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, *55*, 525–534. https://doi.org/10.1177/2F0013164495055004001

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *The Journal of Experimental Education*, *70*, 80–93. https://doi.org/10.1080/00220970109599499

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 147–161. https://doi.org/10.1016/j.wocn.2018.07.008

Verbeke, G., Molenberghs, G., Fieuws, S., & Iddi, S. (2018). Mixed models with emphasis on large data sets. In D. Speelman, K. Heylen, & D. Geeraerts (Eds.), *Mixed-effects regression models in linguistics* (pp. 11–28). Cham, Switzerland: Springer.

Voeten, C. C. (2020). buildmer: Stepwise elimination and term reordering for mixed-effects regression (R package version 1.5) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/buildmer

Wulff, S., Gries, St. Th., & Lester, N. (2018). Optional that in complementation by German and Spanish learners. In A. Tyler, L. Huan, & H. Jan (Eds.), *What is applied cognitive linguistics? Answers from current SLA research* (pp. 99–120). Berlin, Germany: De Gruyter Mouton.

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Springer.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. may be found at https://osf.io/xmkpt/?view_only=0d2fb75051e8449c87dd30eded9683ab

Appendix: Accessible Summary (also publicly available at https://oasis-database.org)

Introduction to (Generalized Linear) Mixed-Effects Modeling

What This Research Was About and Why It Is Important

This article is an introduction and exemplification of (generalized) linear mixed-effects modeling to researchers in linguistics and second language acquisition, an important method that allows researchers to extend the range of ‘‘regular’’ regression modeling to data that involve repeated measurements or other kinds of hierarchical structure. The data used to exemplify this method are concerned with whether native and two kinds of nonnative speakers of English use the complementizer *that* in sentences such as *I hope* that/Ø *he is hungry*. Approximately 6,200 sentences were annotated for a variety of characteristics and then analyzed with a mixed-effects model to study in particular the role that verb-specific preferences and speaker/hearer expectations play in realizing or omitting the complementizer. Results show that the main-clause verb (*hope* in the above example) and the degree to which the complement-clause subject (*he* in the above example) is surprising are significantly correlated with complementizer realization.

What the Researcher Did

- The researcher described the general logic of mixed-effects models and its implementation in R.
- The researcher extracted examples of complementation from native and nonnative speaker corpora.
- The researcher annotated the corpora for about a dozen linguistic/contextual characteristics of the sentences with complementation and the circumstances in which *that* was produced.
- The researcher then used a generalized linear mixed-effects model to analyze the data.

What the Researcher Found

- Complementizer realization can be predicted fairly well.
- The control variables known from previous literature had the expected kinds of effects.
- The main-clause verb's preference for (tendency to be used with) a complementizer had the expected effect: Verbs that "like" the complementizer were positively associated with it being used; this effect varied mildly across the L1s of the speakers.
- The degree to which the beginning of the complement clause is predictable to the hearer had the expected effect: Surprising beginnings of the clause were positively associated with the use of the complementizer.

Things to Consider

- (Generalized) linear mixed-effects modeling is a powerful but complex statistical method for linguistic and second language acquisition research.

- Great care is needed both for the preparatory exploration of the data before any regression modeling and also in deciding how exactly the regression modeling process is undertaken.
- Complementizer realization is determined by a variety of processing-related/psycholinguistic factors, some of which differ in their effects between native speakers of English and learners of English with different L1s.

Materials, data, open access article: Materials and data are publicly available at: https://osf.io/xmkpt/?view_only=0d2fb75051e8449c87dd30eded9683ab.
How to cite this summary: Gries, St., Th. (2021). Introduction to (generalized linear) mixed-effects modeling. *OASIS Summary* of Gries (2021) in *Language Learning*. https://oasis-database.org