The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models

Stefan Th. Gries[1]

Abstract

Much statistical analysis of psycholinguistic data is now being done with so-called mixed-effects regression models. This development was spearheaded by a few highly influential introductory articles that (*i*) showed how these regression models are superior to what was the previous gold standard and, perhaps even more importantly, (*ii*) showed how these models are used practically. Corpus linguistics can benefit from mixed-effects/multi-level models for the same reason that psycholinguistics can – because, for example, speaker-specific and lexically specific idiosyncrasies can be accounted for elegantly; but, in fact, corpus linguistics needs them even more because (*i*) corpus-linguistic data are observational and, thus, usually unbalanced and messy/noisy, and (*ii*) most widely used corpora come with a hierarchical structure that corpus linguists routinely fail to consider. Unlike nearly all overviews of mixed-effects/multi-level modelling, this paper is specifically written for corpus linguists to get more of them to start using these techniques more. After a short methodological history, I provide a non-technical introduction to mixed-effects models and then discuss in detail one example – particle placement in English – to show how mixed-effects/multi-level modelling results can be obtained and how they are far superior to those of traditional regression modelling.

Keywords: corpus structure, lexically specific effects multi-level modelling, speaker-specific effects, varying slopes and intercepts, verb-particle construction

[1] Department of Linguistics, University of California, Santa Barbara, CA 93106-3100, USA.
Correspondence to: Stefan Gries, *e-mail*: stgries@linguistics.ucsb.edu

Corpora 2015 Vol. 10 (1): 95–125
DOI: 10.3366/cor.2015.0068

www.euppublishing.com/journal/cor

1. Introduction and motivation

1.1 A bit of methodological history

By their very nature, corpus-linguistic studies have always been based on frequencies of occurrence of linguistic elements as well as frequencies of co-occurrence of linguistic elements with either other linguistic elements or the co(n)textual characteristics of these linguistic expressions. Essentially, this means that all corpus-linguistic analyses, wherever they are located on a purely hypothetical scale from purely qualitative to purely quantitative work, have been at least implicitly based on statements such as:

- $n_x = 0$ – that is, some *x* does not occur in a (part of a) corpus;
- $n_x > 0$ – that is, some *x* occurs in a (part of a) corpus either as the frequently used category of a hapax (i.e., $n_x = 1$) or more than once (i.e., $n_x > 1$);
- $n_x > n_y$ or $n_x = n$ or $n_x < n_y$ – that is, some *x* occurs more/less/equally often than/as some *y* in a (part of a) corpus;
- any of the above pertaining to particular co(n)texts such as the presence of some other linguistic expression *z*, some discourse-contextual feature *z*, . . .

Especially over the last ten years or so, corpus linguists have begun to take this (in some sense obvious) fact into consideration and have followed the general development in linguistics towards more and more sophisticated quantitative methods. One might have expected corpus linguistics to spearhead this development because, following the above logic, corpus linguistics is inherently quantitative. However, it is other disciplines that have been more explicitly driving this change in the methodological landscape. One of these ‘other disciplines’ is psycholinguistics, in which, for several decades, very many statistical analyses were analyses of variance (ANOVAs) of experiments involving fully factorial designs with repeated measures – that is to say, designs in which:

- the number of predictors was usually small (typically, two or three);
- the predictors were all categorical by nature or they were continuous/numeric by nature but factorised/discretised by the researcher (e.g., when frequencies were dichotomised into low *versus* high); and,
- the predictors were completely crossed so that each subject in the experiment saw each combination of levels of the predictors multiple times.

So, one might consider as prototypical an experiment with two predictors A (with levels A_1 and A_2) and B (with levels B_1 and B_2), which give rise to four kinds of experimental stimuli (A_1B_1, A_1B_2, A_2B_1 and A_2B_2), and every subject sees each of these combinations repeatedly and equally many times. This design would then be analysed with repeated-measures ANOVAs, one on averages for the different subjects, one on averages for the different experimental stimuli, so that the fact that all the data points for each subject/item are related to one another is accommodated.

Psycholinguistics is currently undergoing a major change in statistical methodology, one in which the above ANOVA paradigm (a not uncontroversial method in any case; see Baayen, 2008; and Jaeger, 2008) is now being superseded by (generalised) linear mixed-effects modelling or (G)LMM). These models offer a variety of very attractive features: first, and most straightforwardly, they easily integrate numeric predictors (e.g., lengths and frequencies) without factorisation. Second, they are better at handling unbalanced designs, (i.e., designs in which not all experimental situations are equally frequent).

Third and most importantly, they provide ‘the usual’ type of results for all predictors of interest (in the context of (G)LMM, these are referred to as ‘fixed effects’), but they can also address the fact that data points are related because they were provided by the same subject or for the same item. This means that they can increase the precision of the regression results considerably by offering sophisticated ways to partial out, or accommodate, for instance, speaker- or item-specific effects (or other effects extraneous to the specific research question; see Johnson, 2009: 363, for the same argument in the domain of sociolinguistics). In other words, statistical analyses become more precise because, among other things, the idiosyncrasies of particular speakers, particular stimuli, *etc*., do not distort the regression coefficients of the fixed effects, but are ‘relegated to’ the variance captured by what are called ‘random effects’, (i.e., adjustments to regression coefficients that accommodate the way in which each subject/item may be consistently different from the others). It is useful to reiterate for later that, given the prototypical factorial design, these speaker- and item-specific random effects are often fully ‘crossed’ such that each subject sees all stimuli and all stimuli are seen by each speaker.

Given these advantages, these kinds of models are now rapidly becoming mainstream in the psycholinguistic literature. Returning to corpus linguistics, it is instructive to compare its methodological challenges and trajectories to those of psycholinguistics. Two observations are particularly pertinent in this connection. On the one hand, corpus-linguistic observational data are typically much messier and unbalanced than psycholinguistic experimental data because many confounding and moderator variables that psycholinguists can control for (by randomising, blocking, *etc*.) plague corpus-linguistic analyses. On the other hand, while that means that corpus-linguistic statistics would stand to benefit immensely from more advanced statistics in general, and (G)LMM in particular, they are in fact still further behind on a cline of statistical sophistication. Indeed, many practitioners are only beginning to use monofactorial statistics, and fewer have yet made the move towards regression modelling – ‘linear modelling’ for

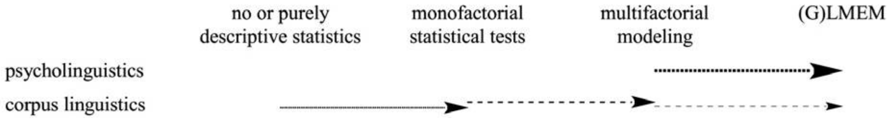

no or purely descriptive statistics
monofactorial statistical tests
multifactorial modeling
(G)LMEM
psycholinguistics
corpus linguistics

Figure 1: Evolution of statistical methods in psycholinguistics and corpus linguistics

numeric dependent variables and 'generalised linear modelling' for ordinal and categorical (binary or polytomous) dependent variables – that would correspond to the previous ANOVA state-of-the-art in psycholinguistics, and only very few people are (already) using (G)LMM. This comparison can be shown as in Figure 1, in which heaviness and solidity of lines represents frequency. Figure 1 is, obviously, a simplification: no goal-directed/-orientated evolution is suggested, nor is it implied that (G)LMEM is the endpoint of evolution (see Section 4 for discussion).

1.2 This paper and its objectives

The general goal of this paper is to help to increase the number of corpus linguists who recognise the problems of the approaches on the left and, thus, decide to move towards the approaches on the right. This kind of recognition in psycholinguistics was hurried along considerably through several publications that not only summarised the advantages of (G)LMM – most notably Baayen (2008), and papers in special issue of JML such as Baayen *et al*. (2008), Quené and van den Bergh (2008), and in particular Jaeger (2008) – but also provided readers with relatively concrete instructions on how to perform such analyses themselves; in a sense, the above publications became *the* go-to articles for practitioners.

However, in spite of the indubitable quality of the above publications and the mark they have left on psycholinguistics, there are two reasons why I think corpus linguistics could still potentially benefit considerably from this paper. First, most of these papers are psycholinguistic in nature and involve experimental data of the kind prevalent in psycholinguistic analyses, which could, understandably, make it more difficult to a core corpus linguist to translate their messages into 'his or her language'.

Second, this also means that these papers – as well as, it seems, most other publications on (G)LMM – only focus on statistical designs in which the random effects are crossed (along the lines discussed above). However, a second domain in which (G)LMM is extremely useful is one that characterises the vast majority of corpus-linguistic studies and that is routinely ignored (and that pertains to most of my own earlier work, too): random effects can be not just crossed but also 'nested' across multiple levels (hence 'multi-level analysis'). For example, in most corpora, speakers/writers are nested into files, which are nested into registers, which are nested into modes. That is, corpora do not usually feature data from each speaker in each register in each mode – rather, a particular speaker was recorded, which was transcribed into one file and one file only, which represents one register and one register only, which represents one mode and one mode only. For instance, the sentence in Example 1 was produced by one speaker (labelled *A*) in one file (S1B-045) representing one sub-register ('spoken public dialog') representing one register ('spoken dialog') representing one mode ('dialog') in the British component of the International Corpus of English (ICE-GB):

(*1*) And all this happened really before you started picking up a camera and becoming a movie maker

All of these characteristics from each of these different levels are hierarchically nested into each other – it is not that, later in this corpus, the same file name is also used for a file with private dialogue, unscripted monologue, printed creative writing, *etc*. This has an extremely important consequence: both in psycholinguistics and in corpus linguistics, the vast majority of datasets violate the assumptions of the simplest test that an analyst might normally think of first – namely, that the data points are independent of one another. In psycholinguistics, data points are dependent on one another because subjects provide multiple responses and items are tested more than once (and these effects are crossed). However, in corpus linguistics, this is even worse: as in psycholinguistics, speakers/writers often provide more than one data point, and we have more than one instance of, say, a constructional choice per verb, but we also have the hierarchically nested/multi-level structure that the corpus comes in: perhaps the speaker (or as a convenient heuristic, the file) is not even the right level of resolution for the current phenomenon – perhaps most of the variability must be explained by looking at registers? And perhaps the frequent distinction between modes – speaking *versus* writing – is actually not relevant for phenomenon *X* (for which we need to look at sub-registers)? In other words, at present, many corpus-linguistic studies make do with a simple chi-squared test or even a more advanced regression, when in fact the data analysed violate the assumption of the independence of data points that these tests routinely come with by ignoring:

(*i*) speaker-/writer-/file-specific effects and lexically specific effects; and,

(*ii*) the multiple levels of structure that corpus data typically come in.

As a consequence, the results that the majority of corpus-linguistic studies report are likely to be very anti-conservative (i.e., too likely to return a significant result) and imprecise (because the results are tainted to an unknown degree by idiosyncrasies from which one can, and should not, generalise) and, just to acknowledge that quite openly, this also applies potentially to several earlier studies of mine. What is needed is an approach that combines the logic of mixed-effects modelling (to deal with the variation resulting from (*i*) and the logic of multi-level modelling (to deal with the variation resulting from (*ii*). Thus, the hopefully not too high-flying goal of this paper is to become for corpus linguistics what the above-quoted papers have become for psycholinguistics: a first go-to resource that explains to corpus linguists what (generalised) linear mixed-effects/multi-level modelling ((G)LMM/MLM) has to offer and that provides them with a concrete example and some instructions on how to perform such analyses. While, given the complexity of these modelling approaches, the coverage can not, of course, be exhaustive, I hope that the attractiveness of the methods to be exemplified below will help make them more common in the discipline and, thus, benefit us all in terms of the greater precision and reliability of the results.

Towards the above objectives, Section 2 provides a brief but necessary introduction to (G)LMM on the basis of very small constructed datasets just to highlight the general logic of the method. More linguistically then, Section 3 discusses a linguistic phenomenon of particle placement, the constituent order alternation exemplified in Example 2, with CONSTRUCTION being a binary response with two levels:

(*2*) a. John picked up the book. CONSTRUCTION: V-Part-DO
b. John picked the book up. CONSTRUCTION: V-DO-Part

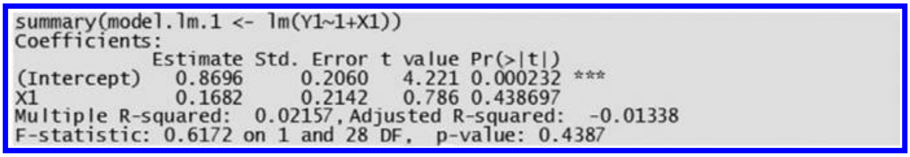
Particle placement has been very thoroughly researched in terms of its fixed effects so it is well-known which determinants have which, and how strong an, effect on the choice of construction in both mono- and multi-factorial analyses (see Gries, 2003; and Szmrecsanyi, 2005, 2006). At the same time, this alternation has also been shown to be subject to exactly the kinds of effects that (G)LMM/MLM deals with much better than traditional multi-factorial regression methods – namely, lexically specific effects as well as sub-register-specific effects (see Diessel and Tomasello, 2005; and Gries, 2006). For these reasons, this alternation is the perfect test bed to showcase (G)LMM/MLM and its advantages: we know what the results should be like so we can focus on the new contributions that the method will make. For many reasons, in this paper I will use the open-source programming language and environment *R* (3.1.1 patched, R Core Team, 2014) for the statistical analyses as well as the packages lme4 (Version 1.1–7, Bates *et al*., 2014) and rms (Version 4.2–0, Harrell, 2014). *R* code and results will be shown `like this in a non-proportional font`; note that the output will often be abbreviated. The paper presupposes no knowledge of these modelling methods but some textbook-level familiarity with basic regression modelling and/or *R* (see, for example, Gries, 2013: Chapter 5) is useful.

2. (G)LMM: a very brief introduction by example

In this section, I will introduce the logic and application of (G)LMM on the basis of several small datasets. Three things are important to bear in mind. First, as will become obvious, the data are simulated and have been designed to have particular immediately obvious characteristics. Second, I am using a linear regression model for this initial explanation. Both of these choices have been made for didactic/expository reasons. I am using simulated data because these data have characteristics that, when plotted, illustrate clearly the patterns that (G)LMEM are particularly good at detecting, and this is something that authentic data would make much more difficult. Second, I am not using a logistic/multinomial model, (i.e., a model with a categorical response), for this example even though such models are more typical of corpus-linguistic applications. This is because these latter kinds of models involve additional conceptual steps (e.g., the use of link functions) that make the exposition less transparent than desirable. Third, the examples in this section are clearly not representative of the real-life applications and challenges of (G)LMM because, as will be seen presently, my examples involve only three speakers (to facilitate plotting, *etc.*) whereas real-life applications will be more complex. However, the example in Section 3 is a fully fledged one that mirrors what corpus linguists would actually study much more accurately and I recommend Gelman and Hill (2006: Sections 11.5 and 12.7) for detailed discussion of when (G)LMMs are most effective.

2.1 Introduction: the traditional approach

Imagine a dataset in which you try to model a numeric response y_1 as a function of a numeric predictor x_1. You have thirty data points – ten from each of three speakers (labelled just 1, 2 and 3). Given the current state of the art in corpus linguistics, it is reasonable to expect this to be analysed with a linear regression and summarised as shown in Figure 2 (in which the plotted numbers represent the data points from the speakers); the regression equation in the first line is to be read as ‘use a linear model (lm) to fit the response y_1 (Y1) as a function of (~) of an intercept (1) and (+) a predictor x_1 (X1).[2]



```
summary(model.lm.1 <- lm(Y1~1+X1))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8696     0.2060   4.221 0.000232 ***
X1            0.1682     0.2142   0.786 0.438697
Multiple R-squared:  0.02157,Adjusted R-squared:  -0.01338
F-statistic: 0.6172 on 1 and 28 DF,  p-value: 0.4387
```

[2] Mentioning the intercept in this regression equation, as I do here with 1+, is redundant, strictly speaking; I only do this here for maximal explicitness and expository clarity.

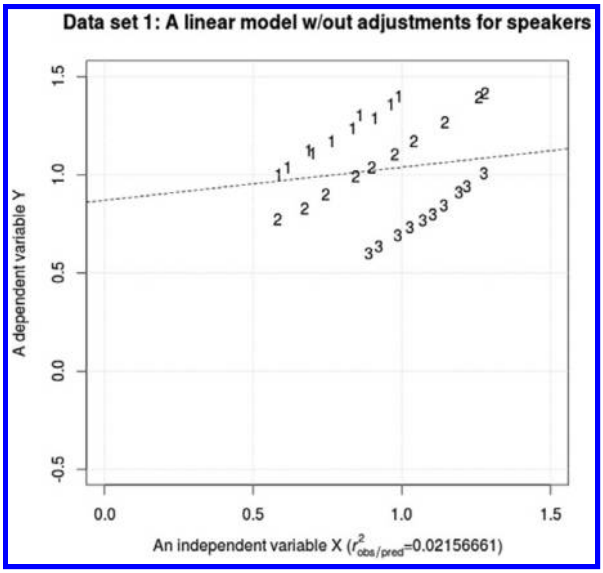


Figure 2: Results of a linear regression on Dataset 1

The regression results and the regression line indicate that there is a positive correlation (for every one-unit increase of x_1, y_1 increases by 0.1682), but the correlation between x_1 and y_1 is not significant (multiple $R^2 = 0.022$, $p = 0.4387$). To any human reader, however, it seems obvious that there is a very strong and very similar correlation between x_1 and y_1, just at different levels of x_1, and this will be the topic of the next section.

2.2 Varying intercepts

In this case, even the simplest (G)LMM can help because, rather than being restricted to one intercept for all three speakers (0.8696 in the above linear model), it can let every speaker get his or her own intercept and then adjusts the slope to make best use of the different intercepts (1|SPEAKER1 means ‘fit a separate intercept (1) for each of the three different speakers included in the variable/vector SPEAKER1’):[3]

[3] One might wonder why SPEAKER1 is not included as a ‘regular’ fixed effect. The reason for this is that the goal of analysis is not to generalise to future results for the same three speakers but to results that would be obtained from different speakers from the population for which the three speakers of SPEAKER1 are (it is to be hoped) representative.

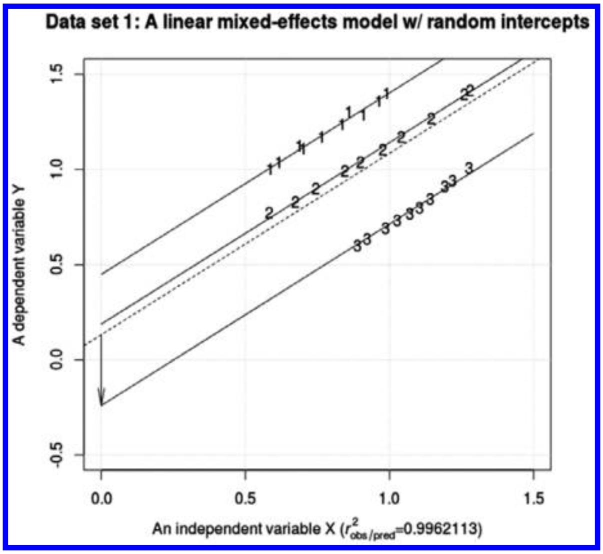
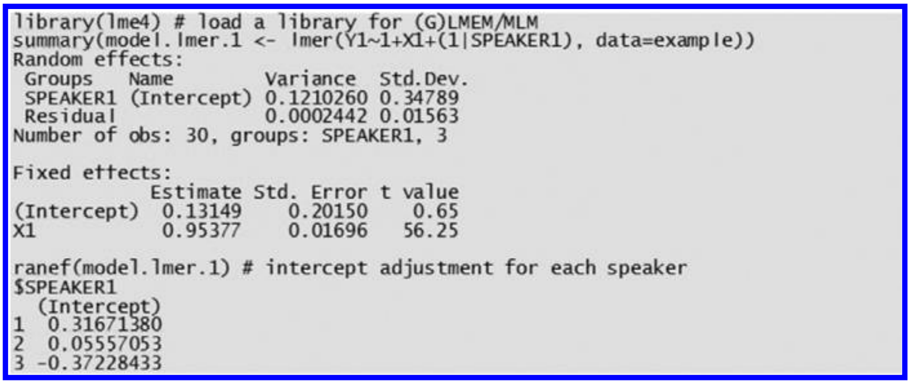


Figure 3: Results of a (G)LMM with varying intercepts on Dataset 1



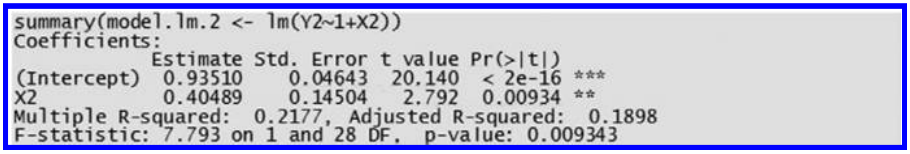
```
library(lme4) # load a library for (G)LMEM/MLM
summary(model.lmer.1 <- lmer(Y1~1+X1+(1|SPEAKER1), data=example))
Random effects:
 Groups   Name        Variance  Std.Dev.
 SPEAKER1 (Intercept) 0.1210260 0.34789
 Residual             0.0002442 0.01563
Number of obs: 30, groups: SPEAKER1, 3

Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.13149    0.20150    0.65
X1           0.95377    0.01696   56.25

ranef(model.lmer.1) # intercept adjustment for each speaker
$SPEAKER1
  (Intercept)
1  0.31671380
2  0.05557053
3 -0.37228433
```

Figure 3 shows these results visually. Specifically, the dotted line is the overall regression line (with the intercept of ≈0.13 and a slope of ≈0.95 shown in the output, above). The three solid lines are the regression lines representing how the overall intercept is adjusted for each speaker while keeping the same slope (0.95377) for all speakers. For example, the arrow shows how the intercept of the overall regression line is adjusted for speaker 3 by ≈−0.37 (i.e., downwards).

The difference in the results is striking in both the numeric output and the plot. Again, there is a positive correlation, but now for every one-unit increase of x_1, y_1 increases by more than 5.5 times as much as before (0.95377 instead of 0.1682), the effect of x_1 is highly significant ($t = 56.25$ and, not shown, $\chi^2 = 126.9$, $df = 1$, $p < 10^{-10}$) and, now that the different speakers are not forced to share the same intercept anymore, the model has a very high degree of explanatory power ($R^2_{\text{conditional}} = 0.998$). Thus, by accounting for the fact that the data points of each speaker are not independent and modelling them as speaker-specific results – giving them separate intercepts – a non-significant result of a model that anyway violated its assumptions suddenly turned into a highly significant result with nearly perfect predictive accuracy.

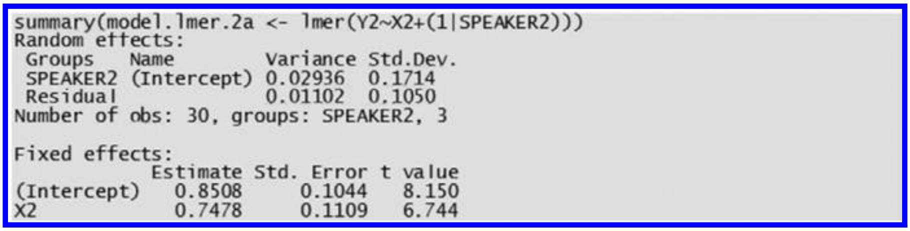
2.3 Varying slopes

Imagine now a dataset with the same structure (two variables, y_2 and x_2), three speakers and thirty data points. As Figure 4 indicates, the 'regular' linear regression model from the first block of code is significant ($p \approx 0.009$) but can again not explain the data well (multiple $R^2 = 0.2177$; see the left-hand panel).



```
summary(model.lm.2 <- lm(Y2~1+X2))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.93510    0.04643  20.140  < 2e-16 ***
X2            0.40489    0.14504   2.792  0.00934 **
Multiple R-squared:  0.2177, Adjusted R-squared:  0.1898
F-statistic: 7.793 on 1 and 28 DF,  p-value: 0.009343
```

The (G)LMM with varying intercepts in the following block of code does better ($R^2_{\text{marginal}} = 0.376$, $R^2_{\text{conditional}} = 0.83$) but is still not a really good model of the data (see the right-hand panel).[4]



```
summary(model.lmer.2a <- lmer(Y2~X2+(1|SPEAKER2)))
Random effects:
 Groups   Name        Variance Std.Dev.
 SPEAKER2 (Intercept) 0.02936  0.1714
 Residual             0.01102  0.1050
Number of obs: 30, groups: SPEAKER2, 3

Fixed effects:
            Estimate Std. Error t value
(Intercept)   0.8508     0.1044   8.150
X2            0.7478     0.1109   6.744
```

Given the plot, it is clear that this dataset has been constructed such that the three speakers have the same intercept (approximately 0.8) but what seem to be different slopes. Figure 5 exemplifies the corresponding (G)LMM analysis (in which (0+X2|SPEAKER2) means 'do not fit separate intercepts for each speaker, but separate slopes for X2 for each speaker') and also

[4] The R^2-values reported here summarise the predictive power of the fixed effects only (R^2_{marginal}) and of both fixed and random effects ($R^2_{\text{conditional}}$) and were computed following Nakagawa and Schielzeth (2013); see also Johnson (2014).

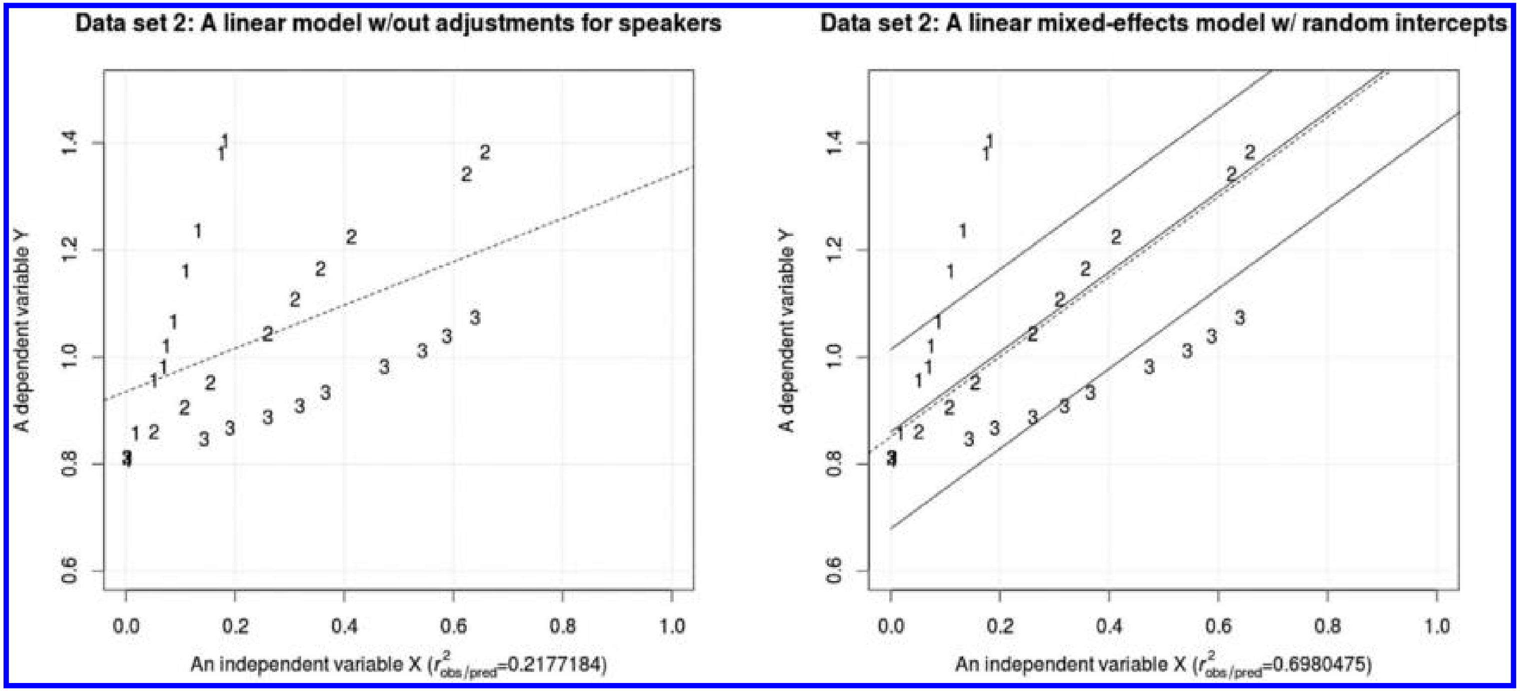


Figure 4: Analyses of Dataset 2: results of a linear regression (left-hand panel); results of a (G)LMM with varying intercepts (right-hand panel)

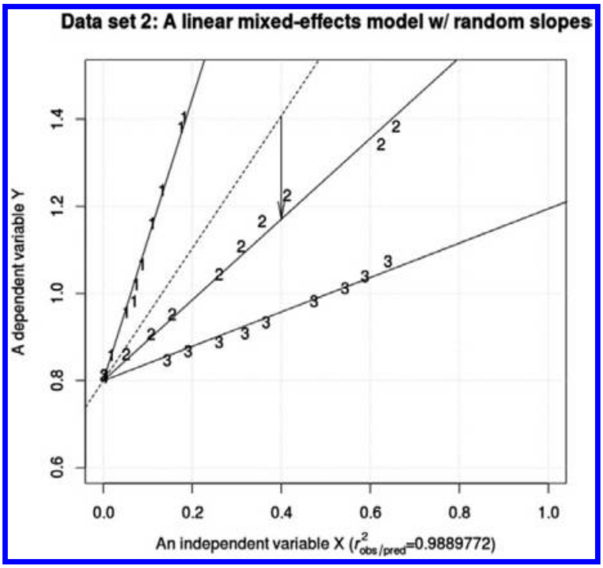
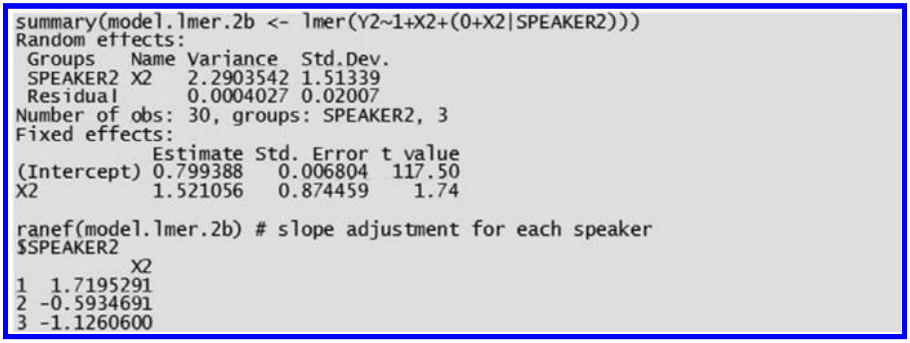


Figure 5: Analysis of Dataset 2: results of a (G)LMM with varying slopes

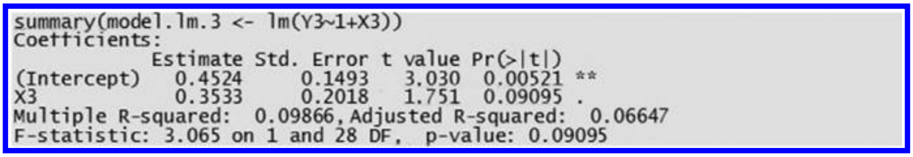
visualises these results. Again, the dotted line is the overall regression line (with the intercept of ≈0.8 and a slope of ≈1.52 shown in the output below). The three solid lines are the regression lines representing how the overall slope of the regression line is adjusted for each speaker. For example, the arrow shows how the overall regression line is adjusted for speaker 2 by ≈–0.59 (i.e., downwards).


```
summary(model.lmer.2b <- lmer(Y2~1+X2+(0+X2|SPEAKER2)))
Random effects:
 Groups   Name Variance  Std.Dev.
 SPEAKER2 X2   2.2903542 1.51339
 Residual      0.0004027 0.02007
Number of obs: 30, groups: SPEAKER2, 3
Fixed effects:
            Estimate Std. Error t value
(Intercept) 0.799388   0.006804  117.50
X2          1.521056   0.874459    1.74

ranef(model.lmer.2b) # slope adjustment for each speaker
$SPEAKER2
         X2
1  1.7195291
2 -0.5934691
3 -1.1260600
```

This result is interesting in how it differs from that of Dataset 1. There, it was the (G)LMM that showed that a result that seemed insignificant after a traditional linear regression was in fact significant once speaker idiosyncrasies were taken into account. Here, it is the other way round: while the (G)LMM with one joint intercept but a different slope for each speaker accounts for the data nearly perfectly ($R^2_{\text{conditional}} \approx 0.999$), the overall effect for x_2 (the slope of 1.521) is actually not significant anymore ($t = 1.74$ and, not shown, $\chi^2 = 2.767$, $df = 1$, $p = 0.096$) and has virtually no explanatory power ($R^2_{\text{marginal}} \approx 0.042$). Thus, (G)LMM can make results more precise in both directions: finding significant results where regular regression does not and finding that results are not significant even though a regular regression returns a significant result.

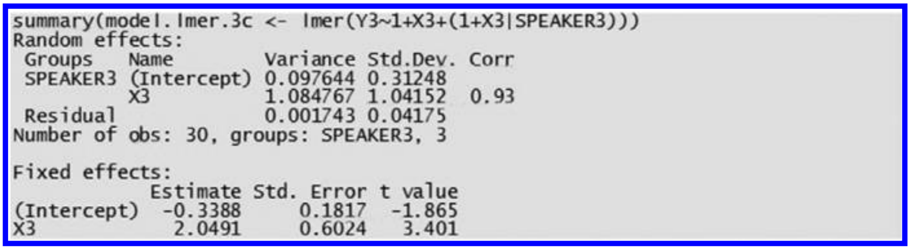
2.4 Varying intercepts and slopes

Imagine, finally, yet another dataset with, again, the same structure (two variables y_3 and x_3, three speakers and thirty data points). As before, a linear regression indicates a positive correlation (for every one-unit increase of x_3, y_3 increases by 0.3533) but the correlation between x_3 and y_3 is not significant (multiple $R^2 \approx 0.099$, $p \approx 0.091$) and, as before, it does not take much training to see in Figure 6 that the linear regression line does not account for the data well.



```
summary(model.lm.3 <- lm(Y3~1+X3))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4524     0.1493   3.030  0.00521 **
X3            0.3533     0.2018   1.751  0.09095 .
Multiple R-squared:  0.09866,Adjusted R-squared:  0.06647
F-statistic: 3.065 on 1 and 28 DF,  p-value: 0.09095
```

To save space, I do not provide graphs and results for (G)LMM analyses with varying intercepts and varying slopes and only wish to observe that while both these models already do much better than the 'regular' linear regression, they are, again, not convincing when one plots the resulting regression lines. Rather, what is needed here is a (G)LMM that has separate intercepts *and* separate slopes for each speaker (as determined by model comparison tests (not shown, $p_{\text{varying intercepts}} < 10^{-11}$ and $p_{\text{varying slopes}} < 10^{-15}$); in the code quoted below (1+X3|SPEAKER3) means 'fit separate intercepts (1) and (+) slopes of x_3 for each speaker (X3|SPEAKER3)'.



```
summary(model.lmer.3c <- lmer(Y3~1+X3+(1+X3|SPEAKER3)))
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 SPEAKER3 (Intercept) 0.097644 0.31248
          X3          1.084767 1.04152  0.93
 Residual             0.001743 0.04175
Number of obs: 30, groups: SPEAKER3, 3

Fixed effects:
            Estimate Std. Error t value
(Intercept)  -0.3388     0.1817  -1.865
X3            2.0491     0.6024   3.401
```

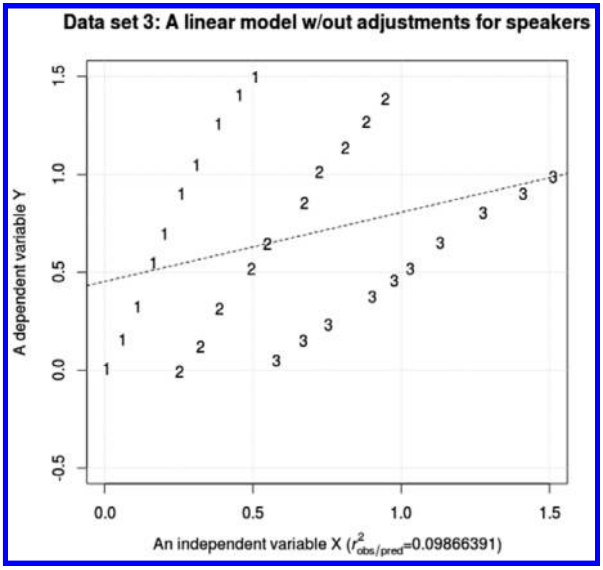


Figure 6: Results of a linear regression on Dataset 3

This model results in a virtually perfect fit ($R^2_{\text{marginal}} = 0.872$, $R^2_{\text{conditional}} = 0.998$), with an effect of x_3 on y_3 that is significant ($t = 3.401$ and, not shown, $\chi^2 = 56.203$, $df = 1$, $p < 10^{-13}$) but is again very different – namely, nearly six times as strong as the linear regression would make us believe (2.0491 rather than 0.3533). Figure 7 visualises the result in the same way as before; the left and right arrows show the adjustment to Speaker 3's intercept and slope.

To conclude, while this section could only scratch the surface and dealt, merely, with numeric dependent variables (y_1, y_2 and y_3) rather than the categorical dependent variables that are more frequent in corpus linguistics, I hope it has become clear how much more precision and reliability (G)LMM has to offer, not to mention the fact that the regular fixed-effects regression would really not even have been permitted in each case given the dependence of the data points. Again, they can protect both against statistical Type I and Type II errors and the better regression coefficients that result allow for better explanation of the phenomena under investigation. In the following section, I discuss the example of particle placement, which will involve a binary categorical dependent variable and add the multi-level perspective that corpus data routinely require. In spite of the didactic nature of this paper, given considerations of space, I cannot discuss all aspects of regression modelling in detail here; instead, I will provide some guidance on the general logic of the overall process and the multi-level perspective as well as help with regard to the interpretation of the results; the discussion here is modelled after Gries's (2013) characterisation of regression modelling.

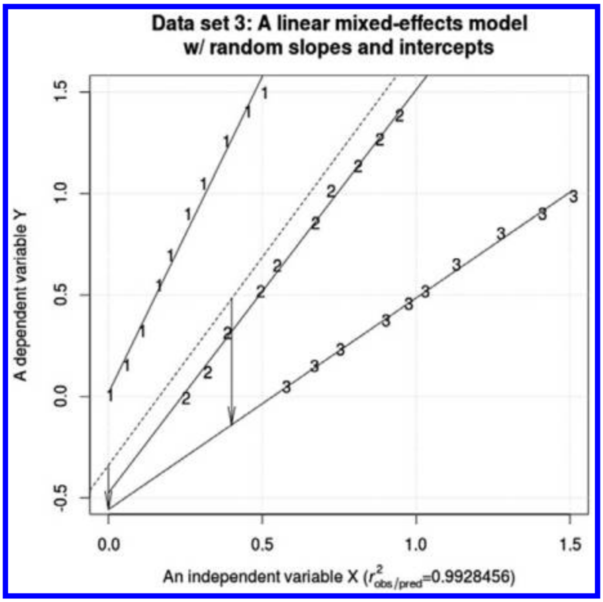


Figure 7: Analysis of Dataset 3: results of a (G)LMM with varying slopes *and* intercepts

3. A ME/MLM analysis of particle placement

3.1 The data

The data studied here are those from Gries (2006: Section 4): 1,192 instances of CONSTRUCTION: V-DO-Part and 1,129 instances of CONSTRUCTION: V-Part-DO from the ICE-GB. Since the main point of this paper is didactic, each of these instances was annotated for only two fixed effects:

- TYPE: the type of head of direct object: 'lexical' or 'non-lexical';
- LOGLENGTH: the logged length of the DO in words.[5]

More importantly for our purposes is the annotation of examples for random effects and hierarchical information. Thus, every instance was also annotated for the variables represented in Table 1. The left three variables reflect the hierarchically nested structure of the ICE-GB; the right two variables represent lexically specific effects.

Thus, Example 1 ('. . . before you started picking up a camera. . . '), would be annotated as in Table 2.

[5] We will ignore here the fact that these two variables are correlated. Regression modelling with residualised versions of these predictors had no impact on whether the main interaction was significant or not and the overall shape of the curve of predicted probabilities.

Nested (from right into left)			Crossed	
Mode	Register	Sub-register	Verb	Particle
Spoken	Dialogue	Private, public	*take*	*up*
	Monologue	Scripted, unscripted	*put*	*out*
	Mixed	Broadcast	*bring*	*in*
Written	Printed	Academic, creative, instructional non-academic, persuasive, reportage	*get pick*	*off down*
	Non-printed	Letters, non-professional	. . .	. . .

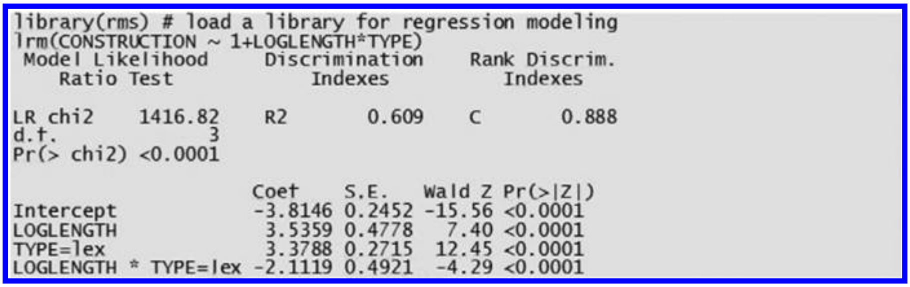
Table 1: Random effects annotated in the particle placement data

Construction	Type	LogLength	Mode	Register	Subregister	Verb	Particle
V-Part-DO	lex	0.6931472	spoken	dialogue	public	*pick*	*up*

Table 2: The annotation of Sentence 1

3.2 A ‘regular’ fixed-effects only logistic regression

If this dataset were to be analysed statistically, the analysis most likely to be found would be a binary logistic regression without random effects (BLR) that models, or tries to predict, the probability of the dependent variable Construction: V-Part-DO (the second level of Construction alphabetically) on the basis of three predictors: the numeric independent variable LogLength, the categorical independent variable Type, and the interaction of the two, which allows the regression to ‘consider’ the possibility that the effect of LogLength might not be the same for Type: lexical and Type: non-lexical, and this would be represented in *R* by a regression equation CONSTRUCTION ~ (‘as a function of’) LOGLENGTH*TYPE; these would be its results:



```
library(rms) # load a library for regression modeling
lrm(CONSTRUCTION ~ 1+LOGLENGTH*TYPE)
 Model Likelihood     Discrimination    Rank Discrim.
    Ratio Test           Indexes           Indexes

LR chi2    1416.82    R2       0.609    C       0.888
d.f.             3
Pr(> chi2) <0.0001

                      Coef    S.E.   Wald Z Pr(>|Z|)
Intercept             -3.8146 0.2452 -15.56 <0.0001
LOGLENGTH              3.5359 0.4778   7.40 <0.0001
TYPE=lex               3.3788 0.2715  12.45 <0.0001
LOGLENGTH * TYPE=lex  -2.1119 0.4921  -4.29 <0.0001
```

In a nutshell, the three predictors, the two fixed-effect independent variables and their interaction, are significantly correlated with the choice of construction ($p < 0.0001$); the correlation is strong (Nagelkerke $R^2 = 0.609$), and the classification accuracy is high ($C = 0.888$ and 79.02 percent of all constructional choices are predicted correctly), as can be seen from the code available from my website.[6] A few brief comments on what the coefficients mean: these coefficients reflect predicted probabilities of CONSTRUCTION: V-Part-DO (after the ilogit transformation):[7]

- the value for the intercept indicates the predicted probability of CONSTRUCTION: V-Part-DO when LOGLENGTH = 0 and TYPE: NON-LEXICAL: ilogit(–3.8146) = 0.0216;
- the value for LOGLENGTH, 3.5359, indicates how the predicted probability of CONSTRUCTION: V-Part-DO changes for every unit-increase of LOGLENGTH (when TYPE: NON-LEXICAL); since the sign is positive, that means, the longer the DO, the more likely CONSTRUCTION: V-Part-DO becomes;
- the value for TYPE: LEXICAL, 3.3788, indicates how the predicted probability of CONSTRUCTION: V-Part-DO changes when the DO type changes from NON-LEXICAL to LEXICAL (when LOGLENGTH: = 0); since the sign is positive, this means that if the DO is lexical, CONSTRUCTION: V-Part-DO becomes more likely;
- the value for the interaction, –2.1119, indicates that the above-mentioned effect that LOGLENGTH has when TYPE: NON-LEXICAL becomes weakened when TYPE: LEXICAL.

If researchers are already so advanced as to use multi-factorial modelling, then this is probably the most widely used kind of analysis of such data. However, it is at least incomplete, if not inappropriate, because it pretends that the 2,321 data points are all independent of one another, which we know they are not: they exhibit inter-relations because they were produced by fewer than 2,321 speakers, because of the lexical items in the verb-particle constructions, and because of the levels of corpus sampling. Thus, let us now turn to the better kind of analysis.

3.3 A (G)LMM/MLM analysis

While (G)LMM/MLM is currently a hot topic in linguistics, there are still many methodological questions that are hotly debated, answered differently in every new article one reads, or answered in some references and glossed over in others (see Gries, 2013: 335–). I, therefore, do not lay claim to

[6] See: http://tinyurl.com/stgries/research/overview-research.html

[7] The ilogit, or inverse logit, transformation transforms a value y from the range $-\infty \leq y \leq \infty$ to the range of a probability scale $0 \leq \text{ilogit}(y) \leq 1$: $1/(1+exp^{-y})$.

presenting the universally accepted best possible analysis of the dataset. Rather, I will follow a stepwise model selection procedure outlined in Zuur *et al*. (2009: Chapter 5), which can be summarised as follows:

(*i*) begin with a model that contains the most comprehensive fixed-effects structure that can be fit given the variables to be explored and find the optimal random-effects structure (varying intercepts for one or more predictors and/or varying slopes for one or more predictors); and,

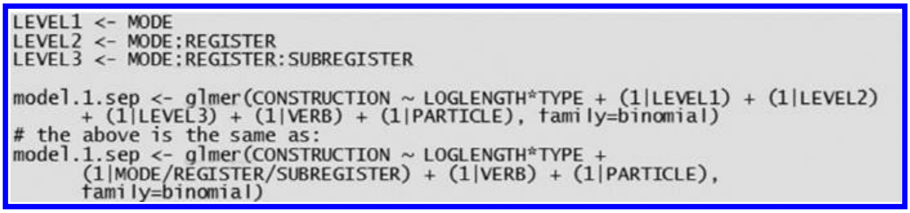
(*ii*) once the optimal random-effects structure has been found, find the optimal fixed-effects structure.

In both these steps, 'optimal' means according to some criterion such as significance testing/*p*-values or information criteria. With *p*-values this would mean that the final model, *m*, contains (*i*) only random effects that make *m* significantly better than if these were not in *m* and (*ii*) only fixed effects predictors – again, independent variables and their interactions – that make *m* significantly better than if these were not in *m* or that are required for higher-level interactions.

3.3.1 Step I: finding the optimal random-effects structure

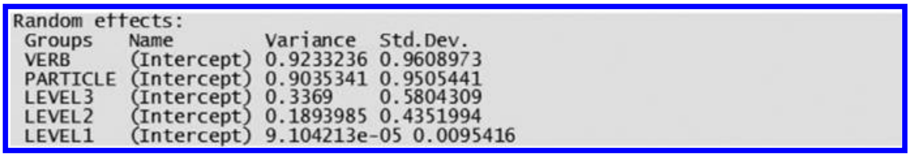
As a first step, we define the three new variables that most explicitly reflect the nested structure of the data: LEVEL1, LEVEL2 and LEVEL3. Then we fit a first model that, as above, contains all fixed effects – LOGLENGTH, TYPE, and their interaction LOGLENGTH:TYPE – as well as varying intercepts for each level of corpus sampling – (1|LEVEL1), (1|LEVEL2), (1|LEVEL3) – and for each verb (1|VERB) and each particle (1|PARTICLE). Given the logic outlined in Sections 2.2 and 2.4, that means that the baseline probability of CONSTRUCTION: V-Part-DO can be different for each of these random effects; in other words, the model allows every corpus part (at each level of corpus organisation), every verb and every particle to have a different baseline 'preference' for CONSTRUCTION: V-Part-DO.[8]

[8] As indicated, there is currently a debate concerning what the maximal random-effects structure is with which one should begin a model selection process. One currently influential paper is Barr *et al*. (2013), who recommend that the most comprehensive random-effects structure possible be used (i.e., random intercepts and slopes for all predictors, independent variables and their interactions). Since this would blow up the model equations beyond what can be reasonably dealt with in an introductory paper, I kept matters 'simple' and did not include varying slopes in any of the random effects specifications. It is also for the sake of simplicity that I have not mentioned issues such as how to deal with collinearity of predictors, non-significant predictors one nevertheless wishes to retain in the model, or how *p*-values should be adjusted when testing on the boundary, *etc*.; all these things are discussed in references that I recommend in the final section of this paper.



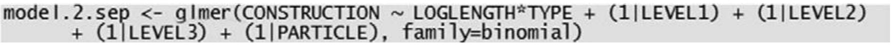
```
LEVEL1 <- MODE
LEVEL2 <- MODE:REGISTER
LEVEL3 <- MODE:REGISTER:SUBREGISTER

model.1.sep <- glmer(CONSTRUCTION ~ LOGLENGTH*TYPE + (1|LEVEL1) + (1|LEVEL2)
      + (1|LEVEL3) + (1|VERB) + (1|PARTICLE), family=binomial)
# the above is the same as:
model.1.sep <- glmer(CONSTRUCTION ~ LOGLENGTH*TYPE +
      (1|MODE/REGISTER/SUBREGISTER) + (1|VERB) + (1|PARTICLE),
      family=binomial)
```

The summary of the random effects looks like this: the larger the values for the variances/standard deviations, the more variability is located 'within' that random effect; this suggests that, in this case, the (crossed) random effects of lexical items contribute more than the (nested) random effects of the corpus organisation.



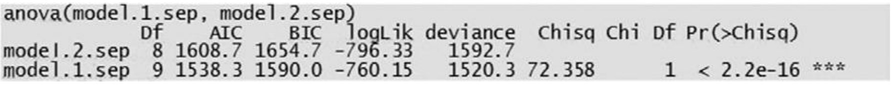
```
Random effects:
 Groups   Name        Variance     Std.Dev.
 VERB     (Intercept) 0.9233236    0.9608973
 PARTICLE (Intercept) 0.9035341    0.9505441
 LEVEL3   (Intercept) 0.3369       0.5804309
 LEVEL2   (Intercept) 0.1893985    0.4351994
 LEVEL1   (Intercept) 9.104213e-05 0.0095416
```

To trim down the random-effects structure, let us first test whether the random effect with the largest amount of variance can be omitted: if that one did not make a significant contribution, this would suggest that the other random effects would not either. To that end, we first fit a new, smaller model that is just like model.1.sep but does not contain (1|VERB).



```
model.2.sep <- glmer(CONSTRUCTION ~ LOGLENGTH*TYPE + (1|LEVEL1) + (1|LEVEL2)
      + (1|LEVEL3) + (1|PARTICLE), family=binomial)
```

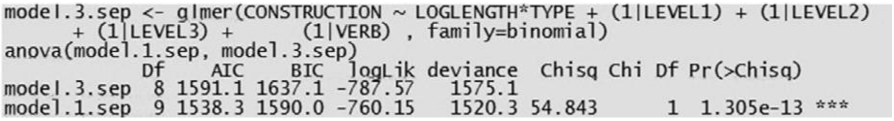
After that, we compare that new model *without* the random effect (1|VERB), model.2.sep, to the old model *with* that random effect, model.1.sep, with the anova function:[9]



```
anova(model.1.sep, model.2.sep)
            Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
model.2.sep  8 1608.7 1654.7 -796.33   1592.7
model.1.sep  9 1538.3 1590.0 -760.15   1520.3 72.358      1  < 2.2e-16 ***
```

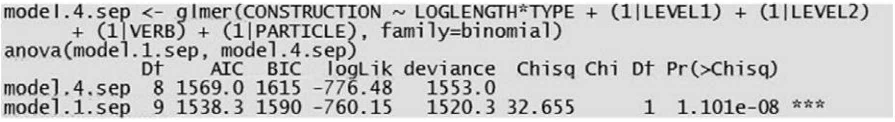
[9] Just like most other aspects of (G)LMEM, this use of anova to determine whether random effects can/should be included in a model (following Zuur *et al*., 2009) is not uncontroversial. For example, Gelman and Hill (2006: 271, original emphasis) state that it is '*not* appropriate to use statistical significance as a criterion for including particular group indicators.' However, just like Zuur *et al*., Baayen (2008) uses anova to determine which of two models with identical fixed effects but different random effects is better (e.g., in Sections 7.1 and 7.5.4); the same strategy is followed by West *et al*. (2007), Larson-Hall (2010: 263), Gałecki and Burzykowski (2013: 298), and it is used in the mixed-effects modelling book by Bates (2010), the main creator of the *R* package lme4.

The difference between the models is significant, which means (1|VERB) needs to stay in our model or, put differently, we need to stick with model.1.sep and cannot use the simpler alternative of model.2.sep because distinguishing different baseline constructional preferences for verbs is significantly necessary (and also supported by the smaller *AIC*-value of the first model; cf. Gries, 2013: 261). What about the random effect with the second largest variance, (1|PARTICLE)? A similar model comparison shows that (1|PARTICLE) is also required:



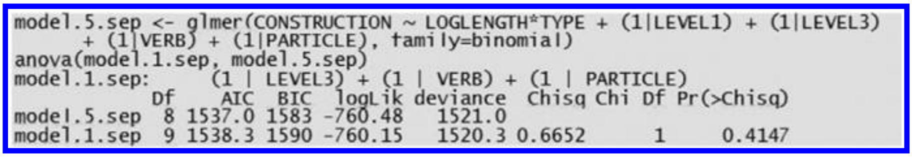
```
model.3.sep <- glmer(CONSTRUCTION ~ LOGLENGTH*TYPE + (1|LEVEL1) + (1|LEVEL2)
      + (1|LEVEL3) +      (1|VERB) , family=binomial)
anova(model.1.sep, model.3.sep)
            Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
model.3.sep  8 1591.1 1637.1 -787.57    1575.1
model.1.sep  9 1538.3 1590.0 -760.15    1520.3 54.843      1  1.305e-13 ***
```

The random effect with the next smaller variance is LEVEL3, i.e., SUBREGISTER. A model comparison analogous to the one above shows that this, too, needs to stay in the model: distinguishing different constructional preferences for sub-registers is significantly necessary:



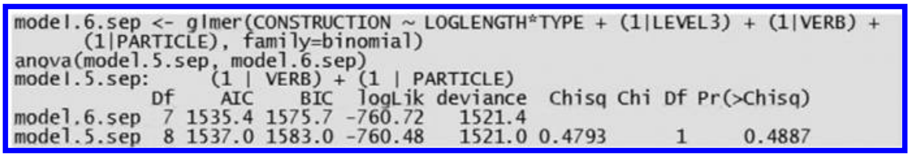
```
model.4.sep <- glmer(CONSTRUCTION ~ LOGLENGTH*TYPE + (1|LEVEL1) + (1|LEVEL2)
      + (1|VERB) + (1|PARTICLE), family=binomial)
anova(model.1.sep, model.4.sep)
            Dt    AIC  BIC  logLik deviance  Chisq Chi Dt Pr(>Chisq)
model.4.sep  8 1569.0 1615 -776.48    1553.0
model.1.sep  9 1538.3 1590 -760.15    1520.3 32.655      1  1.101e-08 ***
```

Now what about LEVEL2, (i.e., REGISTER)? Here, for the first time, we see we need to abandon our first model, model.1.sep, because model.5.sep, the model in which we do not care about registers, is *not* significantly worse than the model in which we do (and probably because the sub-registers already account for most of the variability registers would account for, though this need not always be the case). Thus, model.5.sep becomes our new reference model, which is again also supported by the fact that the *AIC*-value of model.5.sep, is smaller than that of model.1.sep).



```
model.5.sep <- glmer(CONSTRUCTION ~ LOGLENGTH*TYPE + (1|LEVEL1) + (1|LEVEL3)
      + (1|VERB) + (1|PARTICLE), tamily=binomial)
anova(model.1.sep, model.5.sep)
model.1.sep:      (1 | LEVEL3) + (1 | VERB) + (1 | PARTICLE)
            Df    AIC  BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
model.5.sep  8 1537.0 1583 -760.48    1521.0
model.1.sep  9 1538.3 1590 -760.15    1520.3 0.6652      1     0.4147
```

Can we then also assume that the difference between speaking and writing (LEVEL1 or MODE) plays no significant role? We generate a sixth model and compare it to what is now our new reference model, model.5.sep. The answer is 'yes':

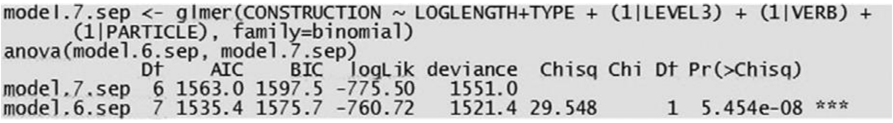


```
model.6.sep <- glmer(CONSTRUCTION ~ LOGLENGTH*TYPE + (1|LEVEL3) + (1|VERB) +
      (1|PARTICLE), family=binomial)
anova(model.5.sep, model.6.sep)
model.5.sep:      (1 | VERB) + (1 | PARTICLE)
            Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
model.6.sep  7 1535.4 1575.7 -760.72   1521.4
model.5.sep  8 1537.0 1583.0 -760.48   1521.0 0.4793      1     0.4887
```

With all this work, we have now completed Step I of Zuur *et al.*'s model selection strategy. With the caveats mentioned in Footnotes 4 and 5, we have now identified the optimal random-effects structure, which turns out to be much more complex than corpus-linguistic studies usually assume. Whatever the results of the fixed-effects part of the analysis in the next section will be, we have already established that it is very much necessary to incorporate verb- and particle-specific effects into the statistical analysis, and on top of that we have also already learned that the statistical analysis benefits significantly from distinguishing the thirteen different sub-registers of the ICE-GB whereas distinguishing the registers or speaking from writing does not do anything. Let us now move on to finding the optimal fixed-effects structure.

3.3.2 Step II: finding the optimal fixed-effects structure

We proceed according to the same logic as before, just with fixed-effects this time: we create a new model that is the same as an old reference model, but delete – this time fixed – effects from it, where we follow the usual rule that effects can only be deleted if they are neither significant themselves nor participate in a significant higher-order interaction (see Gries, 2013: 260). In this case, we only have two fixed-effects independent variables and their one interaction so we test whether this interaction can be deleted:

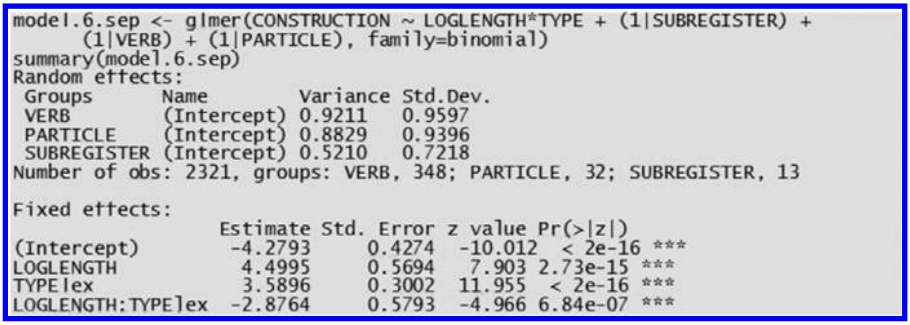


```
model.7.sep <- glmer(CONSTRUCTION ~ LOGLENGTH+TYPE + (1|LEVEL3) + (1|VERB) +
      (1|PARTICLE), family=binomial)
anova(model.6.sep, model.7.sep)
            Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
model.7.sep  6 1563.0 1597.5 -775.50   1551.0
model.6.sep  7 1535.4 1575.7 -760.72   1521.4 29.548      1  5.454e-08 ***
```

The interaction LogLength:Type cannot be deleted with a significant loss in explanatory power, which means no more model simplification can be attempted; we already have the right fixed-effects structure.

3.3.3 Interpreting the final (G)LMM/MLM model

First, to get a slightly more interpretable output, we switch from the variable name LEVEL3 back to that of SUBREGISTER, refit the model with this, and explore the summary output. At the top, we find the three remaining random effects and the amounts of variance they account for.



```
model.6.sep <- glmer(CONSTRUCTION ~ LOGLENGTH*TYPE + (1|SUBREGISTER) +
      (1|VERB) + (1|PARTICLE), family=binomial)
summary(model.6.sep)
Random effects:
 Groups      Name        Variance Std.Dev.
 VERB        (Intercept) 0.9211   0.9597
 PARTICLE    (Intercept) 0.8829   0.9396
 SUBREGISTER (Intercept) 0.5210   0.7218
Number of obs: 2321, groups: VERB, 348; PARTICLE, 32; SUBREGISTER, 13

Fixed effects:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -4.2793     0.4274 -10.012  < 2e-16 ***
LOGLENGTH           4.4995     0.5694   7.903 2.73e-15 ***
TYPElex             3.5896     0.3002  11.955  < 2e-16 ***
LOGLENGTH:TYPElex  -2.8764     0.5793  -4.966 6.84e-07 ***
```

Below that, we find a ‘regular’ regression output with coefficients and their significance tests. Since these are on a log odds scale, which is not easy to interpret, we explore a plot of predicted probabilities such as Figure 8: LogLength is on the *x*-axis, the predicted probability of Construction: V-Part-DO is on the *y*-axis, and the two lines (with shaded confidence intervals) reflect the effect of the interaction LogLength:Type on the predicted probabilities (fixed-effects only).

That is to say, we can see that, the longer the DO, the more V-Part-DO is predicted (which is what every study of particle placement found); but, as with the traditional binary logistic regression, we can also see that this effect is different for the two types: when the DO head is non-lexical, very short DOs strongly prefer V-DO-Part, but V-Part-DO becomes more preferred quickly; but when the DO head is lexical, V-Part-DO is used with longer DOs and becomes more preferred more slowly.

Since these effects are similar to that of the BLR in Section 3.2, the question arises how good this (G)LMM/MLM result is especially when compared to the BLR? Put differently, how much do we benefit from having done the (G)LMM/MLM analysis? It turns out that this latter model does a very good job: $R^2_{\text{marginal}} = 0.57$ and $R^2_{\text{conditional}} = 0.748$, the classification accuracy is 88 percent (compared to 79.02 percent), and the *C*-value is 0.955 (compared to 0.888). In other words, not only did we do the right analysis – because the data violated the assumptions of regular logistic regressions – but we are also rewarded with a model that is more precise and, therefore, highly significantly better than the previous one ($p_{\text{binomial}} < 10^{-29}$). In the next and final section of this case study, let us therefore explore how the (G)LMM/MLM analysis differs from the BLR.

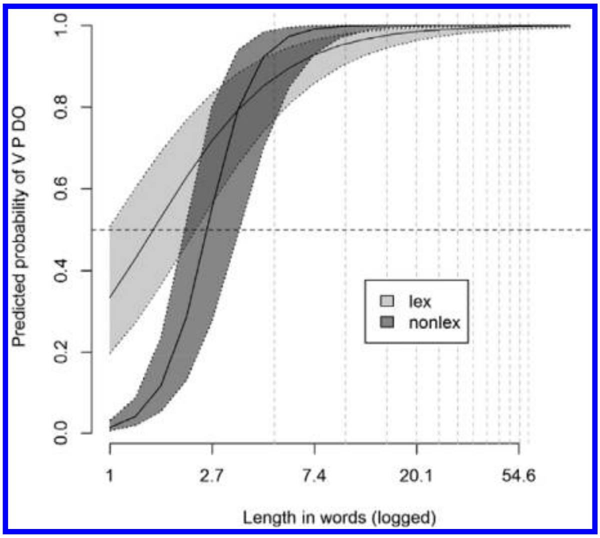


Figure 8: Predicted probabilities of CONSTRUCTION: V-Part-DO, given LOGLENGTH:TYPE

3.3.4 The differences between the BLR and the (G)LMM/MLM approach

As a first general step, let us see where the two kinds of models differ in their classifications, which is represented in Figure 9. While both constructions are about equally frequent in the data, Figure 9 shows that the (G)LMM/MLM approach is particularly better than the BLR in predicting V-DO-Part correctly: of the 1,192 instances of V-DO-Part in the data, the (G)LMM/MLM classifies 999 (83.8 percent) correctly, while the BLR classifies only 816 (68.5 percent) correctly – for the V-Part-DO constructions, the classifications of both models are much more similar (92.4 percent and 90.2 percent, respectively).

However, this is a rather coarse resolution and it is more interesting to see, for all the effects included in the model, where the (G)LMM/MLM approach fares better than the BLR. One of many ways to explore this is to compute for each verb, particle and sub-register, how much more often the (G)LMM/MLM approach makes the correct classification compared to the BLR approach (in percent). For verbs and particles, these results are shown in the left- and right-hand panel respectively of Figure 10, in which the (logged) frequency of occurrence of a verb/particle in verb-particle constructions is on the *x*-axis and the improvement of the (G)LMM/MLM approach in percent is on the *y*-axis.

First, the graph clearly shows yet again how important it is to use lexically specific effects in our corpus-linguistic analyses: there are many verbs and particles whose classification accuracy is improved by 20 percent,

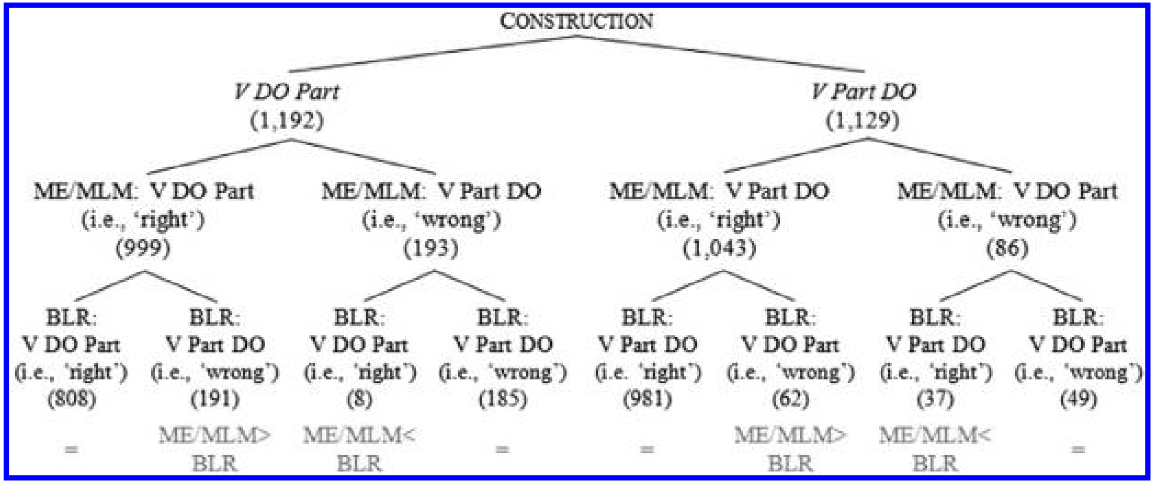

CONSTRUCTION
V DO Part (1,192)
V Part DO (1,129)
ME/MLM: V DO Part (i.e., 'right') (999)
ME/MLM: V Part DO (i.e., 'wrong') (193)
ME/MLM: V Part DO (i.e., 'right') (1,043)
ME/MLM: V DO Part (i.e., 'wrong') (86)
BLR: V DO Part (i.e., 'right') (808)
BLR: V Part DO (i.e., 'wrong') (191)
BLR: V DO Part (i.e., 'right') (8)
BLR: V Part DO (i.e., 'wrong') (185)
BLR: V Part DO (i.e. 'right') (981)
BLR: V DO Part (i.e., 'wrong') (62)
BLR: V Part DO (i.e., 'right') (37)
BLR: V DO Part (i.e., 'wrong') (49)
=
ME/MLM> BLR
ME/MLM< BLR
=
=
ME/MLM> BLR
ME/MLM< BLR
=

Figure 9: A comparison of the overall accuracy of the BLR and the (G)LMM/MLM. For V DO Part, the ME/MLM makes 191–8 = 183 more correct predictions; for V Part DO, the ME/MLM makes 62–37 = 25 more correct predictions

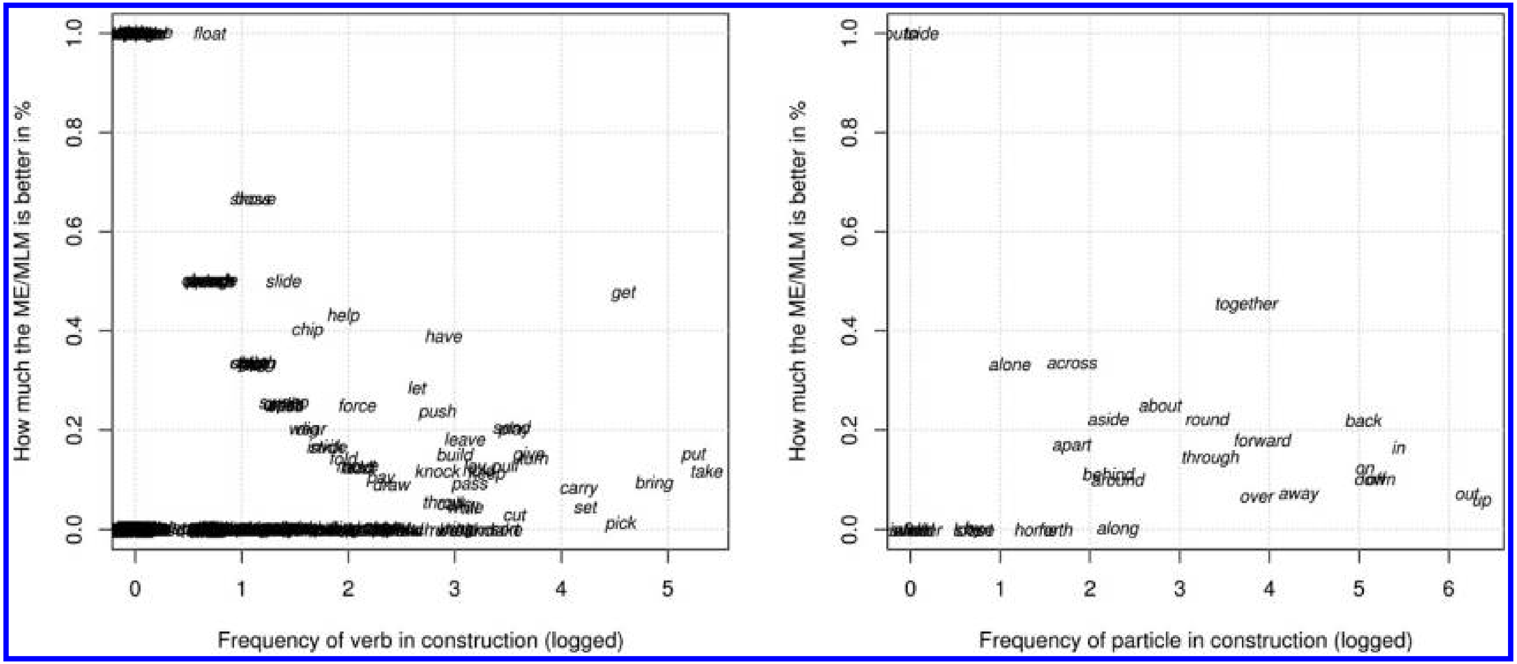


Figure 10: Classification improvements of the (G)LMM/MLM over the BLR per verb (left-hand panel) and per particle (right-hand panel)

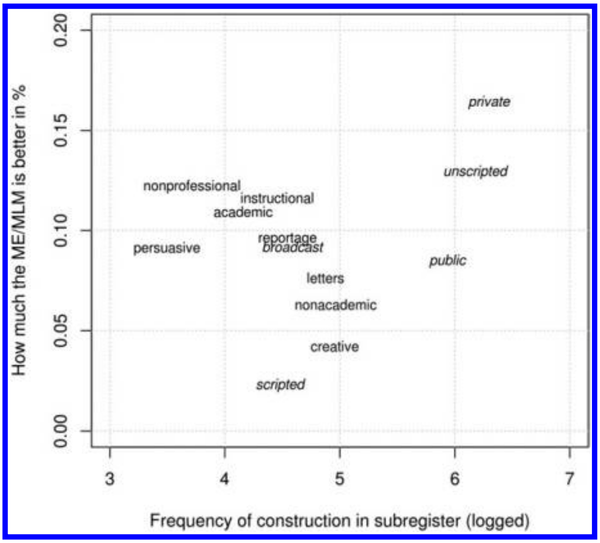


Figure 11: Classification improvements of the (G)LMM/MLM over the BLR per sub-register. Spoken sub-registers are plotted in italics

or even much more, when the more appropriate (G)LMM/MLM approach is used. Also, it is worth anticipating a likely objection: note that the degree to which the (G)LMM/MLM approach is better is not a straightforward function of the frequency of the verbs/particles. Thus, the argument 'Once we have more data everything will be classified better even with BLRs' is flawed: even for some medium- or even higher-frequency verbs and particles (such as *get*/*have* and *together*/*back*), substantial improvements in classification accuracy are obtained.

What about the sub-registers? A similar plot can be drawn and is represented in Figure 11.

Again, it clearly defies what might seem to be straightforward expectations: increased classification accuracy is neither straightforwardly correlated with sample size nor with corpus linguists' most cherished contrast: speaking *versus* writing. However, it is clear that the sub-registers differ strongly from one another (which is why the random effect of SUBREGISTER had to be included) and that they differ strongly in terms of how much they benefit from their idiosyncrasies being taken into account. Funnily enough, it is the sub-register 'creative' that benefits nearly the least from its particular characteristics being taken into account, but one would probably need a real (Biberian) multi-dimensional analysis to try to make sense of the sub-register effects here.

The final way to explore the data further, which is to be exemplified here, would be to look at how the significant interaction of the fixed effects, LOGLENGTH:TYPE, plays out in the different sub-registers (averaging over all verbs and particles). This is represented in Figure 12, which is a finer

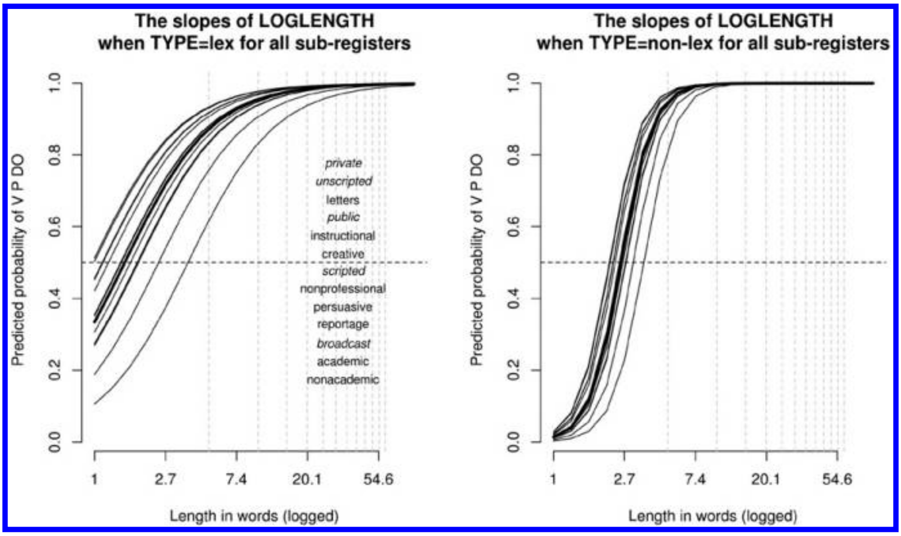


Figure 12: The interaction of LogLength:Type per sub-register: lexical DOs are shown in the left-hand panel; non-lexical ones in the right-hand panel. The order of sub-registers plotted into the left-hand panel corresponds to the order of the curves

resolution of Figure 8 in that each sub-register's predicted-probability curve is shown separately. Again, it would take more detailed analysis than can be provided in this methodological how-to paper, but it is clear that, especially with (the more frequent) lexical DOs there is a huge degree of variation that a BLR would just miss.

4. Conclusion

Mixed-effects modelling, a method of doing regression analyses in which idiosyncrasies of sampled units – speakers, words, *etc.* – can be accounted for elegantly, is a tool that is becoming more and more frequently used, especially in psycholinguistic analyses of (mostly) experimental but also observational data. Given the fact that corpus-linguistic data are much more unbalanced and messier than experimental data, it is time that corpus linguists avail themselves of that same family of methods. Not only would this end the way in which corpus linguists nearly always violate basic assumptions of our statistical tests – because we would finally take into consideration that our data points are not, usually, independent – but it would also allow us to, again finally, take more seriously the exploration of corpus data on the multiple levels of sampling that corpora come in. The multi-level aspects of mixed-effects modelling in the previous literature on this method in psycholinguistics or sociolinguistics has hardly ever been discussed at all in corpus linguistics and is exemplified here for the first time. We know for a fact that there is variability in the corpus – just not on which level(s) – but just pretending 'It's speaking *versus* writing' or 'It's register' does not help: we need to check all those levels at the same time and the technique(s) discussed in this paper, while no cure-all, can help (see Gries and Deshors, forthcoming, for a fully fledged application to learner corpus data).

The potential that a more widespread adoption of these methods have can hardly be overstated. How many papers are there in which authors study a corpus and distinguish between speaking and writing without ever exploring the other levels of resolution on the corpus they could have adopted? (See Gries, 2006, for much discussion.) How many papers are there in which authors study a corpus and do not take into consideration that the results might be skewed because some speakers provide more data than others (see Johnson, 2009, for a nice analysis of simulated data demonstrating the effects this can have)? If we want to avoid violating the basic assumptions of our data, avoid missing significant effects that are actually there (recall Section 2.2 where a regular regression failed to see the strong correlation that a (G)LMM/MLM could detect) and if we want to avoid assuming significant results that are actually not there (recall Section 2.3 where a regular regression returned a significant correlation that a (G)LMM/MLM showed did not exist), then we need to make sure that we use the best statistical methods available and that we apply them correctly (by making sure that we bear in mind, and also test the assumptions, that these best practices rely on, using both statistical and visualisation tools).

However, this is not to say that the current kind of (G)LMM/MLM is the final one method we will ever need. Not only is this technique still undergoing development both conceptually and in terms of its implementation in statistical software, but it also suffers from the same kind of 'problem' from which most regression methods suffer: the fact that correlation is only a necessary condition, but not proof, of ~~correlation~~ causation. In addition, researchers always need to battle over all the other issues that pose risks to regression modelling (collinearity, *etc*.). Thus, (corpus) linguists would be well advised to keep their eyes and minds open for:

- Resolutions of currently debated questions with regard to (G)LMM/MLM such as the relevance of maximal random-effects structures, computation of *p*-values, *etc*.;
- Improvements of (G)LMM/MLM such as multi-model inferencing (see Burnham and Anderson, 2002, in general, and Kuperman and Bresnan, 2012, for an application in linguistics);
- Other different approaches that have characteristics very relevant to empirical linguistics such as structural equation modelling (which targets causal relations) in mind in order to make sense of their complicated data or Bayesian networks (see Theijssen *et al*., 2013).

Regardless of current reservations and future improvements of (G)LMM/MLM, as well as future developments in statistical methodology relevant to linguistics, if this paper provides corpus linguists with a starting point to delve into this area (more) deeply – useful references to explore include Faraway (2006), Gelman and Hill (2006), West *et al*. (2007), Gałecki and Burzykowski (2013), Crawley (2013) or Finch *et al*. (2014) to name but a few – then it has fulfilled its main goal.

Acknowledgments

I would like to thank Antti Arppe, Sandra C. Deshors, members of the audiences at ICAME 35 and AACL 2014, the reviewers, and the editor of *Corpora* for feedback and discussion. The usual disclaimers apply.

References

Baayen, R.H. 2008. Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge: Cambridge University Press.

Baayen, R.H., D.J. Davidson and D.M. Bates. 2008. 'Mixed-effects modeling with crossed random effects for subjects and items', Journal of Memory and Language 59 (4), pp. 390–412.

Barr, D.J., R. Levy, C. Scheepers and H.J. Tily. 2013. 'Random effects structure for confirmatory hypothesis testing: keep it maximal', Journal of Memory and Language 68 (3), pp. 255–78.

Bates, D.M. 2010. lme4: Mixed-effects Modeling with R; Chapter 2: Models with Multiple Random-effects Terms. Unpublished manuscript. 17 Feb 2010.

Bates, D.M., M. Maechler, B. Bolker, S. Walker, R.H. Bojesen Christensen, H. Singmann and B. Dai. 2014. lme4: Linear Mixed-effects Models Using Eigen and S4. R package version 1.1–7. Available online, at: http://CRAN.R-project.org/package=lme4

Burnham, K.P. and D.R. Anderson. 2002. Model Selection and Multimodel Inference: A Practical Information-theoretic Approach. (Second edition.) Berlin and New York: Springer.

Crawley, M.J. 2013. The R Book. Chichester: John Wiley and Sons.

Diessel, H. and M. Tomasello. 2005. 'Particle placement in early child language: a multifactorial analysis', Corpus Linguistics and Linguistic Theory 1 (1), pp. 89–112.

Faraway, J.J. 2006. Extending the Linear Model with R: Generalized Linear, Mixed-effects and Non-Parametric Regression Models. Boca Raton, Florida: Chapman and Hall/CRC.

Finch, W.H., J.E. Bolin and K. Kelley. 2014. Multilevel Modeling Using R. Boca Raton, Florida: Chapman and Hall/CRC.

Gałecki, A. and T. Burzykowski. 2013. Linear Mixed-effects Models Using R: A Step-by-Step Approach. Berlin and New York: Springer.

Gelman, A. and J. Hill. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press.

Gries, St.Th. 2003. Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement. London and New York: Continuum Press.

Gries, St.Th. 2006. 'Exploring variability within and between corpora: some methodological considerations', Corpora 1 (2), pp. 109–51.

Gries, St.Th. 2013. Statistics for Linguistics with R: A Practical Introduction. (Second edition.) Berlin and New York: De Gruyter Mouton.

Gries, St.Th. and S.C. Deshors. Forthcoming. 'EFL and/vs. ESL? On how to properly bridge the paradigm gap', International Journal of Learner Corpus Research 1 (1).

Harrell, F.E. Jr. 2014. rms: Regression Modeling Strategies. R package version 4.2–0. Available online, at: http://CRAN.R-project.org/package=rms

Jaeger, T.F. 2008. 'Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models', Journal of Memory and Language 59 (4), pp. 434–46.

Johnson, D.E. 2009. 'Getting off the GoldVarb standard: introducing Rbrul for mixed-effects variable rule analysis', Language and Linguistics Compass 3 (1), pp. 359–83.

Johnson, P.C.D. 2014. 'Extension of Nakagawa and Schielzeth's R2GLMM to random slopes models', Methods in Ecology and Evolution 5 (9), pp. 944–6.

Kuperman, V. and J. Bresnan. 2012. 'The effects of construction probability on word durations during spontaneous incremental sentence production', Journal of Memory and Language 66 (4), pp. 588–611.

Larson-Hall, J. 2010. A Guide to Doing Statistics in Second Language Research Using R. Available online, at: http://cw.routledge.com/textbooks/9780805861853/guide-to-R.asp

Nakagawa, S. and H. Schielzeth. 2013. 'A general and simple method for obtaining R^2 from generalized linear mixed-effects models', Methods in Ecology and Evolution 4 (2), pp. 133–42.

Quené, H. and H. van den Bergh. 2008. 'Examples of mixed-effects modeling with crossed random effects and with binomial data', Journal of Memory and Language 59 (4), pp. 413–25.

R Core Team. 2014. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. Available online, at: http://www.R-project.org/

Szmrecsanyi, B. 2005. ‘Language users as creatures of habit: a corpus-linguistic analysis of persistence in spoken English’, Corpus Linguistics and Linguistic Theory 1 (1), pp. 113–50.

Szmrecsanyi, B. 2006. Morphosyntactic Persistence in Spoken English: A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis. Berlin and New York: Mouton de Gruyter.

Theijssen, D., L. ten Bosch, L. Boves, B. Cranen and H. van Halteren. 2013. ‘Choosing alternatives: using Bayesian networks and memory-based learning to study the dative alternation’, Corpus Linguistics and Linguistic Theory 9 (2), pp. 227–62.

West, B.T., K.B. Welch and A.T. Gałecki. 2007. Linear Mixed Models: A Practical Guide Using Statistical Software. Boca Raton, Florida: Chapman and Hall/CRC.

Zuur, A.F., E.N. Ieno, N. Walker and A.A. Saveliev. 2009. Mixed Effects Models and Extensions in Ecology with R. Berlin and New York: Springer.