Mixed-effect models

R. Harald Baayen

Abstract

Mixed-effect modeling is recommended for data with repeated measures, as often encountered in designed experiments as well as in corpus-based studies. The mixed-effect model provides a flexible instrument for studying data sets with both fixed-effect factors and random-effect factors, as well as numerical covariates, that allows conclusions to generalize to the populations sampled by the random-effect factors. Mixed-effect models can straightforwardly incorporate two or more random-effect factors. By providing shrinkage estimates for the effects associated with the units sampled with a given random-effect factor, the mixed model provides enhanced prediction accuracy. Mixed-effect models also make available enhanced instruments for modeling interactions of random-effect and fixed-effect predictors. As mixed-effects models do not depend on prior aggregation, they also offer the researcher the possibility to bring longitudinal effects into the statistical model.

Keywords: mixed-effect models, generalized linear mixed-effect models, fixed-effect factor, random-effect factor, shrinkage, longitudinal effects, interactions.

Index Terms: mixed-effect models, generalized linear mixed-effect models, fixed-effect factor, random-effect factor, shrinkage, longitudinal effects, interactions.

1. Introduction

Consider an experiment in which the duration of the first vowel in a word is studied. It is expected that this duration is determined in part by the number of syllables following in the same word, in part by whether the vowel is syllable final, in part by the position of the word in the sentence, by the speech rate, and possibly by the frequency of the word. If our interest is in the generality of vowel shortening, different vowels will be studied, in different words, and produced by different speakers. For this type of experiment, mixed models are an excellent choice.

In this example, the factor syllable `Position` (with levels *final* and *non-final*) is a fixed-effect factor, as its two levels exhaust all possible values that the predictor `Position` can take. By contrast, the factor `Speaker` is a random-effect factor, as its levels, identifiers for the different speakers, are randomly sampled from a much larger population of speakers. `Word` is another random-effect factor, as the words sampled for the experiment represent only a small proportion of the words known to the speakers.

Classical analysis of variance and regression analysis run into problems for data sets combining fixed and random-effect factors, especially when more than one random-effect factor has to be brought into the analysis. Often, researchers aggregate their data to obtain means or proportions for subjects (averaging over items) or for items (averaging over subjects).

In psycholinguistics, the work by Clark (1973) and Forster & Dickinson (1976) led to the practice of averaging both over subjects and over items, with an effect accepted as significant only if it reaches significance both 'by subjects' and 'by items'. Mixed-effect models provide the researcher with a more sophisticated tool for analyzing repeated measures data that is both more flexible, more powerful, and more insightful.

2. Basic Concepts

Let X_1 denote the fixed-effect factor `Position` and let X_2 represent the covariate `Frequency` of occurrence. Suppose that 10 vowels are selected, and that the question of interest is whether the duration of the i-th vowel, Y_i, can be predicted from `Position` (*final* versus *non-final*) and `Frequency`. The linear model decomposes the dependent variable into a weighted sum

$$Y_k = \beta_0 + \beta_1 X_{1k} + \beta_2 X_{2k} + \beta_{12} X_{1k} X_{2k} + \epsilon_k, \quad k = 1, 2, \ldots, 10. \tag{1}$$

Fixed-effect factors are coded numerically using dummy coding, such that a factor with n levels contributes $n - 1$ predictors to the model. Of the many ways in which factors can be coded numerically, *treatment coding* is the most straightforward and the most easy to interpret, especially in the case of analysis of covariance. One level of the factor is selected as default or reference level. Although the selection of the reference level can be guided by theoretical considerations, technically, any level can serve as reference level. For the two-level factor `Position`, treatment coding adds one extra predictor, X_1 in (1), consisting of

ones and zeroes. Observations for the reference level, say *non-final*, are assigned a zero, and observations for the other, contrasting level (*final*) are assigned a one. As a consequence, the β weight for `Position` represents the *difference* (or contrast) between the group mean for the syllable-final vowels and the group mean for the vowels followed by a non-empty coda. This β weight, although technically a slope for a 'degenerate' numerical predictor (consisting only of zeroes and ones), is referred to as a contrast coefficient.

The model defined in (1) includes an interaction term for `Position` by `Frequency`. This interaction allows for the possibility that two different regression lines are required for `Frequency`, one for non-final vowels and a different one for final vowels. As a consequence, two intercepts and two slopes have to be defined. With treatment coding, the regression line for the reference level (*non-final*) is specified by the intercept β_0 and the slope for frequency β_2. The coefficients of the regression line for *final* vowels is obtained by *adjusting* these slopes and intercepts (by β_1 and β_{12}) respectively (see Table 1) to make them precise for the data points with the *final* vowels. In summary, for a fixed-effect factor, one level is selected as the baseline, and coefficients are invested to adjust slopes and intercepts for the other levels of the factor.

When dealing with a random-effect factor, it does not make sense to select one — arbitrary — level (e.g., a given speaker, or a specific word) as reference level: Such a reference level is unlikely to be representative of the population sampled. Therefore, mixed models dispense with fixing a reference level and contrasts for random-effect factors. Instead, the β coefficients for the intercept, covariates, and fixed-effect factors are taken to represent the

β_0	the intercept (group mean) for the reference level *non-final*
$\beta_0 + \beta_1$:	the intercept (group mean) for *final*
β_2	the slope for frequency for *non-final* vowels
$\beta_2 + \beta_{12}$:	the slope for frequency for *final* vowels

Table 1: Treatment coding in analysis of covariance: the contrast coefficients β_1 and β_{12} specify the differences in intercept and slope between the *non-final* and *final* vowels.

population average for each of the populations sampled by the random-effect factors. For any given random-effect factor, adjustments are implemented to allow precise predictions for the individual units sampled, such as the individual speakers in an experiment or corpus. These adjustments (technically referred to as Best Linear Unbiased Predictors or BLUPs) are assumed to follow a normal distribution with mean zero and some unknown standard deviation (to be estimated from the data). Instead of investing $n-1$ coefficients for a simple main effect for a random-effect factor with n levels (e.g., n speakers), only one parameter is invested, a standard deviation characterizing the spread of the adjustments.

By way of example, consider a data set in which vowels are elicited in m words from n speakers, and that a simple main effects model is appropriate. A first model,

$$Y_{ij} = [\beta_0 + b_{0i}] + [\beta_1 + b_{1i}]X_{1j} + [\beta_2 + b_{2i}]X_{2j} + \epsilon_{ij}, \quad i = 1, 2, \ldots, n; j = 1, 2, \ldots, m, \qquad (2)$$

$$b_{0i} \sim \mathcal{N}(0, \sigma_1),\ b_{1i} \sim \mathcal{N}(0, \sigma_2),\ b_{2i} \sim \mathcal{N}(0, \sigma_3),\ \epsilon_{ij} \sim \mathcal{N}(0, \sigma),$$

calibrates the model, for each speaker i, for that speaker's speech rate (through the adjustments b_{0i} to the intercept β_0), as well as for that speaker's sensitivity to the position of

the vowel (through the adjustments b_{1i} to the contrast coefficient β_1) and for that speaker's specific sensitivity to frequency of occurrence (through the adjustments b_{2i} to the slope β_2). Each of the sets of adjustments $b_{\cdot i}$ is assumed to be normally distributed with zero mean. In other words, a random-effect factor (whether speaker, word, text, or syllable) is represented as a source of random variation around the population parameters $\{\beta\}$. This is the sense in which a random-effect factor is 'random'.

Model (2) is incomplete, in that it does not take into account that the words in which the vowels are embedded are repeated across speakers. To incorporate word as a second random-effect factor, (2) has to be modified as follows,

$$Y_{ij} = [\beta_0 + b_{0i} + b_{0j}] + [\beta_1 + b_{1i} + b_{1j}]X_{1j} + [\beta_2 + b_{2i}]X_{2j} + \epsilon_{ij}, \tag{3}$$

$$i = 1, 2, \ldots, n; \;\; j = 1, 2, \ldots, m;$$

$$b_{0i} \sim \mathcal{N}(0, \sigma_1),\; b_{1i} \sim \mathcal{N}(0, \sigma_2),\; b_{2i} \sim \mathcal{N}(0, \sigma_3),$$

$$b_{0j} \sim \mathcal{N}(0, \sigma_4),\; b_{1j} \sim \mathcal{N}(0, \sigma_5),\; \epsilon \sim \mathcal{N}(0, \sigma),$$

with crossed random effects for speaker and word. Adjustments to the intercept are often referred to as random intercepts. Similarly, adjustments to slopes are known as random slopes. In the case of adjustments to a contrast coefficient, one can speak of random contrasts. In (3), there are by-speaker random intercepts (b_{0i}) as well as by-verb random intercepts (b_{0j}). Likewise, there are both by-speaker and by-verb random contrasts (b_{1i}, b_{1j}). The model includes random slopes for frequency only for speaker (b_{2i}). It is not possible to include as well by-verb random slopes for frequency, as this would lead to an unsolvable confound with

frequency itself, which is a by-verb property. In other words, it is only possible to include by-subject random slopes and contrasts for item properties, and by-item random slopes and contrasts for subject properties. For instance, subjects may require adjustments to the slope of the frequency effect, while words may require adjustments to the slope of the effect of aging (see, e.g., Baayen & Milin, 2010).

Whenever in addition to random intercepts, one or more random slopes (or contrasts) are associated with a given random-effect factor, the possibility arises that the random intercepts and random slopes (or contrasts) are correlated. Assuming multivariate normality, the full specification of the random effects for (3) is therefore given by the matrices

$$M_{\text{speaker}} = \begin{bmatrix} \sigma_1 & r_{12} & r_{13} \\ r_{21} & \sigma_2 & r_{23} \\ r_{31} & r_{32} & \sigma_3 \end{bmatrix}, \qquad M_{\text{word}} = \begin{bmatrix} \sigma_4 & r_{45} \\ r_{54} & \sigma_5 \end{bmatrix}, \tag{4}$$

where $r_{kl} = r_{lk}$ specifies the correlation of the adjustments k and l estimated for the population of speakers or the population of words. In other words, the adjustments for a given random-effect factor are assumed to be multivariate normal with zero means and unknown standard deviations and correlations.

3. Advantages of mixed-effects models

Mixed-effect models offer many advantages compared to the classical linear model using dummy coding for random-effect factors. First, a fitted mixed model provides straightforward predictions for unseen levels of random-effect factors. For an unseen speaker and an unseen

word, all $b_{\cdot\cdot}$ are set to zero, and predictions based on model (3) for a given position X_1 and frequency X_2 reduce to

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \tag{5}$$

For a specific speaker i that contributed observations to the data and an unseen word, more precise predictions can be obtained using the by-subject random-effect adjustments:

$$Y_i = [\beta_0 + b_{0i}] + [\beta_1 + b_{1i}]X_1 + [\beta_2 + b_{2i}]X_2. \tag{6}$$

Similarly, when the identity of the word is known, even more precise predictions are available by adding in the by-word random intercepts and slopes. For comparison: The classical linear model only provides predictions for the subjects and items sampled in the data, and for models with many interactions involving subjects and items may not even be able to estimate all relevant coefficients.

Second, the mixed-effect model allows for fine-grained hypotheses about the random-effects structure of the data. For every data set, it is an empirical question whether all the terms in matrices such as shown in (4) contribute to a significantly better fit of the model to the data. The possibility to include or exclude correlation parameters is not available in the classical linear model, but turns out to be an important tool for understanding, for instance, individual differences between the subjects participating in experiments. In chronometric studies, for instance, one may find that subjects with a large positive adjustment to the intercept reveal a large negative adjustment to the slope of frequency of occurrence. Such a negative correlation suggests that slow responders (with large intercepts) carry the frequency effect (see, e.g., Baayen & Milin, 2010, for examples).

Third, mixed-effect models are better able to directly model heteroskedasticity. A fundamental assumption of the linear model is that the residual errors have the same variance across all conditions in the data. In many actual data sets, this assumption of homoskedasticity is violated. For instance, the duration of a vowel might be more variable for a sample of nonnative speakers than for a sample of native speakers. Given a fixed-effect factor distinguishing between native and nonnative speakers, each set of speakers can be assigned its own standard deviation for the by-subject random intercepts, thereby modeling the heteroskedasticity directly (instead of correcting p-values post-hoc for non-sphericity).

Fourth, mixed-effect models can handle autocorrelational structure in data elicited from subjects over time, whether obtained from a stretch of speech or in an experimental context. Human behavior is consistent over time, and this often gives rise to autocorrelations in language data. For instance, although there are fluctuations in speech rate, the speech rate at time t is likely to be very similar to the speech rate at the immediately preceding timesteps $t-1$, $t-2$, ... If the sequence of responses elicited from a given subject constitutes an autocorrelated time series, then it is essential to bring this autocorrelation into the model. If ignored, the residual errors will enter into autocorrelations, violating the assumption of independence of the residual errors, and giving rise to suboptimal conclusions about significance. The simplest way in which autocorrelations can be brought into a mixed model is by including as a separate predictor the response at the preceding point in time. For detailed discussion, the reader is referred to Baayen & Milin (2010).

Fifth, the estimates provided by the mixed-effect model for the adjustments to the pop-

ulation parameters (the BLUPs) are *shrinkage estimates*. A danger inherent in fitting a statistical model to the data is overfitting. By way of example, consider a sample of subjects for which speech rate is recorded. Some subjects will have a faster speech rate than others. The more extreme the speech rate of a given subject is, the less likely it is that in a replication study the speech rate of that subject will be equally extreme (or even more extreme). It is much more likely that in the replication study the speech rate of this subject will have 'regressed' or 'shrunk' towards the mean. Mixed models anticipate this regression towards the mean and implement estimates for the BLUPs that shrink the adjustments in the direction of the mean. As a consequence, predictions for replication studies with the same subjects or items will be more precise.

Sixth, more than two random-effect factors can be included in the model. Returning to the above example, one possible design is to embed the same vowel in different carrier words. In such a design, vowels are repeated independently of the words, and hence the vowel should be considered as a potential third random-effect factor.

Finally, mixed-effect models tend to be better able to detect effects as significant, without giving rise to inflated Type I error rates (see, e.g., Baayen, Davidson & Bates, 2008).

4. Generalized linear mixed models

Thus far, we have considered a dependent variable, duration, that is real-valued, and for which a model assuming homoskedastic Gaussian errors is reasonable. Two commonly

encountered dependent variables require special attention. First, instead of being continuous, the outcome of an experimental observation can be binary: true versus false, correct versus incorrect, success versus failure, present versus absent, etc. For binary dependent variables, the traditional approach is to aggregate over trials (by subjects, or by items) to obtain proportions. Subsequently, analysis of variance or multiple regression is applied with these proportions as dependent variable. Three problems arise with this kind of analysis. First, instead of the variance being independent of the mean, the variance changes systematically with the mean, reaching a maximum when the proportion equals 0.5. Second, proportions are bounded between 0 and 1, but the linear model assumes the dependent variable can assume any real value. The generalized linear model deals with these problems by taking as dependent variable not the proportion P,

$$P = \frac{\#\text{ successes}}{\#\text{ successes} + \#\text{ failures}}, \tag{7}$$

but the log odds ratio (or logit)

$$L = \log \frac{\#\text{ successes}}{\#\text{ failures}}. \tag{8}$$

The log odds ratio ranges from minus infinity to plus infinity, and thus circumvents the problem with the boundedness of proportions. (An alternative to the logit link function that can be attractive for researchers familiar with signal detection theory is the probit link function.) The generalized linear model also implements different options for how the variance changes with the mean. For binary dependent variables, the appropriate variance function is that of a binomial random variable. Given the log odds (or logit) as *link function*

and binomial variance, it becomes possible to obtain for each individual observation a good estimate of the probability of a success (or a failure).

When the dependent variable records counts observed for a fixed window in time, such as the number of segment deletions in text chunks of a fixed length, the problem arises that the variance increases with the mean, again violating homoskedasticity. The solution here is to use the log as the link function, and to assume that the variance function is that of the Poisson distribution.

The generalized linear model has been extended to incorporate random-effect factors in addition to fixed-effect factors. Crucially, generalized linear mixed-effect models, or GLMMs, do not require any prior aggregation into proportions, as the ambition is to provide estimates of the likelihood of a success (or failure), or the rate at which a phenomenon occurs (in the case of count data), for each individual observational unit.

5. Significance in mixed-effect models

The significance of covariates and fixed-effect factors can be evaluated in two ways. One option is to test whether slopes or contrasts are significantly different from zero. For non-Gaussian GLMMs, evaluation is based on Z-scores and associated p-values. For Gaussian models, the relevant t-tests run into the problem that there is no good analytical solution for the appropriate degrees of freedom. For large datasets, the upper bound for the degrees of freedom, the number of observations minus the number of fixed-effect parameters, often

provides a good approximation. Informally, an absolute t-value exceeding 2 is a robust indicator of significance for $\alpha = 0.05$.

As an alternative to the t-test, a Bayesian method estimating the posterior distribution of the parameters can be used to obtain 95% credible intervals for the coefficients, as well as estimates of the probability of values more extreme than those actually observed. For data sets with at least several hundreds of observations, these probabilities are very similar to the probabilities obtained with the t-test based on the upper bound for the degrees of freedom. For smaller samples, the Bayesian probabilities are more precise.

A second option for evaluating significance of a predictor is to compare a model with and a model without a given predictor in order to ascertain whether the parameters invested for this predictor lead to a non-trivial increase in goodness of fit, using a likelihood ratio test. Both models should have been fitted with maximum likelihood rather than with the default, relativized maximum likelihood.

In order to gauge whether random intercepts, random contrasts, or random slopes are justified by a significant increase in goodness of fit, a likelihood ratio test should be used, but now the models involved in the comparison should be fitted with relativized maximum likelihood, which ensures optimal estimation of the random effects in the model.

The test statistic used by the likelihood ratio test is two times the difference between the log likelihood of the model with more parameters and the log likelihood of the model with fewer parameters. This test statistic follows a chi-squared distribution with as degrees of freedom the difference in the number of parameters. For this test to be precise, the models

entering into the comparison should be nested, i.e., the full set of parameters of the model with fewer parameters should be a subset of the set of parameters of the model with more parameters.

6. Working with mixed models

When working with mixed models, several questions may arise. First, there are cases where it is not immediately self-evident whether a factor is to be modeled as fixed or random. Consider an experiment targeting the duration of English front high and mid vowels. Let `Vowel` denote the pertinent factor with as its four levels the four targeted vowels. Is `Vowel` fixed or random? English has 14 vowels, so we are dealing with a sample of vowels. On the other hand, the population of vowels is quite small. In this example, `Vowel` is best modeled as a fixed-effect factor. The front high and mid vowels do not constitute a random sample from the population of vowels. The focus of the study is on specifically the four high and mid front vowels, with no aims to generalize beyond these four vowels to, e.g., back vowels or diphthongs.

Second, for a classical linear model fitted to a data set, an R-squared (or adjusted R-squared) value is generally reported. This R-squared specifies the proportion of the variance accounted for by the model. For mixed models, an R-squared is often not reported, because it is no longer a good measure for understanding the contribution of the linguistic variables to explaining the variance: Parts, often very substantial parts, of the variance are explained by

the random-effect factors. In chronometric studies, for instance, linguistic predictors sometimes contribute less than 1% to the R-squared (Baayen, 2008). If required, the R-squared can be calculated by squaring the correlation coefficient for the observed and expected values of the dependent variable in the case of Gaussian and Poisson models, and the index of concordance (Harrell, 2001) for binomial models.

7. Selected studies using mixed models

Mixed-effect models are a relatively recent development in statistics, and do not have a long history of use in language studies. In psycholinguistics, mixed-effect models are rapidly becoming the new standard for data analysis with repeated measures. Quené and van den Bergh (2008), Baayen et al. (2008), and Jaeger (2008), all in a special issue in the Journal of Memory and Language, provide non-technical introductions, with Quené and van den Bergh discussing an example from phonetics, Baayen et al. presenting simulations of data sets as encountered in psycholinguistics, and Jaeger focusing on generalized linear mixed-effect models for binary data. Chapters 1 and 4 of Pinheiro and Bates (2000) are also highly recommended for introductory reading. Examples of psycholinguistic studies of auditory comprehension using mixed-models are Baayen, Wurm, and Aycock (2007), Ernestus and Baayen (2007), and Balling and Baayen (2008). For application of mixed-models to corpus-based data, see Ernestus, Lahey, Verhees, & Baayen (2006), Janda, Nesset & Baayen (2010), and Keune, Ernestus, Van Hout, & Baayen (2005).

8. Example code for mixed-effect modeling using R

Mixed models are implemented in a range of software packages (e.g., SPSS, SAS, MLwiN, ASReml, S-Plus) and can be programmed within WinBUGS as well. Open-source software for carrying out mixed-effect modeling is available in `R` (the de-facto standard in statistical computing, freely available at `www.r-project.org`) using the `lme4` package by Bates & Maechler (2009).

Given a factor X_1 and a centered covariate X_2, and Speaker and Word as crossed random effects, the following sequence of nested models with increasingly complex random-effects structure is specified in `R` as follows:

$$\text{Model0} = \text{lmer}(\text{Duration} \sim X_1 + X_2 + (1|\text{Speaker}) + (1|\text{Word})) \quad (9)$$

$$\text{Model1} = \text{lmer}(\text{Duration} \sim X_1 + X_2 + (1|\text{Speaker}) + (0 + X_2|\text{Speaker}) + (1|\text{Word})) \quad (10)$$

$$\text{Model2} = \text{lmer}(\text{Duration} \sim X_1 + X_2 + (1 + X_2|\text{Speaker}) + (1|\text{Word})) \quad (11)$$

$$\text{Model3} = \text{lmer}(\text{Duration} \sim X_1 + X_2 + (1 + X_2|\text{Speaker}) + (1 + X_1|\text{Word})) \quad (12)$$

Model0: random intercepts for speaker and word; Model1: random intercepts for speaker and word, for speaker, independent random slopes for X_2; Model2: random intercepts for speaker and word, by-subject random slopes for X_2, and a correlation parameter for the by-subject slopes and intercepts; Model3: as Model2, but with heteroskedastic variance for the levels of X_1. A likelihood ratio test comparing these four models is carried out in `R` with

anova(Model0, Model1, Model2, Model3)

A generalized mixed model for binomial data (e.g., the presence or absence of a segment in the speech signal) with two predictors and random intercepts for speaker and word is specified as follows,

$$\text{Model4} = \text{lmer}(\text{Present} \sim X_1 + X_2 + (1|\text{Speaker}) + (1|\text{Word}),\ \text{family}=\text{“binomial”}) \quad (13)$$

and a generalized mixed model for count data (e.g., the count of segment deletions observed in a 5 minute interview) with random intercepts for speaker and word with the same predictors would be

$$\text{Model5} = \text{lmer}(\text{Count} \sim X_1 + X_2 + (1|\text{Speaker}) + (1|\text{Word}),\ \text{family}=\text{“poisson”}) \quad (14)$$

References

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

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

Baayen, R. H. and Milin, P. (2010). Analyzing reaction times. Manuscript submitted to *International Journal of Psychological Research*.

Baayen, R. H., Wurm, L. H., & Aycock, J. (2007). Lexical dynamics for low-frequency complex words. A regression study across tasks and modalities, *The Mental Lexicon*, 2:

419–463.

Balling, L., & Baayen, R. H. (2008). Morphological effects in auditory word recognition: Evidence from Danish, *Language and Cognitive Processes*, 23: 1159–1190.

Bates, D. & Maechler, M. (2009). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-32 (`http://CRAN.R-project.org/package=lme4`).

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.

Ernestus, M., & Baayen, R. (2007). Paradigmatic effects in auditory word recognition: The case of alternating voice in Dutch, *Language and Cognitive Processes*, 22: 1–24.

Ernestus, M., Lahey, M., Verhees, F., & Baayen, R. H. (2006). Lexical frequency and voice assimilation, *Journal of the Acoustical Society of America*, 120: 1040–1051.

Forster, K. I. & Dickinson, R.G. (1976). More on the language-as-fixed effect: Monte-Carlo estimates of error rates for F_1, F_2, F′, and *min*F′, *Journal of Verbal Learning and Verbal Behavior*, 15: 135–142.

Harrell, F. (2001). *Regression modeling strategies*. Springer, Berlin.

Jaeger, F. (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language*, 59: 434–446.

Janda, L. A., Nesset, T. & Baayen, R.H. (2010). Capturing Correlational Structure in Russian Paradigms: a Case Study in Logistic Mixed-Effects Modeling, *Corpus Linguistics and Linguistic Theory* (in press).

Keune, K., Ernestus, M., Van Hout, R., & Baayen, R. (2005). Social, geographical, and

register variation in Dutch: From written 'mogelijk' to spoken 'mok'. *Corpus Linguistics and Linguistic Theory*, 1: 183–223.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Quené, H., & Bergh, H. van den. (2008). Examples of mixed-effects modeling with crossed random e effects and with binomial data. *Journal of Memory and Language*, 59: 413–425.

Tabak, W., Schreuder, R., and &, Baayen, R. H. (2010). Producing inflected verbs: A picture naming study, *The Mental Lexicon*, in press.