Capturing Correlational Structure in Russian Paradigms:
a Case Study in Logistic Mixed-Effects Modeling

Laura A. Janda[1], Tore Nesset[2] & R. Harald Baayen[3]
University of Tromsø[1,2] and University of Alberta[3]
December 9, 2008

Abstract

This study addresses the statistical analysis of a phenomenon in Russian verbal paradigms, a suffix shift that is spreading through the paradigm and making it more regular. A problem that arises in the analysis of data collected from the Russian National Corpus is that counts documenting this phenomenon are based on repeated observations of the same verbs, and, moreover, on counts for different parts of the paradigms of these same verbs. Unsurprisingly, individual verbs display consistent (although variable) behavior with respect to the suffix shift. The non-independence of the elementary observations in our data has to be taken into account in the statistical evaluation of the patterns in the data. We show how mixed-effects modeling can be used to do this in a principled way, and that it is also necessary to do so in order to avoid anti-conservative evaluation of significance.

1 Introduction

A group of Russian verbs is undergoing a diachronic change in which the suffix -*a* is being replaced by the productive suffix -*aj*. The Russian suffix shift is recognized in reference works such as Zaliznjak (1977) and Švedova (1980), and has been investigated in the contexts of language acquisition, psycholinguistics, stylistic variation, sociolinguistics and dialectology (cf. e.g. Andersen, 1980; Gagarina, 2003; Gor and Chernigovskaya, 2001, 2003a,c,b; Kiebzak-Mandera et al., 1997; Krysin, 1974; Tkachenko and Chernigovskaya, 2006). The suffix shift is evident in present tense, imperative, present active participle and gerund paradigm slots, where the -*a* suffixed forms show suffix truncation usually accompanied by alternation of the root final consonant, whereas the -*aj* suffixed forms lack such alternations. Russian suffix shift thus yields a regularization among verbs comparable to the shift of English verbs from the weak to strong pattern. Table 1 presents the relevant forms of *maxat′* 'wave', showing that the -*a* suffixed forms have a $x \sim \check{s}$ alternation, while the -*aj* suffixed forms preserve both the suffix and the *x* throughout the present paradigm (phonemically, there is a /j/ between the vowels in the orthographic sequences in the 2sg, 3sg, 1pl and 2pl forms). This table also includes the infinitive and masculine singular past forms in addition to the forms relevant to the suffix shift.

[Table 1 about here.]

Corpus data show that the Russian suffix shift is not taking place uniformly, but is dependent upon two factors: paradigm slot and root final consonant. Verbs undergoing the

Russian suffix shift have root final consonants with three different places of articulation: labial, which most favors the innovative *-aj* suffix; dental, which most favors the conservative *-a* suffix; and velar, which is intermediate in implementation of the suffix shift. Turning to the paradigm slots, the gerund appears to be the most innovative in replacing *-a* with *-aj*, and other relevant forms follow a cline, ending with the 3sg present as the most conservative form, resisting suffix shift by maintaining *-a* most strongly. This cline and its theoretical implications for paradigm structure are discussed in detail in Nesset & Janda (in prep.).

The issue addressed in this study is what the best way is to analyse counts of *-a* and *-aj* in the Russian National Corpus (`www.ruscorpora.ru`), obtained for a number of different verbs with varying root-final consonants across different paradigm slots. Although a straightforward examination of the probabilities of the two suffixes aggregated over verbs suggests a clear pattern, a statistical evaluation of this pattern requires that we take into account the fact that the presence of repeated observations for these verbs renders inappropriate common tests that presuppose the independence of the elementary observations. The solution we explore is to use logistic mixed-effects modeling, which allows us to bring under control strong verb-specific trends that are present in our data.

2 Analysis

Our data set comprises 37 verbs. The final consonant of the root is a dental for 11 verbs, a labial for 9 verbs, and a velar for 17 verbs. For each verb, counts are available of how often the rival suffixes *-a* and *-aj* are attested in the Russian National Corpus, for each of six slots in the Russian verbal paradigm: the first and second person (singular and plural), the third person singular, the third person plural, the imperative, the gerund and the active participle.

A barplot of the counts for *-a* (black) and *-aj* (white) shows that the extent to which *-a* is favored over *-aj* varies considerably across paradigm slots (see Figure 1). The third person singular favors *-a* most, while the gerund shows roughly equal counts for the two suffixes.

[Figure 1 about here.]

Although Figure 1 invites the use of a chi-squared test to evaluate whether there are significant differences in the use of the two suffixes across paradigm slots and place of articulation, the chi-squared test is not optimal. First, given the large differences visible in Figure 1, and given the large number of observations involved, 11460, a chi-squared test is certain to argue against the possibility that the likelihood of *-a* and *-aj* would not vary significantly for the different paradigm slots. Second, the chi-squared test is inappropriate as the observations underlying the counts are not independent. It is not the case that the 11460 observations summarized in Figure 1 represent 11460 different verbs sampled randomly from some population. To the contrary, there are only 37 verbs underlying our counts, with the number of observations for a given verb ranging from just 2 to no less than 1343. We have to take the possibility that the different verbs have their own consistent preferences into consideration.

In what follows, we investigate these data with the help of logistic regression (see, e.g., Jaeger, 2008; Baayen, 2008; Bresnan et al., 2007). Logistic regression provides us with the means of estimating the likelihood of *-a* (and *-aj*), albeit indirectly, by transforming the counts of *-a* and *-aj* into a log odds ratio, the log of the ratio of 'successes' (*-a*) and 'failures' (*-aj*). We therefore begin with a graphical exploration of this log odds ratio as a function of paradigm slot, coding the log odds by hand (and backing off from zero by adding one to all counts before taking the log odds).

Figure 2 presents a dotplot that summarizes, for each verb, the log odds for each of the six paradigm slots. Two things are noteworthy. First, some verbs show substantial variability in the extent to which they favor *-a* over *-aj* (e.g., *žaždat′*, 'thirst') while for others (e.g., *bryzgat′*, 'spatter') this variation is much reduced.

For some verbs (e.g., *krapat′*, 'sprinkle'), it seems as if there is no variation at all, but this is due to the presence of zero counts (for *krapat′*, 'sprinkle', nonzero counts are available only for the third person singular). As a consequence, the log odds defaults to $\log(1) = 0$ (recall that we add one to all counts before taking the log odds). Second, Figure 2 also clarifies that the verbs differ substantially in their overall preference for *-a*. The verb *pryskat′*, 'spray', clearly favors *-aj*, whereas a verb such as *dremat′*, 'doze', favors *-a*. When we model the probability of *-a* and *-aj*, we will therefore have to take into account verbs that have different individual overall preferences, as well as individual specific preferences depending upon which paradigm slot is considered.

[Figure 2 about here.]

Within the framework of mixed-effects modeling, we take these two verb-specific preferences into account by means of random intercepts for verbs combined with by-verb random contrasts for paradigm slot. The random intercepts allow us to model the verb's overall preferences as adjustments with respect to the population mean. The random contrasts provide the opportunity for fine-tuning the contrast coefficients for paradigm slot. The contrast coefficients for paradigm slot estimate the differences between a given paradigm slot and a reference paradigm slot, in our analysis, 'a' (the active participle). The contrasts that we estimate represent the average contrasts expected for some unseen, new verb. To make these general contrasts precise for the 37 individual verbs in our sample, we need verb-specific adjustments for each paradigm slot for each verb. These are the abovementioned random contrasts.

Before we fit a mixed model to the data, we should consider whether we need a parameter in our model that captures potential correlational structure involving the random intercepts and the random contrasts. For each verb, we have one intercept and five random contrasts ('f' versus 'a', 'g' versus 'a', 'i' versus 'a', 'p' versus 'a' and 's' versus 'a'). Since all these adjustments are measured on the same verb, they might be correlated. As a next step, we therefore graphically examine our data for the presence of such correlational structure by means of a pairs plot. Figure 3 plots the pairwise correlations for the log odds across each of the six paradigm slots. Dots represent verbs. With only one exception ('a' and 'g'), the log odds in one paradigm slot enter into strong correlations with the log odds in other paradigm slots. This indicates that we will need a model with a non-trivial random effects structure.

[Figure 3 about here.]

Figure 4 is similar to Figure 3, but instead of considering the log odds for all six levels of paradigm slot, we consider the reference factor level, 'a', and the contrasts between the other factor levels and this reference level, as we will be using contrast coding for the handling of our factorial predictors. Because we are now dealing with differences with respect to 'a', the correlations in the top row of Figure 4 change sign. This is the form in which the tight correlational structure in our data will be captured by our mixed-effects model.

[Figure 4 about here.]

We now fit a mixed model to the data with the log odds modeled as a function of two main effects, paradigm slot and place of articulation of the root final consonant, and with correlated random intercepts and random contrasts for paradigm. In `R`, using the `lmer()` function in the `lme4` package (Bates and Sarkar, 2007), we proceed as follows, assuming that the data are available in a data frame named `dat` that is in long format (with a given row in the data frame specifying the counts for *-a* (`a`) and *-aj* (`aj`) for each unique combination of `Verb`, `Paradigm` slot, and `Place` of articulation of the root final consonant):

```
dat.lmer = lmer(cbind(a, aj) ~ Paradigm + Place + (1+Paradigm|Verb),
             data=dat, family="binomial")
```

The algorithm now takes care of backing off from zero, all we need to do is provide it with the raw counts for each verb, supplied to `lmer()` as paired counts (`cbind()` binds vectors column-wise). The random intercepts in our model (the '1' in `(1+Paradigm|Verb)`) take the verb-specific preferences for *-a* as compared to the population average into account. The random contrasts for verb (specified by `Paradigm` in `(1+Paradigm|Verb)` model the verb-specific preferences for *-a* across paradigm slots. Correlation parameters (specified in `(1+Paradigm|Verb)` by specifying both intercept and `Paradigm` before `|Verb`) are essential to do justice to the substantial non-independence that we observed for the verb-specific intercepts and contrasts (as shown in Figure 4).

[Table 2 about here.]

Table 2 lists the estimates of the coefficients, together with their Z-statistics. Of the five contrasts pitting paradigm slots against the reference level of the active participle, three are significant, namely *g-a*, *p-a* and *s-a* (see the column labeled p-value). The two contrasts comparing labial and velar place of articulation with dental place of articulation also reach significance. Figure 5 presents the estimated probabilities of *-a* for each of the levels of paradigm slot and place of articulation. We see that within the set of different paradigm slots, the gerund reveals an exceptional preference for *-aj*. Comparing the right with the left panel, we observe that the differences in the probabilities of *-a* vary more substantially with place of articulation than with paradigm slot.

[Figure 5 about here.]

The random effects structure of our model is summarized in Table 3. To capture the correlational structure in the paradigm we need no less than 21 parameters (6 standard deviations and 15 correlations). Figure 6 visualizes this correlational structure by plotting the estimated by-verb adjustments to the population intercepts and contrasts (the so-called best linear unbiased predictors, BLUPs). A comparison of Figure 6 with Figure 4 shows that the model captures succesfully the interdependencies between the counts of *-a* and *-aj* across the different paradigm slots.

[Table 3 about here.]

[Figure 6 about here.]

The question that we have to address at this point is whether the large number of parameters for the random effects structure is justified. As a first step, we consider Akaike's information criterion (AIC), a measure of goodness of fit. When comparing models, the smaller the AIC, the better the fit is. For a logistic model without any random effects structure, i.e., a model ignoring the verb altogether, the AIC equals 5155. When we bring into the model random intercepts for verb, the AIC reduces to 1397. Further inclusion of random contrasts and correlation parameters for paradigm slot results in the smallest AIC, 525. A likelihood ratio test provides further confirmation that the complex random effects structure of our model is justified compared to a model with only by-verb random intercepts ($X^2_{(20)} = 911.96, p < 0.0001$). It is noteworthy that the `lmer()` function does not allow a model to be fit to the data in which the correlation parameters of the random effects structure are set to zero. Once by-verb adjustments for paradigm slot are taken into account, the correlation parameters must be taken into account as well.

[Figure 7 about here.]

Figure 7 graphs the log odds ratios for the different paradigm slots as estimated by our model against the corresponding log odds ratios in the data, aggregated over verbs (compare Figure 1 for the corresponding barplot of observed counts). The fitted log odds ratios are larger than the observed logits. This is because the fitted logits are tailored to the set of verbs with a root-final dental. As expected, the two sets of log odds ratios are highly correlated. Nevertheless, they are not identical. This is because the mixed model takes the differences between verbs into account, both with respect to the number of observations the verbs contribute (ranging from 2 to 1343), as well as with respect to their idiosyncratic preferences for *-a* versus *-aj*.

While the corrected estimates are very similar to the group averages, as shown in Figure 7, the estimates of the standard errors for the estimates and the corresponding probabilities are very different. This can be seen in Table 4, which lists standard errors and p-values for logistic models with and without random effects structure for the verbs. The estimated standard errors for the model with random effects are sometimes more than 10 times those estimated by the model without random effects structure. Unsurprisingly, the p-values for the latter model are consistently substantially reduced compared to the model that includes the random effects structure. This illustrates that by ignoring the random effects structure in the data, we run the risk of obtaining highly anti-conservative p-values.

[Table 4 about here.]

The distribution observed in Figure 5 corresponds neatly to the cline that we would predict on a priori grounds given what one would expect for the internal structure of the verbal paradigm. We can make the following assumptions, all justifiable on independent grounds (cf. Janda 1995 for relationships between markedness and prototypicality for such categories, and Bybee 1985: 50-52 for further discussion of 3sg as a prototypical verbal form):

- Finite forms are more prototypical than non-finite forms;
- indicative is more prototypical than imperative;
- singular is more prototypical than plural; and
- third person is more prototypical than first and second person.

This yields the following expected hierarchy, ranging from most prototypical (finite indicative, third person singular) to least prototypical (non-finite forms):

3sg	>	3pl	>	1&2person	>	imperative	>	gerund/active participle
‘s’		‘p’		‘f’		‘i’		‘g’, ‘a’
1		2		3		4		5, 6

The order in the hierarchy (‘s’:1, ..., ‘g’, ‘a’:5, 6) receives confirmation from the ranking of the probabilities in Figure 5: Spearman’s ρ is estimated at -0.8286, $p = 0.0292$, one-tailed test. We note that the expected hierarchy recieves this support in an analysis that does not just consider the simple group probabilities, but also takes individual variation attached to verbs and paradigm slots into account, and is therefore appropriately conservative.

The only paradigm slot in the hierarchy that does not follow the anticipated distribution is the active participle, which reveals a probability that is most similar to the 1&2 person finite indicative forms, instead of being most similar to the probability of the gerund (from which it differs significantly, see Table 2). Interestingly, a closer look at this participle reveals that it has close ties to indicative forms that have probably influenced its behavior. Although in terms of the abstract semantics of paradigm categories the participle is, of course, a non-finite form, it is formally closely related to the 3pl form: The suffix of the active participle always contains the same vowel as the 3pl form. Our data indicate that, apparently, it is this formal relationship that overrides the semantic hierarchy that we initially expected, by placing the active participle on the periphery of the finite indicative forms. For a more detailed discussion of this cline the reader is referred to Nesset & Janda (in prep.).

3 Conclusions

A pervasive characteristic of language is that it is a system in which there are large numbers of interdependencies. When analyzing quantitative linguistic data, it is essential that these interdependencies are brought into the statistical model. Failure to do so may give rise to anti-conservative analyses and may cause the researcher to draw incorrect conclusions from the data. We have shown how mixed-effects modeling can serve as a tool that may help us to better model the complex interdependencies in language data. Our present example addressed dependencies in paradigm structure, but similar dependencies may arise whenever multiple data points are collected, for given linguistic units (verbs, constructions, multiword units, etc.) as well as for speakers and writers (in corpus linguistics) and informants (in sociolinguistics).

A direction for further research on the emergence of *-aj* in Russian is to also bring the identity of the speakers/writers in the Russian National Corpus into the model, as different language users may have have their own preferences or dispreferences for *-aj*. Ultimately, a full understanding of the data will require the joint efforts of corpus linguists and sociolinguists.

References

Andersen, H. (1980). Russian conjugation: Acquisition and evolutive change. In Traugott, E. C., editor, *Papers from the 4th international conference on historical linguistics*, pages 285–301. John Benjamins, Amsterdam.

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

Bates, D. and Sarkar, D. (2007). *lme4: Linear mixed-effects models using S4 classes.* R package version 0.9975-13.

Bresnan, J., Cueni, A., Nikitina, T., and Baayen, R. H. (2007). Predicting the dative alternation. In Bouma, G., Kraemer, I., and Zwarts, J., editors, *Cognitive Foundations of Interpretation*, pages 69–94. Royal Netherlands Academy of Arts and Sciences.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form.* Benjamins, Amsterdam.

Gagarina, N. (2003). The early verb development and demarcation of stages in three Russian-speaking children. In Bittner, D., Dressler, W. U., and Kilani-Schoch, M., editors, *Development of Verb Inflection in First Language Acquisition: A Cross-Linguistic Perspective*, pages 131–170. Mouton de Gruyter, Berlin – New York.

Gor, K. and Chernigovskaya, T. (2001). Rules in the processing of Russian verbal morphology. In Zybatow, G., Junghanns, U., Melhorn, G., and Szucsich, L., editors, *Current Issues in Formal Slavic Linguistics*, pages 528–536. Peter Lang, Frankfurt am Main.

Gor, K. and Chernigovskaya, T. (2003a). Formal instruction and the acquisition of verbal morphology. In Housen, A. and Pierrard, M., editors, *Current Issues in Instructed Second Language Learning*, pages 103–136. Mouton de Gruyter, Berlin and New York.

Gor, K. and Chernigovskaya, T. (2003b). Generation of complex verbal morphology in first and second language acquisition: Evidence from Russian. *Nordlyd*, 31(6):819–833.

Gor, K. and Chernigovskaya, T. (2003c). Mental lexicon structure in L1 and L2 acquisition: Russian evidence. *GLOSSOS*, 4:1–31.

Jaeger, F. (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language*, X:in press.

Janda, L. A. (1995). Unpacking markedness. In Casad, E., editor, *Linguistics in the Redwoods: The expansion of a new paradigm in Linguistics*, pages 207–233. Mouton de Gruyter, Berlin.

Kiebzak-Mandera, D., Smoczynska, M., and Protassova, E. (1997). Acquisition of Russian verb morphology: the early stages. In Dressler, W., editor, *Studies in Pre- and Protomorphology*, pages 101–114. Verlag der Österreichischen Akademie der Wissenschaften, Wien.

Krysin, L. P. (1974). *Russkij jazyk po dannym massovogo obsledovanija*. Nauka, Moscow.

Nesset, T. and Janda, L. A. (in preparation). Paradigm structure: evidence from Russian suffix shift.

Tkachenko, E. and Chernigovskaya, T. (2006). Focus on form in the acquisition of inflectional morphology by L2 learners: Evidence from Norwegian and Russian. paper presented at The Second Biennial Conference on Cognitive Science, St. Petersburg, June 9–13, 2006.

Švedova, N. J. (1980). *Russkaja Grammatika (vol. 1)*. Nauka, Moscow.

Zaliznjak, A. A. (1977). *Grammatičeskij slovar' russkogo jazyka*. Izdatel'stvo Russkij Jazyk, Moscow.

List of Figures

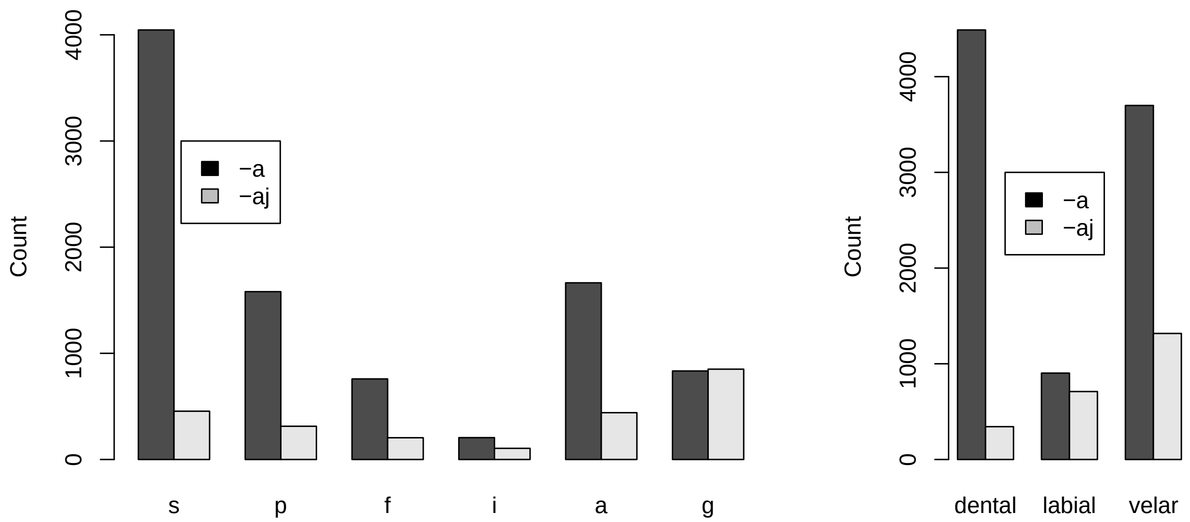

Figure 1: Counts of -*a* (black) and -*aj* (white) realizations for six paradigm slots (left) and place of articulation of the final consonant of the root (right). a: active present participle, p: third person plural, s: third person sigular, f: first/second person, i: infinitive, g: gerund.

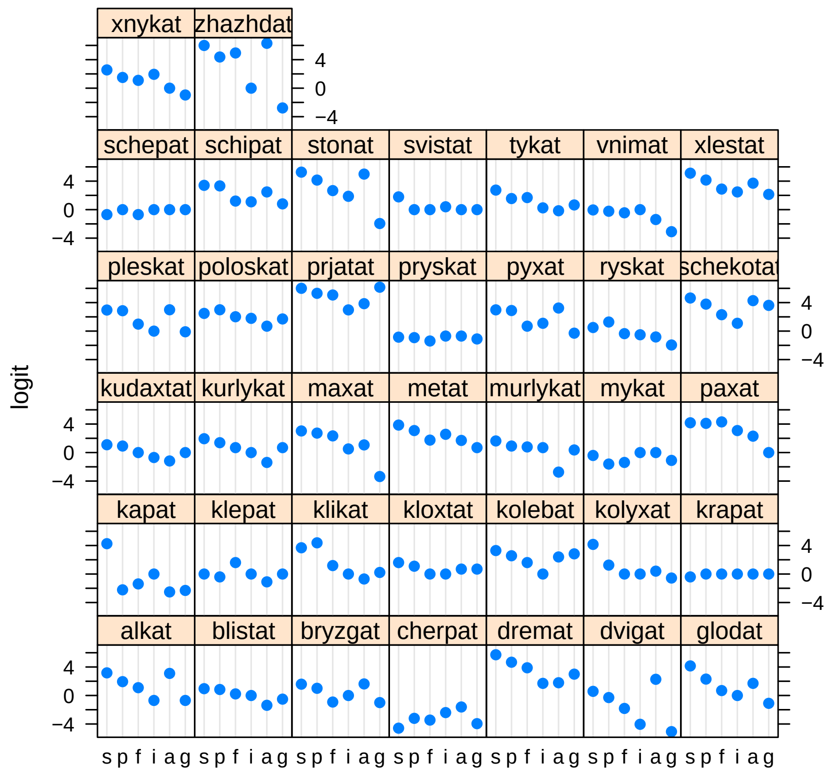

Figure 2: The log odds (of -*a* versus -*aj*) for each of the six paradigm slots ('s': third person singular, 'p': third person plural, 'f': first and second person, 'i': infinitive, 'a': active participle, 'g': gerund). Log odds were calculated after backing off from zero by adding 1 to all counts. A log odds greater than zero indicates a preference for -*a*, a log odds smaller than zero a preference for -*aj*.

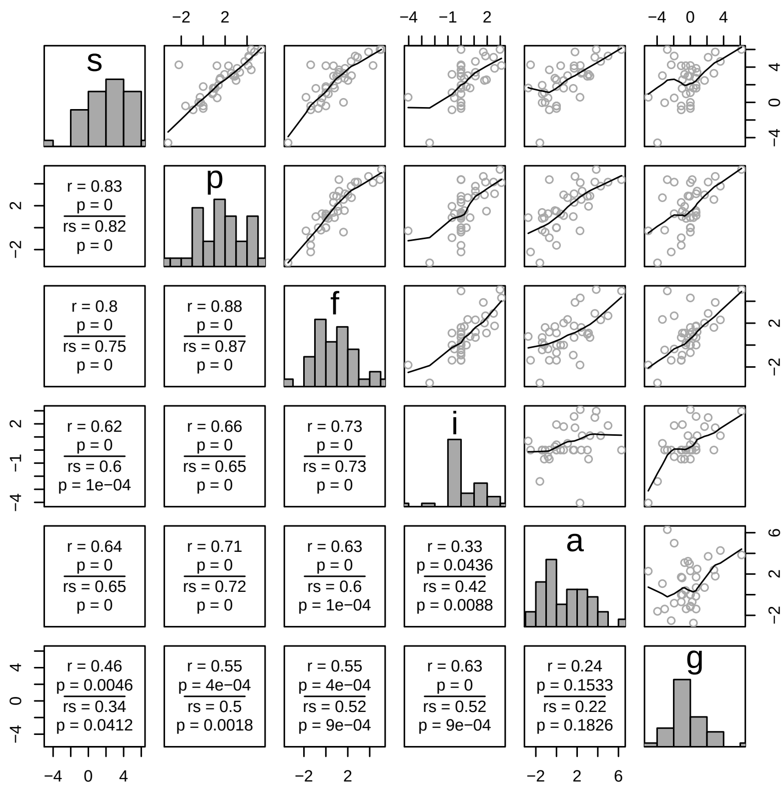

Figure 3: Pairwise correlations for the log odds for the six paradigm slots. Dots represent verbs. The lower half summarizes Pearson (above the line) and Spearman (below the line) correlation coefficients and the associated p-values.

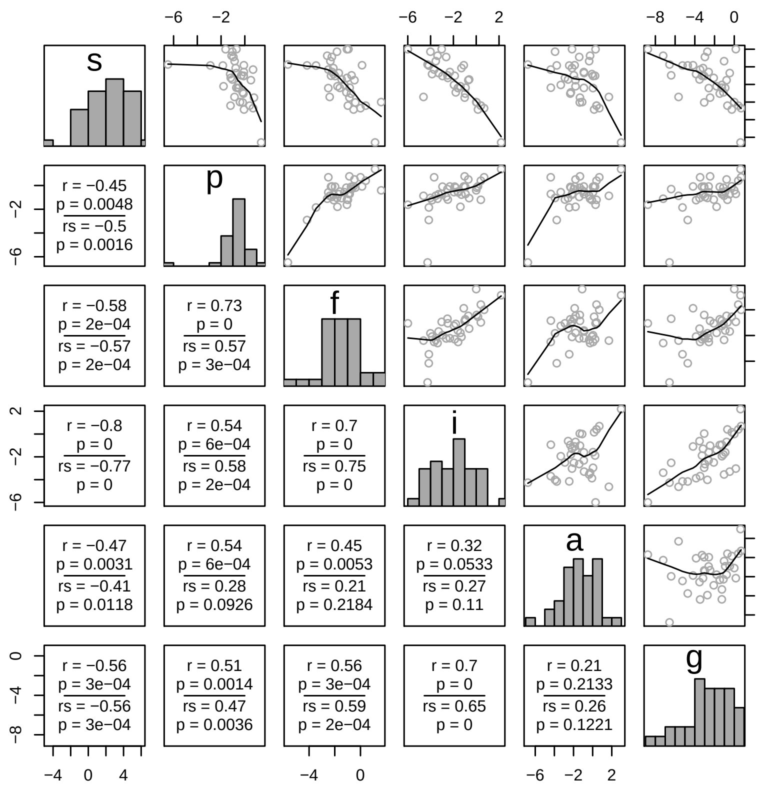

Figure 4: Pairwise correlations for the log odds for the reference level ('a', active participle) and the contrasts with the five remaining levels of paradigm slot. Dots represent verbs.

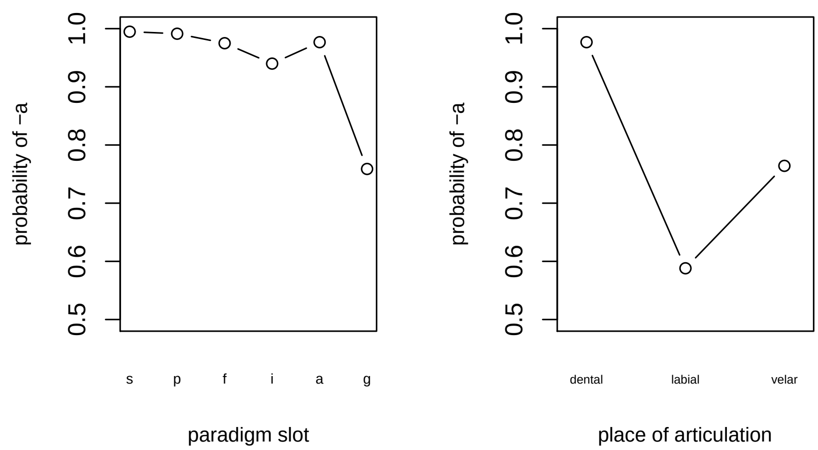

Figure 5: Probabilities of -*a* for paradigm slots (left) and place of articulation of the final consonant of the root (right) as predicted by a mixed-effects logistic model. The probabilities shown in the left panel are adjusted to dental place of articulation. The probabilities in the right panel are adjusted for the active participle ('a').

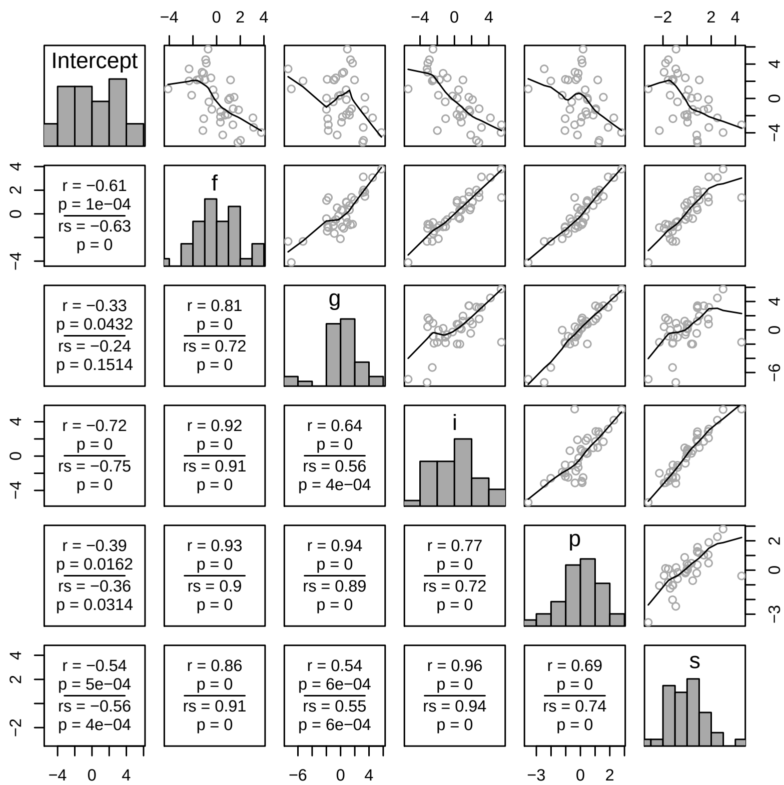

Figure 6: Pairs plot for the by-verb adjustments (BLUPs) to the intercept and contrast coefficients for paradigm slot as estimated by the logistic mixed-effects model.

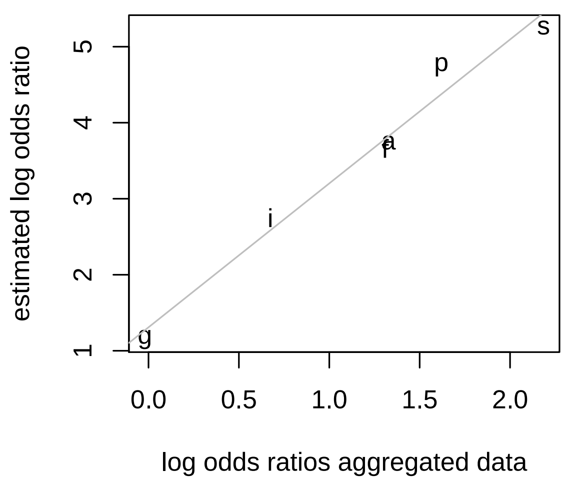

Figure 7: The log odds ratios for the data aggregated by paradigm slot (compare Figure 1) compared to the corresponding log odds ratios as estimated by the mixed model.

List of Tables

Table 1: Forms of *maxat′*, 'wave'.

	forms suffixed with -*a*	form suffixed with -*aj*
infinitive	*maxat′*	*maxat′*
masculine sg past	*maxal*	*maxal*
1sg present	*mašu*	*maxaju*
2sg present	*mašeš′*	*maxaeš′*
3sg present	*mašet*	*maxaet*
1pl present	*mašem*	*maxaem*
2pl present	*mašete*	*maxaete*
3pl present	*mašut*	*maxajut*
imperative	*maši(te)*	*maxaj(te)*
present active participle	*mašuščij*	*maxajuščij*
gerund	*maša*	*maxaja*

Table 2: Coefficients of the mixed-effects model and associated Z-statistics.

	Estimate	Standard Error	z-value	p-value
a, dental (intercept)	3.738	0.915	4.085	<0.0001
f - a (contrast)	−0.075	0.483	−0.154	0.8774
g - a (contrast)	−2.591	0.707	−3.663	0.0002
i - a (contrast)	−0.988	0.693	−1.425	0.1542
p - a (contrast)	0.998	0.366	2.731	0.0063
s - a (contrast)	1.513	0.432	3.502	0.0005
labial - dental (contrast)	−3.382	1.164	−2.906	0.0037
velar - dental (contrast)	−2.562	0.967	−2.649	0.0081

Table 3: The random effects structure: The column labeled Standard Deviation lists the standard deviations of the by-verb adjustments to the intercept and the contrast coefficients for paradigm slot. The correlation matrix to its right summarizes the pairwise correlations between all six sets of adjustments.

	Standard Deviation	Correlations				
		Intercept	f	g	i	p
Intercept	3.366					
f	2.229	-0.654				
g	3.519	-0.398	0.766			
i	3.323	-0.757	0.879	0.638		
p	1.627	-0.472	0.919	0.893	0.777	
s	2.042	-0.609	0.838	0.535	0.950	0.676

Table 4: Standard errors and p-values as estimated in logistic models with (first two columns) and without (second two columns) random effects structure for the verbs.

	mixed model		model without random effects	
	Standard Error	p-value	Standard Error	p-value
a, dental (intercept)	0.9149	0.0000	0.0746	0.0000
f - a (contrast)	0.4829	0.8774	0.1038	0.0003
g - a (contrast)	0.7074	0.0002	0.0786	0.0000
i - a (contrast)	0.6931	0.1542	0.1411	0.0029
p - a (contrast)	0.3655	0.0063	0.0889	0.0000
s - a (contrast)	0.4319	0.0005	0.0793	0.0000
labial - dental (contrast)	1.1637	0.0037	0.0832	0.0000
velar - dental (contrast)	0.9671	0.0081	0.0681	0.0000