Making choices in Slavic: Pros and cons of statistical methods for rival forms

The anonymous six

September 14, 2012

Abstract

Sometimes languages present speakers with choices among rival forms, such as Russian остричь vs. обстричь 'cut hair' and проникнув vs. проникши 'having penetrated'. The choice of a given form is often influenced by various considerations involving the meaning and the environment (syntax, morphology, phonology) and rival forms can often simultaneously compete in some environments while showing strong tendencies to prefer one form over the other in other environments. Understanding the behavior of rival forms is crucial to understanding the form-meaning relationship of language, yet this topic has not received as much attention as it deserves. Given the variety of factors that can influence the choice of rival forms, it is necessary to use statistical models in order to accurately discover which factors are significant and to what extent. The traditional model for this kind of data is logistical regression, but recently two new models, called "tree & forest" and "naive discriminative learning" have emerged as alternatives. We compare the performance of logistical regression against the two new models on the basis of four datasets reflecting rival forms in Russian. We find that the three models generally provide converging analyses, with complementary advantages. After identifying the significant factors for each dataset, we show that different sets of rival forms occupy different regions in a space defined by variance in meaning and environment.

1 Introduction

This article focuses on statistical analysis of rival forms in language. Rival forms exist when a language has two (or more) forms that express a similar meaning in similar environments, giving the speaker a choice of options. The choice made between rival forms is often influenced by a range of factors such as the syntactic, morphological, and phonological environment. We will commence by examining the place of rival forms in the form-meaning relationship.

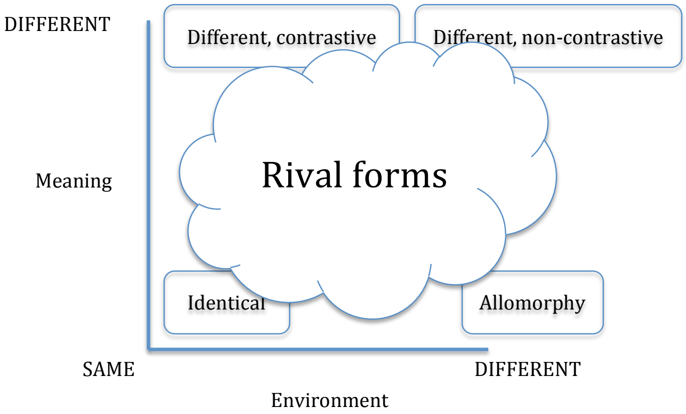
The form-meaning relationship is essential to language, yet highly complex, both in terms of the relationship itself, and in terms of the environments in which this relationship obtains. We can think of this relationship as a three-dimensional space, with form, meaning, and environment as the three axes that define this space. Each axis has a continuum of values that range from perfect identity (when the form, meaning, and environment are exactly the same) to contrast (when the form, meaning, and environment are entirely different). At these two extremes we have trivial cases of either identical items (with identical meanings found in identical environments), or different items (with different meanings found in different environments). However, each axis captures a gradient that also includes variants lying between identity and difference, involving near-identity, similarity, overlap, and varying degrees of contrast, fading out into mere (non-contrastive) difference. If we choose to look only at cases showing difference in form, then meaning and environment yield a two-dimensional space, as visualized in Figure 1.

In addition to the labels at the four corners of Figure 1, synonymy lies along the bottom horizontal axis of the space. Whereas strictly speaking synonyms should have the "same" meaning, in reality even the best of synonyms are usually near-synonyms, with slightly different shades of meaning. Thus synonymy is a gradient phenomenon, with some synonyms overlapping nearly entirely in terms of both meaning and environment, but others showing some deviation.[1] The space in the center of Figure 1 is labeled "Rival forms" and includes relationships involving near-synonymy and partial synonymy as well as various degrees of overlap in terms of environments.

Linguists tend to focus on the four corners of this space, which we can

[1] This article does not address antonyms, which are actually very similar to synonyms, providing contrast in only one (or a few) parameters, but usually found in the same environments and thus located along the leftmost vertical axis of Figure 1.

Figure 1: The space defined by variance in meaning and environment



illustrate with Russian verbal prefixes and environments involving syntactic, morphological (word-formation), and phonological factors. Let's begin at the origin, where the environment and meaning are the same, and go clockwise around the corners from there. For example, if we have two attestations мать остригла волосы ребенку and мать обстригла волосы ребенку 'the mother cut the child's hair', we have the same meaning and the same environment (in terms of word-formation and syntax), and the variant forms о- and об- are performing an identical role. If we change the meaning but keep the word-formation and syntactic environment the same we can get contrasting meanings of the prefixes во- and при- as in мать вошла в церковь 'mother entered (into) the church' and мать пришла в церковь 'mother came to church', where the former phrase emphasizes the church as a building and the latter one refers to a functional relationship (it is most likely that mother in this phrase is attending a service or other meeting). The fact that во- and при- can occur in some of the same environments makes it possible for their meanings to be used contrastively. Next is a case where both the meaning and the environment (in terms of syntax) are different, as in мать вошла в церковь 'mother entered (into) the church' and мать вышла из церкви 'mother exited (from) the church', where the prefixes во- and вы- are simply different in both their meaning and their distribution. In the last corner we find allomorphy, traditionally defined as a relationship of different forms that share a meaning but appear in complementary distribution (Bauer 2001: 14; Booij 2005: 172; Haspelmath 2002: 27; Matthews 1974: 116). Here we have phonologically conditioned examples like мать вошла в церковь ‘mother entered (into) the church (walking)’ and мать вбежала в церковь ‘mother entered (into) the church (running)’, where во- and в- are allomorphs and their different distribution is conditioned by the phonological shape of the root to which they are attached. Here the environment is phonological rather than involving word-formation or syntax.

The space between the four points in Figure 1 has not been thoroughly explored by linguists, yet arguably contains many of the most interesting form-meaning-environment relationships found in language. Although rival forms have received some attention in the literature (cf. Riddle 1985 and Aronoff 1976 on rival affixes in English word-formation, such as -ity and -ness), this is an understudied topic. More empirical studies are needed. The present article is an attempt to fill this need.

We examine four case studies of rival forms: 1) грузить 'load' and its prefixed perfective forms which appear in two rival constructions, 2) the prefixes пере- vs. пре-, 3) the prefixes о- vs. об-, and 4) the use of -ну vs. Ø forms of verbs like (об)сохнуть 'dry'. Although this is primarily a methodological article, the case studies all relate to the topic of this special issue, namely the understanding of time in Russian since they involve rival forms of Russian verbs associated with perfectivizing prefixes and the -ну suffix. Each case study is supported by an extensive dataset and a variety of statistical models are applied in order to discover the complex structures in the form-meaning-environment relationships. Section 2 provides a general discussion of the range of options for statistical analysis and problems posed by various datasets. The studies are presented in Section 3, which relates each case study to the parameters in Figure 1 and also states the linguistic objective of each study. The results of the analyses are summarized in the conclusions in Section 4. All the datasets and the code used for their analyses are available at this site: `laura's-website`. All analyses are performed using the statistical software package R (2011), which is available for free at www.r-project.org.

2 Options for statistical analysis

This section presents the three statistical models that we compare: the logistic regression model, the tree & forest model (combining classification trees with random forests), and the naive discriminative learning model.

Despite the variety of data represented in our four case studies, they share a similar issue: each one presents a pair of rival forms and their distribution with respect to an array of possible predicting factors. If we call the rival forms X vs. Y, then we can define a factor, say Realization, that has as its levels the rival forms (X and Y). Given the semantic and environmental predictors A, B, C, D etc., we can restate all of the case studies in terms of questions like these:

1. Which combinations of values for A, B, C, D predict the value of the response variable "Realization"?

2. How do the predictors rank in terms of their relative strength or importance?

3. If we build a model that optimizes the use of the predictors to predict the response (X vs. Y), how accurate is that model, how well does it capture valuable generalizations without being overly affected by low level variation that is merely "noise"?

We can think of these questions as being parallel to many other types of questions one might ask in many non-linguistic situations such as:

- Predicting whether patients will get cancer (X = yes vs. Y = no) given possible predictors such as age, body mass index, family history, smoking history, alcohol use, diet, exercise, etc.

- Predicting which candidate voters select (X = democrat vs. Y = republican) given possible predictors such as age, race, religion, income, education level, region, etc.

- Predicting which product (X = name brand vs. Y = generic brand) consumers will select given possible predictors such as price, volume, advertising, packaging, etc.

The popular method statisticians apply to such situations with a binary response variable is logistic regression (cf. Baayen 2008: Chapter 6). The first subdiscipline in linguistics to make use of logistic models is sociolinguistics (Cedergren & Sankoff, 1974, see also Tagliamonte & Baayen, 2012). More recently, this type of modeling has also been applied to lexical choices (Arppe, 2008) and grammatical constructions (Bresnan, Cueni, Nikitina, & Baayen, 2005). The strategy of a regression model is to model the functional relation between the response and its predictors as a weighted sum quantifying the consequences of changing the values of the predictors. For factorial predictors (such as perfective versus imperfective), the model specifies the change in the group means when going from one factor level (e.g. perfective) to the other (imperfective). For numerical predictors, the model specifies the consequences of increasing the predictor's value by one unit. The goal of a logistic regression model is to predict the probability that a given response value (X, or alternatively, Y) will be used. It does so indirectly, for mathematical reasons, by means of the logarithm of the odds ratio of X and Y. The odds ratio is the quotient of the number of observations supporting X and the number of observations supporting Y. The log of the odds ratio is negative when the count for Y is greater than the count for X. It is zero when the counts are equal. It is positive when the counts for X exceed the counts for Y.

Fitting a logistic regression model to the data amounts to finding the simplest yet adequate model for the data. A model is simpler when it has fewer predictors. A model is more adequate when its predictions approximate the observations more closely. Typically, one will have to find a balance between the two, by removing predictors that do not increase the goodness of fit, and by adding in predictors that make the model more precise. How to find the best model is an active area of research. In the present study, we use a hypothesis-driven search for the best model.

An important concept in statistical modeling is that of an interaction between predictors. Consider two predictors, for instance, Animacy (with levels animate and inanimate) and aspect (with levels perfective and imperfective). There is no interaction when a change in animacy (or a change in aspect) is the same for all the levels of the other factor. However, when the likelihood of response X increases when changing from animate to inanimate for perfective verbs, but decreases (or increases less) for imperfective verbs, then an interaction of Animacy by Aspect is at issue. Adding in interaction terms may substantially increase the goodness of fit of a model.

The output of a logistic regression model gives us information that addresses all three questions stated above:

1. We can discover which of the predictors predict the value of the response variable by checking whether a change in the value of a given predictor implies a significant change in the value of the response. In the case of logistic regression, this implies a significant change in the value of the log-odds, which translates into a significant change in the probability of, e.g., the response value X.

2. Information about the relative strength and importance of a predictor can be obtained by inspecting both the magnitude of its effect on the response, and by considering the extent to which adding the predictor to the model increases its goodness of fit. This is typically accomplished with the AIC measure (Akaike's information criterion). Lower values of AIC indicate a better model fit.

3. It is possible to evaluate the accuracy of the model by comparing its predictions (whether the response has as its value X or Y) with the actual observed values. Accuracy measures can be imprecise, however, because the model delivers probabilities whereas the observations are categorical (X or Y). One can posit that a probability of X greater than or equal to 0.5 is an X response, and a probability of X less than 0.5 a Y response. But this procedure makes it impossible to see that the model might be correctly predicting differences in probability below (or above) 0.5. For instance, changing from inanimate to animate might raise the probability of an X response from 0.6 to 0.8. The accuracy measure cannot inform us about this. A second measure, C, the index of concordance, has been developed that does not have this defect, and therefore provides a more precise measure of how well the model performs. For a model to be considered a good classifier, the value of C should be at least 0.8.

Most readers who are not already proficient with statistics are likely to express frustration at this point, since the tasks of designing an optimal logistic regression model and then interpreting the output are rather daunting. In fact, guidelines and principles for finding the optimal model are an active area of research, with on the one hand computer scientists proposing algorithms that will find the best fitting model on the one hand, and researchers preferring hypothesis-driven model selection on the other hand. The goal of this article is to illustrate logistic modeling, but to complement it with two alternative models that are more straightforward to use, and that sometimes yield results that are more intuitive in their interpretation. The two alternatives we present here are: 1) classification trees and random forests (henceforth "tree & forest"; cf. Strobl et al. 2009) and 2) naive discriminative learning (Baayen 2011). Both alternatives eliminate the step of searching for an optimal regression model: They arrive at their optimal solutions on their own. Especially in the case of the "tree & forest" method, the output is often easier to interpret as well: The classification tree is an entirely intuitive diagram of the outcomes that are predicted and yielded by various combinations of predictor values.

Logistic regression modeling is a very powerful tool when the data do not violate their underlying mathematical assumptions. One such assumption is that when testing for interactions between two factors, all combinations of factor levels should be attested. For linguistic datasets, this condition is not always satisfied, often because the grammar does not allow for certain combinations. For instance, in the -ну vs. Ø dataset, there are no unprefixed past gerunds. An advantage of classification trees & random forests and naive discriminative learning is that they do not impose distributional constraints, and are thus better suited for many types of datasets involving naturalistic data on rival linguistic forms.

In the R programming environment, all three types of models use the same basic format for the formula that relates the rival forms to the predictors. This formula places the predicted variable to the left of a tilde ∼ and places the predictors to the right, separated by plus "+" signs.[2] Our abstract and hypothetical examples above would be rendered by these formulas (using "Response" to refer to X vs. Y):

1. rival linguistic forms:
   ```
   Response ∼ A + B + C + D
   ```

2. cancer prediction:
   ```
   Response ∼ Age + BodyMassIndex + FamilyHistory +
   + SmokingHistory + AlcoholUse + Diet + Exercise
   ```

3. voter choice prediction:
   ```
   Response ∼ Age + Race + Religion + Income +
   + EducationLevel + Region
   ```

[2] The plus sign does should be read as "and" and not as summation. It is only in the case of logistic models that the plus sign can be interpreted as summation, but then it indicates that the response is modelled as a weighted sum of the predictor values.

4. consumer choice prediction:

    ```
    Response ~ Price + Volume + Advertising +
    + Packaging
    ```

While both the tree & forest model and naive discriminative learning are non-parametric classification models (as opposed to the parametric logistic model), they work on different principles and this has implications for the kinds of datasets that can be modeled and the results of analysis. The tree & forest model uses recursive partitioning to yield a classification tree that provides an optimal partitioning of the data, giving the best "sorting" of observations separating the response outcomes. It can literally be understood as an optimal algorithm for predicting an outcome given the predictor values.

Naive discriminative learning provides a quantitative model for how the brain makes the choice between rival forms and constructions. This type of model makes use of a two-layer network, the weights of which are estimated using the equilibrium equations of Danks (2003) for the Rescorla-Wagner equations (Wagner & Rescorla 1976) that summarize and bring together a wide body of results on animal and human learning. The basic idea underlying this model is best explained by an example. Consider English scrabble, and imagine a situation in which one has a Q, an A, but no U. In that case, knowledge of the legal English scrabble word qaid will increase the chances of playing the Q. The letter combination QA, although very infrequent, is an excellent cue for the word qaid. The greater the number of words with a given form pattern, the less good that form pattern will be as a cue to the meaning of any specific word with that pattern. Naive discriminative learning estimates from (corpus) data the strengths with which form cues support a given meaning. Baayen et al. (2011) showed that a simple naive discrimination network can account for a wide range of empirical findings in the literature on lexical processing. Baayen (2011) used a discrimination network to model the dative alternation in English (Bresnan et al., 2006), and showed that such a network performed with accuracy on a par with that of other well-established classifiers. This shows that human probabilistic behavior can be understood as arising from very simple learning principles in interaction with language experience as sampled by corpus data. The naive discriminative learning model can be pitted against naturalistic datasets in order to ascertain to what extent human learning (under ideal conditions) and statistical learning (using computational algorithms with no cognitive plausibility) converge.

Both the tree & forest model and naive discriminative learning provide a mechanism for validating the model. These validation techniques assess how the results of a statistical analysis will generalize to an independent dataset. Ideally one would build a statistical model for a given phenomenon based on one dataset (the training dataset) and then test the performance of that model using a second, independent dataset (the validation dataset). In this way one can avoid circular reasoning that would result from building and validating the model on the same dataset (since of course the model will perform best if we ask it to predict the outcomes of the data that were the input for its design). These techniques also protect against overfitting the data. Overfitting occurs when the model reflects variation that is characteristic of the particular sample of data, and this interferes with how the model reflects the generalizations that are relevant to the phenomenon under study in the population from which the data were sampled. In other words, any given sample might misrepresent the relationship between the rival outcomes and possible predictors due to chance variation, and ideally this problem would be solved by using two samples, a training dataset and and independent, new "validation" dataset. Statisticians have designed a variety of validation techniques in order to address the gap between the ideal situation and the limitations of reality. In many cases it is not really possible (or at least extremely difficult) to get two large independent samples of the relevant data. Linguists face this problem, for example, due to limits on corpus data: the size of any given corpus is finite, and once all the relevant data from a given corpus has been mined out, it is not possible or very difficult to get a second independent dataset that would be an equivalent sample in terms of size and sources.

The basic idea underlying the validation techniques is to use part of the available data for fitting (or training) the model, and the remaining part of the data to test the predictions of the model on.

In the tree & forest model, bootstrap samples are used. A bootstrap sample is a sample, drawn with replacement, of size N drawn from a dataset with N observations. As a consequence of replacement, some observations are sampled more than once, and others are not sampled at all. The data points sampled at least once constitute the in-bag observations on which we base learning, the data points that are not sampled constitute the out-of-bag observations, which we will predict.

Naive discriminative learning uses a ten-fold cross-validation. This validation technique partitions the data into ten subsamples. Nine of the subsamples serve collectively as the training dataset (the in-bag observations), while the remaining subsample is used as a validation dataset (the out-of-bag observations on which we test our predictions). This process is repeated ten times, so that each of the ten subsamples has been used once as a validation dataset.

One thing to remember with both the random forest and naive discriminative learning models is that because randomization is used in the calculations, some of the output can differ slightly each time these analyses are run. In fact, it is always a good idea to run the validation procedure several times, to make sure that a particular result does not depend on how the data happened to be sampled.

We will take up each dataset in turn, motivate our choice for the optimal statistical model, and detail its interpretation. In addition to this primary goal of alternative models and their interpretation, our secondary goal is to show how statistical models can help us to explore and understand the structure of naturalistic datasets such as the ones presented here. More specificall, we will use statistical models as a sensitive multi-purpose tool for ferreting out the relationships between rival forms and their predictors.

3 Analyses

The analyses are presented according to the relative complexity of the data, starting with the most straightforward dataset. Each subsection below presents a dataset by stating its name, source, overall size, rival forms, and values for predictors. We then present the optimal statistical model and compare it with other possible models and briefly discuss the results and what they tell us about the rival forms and their behaviors. The first dataset is the one with the грузить 'load' data ("LOAD"), which is relatively simple because it has few predictors, each with few levels. This dataset is amenable to analysis by all three of the methods we present in this article, yielding very similar results for all three. Thus the LOAD dataset is a natural point of departure, and will be presented first. We give a relatively detailed explanation of how to interpret the results of the three types of models for the LOAD data and more abbreviated notes on the results for the remaining datasets. Some additional details are available in the annotations to the R script at `laura's-url`.

3.1 Грузить 'load' and its perfectives in the theme-object vs. goal-object constructions

The objective of this case study is to show that so-called "empty" perfectivizing prefixes are actually distinct since they can show unique patterns of preference for grammatical constructions. When prefixes are used to form perfective partner verbs, it is traditionally assumed that the prefixes are semantically "empty" (Šaxmatov 1952, Avilova 1959 & 1976, Tixonov 1964 & 1998, Forsyth 1970, Vinogradov 1972, Švedova et al. 1980, Čertkova 1996; however note that some scholars have opposed this tradition, especially van Schooneveld 1958 and Isačenko 1960). Грузить 'load' provides an ideal testing ground for the "empty" prefix hypothesis, since a) this verb has three supposedly empty prefixes in the partner perfective verbs загрузить, нагрузить, and погрузить all meaning 'load (perfective)'; and b) all four verbs (imperfective грузить and all three perfectives) can appear in two competing constructions, the theme-object construction in грузить ящики на телегу 'load boxes onto the cart', and the goal-object construction грузить телегу ящиками 'load the cart with boxes'.

The point is to show that the prefixes provide different environments for the constructions and because prefixes do not behave identically they are therefore not identical in function or meaning. We discover that нагрузить strongly prefers the goal-object construction, погрузить almost exclusively prefers the theme-object construction, whereas загрузить has a more balanced distribution. Thus one can say that each prefix has a unique characteristic preference pattern. Our analysis shows that this is a robust finding, even when we take into account relevant additional environmental variation, namely the use of the prefixes in constructions with passive participles, as in Ирина Владимировна шла нагружённая сумками и сумочками 'Irina Vladimirovna walked along, loaded with bags and pouches', and the use of reduced constructions where one of the participants is missing, as in Мужики грузили лес и камень 'The men loaded timber and rock' (where the goal argument is not mentioned).

Table 1 provides a description of the dataset. The aim of a statistical model for this dataset is to predict the CONSTRUCTION based on the predictors VERB, REDUCED, and PARTICIPLE. This prediction can be modeled using all three kinds of models considered here: logistic regression, tree & forest, and naive discriminative learning.

Dataset and R script:	datLOAD.csv; LOAD.R
Source of dataset:	Russian National Corpus (www.ruscorpora.ru)
Size of dataset:	1920 rows (observations), each representing an example sentence containing грузить, нагрузить, загрузить or погрузить ‘load’
Rival forms:	theme-object construction vs. goal-object construction, represented as CONSTRUCTION with values: theme, goal
Predictors:	
VERB:	zero (for the unprefixed verb грузить ‘load’), na-, za-, and po-
REDUCED:	yes (construction is reduced) or no (full construction)
PARTICIPLE:	yes (passive participle) or no (active form)

Table 1: Description of the Грузить ‘load’ dataset

3.1.1 Logistic regression

The optimal logistic regression model for this dataset includes all three predictors as main effects, plus an interaction between the verb and participle predictors.[3] The formula for this model is (the asterisk "*" tells R to include not only VERB and PARTICIPLE as main effects, but also their interaction)[4]:

```
CONSTRUCTION ~ VERB + REDUCED + PARTICIPLE + VERB*PARTICIPLE
```

The linear model yields the estimates for the coefficients shown in Table 2. This table may seem rather daunting, but the basic ideas underlying these numbers are straightforward. The first column, labeled 'Estimate', presents the estimated coefficient. To interpret the values of the coefficients, recall that a logistic model estimates how the log of the odds ratio depends on the predictors. For an odds ratio, we need to know what R considers to be a success and what it takes to be a failure. By default, R will order the levels of the response alphabetically, and take the second one to be a success. For the present data, this means that the theme construction is a success, and that

	Estimate	Std. Error	Wald Z	p-value
Intercept	-0.946	0.202	-4.679	0.0000
VERB=po	6.714	1.022	6.570	0.0000
VERB=za	1.092	0.245	4.455	0.0000
VERB=zero	2.334	0.245	9.539	0.0000
PARTICIPLE=yes	-4.186	1.022	-4.096	0.0000
REDUCED=yes	-0.889	0.175	-5.085	0.0000
VERB=po, PARTICIPLE=yes	3.895	1.598	2.438	0.0148
VERB=za, PARTICIPLE=yes	1.409	1.077	1.308	0.1910
VERB=zero, PARTICIPLE=yes	-1.772	1.441	-1.229	0.2190

Table 2: Coefficients for logistic regression model of LOAD data

[3] This logistic regression model is also presented in Sokolova et al. forthcoming.

[4] Note that because any predictor that is present in an interaction is also automatically considered as a main effect, this formula can be rendered more succinctly as: `CONSTRUCTION ~ VERB*PARTICIPLE + REDUCED`. The LOAD.R script tracks how this formula was arrived at through successive iterations, gradually increasing the number of predictors and comparing the results. Further interactions were not found to be statistically significant.

the model is ascertaining how the log of the number of theme constructions divided by the number of goal constructions depends on the predictors.

The list of estimates for the coefficients begins at the Intercept. The way in which R by default deals with factors is to take one factor level as point of reference. For this particular factor level, e.g., `no` for the factor REDUCED, the group mean is calculated. For the other factor level (yes), the difference between its group mean and the group mean for `no` (the reference level) is calculated. All group means are on the logit scale.

R chooses as values at the Intercept those that come first alphabetically (unless the user specifies otherwise). Thus the Intercept here involves these values for the three predictors: VERB=na, PARTICIPLE=no, REDUCED=no. The intercept has the value -0.9465, indicating that for the subset of data for which VERB=na, PARTICIPLE=no, and REDUCED=no, the theme construction is used less often than the goal construction (the odds ratio is less then one, and the log of a number between 0 and 1 is negative). When we change to another group mean, for VERB=na, PARTICIPLE=no, and REDUCED=yes, the group mean is $-0.9465 - 0.8891 = -1.8356$, indicating that for REDUCED observations, the theme construction is an even smaller minority.

The interpretation of VERB and PARTICIPLE requires special attention, because these two predictors enter into an interaction. The interaction introduces additional adjustments that have to be applied when the factors involved in the interaction both have values that differ from the reference values. The eight group means can be constructed from the estimates of the coefficients as follows:

VERB=na, PARTICIPLE=no:	-0.9465
VERB=po, PARTICIPLE=no:	-0.9465+6.7143
VERB=za, PARTICIPLE=no:	-0.9465+1.0920
VERB=zero, PARTICIPLE=no:	-0.9465+2.3336
VERB=na, PARTICIPLE=yes:	-0.9465 - 4.1862
8> VERB=po, PARTICIPLE=yes:	-0.9465+6.7143+3.8913-4.1862
VERB=za, PARTICIPLE =yes:	-0.9465+1.0920+1.4087-4.1862
VERB=zero, PARTICIPLE=yes:	-0.9465+2.3336-1.7717-4.1862

Thus, for VERB=zero, PARTICIPLE=yes, REDUCED=no, the model predicts a log odds ratio equal to -4.5708, which converts (with the plogis function) to a proportion of 0.0102. This compares well with the observed counts, 90 for goal and 1 for theme (proportion for theme: 0.0110).

The second column in Table 2 presents a measure of how uncertain the model is about the estimate for the coefficient. The greater this measure, the standard error, the more we should be on guard. The third column is obtained by taking the values in the first column and dividing them by the values in the second column, resulting in so-called Z scores. These Z scores follow a standard normal distribution, and the final column with p-values presents a measure of how surprised we should be that the scores are as big as they are. More specifically, p-values evaluate how surprised we should be to observe a coefficient with as large (or as small, when negative) a value as actually observed, where we evaluate surprise against the possibility that the predictor is not associated with the response at all, i.e., that the values of the predictors and the response are random. The standard cutoff for recognizing statistical significance in our field is $p = 0.05$, but it should be kept in mind that for large datasets, and for data with much better experimental control than we usually have in language studies, the cutoff-value can be set much lower. The values for the first six lines in the table are all < 0.0001. For the intercept, the small p-value indicates that the group mean for VERB=na, REDUCED=no, PARTICIPLE=no has a log odds that is significantly below 0. Translated into proportions, this means that the proportion of the theme construction is significantly below 50%. For the other terms with small p-values, we have good evidence that the differences in group means are significant.

The interaction of VERB and PARTICIPLE gets lower marks, since only one of the three coefficients has a p-value below 0.05. This raises the question of whether the interaction is really needed. The problem here requires some care. The table of coefficients only lists three corrections on differences between group means (the interaction terms), while there are in all $\binom{4}{2} = 6$ pairwise comparisons (e.g., VERB=po versus VERB=zero is missing). As a consequence, we may be missing out on the most striking group difference. Furthermore, when multiple coefficients are evaluated with p-values, there is an increased probability of getting a low p-value by chance. This can be corrected for by applying the Bonferroni correction, which works as follows for the present example. We have 3 coefficients for the interaction, and our significance level (alpha) is 0.05. We divide alpha by 3, resulting in 0.0167. Any coefficient with a p-value less then 0.0167 is certain to be significant. So we now know that the interaction captures at least one significant contrast.

A second way of evaluating the interaction is to compare a model without the interaction with a model that includes the interaction. We can do this with an analysis of deviance test, which will evaluate whether the extra coefficients required for the interaction buy us a better fit to the data. In fact, we can apply this approach to a sequence of models, each one having one more predictor than the previous one. If we start with a model with just an intercept (the grand mean, model 1), and then add in first VERB, then PARTICIPLE, then REDUCED, and finally the interaction of VERB by PARTICIPLE (model 5), we obtain Table 3.

	Resid. Dev	Df	Deviance	p-value	Reduction in AIC
Intercept	2645.16				
Verb	1305.31	3	1339.85	0.0000	1333.8
Participle	950.73	1	354.58	0.0000	352.6
Verb:Participle	933.48	3	17.25	0.0006	11.2
Reduced	906.69	1	26.80	0.0000	24.8

Table 3: Model comparison statistics for the LOAD data

The column named Resid. Dev lists the residual deviance, the unexplained variation in the data. As we include more predictors, the residual deviance decreases. The column labeled Df specifies how many coefficients were required to bring the residual deviance down. How much the deviance was reduced is given by the column labeled Deviance. The column with p-values shows that each reduction in deviance is significant. Finally, the last column lists the reduction in Akaike's information criterion (AIC), a measure of goodness of fit that punishes models for having many coefficients. The reduction in AIC accomplished by a predictor is an excellent guide to its importance. Here, we see that VERB is most important, followed by PARTICIPLE, followed by REDUCTION, followed by the interaction of VERB by PARTICIPLE.

The C value (concordance index; this is one of the statistics yielded by the logistic regression — see the R code and output on laura's-url) of 0.96 tells us that the fit of the model is excellent. The accuracy of the model is 89%, where we judge the model to make a correct prediction if the estimated probability for the theme construction is greater than or equal to 0.5 and the theme construction was actually observed.

3.1.2 Tree & Forest

The tree & forest analysis gives entirely parallel results. Here our formula is:

```
CONSTRUCTION ~ VERB + REDUCED + PARTICIPLE
```

In tree & forest analysis we can skip the tedium of testing different model equations. We don't have to worry about how many predictors we put in, nor do we have to specify interactions. Both the classification tree and the classification forest will eliminate any predictors that are not significant and interactions are taken into account automatically, as described below.

Figure 1 summarizes graphically the results of the recursive partitioning tree. The first split is on VERB, distinguishing po (for which the theme is almost always used) from the other three cases for which the theme is less probable. The p-value in the oval presents a measure of surprise for how well separable the theme and goal realizations are given information about the level of VERB. The algorithm considers all possible splits, not only for VERB, but also for PARTICIPLE and REDUCED, and chooses the predictor (and the combination of levels of that predictor) that separates the theme and goal constructions best. The choice of the best splitting criterion is made locally. The algorithm does not look ahead to see whether an initial less good split might be offset by later greater gains. As a consequence, the predictor providing the first split often is one of the most important predictors, but it is not necessarily true that it is the most important predictor.

Once a split has been made, the same procedure (finding the locally best splitting criterion, if any) is applied to both subsets (in the present case, po versus na, za, zero). In this way, the dataset is recursively partitioned into increasingly smaller subsets that are more homogeneous with respect to the choice between theme and goal. If we go to the right branch of the tree and look for the strongest factor within that branch, which is REDUCED (also with $p < 0.001$), we find a split with yes on the right and no on the left. Within these new subsets, further significant splits are not detected, which is not surprising as choice behavior is nearly categorical here. In the left branch of the tree, further splits are made on PARTICIPLE, followed by VERB and REDUCED. The algorithm stops partitioning either when there is no further gain in separability or when there are too few data points to allow for a meaningful split.

The bargraph below each terminal node represents the percentage of goal (light grey) vs. theme (dark grey) outcomes, and "n = " indicates the total

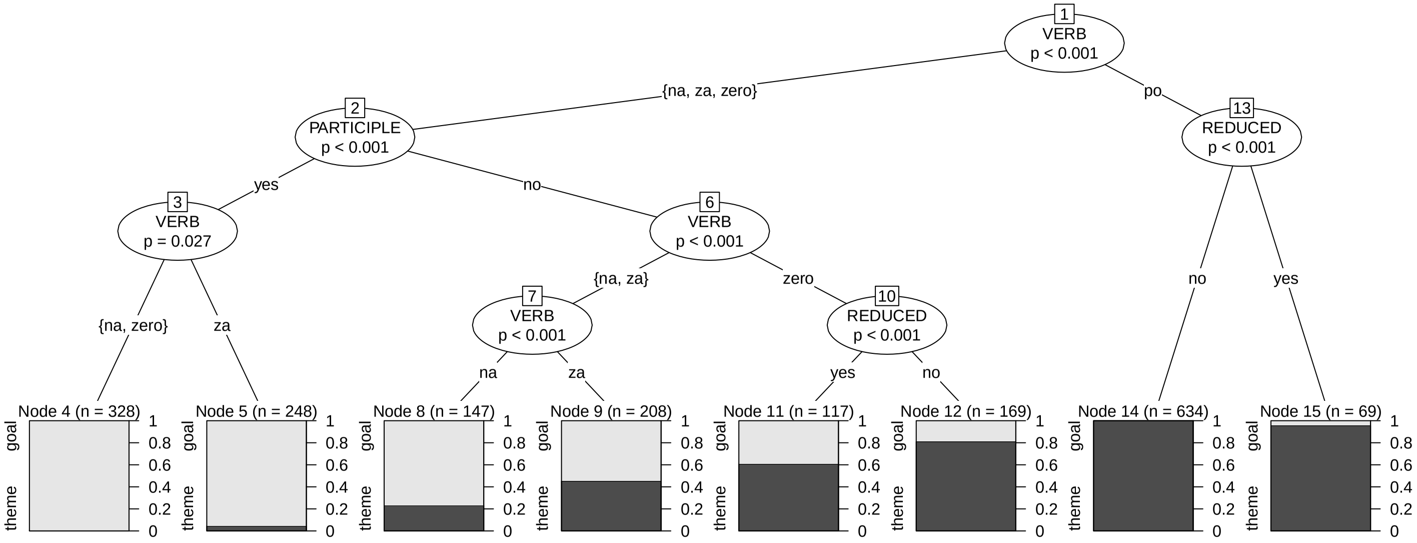


Figure 1: Recursive partitioning tree for the LOAD data.

number of datapoints in that node. So, for example, node 4 contains all of the examples that involve a (past passive) participle form of either нагрузить or грузить; there are 328 examples of that type, and 326 (99.4%) of those have the goal construction, whereas 2 (0.6%) have the theme construction. To take another example, Node 9 shows us the results for active forms of загрузить: there are 208 such examples, of which 114 (54.8%) have the goal construction, but 94 (45.2%) have the theme construction.

In a classification tree we see an interaction any time that the left branch of the tree is different from the right branch, and/or the barplots below the terminal nodes are showing different patterns. Therefore, the classification tree shows us that there is in fact a complex interaction among the three factors. Within the framework of a logistic regression model, one would have to include a VERB by REDUCED by PARTICIPLE interaction, which would result in a large number of coefficients and no noticeable improvement in goodness of fit. A classification tree makes no statement about main effects, i.e., it does not provide information about the effect of a given predictor with all other predictors held constant. For such global statements, a logistic model should be used. This having been said, it is clear that the classification tree gives us a description of what is going on in the data, in a way that is visually much more tractable and intuitive than the tables of figures we receive as output in the regression model.

However, a classification tree makes its splits based on local best performance, as mentioned above. Working with look-aheads would make the procedure computationally intractable. In order to obtain a tree-based model that avoids the risk of overfitting due to local optimization, it is useful to complement the classification tree with a random forest. The random forest technique constructs a large number of bootstrap samples and builds a recursive partitioning tree for each of them. In order to obtain predictions from this forest of trees, votes are collected from the individual trees on what they, based on their training data, believe the response (e.g., goal versus theme construction) to be. Typically, a random forest makes more precise predictions than a standard classification tree. For the present example, the tree has a classification accuracy of 0.88%, and the forest's accuracy increases, rather atypically, only slightly to 0.89%. For both, $C = 0.96$.

The forest of trees does not provide useful information about how the predictors work together. For that, we have to let ourselves be guided by the classification tree. The forest does provide us with a means for assessing the relative importance of the different predictors in the model. It assesses the importance of a predictor, say, VERB, by randomly permuting the values of VERB (na, po, za, zero) so that the relation between VERB and construction is destroyed. If a predictor is truly associated with the response (theme versus goal), then this procedure will cause the classification accuracy of the tree to plummet. If a predictor is not predictive at all, permuting it shouldn't matter, and classification accuracy should stay about the same. A measure of variable importance can therefore be defined as the reduction in classification accuracy under random permutation.

For the present data, the variable importances are 0.003 for REDUCED, 0.073 for PARTICIPLE, and 0.338 for VERB. VERB is the strongest predictor, since a model excluding VERB is 33.8% worse than one that includes it. PARTICIPLE comes next, and its removal damages the model by 7.3%. Least important is REDUCED, with a value of only 0.3%. In comparison with the regression model, the random forest gives us comparable values for concordance, with $C = 0.96$, and an accuracy of 0.89%.

Trees & Forest is often an excellent choice for data with factors with few factor levels. When the number of factor levels becomes large (e.g., a factor VERB with 20 different verbs) and especially when there is more than one factor with many factor levels, the technique becomes computationally intractable. For such datasets, mixed logistic regression is the best choice, see section 3.3 for an example.

3.1.3 Naive discriminative learning

Naive discriminative learning can also be used as a classifier for the present dataset. Once again our formula is simply:

$$\texttt{CONSTRUCTION} \sim \texttt{VERB + REDUCED + PARTICIPLE}$$

The naive discriminative learning model yields a matrix of the weights that quantify how strongly the different predictor values are associated with the rival forms goal and theme, presented here in Table 4.

Let's see how to read this table by considering the configuration of predictors VERB=na, PARTICIPLE=no and REDUCED=no. The support for the theme construction is obtained simply by summing the relevant entries in Table 4: -0.25 +0.32+0.22 = 0.29. The support for the goal construction is 0.45 + 0.08 + 0.18 = 0.71. The proportional support for the theme is therefore 0.29/(0.29+0.71) = 0.29. If we look at the data, we find that for this cell of the design, 27 observations support the theme, and 70 the goal,

	goal	theme
PARTICIPLE=no	0.0794	0.3206
PARTICIPLE=yes	0.3590	0.0410
REDUCED=no	0.1757	0.2243
REDUCED=yes	0.2627	0.1373
VERB=na	0.4498	-0.2498
VERB=po	-0.4379	0.6379
VERB=za	0.3189	-0.1189
VERB=zero	0.1076	0.0924

Table 4: NDL weights for the LOAD data

i.e., 28%. This fits well with the proportion predicted by naive discriminative learning (29%). For any other combination of predictors and their values, the calculations proceed in exactly the same way.

From a cognitive processing perspective, the idea is that given a set of cues (VERB=na, PARTICIPLE=no, REDUCED=no), activation propagates over the connections of these cues with the outcomes (the goal and theme constructions). The extent to which a given outcome becomes active is given simply by the sum of the weights on the connections from the active cues to each construction. The construction that receives most support is then the most likely one to be used.

To assess how important a predictor is in our NDL model, we can take the sum of the absolute differences of the relevant weights (for PARTICIPLE: $|0.08 - 0.32| + |0.35 - 0.04| = 0.56$). The resulting values correlate extremely well with the variable importance as assessed by the random forest ($r = 0.9998$). Again, VERB is by far the most important factor, followed by PARTICIPLE, followed by REDUCED. In other words, we get the same results as in both the logistic regression and the tree & forest analyses. The evaluation of the naive discriminative learning model is also comparable, since it provides an excellent fit with $C = 0.96$ and 0.88% accuracy, and these figures remain unchanged under ten-fold cross-validation. This example illustrates that, under ideal learning conditions, human learning and statistical learning can produce nearly identical results.

It should be noted, however, that naive discriminative learning does not supply p-values of any kind. It finds a set of weights that allow it to make excellent predictions given the corpus data on which it is trained.

For ascertaining whether a predictor is statistically significant, the reader is advised to use logistic regression or a conditional inference tree.

3.2 Пере- vs. пре-

This case study addresses the question of whether the variants represent one morpheme or two. Пере- vs. пре- are etymologically related prefixes, but their history and behavior are quite different.[5] In this case пере- is the native Russian variant, whereas пре- is a Church Slavonic borrowing (Vasmer 1971: Vol. 3, 356). Пере- has received much more attention in the scholarly literature (Janda 1986: 134-173; Flier 1985; Dobrušina & Paillard 2001: 76-80; Shull 2003: 113–119). Пре-, by contrast, is normally mentioned only as a Church Slavonic variant (Townsend 2008: 59; 128; but see Soudakoff 1975 who argues that пере- and пре- should be considered distinct morphemes).

Our data explore variation both in terms of meaning and environment, but we consistently find tendencies rather than hard-and-fast rules for the distribution of forms. For example, пере- is usually preferred to express spatial 'transfer', as in перевести 'lead across', whereas пре- predominates in other meanings such as 'superiority', as in преобладать 'predominate', but counterexamples for this tendency are found (препроводить 'convey' as an example of a spatial 'transfer' use for пре- and перекричать 'outshout' as an example of 'superiority' with пере-). In terms of environment, the most salient tendencies involve a situation in which there is either prefix stacking or +/- shift in aspect. Prefix stacking occurs when a given verb contains more than one prefix, and here пре- is more common, as in превознести 'extol' and преподнести 'present with', however examples with пере- are also found, as in переизбрать 're-elect' and перенаселить 'overpopulate'. Whereas all prefixes are strongly associated with marking perfective aspect, and thus typically serve to shift the aspect of imperfective base verbs to perfective, пре- commonly fails to effect this shift, as in преследовать 'persecute' (an imperfective verb built from the imperfective base следовать 'follow'). However, пере- can also fail to shift aspect, as in переменять 'change' (imperfective from imperfective base verb менять 'change'),[6] and there are also examples where both пере- and пре- serve the usual role of perfectivizers,

[5] Note that although these prefixes can be added to adjectives and adverbs, this case study focuses exclusively on their use with verbs.

[6] An alternative interpretation is available for this example, since переменять is also the secondary imperfective of переменить 'change'.

as in перетерпеть ‘overcome’ and претерпеть ‘undergo, endure’ which are both perfective verbs from the imperfective терпеть ‘suffer’. Our analysis reveals the various strengths of the semantic and environmental factors associated with пере- vs. пре- in Russian verbs.

Table 5 provides a description of the dataset. The aim of a statistical model for this dataset is to predict the Prefix from the predictors. There are two things to note about the PERE dataset that distinguish it from the LOAD dataset: 1) this data has a strongly unbalanced distribution, with 1727 examples of пере-, but only 107 examples of пре-; and 2) this dataset includes frequency, which is a numerical, quantitative predictor, as opposed to the other predictors, which are factorial (categorical, or qualitative) predictors (which have discrete levels such as yes vs. no or not stacked vs. stacked).

3.2.1 Logistic regression

The optimal model for this dataset is captured by the following regression equation, which has simple main effects only:

$$\texttt{Prefix} \sim \texttt{ShiftTrans + PrefixStacking + ShiftAspect +}$$
$$\texttt{SemanticGroup + LogFreqPrefVerb}$$

This model specification yields a very large table of coefficients (see Table 6), a straightforward consequence of the large number of levels of the factor SemanticGroup. With the large number of factor levels in this dataset, the table of coefficients becomes less informative. Many of the differences in the group means for different values of ShiftAspect and SemanticGroup are not listed in the table. Two effects are easy to interpret, however. First, the probability of пре- increases with prefixstacking, and second, this probability increases with the frequency of the prefixed verb: In Table 6, both predictors are paired with a positive and significant estimate.

Rather than going through all the contrasts listed in the table of coefficients, we move on to assess the importance of the different predictors, we therefore compare a sequence of nested models, beginning with a model with an intercept only (the grand mean), to which we add successively the predictors ShiftTrans, PrefixStacking, ShiftAspect, SemanticGroup, and LogFreqPrefVerb in this order. The result is shown in Table 7, from which we can read off that Semantic Group is the most important predictor, and ShiftTrans the least important. The classification accuracy of this model is 96%, the index of concordance is $C = 0.95$.

Dataset and R script:	datPERE.csv; PERE.R
Source of dataset:	Russian National Corpus (www.ruscorpora.ru)
Size of dataset:	1836 rows, each representing a verb prefixed by either пере- or пре- that is attested at least once in the Russian National Corpus
Rival forms:	пере- vs. пре-, represented as Prefix with values: pere, pre
Predictors:	
ShiftTrans	comparison of transitivity of base verb and prefixed verb, where “intr” = intransitive, “tr” = transitive, “no” = no existing base verb: intr-intr, intr-tr, no-intr, no-tr, tr-intr, tr-tr
PrefixStacking:	not stacked, stacked
ShiftAspect	comparison of aspect of base verb and prefixed verb, where “imp” = imperfective, “pf” = perfective, “no” = no existing base verb: imp-pf, imp-imp, pf-pf, no-imp, no-pf
FreqBase	frequency of the base verb in the RNC: ranges from 0 to 2694330; this parameter is also available in log-transferred form as LogFreqBase. Frequency distributions have long tails, and without a logarithmic transformation, the highest-frequency words become atypical outliers that may completely distort logistic regression models
FreqPrefVerb	frequency of the prefixed verb in the RNC: ranges from 1 to 34992; this parameter is also available in log-transferred form as LogFreqPrefVerb
PerfectiveType	natural, specialized, not applicable (for imperfective) (cf. Janda 2007 for types of perfectives)
SemanticGroup	meaning of the prefix (cf. Endresen forthcoming and http://emptyprefixes.uit.no/pere_eng.htm): bridge, divide, interchange, mix, overcome-duration, overdo, redo, seriatim, superiority, thorough, transfer, transfer metaphorical, turn over, very (Note: These are the full names as listed under SemanticGroupFullName; in SemanticGroup they are abbreviated)

Table 5: Description of the пере- vs. пре- dataset

	Estimate	Std. Error	Wald Z	p-value
(Intercept)	-2.056	0.683	-3.011	0.0026
ShiftTrans=intr-tr	-0.841	0.615	-1.368	0.1712
ShiftTrans=no-intr	18.152	3540.605	0.005	0.9959
ShiftTrans=no-tr	17.103	3540.605	0.005	0.9961
ShiftTrans=tr-intr	-0.209	0.857	-0.243	0.8077
ShiftTrans=tr-tr	-0.649	0.347	-1.867	0.0619
PrefixStacking=stacked	2.755	0.490	5.620	0.0000
ShiftAspect=imp-pf	-1.485	0.409	-3.634	0.0003
ShiftAspect=no-imp	-20.160	3540.605	-0.006	0.9955
ShiftAspect=no-pf	-18.922	3540.605	-0.005	0.9957
ShiftAspect=pf-pf	-0.612	0.406	-1.507	0.1318
SemanticGroup=div	0.229	0.609	0.377	0.7062
SemanticGroup=intrch	-1.828	0.801	-2.281	0.0225
SemanticGroup=mix	-19.119	4435.633	-0.004	0.9966
SemanticGroup=ovc-dur	-0.795	0.676	-1.175	0.2402
SemanticGroup=overdo	-3.073	0.728	-4.221	0.0000
SemanticGroup=redo	-21.413	1189.419	-0.018	0.9856
SemanticGroup=seria	-19.398	1816.033	-0.011	0.9915
SemanticGroup=super	-0.110	0.690	-0.159	0.8737
SemanticGroup=thorough	-19.391	4849.044	-0.004	0.9968
SemanticGroup=transf	-2.367	0.631	-3.751	0.0002
SemanticGroup=transf-met	0.342	0.547	0.625	0.5318
SemanticGroup=turn	-19.671	5120.003	-0.004	0.9969
SemanticGroup=very	20.187	7565.807	0.003	0.9979
LogFreqPrefVerb	0.360	0.063	5.690	0.0000

Table 6: Coefficients for logistic regression model of the пере- vs. пре- dataset

Interpreting the model using the table of coefficients is difficult, especially because various predictors have many factor levels. One option for further analysis is to simplify a predictor such as SemanticGroup, by collapsing similar levels. However, often the categorization into many factor levels is well motivated, and we therefore now consider the tree & forest method, which provides a simpler guide to the interpretation of the data.

	Resid. Dev	Df	Deviance	p-value	AIC
Intercept	815.70				
ShiftTrans	789.17	5	26.53	0.0001	16.5
PrefixStacking	739.16	1	50.01	0.0000	48.0
ShiftAspect	694.90	4	44.26	0.0000	36.3
SemanticGroup	415.90	13	279.00	0.0000	253.0
LogFreqPrefVerb	379.56	1	36.34	0.0000	34.3

Table 7: Model comparison statistics for the Пере- vs. пре- dataset

3.2.2 Tree & forest

The formula for this analysis is nearly the same as the one for the logistic regression, but it is not necessary (although not harmful either) to log-transform the frequency counts for the base verb and the prefixed verb. Furthermore, we include Perfective Type as a predictor. In the logistic regression, Perfective Type failed to reach significance, and we therefore do not expect to see it emerge in the classification tree.

$$\texttt{Prefix} \sim \texttt{ShiftTrans + PrefixStacking + ShiftAspect +}$$
$$\texttt{PerfectiveType + SemanticGroup + FreqBase + FreqPrefVerb}$$

The recursive partitioning algorithm yields the classification tree shown in Figure 2, and the random forest works out the following variable importances: PerfectiveType: 0.0002, ShiftTrans: 0.0002, FreqBase: 0.0006, FreqPrefVerb: 0.0030, ShiftAspect: 0.0131, PrefixStacking: 0.0175, SemanticGroup: 0.0380.

Notice first of all that the classification tree does not include all of the predictors that appear in the formula: it retains SemanticGroup, PrefixStacking, ShiftAspect, FreqPrefVerb and FreqBase, but excludes ShiftTrans and PerfectiveType. This fits well with the results of the logistic regression, which did not support PerfectiveType at all, and which revealed ShiftTrans to be the least important predictor. As promised above, the classification tree can decide on its own which variables are important and which are not, and it simply ignores the ones that are not important. The variable importance according to the random forest is in agreement with the ranking of variable importance based on the reduction in AIC for the logistic model. Interestingly, the classification forest outperforms the logistic regression model: C =0.98 and accuracy = 96%.

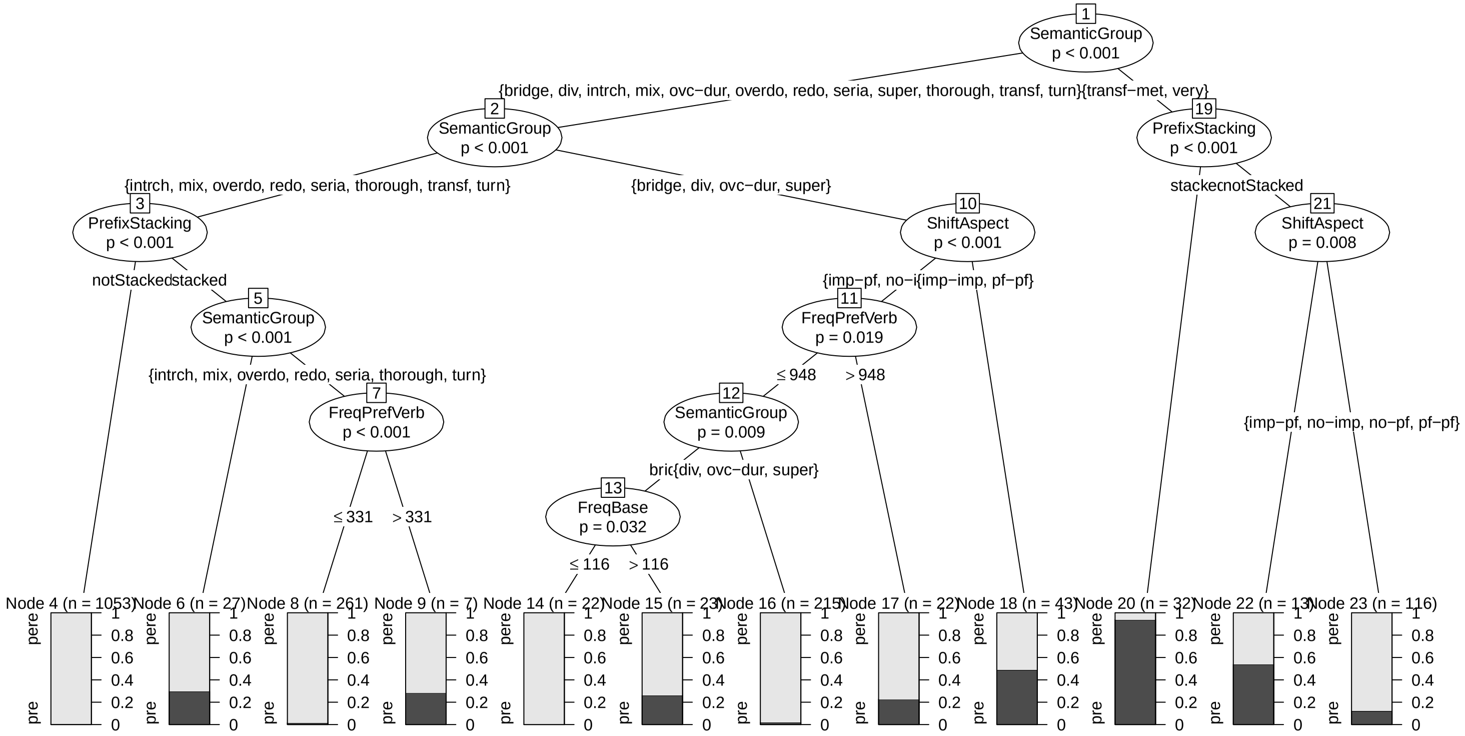


Figure 2: Recursive partitioning tree for the Пере- vs. пре- dataset.

The classification tree guides us towards a more complex interpretation of the data than the logistic regression model, which only detected simple main effects. From Figure 2 we can read off, for instance, that for verbs from the transf-met and very semantic groups, пре- is used almost exclusively when there is no prefix stacking.

3.2.3 Naive discriminative learning

The observations in this dataset are a sample of the experience that the average language user has with the contexts in which the choice between the rival forms пере- vs. пре- arises. Therefore, naive discriminative learning is an appropriate model for this dataset. We are interested in whether naive discriminative learning also provides a good fit to the data, for two reasons. First, if the model provides a good fit, it provides an explanation for how language users, immersed in an environment from which the corpus data are sampled, implicitly absorb and internalize the quantitative forces shaping the use of пере- vs. пре-. Second, the tighter the fit of the model to the data, the more stable we may expect the system to be.

The пере- vs. пре- data are especially interesting from a learning perspective because these data provide information on the frequency with which forms are used. In random forest and logistic regression analyses, as described above, this frequency is taken into account as a property of a given data point, along with other properties such as shifts in aspect or transitivity. Within the naive discriminative learning approach, the frequency of the derived word is not taken into account as a word property, but rather as part of the learning experience. The equilibrium equations that define the weights are calculated from the co-occurrence frequencies of the word's properties. The frequencies of the derived words codetermine these co-occurrence frequencies, and hence are taken into account for the estimation of the model's weights. Predictions of which prefix is most appropriate are derived from the weights on the links from a word's properties (such as aspect or transitivity shifting) to the prefix allomorph.

The model's classification performance, as estimated by the index of concordance C, is 0.97, and its accuracy is at 94%. Under cross-validation, these values decrease to 0.87 and 84% respectively. It should be noted, however, that with 107 rows in the dataset (out of 1834, so 6%), which account for 16% of the occurrences of пере- (649757) vs. пре- (125668), data on пре- are sparse and as a consequence, crucial information about this suffix

	pere	pre
PerfectiveType=natural	0.243	0.019
PerfectiveType=not-applicable	0.274	-0.012
PerfectiveType=specialized	0.025	0.238
PrefixStacking=notStacked	0.438	-0.045
PrefixStacking=stacked	0.104	0.289
SemanticGroup=bridge	0.081	-0.025
SemanticGroup=div	-0.099	0.155
SemanticGroup=intrch	0.192	-0.135
SemanticGroup=mix	0.160	-0.103
SemanticGroup=ovc-dur	0.104	-0.048
SemanticGroup=overdo	0.135	-0.079
SemanticGroup=redo	0.219	-0.163
SemanticGroup=seria	0.175	-0.119
SemanticGroup=super	-0.333	0.389
SemanticGroup=thorough	0.189	-0.133
SemanticGroup=transf	0.218	-0.162
SemanticGroup=transf-met	-0.285	0.341
SemanticGroup=turn	0.189	-0.133
SemanticGroup=very	-0.403	0.459
ShiftAspect=imp-imp	-0.153	0.310
ShiftAspect=imp-pf	0.270	-0.113
ShiftAspect=no-imp	0.013	0.144
ShiftAspect=no-pf	0.222	-0.065
ShiftAspect=pf-pf	0.190	-0.032
ShiftTrans=intr-intr	0.083	0.048
ShiftTrans=intr-tr	0.121	0.010
ShiftTrans=no-intr	0.135	-0.004
ShiftTrans=no-tr	0.105	0.026
ShiftTrans=tr-intr	0.002	0.129
ShiftTrans=tr-tr	0.096	0.035

Table 8: NDL weights for the пере- vs. пре- dataset.

will often be lost in the training sets. Similarly, particular factor levels may not have been realized in an in-bag training set, which has as its consequence that the model has to ignore such ‘unseen’ factor levels altogether.

When we assess variable importance according to NDL, we obtain the following ranking: ShiftTrans: 0.55, PrefixStacking: 0.67, PerfectiveType:

0.72, ShiftAspect: 1.49, SemanticClass: 5.22, which hardly differs from the ranking suggested by the reduction in AIC for the logistic model, as illustrated in Figure 3. What this figure shows very clearly is that the most important predictor is semantic group.

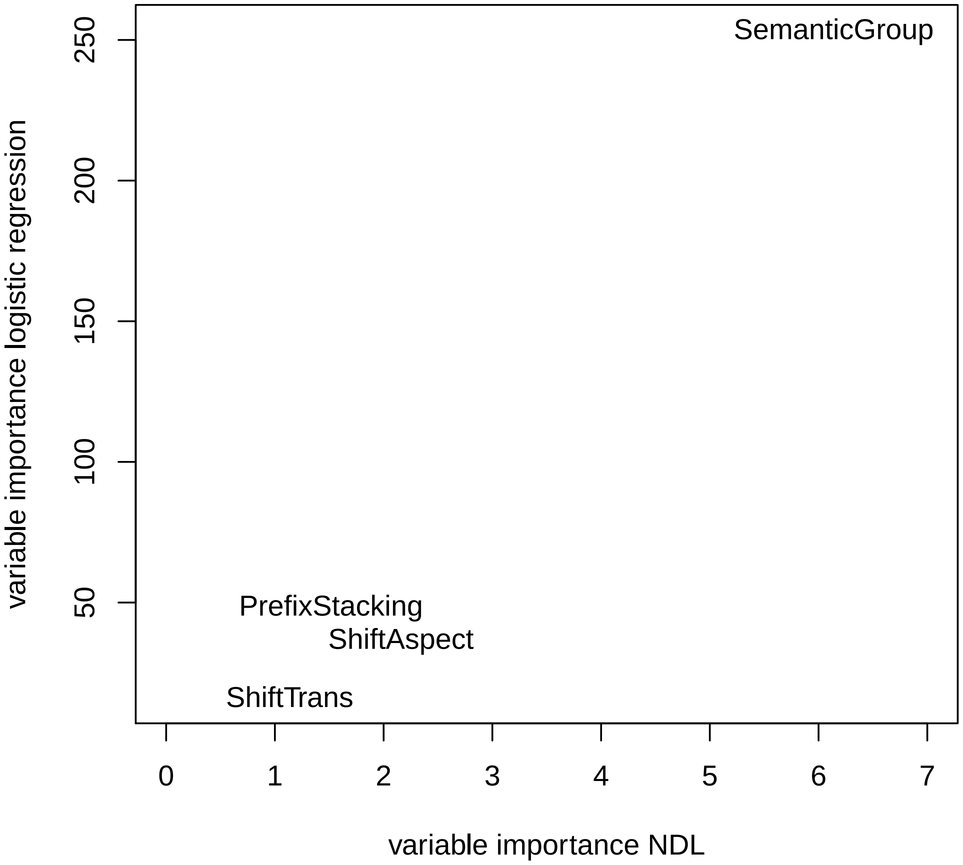


Figure 3: Variable importance according to the logistic regression model and according to naive discriminative learning for the пере- vs. пре- dataset.

To conclude, let's consider again how frequency of occurrence is used by the logistic regression and the classification tree on the one hand, and by naive discriminative learning on the other. The logistic regression tells us that if a prefixed verb has a higher frequency, it is more likely to find пре- than пере-. This is useful information, but unless one believes that speakers have counters in their heads that keep track of how often specific forms have been used, it is information at a high level of abstraction. By contrast, the NDL model undergoes as it were the frequencies with which verbs and their distributional properties occur, and derives its predictions from the resulting discrimination weights. It is conceivable, but at present far from certain, that the naive discrimination model provides a cognitively more plausible assessment of the usage of пере- and пре-.

3.3 O- vs. об-

The objective of this section is to address the controversy concerning the status of o- vs. об- as either a single morpheme or two separate ones. The etymologically related variants o- vs. об- show a complex relationship involving a variety of both semantic and phonological environments (in addition to the phonologically conditioned обо-). While many standard reference works (Zaliznjak & Šmelev 1997: 73; Zaliznjak & Šmelev 2000: 83; Wade 1992: 277; Timberlake 2004: 404; Townsend 1975: 127; Grammatika russkogo jazyka 1952: Vol. 1 589–592; Isačenko 1960: 148), plus several specialized works (Barykina, Dobrovol'skaja, Merzon 1989; Hougaard 1973, and Roberts 1981) treat o- and об- as allomorphs of a single morpheme, some scholars (Alekseeva 1978, Andrews 1984 and Krongauz 1998: 131–148) argue that they have split into two separate morphemes that just happen to share the same forms.

The controversy is well motivated, since the behavior of o- vs. об- covers a large portion of the space depicted in Figure 1. We saw already in the use of остричь vs. обстричь 'cut' that the two variants can sometimes be identical in terms of both meaning and environment. Additionally one can argue on the basis of examples like окружить 'surround' vs. объехать 'ride around' that o- vs. об- are classic allomorphs expressing the same meaning in phonologically complementary (non-sonorant root onset vs. sonorant root onset) environments. However, o- vs. об- can also express a range of meanings: in addition to a meaning that can be captioned as 'around', as in the examples above, there are also so-called factitive uses built from adjectives meaning 'make something be Y' (where Y is the meaning of the base adjective or noun), as in осложнить 'make complicated' (from сложный 'complicated') and обновить ‘renew’ (from новый ‘new’); and these two verbs additionally suggest that phonology is decisive, again with о- associated with a non-sonorant vs. об- associated with a sonorant. However these examples give a mistaken impression: phonology is not an isolated or deciding factor, as we see in онемечить ‘germanify’ (a factitive verb from немецкий ‘German’) which combines о- with a sonorant onset, nor in обгладить ‘smooth’ (a factitive verb from гладкий ‘smooth’) and in обскакать ‘gallop around’, both of which combine об- with a non-sonorant. We thus see a diverse collection of possibilities with the factors of both meaning and environment ranging from “same” to various degrees of “different”. Additionally there is a semantic continuum between ‘around’ and the factitive type, since there are verbs like окольцевать ‘encircle’ that combine the two meanings (which can be interpreted as both a spatial sense of ‘around’ and as a factitive from кольцо ‘ring’). Since existing verbs and corpus data limit our opportunity to study the effects of various factors on the choice of о- vs. об-, we present an experiment using nonce words, which give us more control over the factors. Our analysis addresses differences in meaning and differences in environment, as well as individual preferences of subjects and stems.

The aim of the analysis of this dataset is to predict the choice between о- vs. об-. There is one feature that is relevant only to part of the data: The nonce verbs were presented both as stem-stressed and as suffix-stressed, whereas the nonce adjectives were all stem-stressed. Here, we focus on the subset of the data where stress varies, i.e., the verb data.

This dataset has a feature that we haven’t seen in the previous analyses. In addition to comprising both quantitative (Age) and qualitative (e.g., Manner) predictors, the dataset has two predictors that have large numbers of levels: Stem (46) and Subject (60). For predictors with so many levels, it does not make sense to treat them as standard factors, which typically have only a few values which exhaustively describe all possibilities. In fact, stems and subjects are typically sampled from larger populations of stems and subjects. Under the assumption that stems and subjects are sampled randomly from these populations (an ideal that is often not truly met), these factors are referred to in the statistical literature as random-effect factors, contrasting with fixed-effect factors such as Sex (male versus female) or Voice (active, passive). Subjects and items (stems in the present example) tend to have their own specific preferences or dispreferences for a given choice (see, e.g., Dąbrowska 2008, 2010; Street & Dąbrowska 2010, and Nesset et al. 2010, for examples from linguistics). Individual speakers, for instance, might have a

Names of dataset and R script:	datOB.csv; OB.R
Source of dataset:	Psycholinguistic experiment reported in Baydimirova 2010, Endresen 2011
Size of dataset:	2630 rows, each corresponding to a response from one of sixty subjects
Rival forms:	о- vs. об-, represented as FirstResponse1 with values: O, OB. Subjects were allowed to also make an additional response (in other words, if they first responded O, they were allowed to make a second choice of OB). We represent only the subjects' first response in this dataset.
Predictors:	
Subject	anonymized subject identifier, such as A1, A2, A3, etc.
Stem	the nonce stem tested, such as bukl, chup, dukt, lus, etc.
StimulusType	word class of the stimulus presented to subjects: adjective, verb
Onset	onset consonant(s) of nonce stem: m, n, b, d, etc.
ClusterOnset	whether the onset contained a consonant cluster: yes, no
PossibleWithB	whether the Russian phonotactics allow the combination of b + the given onset1: TRUE, FALSE. Incompatible clusters tested in the experiment are: жр, чт, жг, тк.
Place	place of articulation of the onset: alveopalatal, dental, labial, velar
Manner	manner of articulation of the onset: affricate, fricative, sonorant, stop
StressStimulus	place of stress on stimulus (differentiated only for verbs; all nonce adjectives were stem-stressed): root, suffix, NotRelevant (for adjectives)
Gender (of subject):	male, female
Age (of subject):	ranging from 18 to 59
EducationLevel:	Higher, IncompleteHigher, Secondary
EducationField:	Humanities, Science
SubjectGroup	subjects were grouped according to stimulus type: A (root-stressed verb), B (suffix-stressed verb), C (root-stressed adjective)

Table 9: Description of the о- vs. об- dataset

personal preference for о- or for об-. Although this dataset deals with nonce words, these nonce words will have various likenesses to real words, so we also need to weed out this potential source of extra variation in the data that could obscure the structure we are seeking to find. It will be clear that we need to bring this variability into the model in a principled way. If we fail to do so, substantial correlational structure in the model will not be accounted for, and the p-values obtained will be anti-conservative.

Mixed-effects logistic regression makes it possible to distinguish between variability tied to subjects and items and variability linked to the predictors of primary interest. The tree & forest model, given current implementations and hardware limitations, does not scale up to data with many subjects and many items, so we will not include that model here.

3.3.1 Logistic regression

In order to facilitate the interpretation of the coefficients of the model, we center Age by subtracting from each age value the mean of Age, resulting in the predictor AgeCentered. The best mixed-effects logistic model for the subset of verbs is described by the following formula:

```
FirstResponse ~ ClusterOnset + StressStimulus * AgeCentered +
                Manner + (1|Stem) + (1|Subject)
```

The formula indicates that StressStimulus is taken into account both as a main effect and in an interaction with Age, together with a main effect of ClusterOnset. The last two terms in the formula, `(1|Stem)` and `(1|Subject)`, indicate that Stem and Subject are to be treated as random effect factors. The other predictors are treated as fixed effect factors: they have only a fixed (usually small) number of different levels (values) that are repeatable, in the sense that one can easily build a new dataset with the same factor levels. This is not possible for subjects sampled randomly from a large population of subjects: a new random sample will contain many new subjects, and likely only subjects that have not been seen before. This explains the term 'mixed model': it is a model that 'mixes' fixed-effect and random-effect factors in one and the same analysis (cf. Baayen 2008: Chapter 7).

Table 10 lists the coefficients for the fixed-effect predictors. The intercept represents the group mean (on the logit scale) for ClusterOnset=no, StressStimulus=root, and Manner=affricate, for AgeCentered = 0 (which is equivalent to Age = mean of Age), and its negative value tells us that the

	Estimate	Std. Error	z value	p-value
(Intercept)	-0.430	0.391	-1.101	0.2710
ClusterOnset=yes	-0.596	0.236	-2.532	0.0113
StressStimulus=suffix	1.344	0.404	3.323	0.0009
AgeCentered	0.024	0.022	1.065	0.2869
Manner=fricative	0.149	0.316	0.472	0.6366
Manner=sonorant	1.079	0.348	3.104	0.0019
Manner=stop	-0.124	0.325	-0.382	0.7022
StressStimulus=suffix:AgeCentered	0.255	0.086	2.981	0.0029

Table 10: Coefficients for a mixed-effects logistic regression model for the o- vs. об- dataset

model predicts o- here. All predictors are well-supported by low p-values, where we should keep in mind that for Manner we see that there is one contrast in the group means (those of sonorants and affricates) that reaches significance under the Bonferroni correction (the p-value for this contrast is far below 0.05/3 = 0.0167). Interestingly, when stress is on the suffix, the probability of using об- increases with age. When the stress is on the root, there is no such effect of age.

	logLik	Chisq	Chi.Df	p-value	Reduction in AIC
Subject	-807.13				217.6
Stem	-783.49	47	1	0.0000	45.3
ClusterOnset	-779.65	8	1	0.0056	5.7
StressStimulus	-777.96	3	1	0.0660	1.4
AgeCentered	-776.58	3	1	0.0967	0.8
StressStimulus:AgeCentered	-772.59	8	1	0.0047	6.0
Manner	-762.28	21	3	0.0001	14.6

Table 11: Model comparison statistics for the o- vs. об- dataset

Table 11 lists the statistics for the decrease in AIC (in the column labeled AIC) as the different terms (listed in the rows of this table) are added to the model specification. The first row in this table compares the AIC of a model with Subject to that of a model with only an intercept term. The large decrease in AIC (217.6) indicates that Subject is the most important predictor. The next most important predictor is Stem, which comes with a reduction in AIC of 45.3. The contributions of the linguistic predictors are much smaller. It is clear that ClusterOnset and also the interaction of StressStimulus by AgeCentered contribute to the model fit. It is also clear that Manner is by far the most important linguistic predictor. (The other columns in this table have the following interpretation: logLik is the model's log likelihood, another measure of goodness of fit. Chisq is twice the difference in logLik, which follows a chi-squared distribution with as degrees of freedom the number of additional parameters used by the more complex model. This number is listed in the column labeled Chi.Df. The p-value is derived from these chi-squared statistics.)

The index of concordance for this model is C =0.82, and its accuracy is 74%.

3.3.2 Naive discriminative learning

Naive discriminative learning, using the following model specification,

$$\texttt{FirstResponse} \sim \texttt{ClusterOnset + StressStimulus + Age + Manner +}$$
$$\texttt{Stem + Subject}$$

performs equally well as the mixed-effects model: C =0.82 and an accuracy equal to 75%. It should be noted that naive discriminative learning is defined only for factorial predictors. Since Age is a numerical predictor, it is automatically split on the mean into two subsets, in the present case, subjects older or younger than 24. Table 12 lists the weights for the main predictors, after removal of the weights for the individual stems and subjects. From this table, it is easy to read off that the younger subjects prefer o-, whereas the older subjects prefer об-. In contrast to the mixed-effects logistic regression model, the naive discrimination model supports an unconditioned effect of age. The predictors are ranked according to their variable importances as follows: ClusterOnset: 0.21, Age: 0.22, StressStimulus: 0.26, Manner: 0.52, Stem: 7.66, Subject: 11.16. NDL is in agreement with the mixed-effects logistic model that Manner, Stem, and Subject are the most important predictors.

Although naive discriminative learning works well for this dataset as a statistical classifier, the weights do not have a good interpretation from a learning perspective. From a cognitive perspective, it would be much preferable to train a naive discriminative learning network on the experience

	O	OB
Age in [18,24)	0.19	0.09
Age in [24,59]	0.07	0.20
ClusterOnset=no	0.09	0.20
ClusterOnset=yes	0.18	0.08
Manner=affricate	0.10	0.02
Manner=fricative	0.09	0.05
Manner=sonorant	-0.07	0.20
Manner=stop	0.14	0.00
StressStimulus=root	0.20	0.07
StressStimulus=suffix	0.07	0.21

Table 12: Naive discriminative learning weights (selected) for the о- vs. об- dataset

that speakers have with the о- and об- rival prefixes, and then to use this network to predict what prefix speakers use for nonce verbs. In this respect, the о- vs. об- dataset differs from the грузить 'load' data and the пере- vs. пре- data, which comprise observations from corpora that constitute speakers' experience with the language, and from which we can draw conclusions about what they have learned and what choices they are likely to make.

3.4 -Ну vs. Ø

The objective of this case study is to chart an ongoing language change that serves to support a distinction between inchoative and stative verbs that are undergoing the change as opposed to semelfactive verbs that are not undergoing the change. Inchoative verbs such as (об)сохнуть 'dry' are undergoing a language change in Russian in which some past tense forms are dropping the -ну suffix in favor of unsuffixed (Ø) variants. This language change has been discussed in the scholarly literature (Bulaxovskij 1950, 1954; Černyšev 1915; Dickey 2001; Gorbačevič 1971, 1978; Nesset 1998; Plungian 2000; Rozental' 1977; Vinogradov and Švedova 1964), but only one previous corpus study has been carried out, and that one was based on data from the 1960-1970s (Graudina et al. 1976, 2001, 2007). Table 13 presents the relevant forms (using (об)сохнуть 'dry' to illustrate) and variants arranged according to overall trends identified in our case study. The left-hand side of the table presents forms for which the -ну variant is preferred; forms that prefer the Ø variant are on the right. On the vertical dimension, each side of the table is ordered according to the strength of the preference, with the strongest preference on top.

Since the data in this case study involves primarily inchoative and stative verbs (plus a few transitives like двинуть 'move'), there is no variation along the meaning dimension in Figure 1, but Table 13 gives some indication of the complex relationships among differences in environment, since here we see already an interaction between the grammatical form and the presence vs. absence of a prefix. At least two other environmental factors seem to be involved, namely the phonological shape of the root and the presence vs. absence of the -ся/сь reflexive marker. Verbs with roots ending in a velar fricative like (об)сохнуть 'dry' are generally the most likely to retain -ну, heading a cline that proceeds through velar plosives as in (по)блекнуть 'fade' and then dental fricatives as in (по)гаснуть 'go out', ending with labial plosives which are most likely to prefer Ø as in (по)гибнуть 'perish'. The -ся/сь reflexive marker also has an effect: when the marker is present, the gerund appears in nearly equal numbers with -ну vs. Ø, so forms like проникнувшись and проникшись, both meaning 'having penetrated (intrans.)' are attested approximately equally. However, when -ся/сь is absent, a preference for -ну is maintained, so проникнув is more frequent than проникши 'having penetrated (trans.)'. Our analysis accounts for these and additional factors along the additional diachronic dimension of change.

Like the PERE dataset, NU (Table 14) presents us with very unbalanced data, since there are 31790 observations with Ø, as opposed to only 2289 with -ну. The Period and Genre predictors introduce two new types of data not present in the three datasets analyzed above, namely diachronic data and society-level data. In what follows, we focus on these two predictors.

3.4.1 Logistic regression

We begin with fitting a simple main effects model to the data, using the model equation

$$\texttt{NU} \sim \texttt{Form + Prefix + Genre + Rootfinal + SemClass + SJA +} \\ \texttt{Period.}$$

Table 15 lists the coefficients of this model. Due to the many predictors, and the many factor levels for these predictors, the number of coefficients is

strength of preference	forms prefering -ну	forms prefering Ø
strongest	*unprefixed participle:*	*non-masculine finite past:*
	сохнувший > сохший	(об)сохнула, -о, -и < (об)сохла, -о, -и
	gerund:	*prefixed masculine finite past:*
	обсохнув > обсохши	обсохнул < обсох
		unprefixed masculine finite past:
weakest		сохнул < сох

Table 13: Overall preferences for -ну vs. Ø among inchoative and stative verbs.

Name of dataset:	datNU.csv1
Source of dataset:	Russian National Corpus (www.ruscorpora.ru)
Size of dataset:	34079 rows, each representing an example sentence containing an inchoative verb whose infinitive form ends in -нуть
Rival forms:	-ну vs. Ø, represented as NU with values nu and NoNu
Predictors	
Form (of the verb):	finite (non-masculine past tense forms), (past) gerund, mascsg (masculine past tense form), part (past active participle)
Prefix:	Prefixed, Unprefixed
Period:	1800–1849, 1850–1899, 1900–1949, 1950–1999, 2000–2010
Genre:	church, fiction, massmedia, mix, nonfiction, private (as specified in the Russian National Corpus)
Rootfinal:	type of root-final consonant, levels: dentalfricative, dentalplosive, labialplosive, none, velarfricative, velarplosive
SemClass	designation according stative vs. inchoative and transitive vs. intransitive, levels: InchIntr (inchoative intransitive), StatIntrans (stative intransitive), Transitive
SJA	presence vs. absence of -ся/сь reflexive marker, levels: Sja, NoSja

Table 14: Description of the -Ну vs. Ø dataset

quite large. Most of the *p*-values are small, indicating that many of the listed contrasts are significant. However, the table lists only a small number of the possible comparisons of group means. For instance, for Genre, 'church' is the reference level, and the other genres are compared to this reference level, but not with each other.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.25	0.35	-15.17	0.0000
Formgerund	8.36	0.15	55.41	0.0000
Formmascsg	2.24	0.12	18.91	0.0000
Formpart	3.98	0.12	33.22	0.0000
PrefixUnprefixed	3.08	0.11	27.21	0.0000
Genrefiction	1.04	0.32	3.23	0.0012
Genremassmedia	1.22	0.32	3.77	0.0002
Genremix	1.07	0.46	2.32	0.0203
Genrenonfiction	1.30	0.33	3.94	0.0001
Genreprivat	0.87	0.39	2.21	0.0270
Rootfinaldentalplosive	-10.17	169.96	-0.06	0.9523
Rootfinallabialplosive	-1.49	0.12	-12.58	0.0000
Rootfinalnone	-1.24	0.30	-4.10	0.0000
Rootfinalvelarfricative	-1.10	0.11	-10.22	0.0000
Rootfinalvelarplosive	-0.95	0.09	-10.36	0.0000
SemClassStatIntrans	-0.45	0.10	-4.35	0.0000
SemClassTransitive	2.07	0.09	21.81	0.0000
SJASja	-0.55	0.12	-4.54	0.0000
Period1850-1899	-0.91	0.13	-6.76	0.0000
Period1900-1949	-1.60	0.13	-12.63	0.0000
Period1950-1999	-1.97	0.13	-15.48	0.0000
Period2000-	-1.90	0.13	-14.53	0.0000

Table 15: Table of coefficients for the main effects logistic model for the NU dataset.

To quickly assess all possible pairwise comparisons, while correcting the p-values for the fact that we are performing a large number of comparisons, we can make use of the glht function from the multcomp package (Hothorn et al., 2008).[7] Figure 4 presents, for each pair of group means, the 95% confidence interval for the difference between these group means. For instance, the first row in the plot indicates that when the estimated group mean for 'church' is subtracted from the group mean for 'fiction', a 95% confidence interval (adjusted for multiple comparisons) is obtained that does not straddle zero (indicated by the vertical dashed line). From this, we can conclude that there is a significant difference between the two group means. Figure 4 indicates that there are two other contrasts that are significant, both involving 'church'. All other pairwise comparisons do not support significant differences.

Next, consider the coefficients for Period. The reference level for this factor is 1800–1849, and the four coefficients listed therefore compare later half centuries with the first half of the nineteenth century. First note that all four coefficients are negative. This indicates that at later moments in time, NU was used less often. Also note that the coefficients become more negative as time proceeds. Only for the most recent period, the coefficient is no longer more negative than that of the preceding period. This indicates that NU is used progressively less frequently over the last two hundred years, with this process of attrition possibly coming to a halt in the 21st century. Table 16 lists, for each half-century, the number of occurrences of NoNu and Nu, as well as the proportion of Nu attestations. The proportions show exactly the same pattern as the coefficients of the logistic model, unsurprisingly. A multiple comparisons test (not shown) indicates that all pairwise comparisons of half-centuries are significant, with the exception of the most recent pair (1950–1999 versus 2000–). The index of concordance for this model is 0.95 and its accuracy is 96.3%. A slight improvement (C = 0.955, accuracy =

Period	NoNu	Nu	Proportion
1800-1849	1073	239	0.182
1850-1899	3290	348	0.096
1900-1949	8012	554	0.065
1950-1999	10810	605	0.053
2000-	8605	543	0.059

Table 16: Counts of occurrences of NoNu and Nu, and the proportion of Nu, for 5 successive half-century periods.

[7] In this example, we have made use of Tukey's multiple comparisons method.

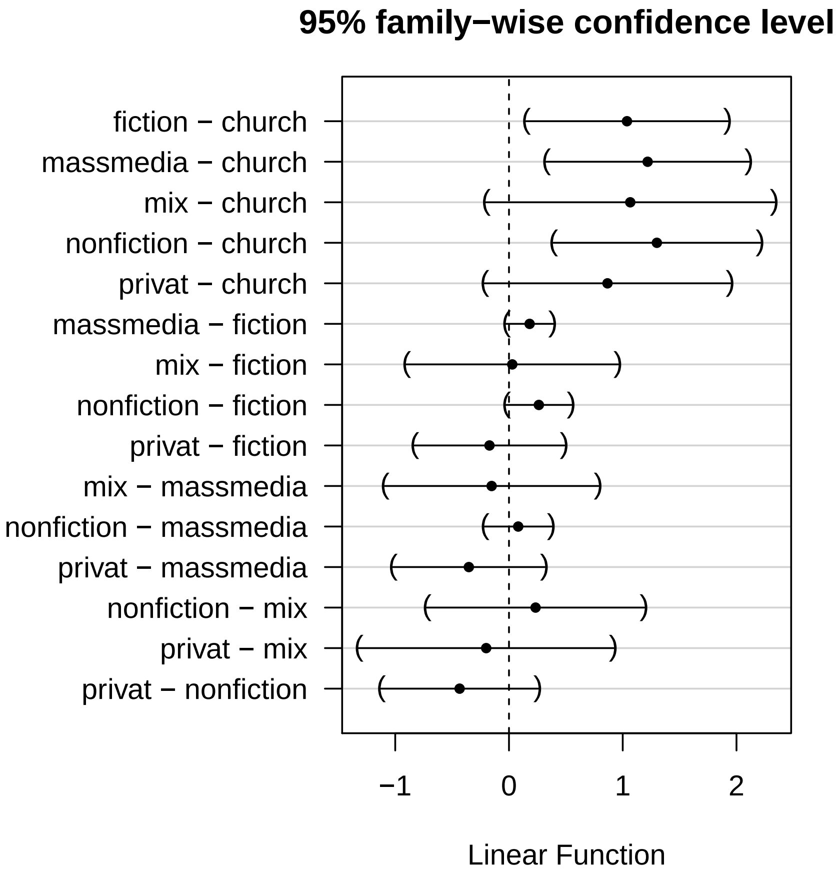


Figure 4: Tukey's all-pair comparisons between group means for Genre.

96.6%) can be obtained by including several interactions, which increases the number of coefficients to no less than 98. As the dataset is large, the small increase in accuracy still amounts to roughly a hundred additional correct classifications. Unfortunately, the model with interactions becomes linguistically uninterpretable.

3.4.2 Tree & Forest

The tree & forest method turns out to support the presence of many highly complex interactions. The classification tree shown in Figure 5, obtained with exactly the same model specification equation as used for the logistic model, represents only the tip of the iceberg by restricting the number of splits to three levels. The tree indicates that there are two conditions in which NU is highly likely to be present: gerunds with no SJA and with no root final plosive, and unprefixed participles. The (full) classification tree has $C = 0.964$ and accuracy = 96.7%. This compares well with the logistic model. For an evaluation of the main trends of individual predictors, the main effects logistic model is useful, for coming to grips with the interactions, the conditional inference tree is a good guide. It should be kept in mind, though, that for the full accuracy of the tree to be achieved, the full tree (not shown) is required. In that tree (as in the logistic model with interactions), many of the minor splits may be due to stochastic variation that comes with sampling data for inclusion in a large text corpus.

3.4.3 Naive discriminative learning

We assess the importance of the different predictors with naive discriminative learning, using the same model specification as for the logistic and tree models. This model, for which C =0.95 and for which accuracy = 0.963, indicates that Form is by far the most dominant predictor, followed at a large distance by Period and Semantic Class (see Figure 6).

Accuracy can be increased by allowing an interaction between Form and Prefix into the model, using the model specification

```
NU ~ Form * Prefix + Genre + Rootfinal + SemClass + SJA +
                          Period.
```

This results in C =0.953 and an accuracy equal to 0.967, indicating an accuracy equal to that of the other two models. The interaction asks the naive discriminative learner to add as independent cues all unique combinations of the levels of Form and the levels of Prefix. Table 17 lists all cues and their association strengths (weights) to NoNu and Nu, ordered by the values for Nu.

According to the recursive partitioning tree, the conditions favoring NU most were gerunds with no SJA, and unprefixed participles with no root-final consonant. From Table 17 we can read off the NDL support for these

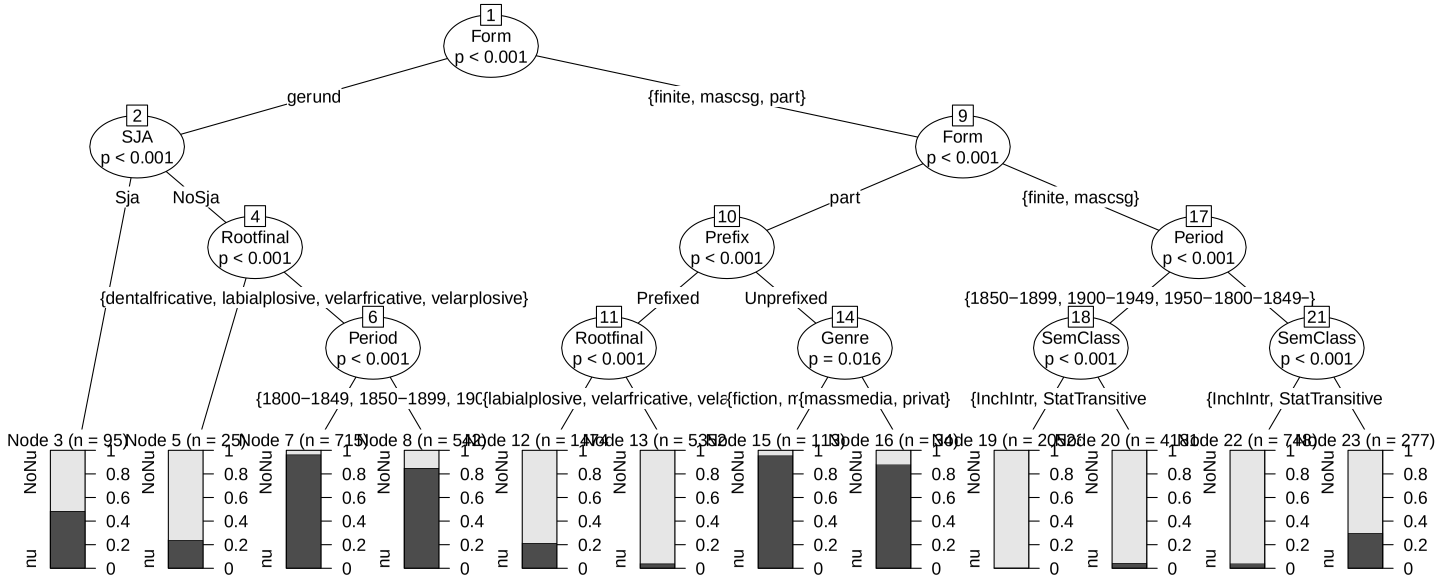


Figure 5: Classification tree for the NU dataset.

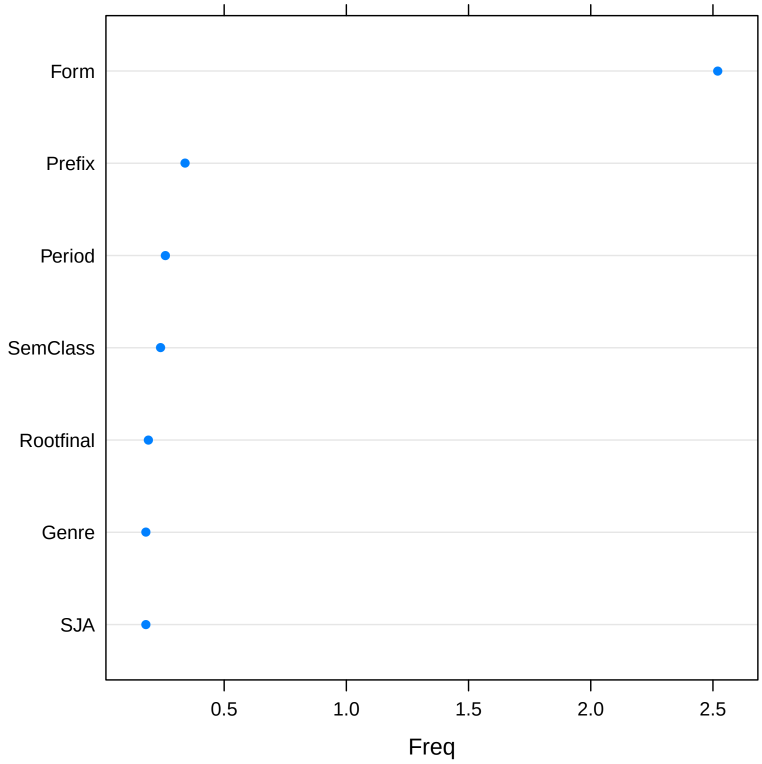


Figure 6: Variable importance for the NU dataset using a simple main effects ndl model.

conditions, Formgerund: +0.326 + NoSJA +0.089 = 0.415 and Rootfinal none: 0.014 + Formpart:PrefixUnprefixed 0.432 = 0.446. We can also clearly see that the support for Nu decreases over time: 0.092 → 0.041 → 0.016 → 0.007 → 0.008.

4 Conclusions

To conclude, we summarize the results in two ways, first focusing in the relative strengths and merits of the three statistical models used to analyze our data and second interpreting the behavior of our rival forms in terms of the relationships between their meanings and the environments they appear

	weight NoNu	weight Nu
Formpart:PrefixPrefixed	0.32	-0.28
Formfinite	0.30	-0.18
Formfinite:PrefixUnprefixed	0.24	-0.17
Formmascsg	0.25	-0.13
Formmascsg:PrefixUnprefixed	0.17	-0.10
Formmascsg:PrefixPrefixed	0.09	-0.04
Formfinite:PrefixPrefixed	0.07	-0.02
PrefixPrefixed	0.24	-0.01
Genrechurch	0.07	0.00
Period1950-1999	0.08	0.01
Period2000-	0.08	0.01
Rootfinallabialplosive	0.06	0.01
Rootfinalvelarfricative	0.06	0.01
Rootfinalnone	0.06	0.01
Period1900-1949	0.07	0.02
SemClassStatIntrans	0.13	0.02
Rootfinalvelarplosive	0.05	0.02
Genreprivat	0.05	0.03
Genremix	0.04	0.03
Genrefiction	0.04	0.03
Genremassmedia	0.04	0.04
SemClassInchIntr	0.11	0.04
Period1850-1899	0.05	0.04
Genrenonfiction	0.03	0.04
Rootfinaldentalfricative	0.02	0.05
Rootfinaldentalplosive	0.02	0.05
SJASja	0.15	0.07
SJANoSja	0.13	0.09
Period1800-1849	-0.00	0.09
SemClassTransitive	0.04	0.11
Formpart	-0.04	0.16
PrefixUnprefixed	0.04	0.17
Formgerund	-0.24	0.33
Formgerund:PrefixPrefixed	-0.24	0.33
Formpart:PrefixUnprefixed	-0.36	0.43

Table 17: NDL weights for NoNu and Nu.

in.

4.1 Pros and cons of the methods

The three statistical techniques that we have explored have different strengths and weaknesses. In what follows, we discuss these by going through a list of issues that arise in statistical modeling of choice data.

1. random-effect factors: The tree & forest method does not scale up for datasets with random-effect factors with many levels. We saw this for the the psycholinguistic study of the distribution of о- vs. об- in nonce words. Here, mixed-effects logistic models are the best choice. Compared to naive discriminative learning, they also provide better insight into the variability associated with, for instance, speakers.

2. interactions: The tree & forest method is able to detect complex interactions that are beyond the means of logistic models. The NU dataset provides an eloquent example of this. Naive discriminative learning can deal with complex interactions, but the weights will often not be easy to interpret.

3. classification accuracy: All three techniques produce probabilities for which rival form is most likely. These predictions can be used to calculate accuracy scores and indices of concordance. Across the four data sets, the different statistical methods provide very similar results, although occasionally, one method may clearly outperform the others. The general convergence, however, is reassuring, for two reasons. First, it shows that we have a good understanding of the quantitative structure of the data. Second, we can use different methods in parallel, combining the strengths of both to compensate for individual weaknesses. For instance, a conditional inference tree can be used to better understand interactions in a logistic model.

4. variable importance: All three methods come with a method for assessing variable importance. Here too, there is remarkable convergence between methods.

5. p-values: Tests of significance are available for the logistic model and for the tree & forest method. Permutation tests providing p-values

could be added to naive discriminative learning, but are currently not implemented. Therefore, naive discriminative is not a good choice for hypothesis testing.

6. cognitive interpretation: the logistic regression and the tree & forest method are statistical techniques using mathematical principles that are probably very different from those used by the brain. Naive discriminative learning, by contrast, is grounded in principles of human learning, and may therefore have increased cognitive plausibility, albeit still at a high level of abstraction.

7. ease of interpretation: Recursive partitioning trees tend to be easy to read and provide straightforward insight into the structure of the data. However, they may become extremely complex, with many levels of branching structure, in which case interpretation becomes bewilderingly fractionated. For simple models with factors with only two or three levels, and simple interactions, the coefficients of logistic models are well-interpretable. But for more complex models, interpretation of the coefficients becomes intractable, in which case the value of the model resides in the measures of variable importance and significance tests that it provides. Interpretation will have to proceed using different means, such as cross-tabulation or recursive partitioning trees. Naive discriminative learning provides weights that have a simple interpretation in terms of positive (or negative) support for a rival form from a given factor level. These weights may be easier to interpret than the coefficients of a logistic model, but, as mentioned above, they do not come with p-values.

8. appropriateness: All three models can be used as statistical classifiers. However, from a cognitive perspective, naive discriminative learning makes sense only when the data can be viewed as a window on a speaker's learning experience. As a consequence, it is not recommended as a model for data spanning a long time period (i.e., more than a century). Human learning is more local, and to properly model actual speakers, one would have to restrict the input data to a time interval that mirrors the average life span of a speaker.

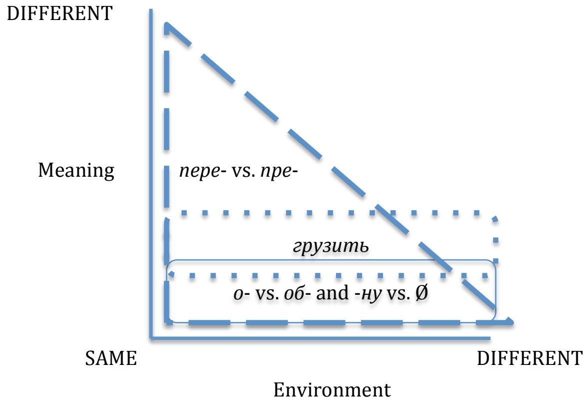
In summary, we recommend the tree & forest method as a highly useful method complementing logistic models. Often, it will be helpful to use both in parallel. Naive discriminative learning is offered as an alternative that is of potential interest from a cognitive perspective. The present study is the first to show that it performs with similar accuracy as the other two methods. It is conceivable that naive discriminative learning may not perform as well as other methods due to other methods using computational resources that are not available to the brain. By way of example, the excellent performance of random forests is due a smart voting scheme that consults hundreds of individual trees grown on parts of the data. It seems unlikely to us that an individual's brain would work along similar lines. On the other hand, within a language group, individual speakers might be comparable to the individual trees in a forest, with the community's consensus on what form to use arising through an implicit social 'voting' scheme driven by optimization of communication. It should therefore be kept in mind that naive discriminative learning represents only low-level learning at the level of the individual, and that the forces shaping a language are much more complex. The vision behind naive discriminative learning, however, is that it would be great to have a computational model that explains how grammar emerges from usage, and our current implementation should be viewed as a very first step in that direction.

4.2 Rival forms and the meaning/environment plane

Where do the rival forms in our case studies fit in the space defined by variance in meaning and environment? Figure 4.2 gives an approximate visualization of their behavior.

For both о- vs. об- and -ну vs. Ø, only differences in environment (including both morphological and phonological environment, but also the environment of Genre for the latter) were considered while meaning was held more or less constant. The region these rival forms occupy is suggested by the thin solid line encircling "о- vs. об- and -ну vs. Ø" in the figure. For both case studies, the rival forms can both compete in the same environment and can also be more (or less) characteristic of different environments, so they occupy a continuum between "same" and "different" on the bottom axis of the figure.

Partially overlapping with о- vs. об- and -ну vs. Ø is грузить, represented by a dotted line. The rival forms in the грузить dataset are near-synonyms that, like the previous two sets, vary in their ability to compete in the same environments while also showing some preferences for different environments.



The remaining case study is пере- vs. пре-, which is represented by a triangle with a dashed line. These rival forms cover a greater portion of the space in the figure because they can both overlap and contrast in terms of both meaning and environment.

In sum, we see that different rival forms show different patterns in terms of variation in meaning and environment. This is a complicated area of linguistics that we are just beginning to explore with the help of appropriate statistical methods.

References

Alexeeva, A. P.: 1978. Iz istorii pristavočnogo glagolnogo slovoobrazovanija (na primere obrazovanij s OB i O). Avtoreferat na soiskanije učenoj stepeni kandidata filologičeskix nauk. Leningrad.

Andrews, E.: 1984. “A Semantic Analysis of the Russian Prepositions/Preverbs O(-) and OB(-).” In The Slavic and East European Journal, Vol. 28, No. 4, pp. 477–492.

Aronoff, M.: 1976. Word formation in generative grammar. Linguistic Inquiry Monograph 1. Cambridge, MA: MIT Press.

Arppe, A.: 2008. “Univariate, bivariate and multivariate methods in corpus-based lexicography. A study of synonymy. Helsinki“.

Avilova, N. S.: 1959. “O kategorii vida v sovremennom russkom literaturnom jazyke.” In Russkij jazyk v nacional’noj škole 4, 21-26.

Avilova, N. S.: 1976. Vid glagola i semantika glagol’nogo slova. Moscow. Nauka.

Baayen, R. H., Milin, P., Filipovic Durdjevic, D., Hendrix, P., & Marelli, M. (2011). “An amorphous model for morphological processing in visual comprehension based on naive discriminative learning”. In Psychological Review 118, 438–482.

Baayen, R. H.: 2011. “Corpus linguistics and naive discriminative learning”. In Brazilian Journal of Applied Linguistics 11, 295–328.

Barykina, A. N., V. V. Dobrovolskaja, and S. N. Merzon: 1989. Izučenije glagol’nyx pristavok. Moskva.

Bauer, L.: 2001. Introducing linguistic morphology. (first ed. 1988) Edinburgh University Press. Bristol.

Baydimirova (Endresen), A.: 2010. Russian aspectual prefixes O, OB, and OBO: A Case Study of Allomorphy. Master’s thesis. University of Tromsø. Available at http://www.ub.uit.no/munin/handle/10037/2767.

Booij, G.: 2005. The grammar of words: an introduction to linguistic morphology. Oxford University Press.

Bresnan, J. A.; Cueni, A.; Nikitina, T. & R. H. Baayen: 2007. “Predicting the dative alternation”. In Cognitive foundations of interpretation, Boume, G., Kraemer, I. and Zwarts, J. (eds.), 69-94. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Bulaxovskij, L. A.: 1950. Istoričeskij kommentarij k russkomu literaturnomu jazyku. Kiev.

Bulaxovskij, L. A.: 1954. Russkij literaturnyj jazyk pervoj poloviny XIX veka. Moscow.

Černyšev, V. I.: 1915. Pravil’nost’ i čistota russkoj reči. Izdanie 2-oe. Tom 2: časti reči. Petrograd.

Čertkova, M. Ju.: 1996. Grammatičeskaja kategorija vida v sovremennom russkom jazyke. Moscow: Moscow State University.

Dąbrowska, E.: 2008. “The effects of frequency and neighbourhood density on adult native speakers’ productivity with Polish case inflections: An empirical test of usage-based approaches to morphology”. In Journal of Memory and Language 58, 931–951.

Dąbrowska, E.: 2010. “Naive v. expert intuitions: An empirical study of acceptability judgments”. In The Linguistic Review 27, 1–23.

Danks, D.: 2003. “Equilibria of the Rescorla-Wagner model”. In Journal of Mathematical Psychology 47(2), 109–121.

Dickey, S. M.: 2001. “Semelfactive” -no and the Western Aspect Gestalt. In Journal of Slavic Linguistics 9.1, 25–48.

Dobrušina, Je. R., Je. A. Mellina and D. Paillard: 2001. Russkije pristavki: mnogoznačnost’ i semantičeskoje edinstvo: Sbornik. Moscow. Russkije slovari.

Endresen, A.: Forthcoming. “Allomorphy via borrowing? The status of the prefixes PRE- and PERE- in Modern Russian”.

Flier, M. S.: 1985. “Syntagmatic Constraints on the Russian Prefix pere- “. In Issues in Russian Morphosyntax. Ed. Flier Michael S. and Richard D. Brecht. Slavica Publishers. Columbus, Ohio.

Forsyth, J. A.: 1970. Grammar of Aspect. Cambridge: Cambridge University Press.

Gorbačevič, K. S.: 1971. Izmenenie norm russkogo literaturnogo jazyka. Leningrad.

Gorbačevič, K. S.: 1978. Variantnost’ slova i jazykovaja norma. Leningrad.

Grammatika russkogo jazyka: 1952. Moscow.

Graudina, L. K., Ickovič, V. A., and L. P. Katlinskaja: 1976. Grammatičeskaja pravil’nost’ russkoj reči. Opyt častotno-stilističeskogo slovarja variantov. Moscow.

Graudina, L. K., V. A. Ickovič and L. P. Katlinskaja: 2001. Grammatičeskaja pravil’nost’ russkoj reči. Moscow.

Graudina, L. K., V. A. Ickovič and L. P. Katlinskaja: 2007. Slovar’ grammatičeskix variantov russkogo jazyka. 3-e izdanie, Moscow.

Haspelmath, M.: 2002. Understanding Morphology. Oxford University Press. London.

Hothorn, T., Bretz, F. and Westfall, P: 2008. Simultaneous Inference in General Parametric Models. Biometrical Journal 50.3, 346–363.

Hougaard, Ch.: 1973. “Vyražaet li o-/ob- soveršaemost’?” In Scando-Slavica. Tomus XIX. Copenhagen. 119–125.

Isačenko, A. V.: 1960. Grammatičeskij stroj russkogo jazyka v sopostavlenii s slovackim. Morfologija. Vol. II. Bratislava: Slovak Academy.

Janda, L. A.: 1986. A Semantic Analysis of the Russian Verbal Prefixes ZA-, PERE-, DO- and OT- (= Slavistische Beiträge, Band 192). Munich: Otto Sagner.

Janda, L. A.: 2007. "Aspectual Clusters of Russian Verbs." In Studies in Language 31.3. 607–648.

Krongauz, M. A.: 1998. Pristavki i glagoly v russkom jazyke: semantičeskaja grammatika. Moscow: Jazyki russkoj kul'tury.

Matthews, P. H.: 1974. Morphology. An introduction to the theory of word-structure. Cambridge University Press.

Nesset, T.: 1998. Russian conjugation revisited: A cognitive approach to aspects of Russian verb inflection. Oslo: Novus Press.

Nesset, T., L. A. Janda and R. H. Baayen: 2010. "Capturing correlational structure in Russian paradigms: A case study in logistic mixed-effects modeling." In Corpus linguistics and linguistic theory; Volume 6 (1).

Nesset, T. and A. Makarova: 2011. "'Nu-drop' in Russian verbs: a corpus-based investigation of morphological variation and change". In Russian Linguistics 35.4, 41–63.

Plungian, V. A.: 2000. Bystro' v grammatike russkogo i drugix jazykov'. In Iomdin, L. L. and L. P. Krysin (eds.), Slovo v tekste i v slovare: sbornik statej k semidesjatiletiju akademika Ju.D. Apresjana. Moscow, pp. 212–223.

Riddle, E. M.: 1985. "A Historical Perspective on the Productivity of the Suffixes -ness and -ity." In: J. Fisiak (ed.) Historical Semantics; Historical Word-Formation. Berlin: Mouton de Gruyter, pp. 435–461.

Roberts, C. B.: 1981. "The origins and development of O(B)- prefixed verbs in Russian with the general meaning 'deceive'". In Russian Linguistics 5.3. 217–233.

Rozental', D. È.: 1977. Praktičeskaja stilistika russkogo jazyka. Moscow.

Šaxmatov, A. A.: 1952. Učenie o častjax reči. Moscow: Učebno-pedagogičeskoe izdatel'stvo.

Shull, S.: 2003. The Experience of Space. The Privileged Role of Spatial Prefixation in Czech and Russian. Munich: Verlag Otto Sagner.

Sokolova, S., L. A. Janda and O. Lyashevskaya: Forthcoming. "The Locative Alternation and the Russian 'empty' prefixes: A case study of the verb gruzit' 'load"'. In: Divjak, D. and St. Th Gries (eds.). Frequency effects in cognitive linguistics (Vol. 2): what statistical effects can(not) explain (Trends in Linguistics Series). Berlin: Mouton de Gruyter.

Soudakoff, D.: 1975. “The Prefixes pere- and pre-: A Definition and Comparison”. In The Slavic and East European Journal, 19.2, Special Issue: Soviet-American Russian Language Contributions (summer 1975), 230–238.

Street, J. and E. Dąbrowska: 2010. “More individual differences in Language Attainment: How much do adult native speakers of English know about passives and quantifiers?”. In Lingua 120, 2080–2094.

Strobl, C., G. Tutz and J. Malley: 2009. “An introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests”. In Psychological Methods, 14.4, 323–348.

Švedova, N. Ju. (ed.): 1980. Russkaja grammatika, Vol. 1. Moscow: Nauka.

Timberlake, A.: 2004. A reference Grammar of Russian. Cambridge University Press.

Tixonov, A. N.: 1964. “Čistovidovye pristavki v sisteme russkogo vidovogo formoobrazovanija”. In Voprosy jazykoznanija 1, 42–52.

Tixonov, A. N.: 1998. Russkij glagol. Moscow: Russkij jazyk.

Townsend, Charles E. (2008) Russian word-formation. Reprint of 1968, 1975. Slavica Publishers.

van Schooneveld, C. H.: 1958. “The so-called ‘préverbe vides’ and neutralization”. In Dutch contributions to the Fourth International Congress of Slavistics. The Hague: Mouton. 159–161.

Vasmer, M.: 1971. Etimologieskij slovar’ russkogo jazyka. Moskva.

Vinogradov, V. V. and N. Ju. Švedova (eds.): 1964. Glagol, narečie, predlogi i sojuzy v russkom literaturnom jazyke XIX veka. Moscow.

Vinogradov, V. V.: 1972. Russkij jazyk. Moscow: Vysšaja škola.

Wade, T.: 1992. A comprehensive Russian grammar. Blackwell Publishers. Oxford, UK & Cambridge, Massachusetts.

Wagner, A. and Rescorla, R.: 1972. “A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement”. In Black, A. H. and Prokasy, W. F. (eds.), Classical Conditioning II, 64–99. Appleton-Century-Crofts.

Zaliznjak, A. A. and A. D. Šmelev: 1997. Lekcii po russkoj aspektologii. Verlag Otto Sagner. München.

Zaliznjak, A. A. and A. D. Šmelev: 2000. Vvedenije v russkuju aspektologiju. Moscow. Jazyki russkoj kultury.