Making choices in Slavic:

Pros and cons of statistical methods for rival forms

Abstract:

(TBA)

1. Introduction

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, meaning and environment yield a two-dimensional space, as visualized in Figure 1.

Meaning

Environment

SAME

DIFFERENT

DIFFERENT

Allomorphy

Different, non-contrastive

Different, contrastive

Identical

Rival forms

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

Linguists tend to focus on the four corners of this space, which we can illustrate with Russian verbal prefixes. Let’s begin at the origin, where the 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, and the variant forms *o*- and *об*- are performing an identical role. If we change the meaning but keep the 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 are different, as in *мать вошла в церковь* ‘mother entered (into) the church’ *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 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 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 environmentally conditioned by the phonological shape of the root to which they are attached.

The space between these four points has not been thoroughly explored by linguists[[1]](#footnote-1), yet arguably contains many of the most interesting form-meaning-environment relationships found in language. This space is labeled “Rival forms” in Figure 1 and includes relationships involving near-synonymy and partial synonymy as well as various degrees of overlap in terms of environments.

We examine five case studies of rival forms. 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 Slavic languages. All five case studies involve rival forms of verbs. Two of the case studies focus on past tense forms of verbs, and the remaining three focus on perfectivizing prefixes that invoke spatial relations in the temporal (aspectual) domain. 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. The studies are presented via brief descriptions in Section 2, which relates each case study to the parameters in Figure 1 and also states the linguistic objective of each study. Section 3 follows with a general discussion of the range of options for statistical analysis and problems posed by various datasets. Section 4 presents the individual analyses, which are then summarized in the conclusions in Section 5. All the datasets and the code used for their analyses are available at this site: URL. All analyses are performed using the statistical software R, which is available for free download (see our website for links and instructions).

2. The datasets

We present five datasets representing different kinds of rival forms found in modern Russian and medieval Czech. Three of these datasets focus on rival Russian verbal affixes, namely the prefixes *о*- vs. *об*-, *пере*- vs. *пре*-, and the suffix -*ну* and its null rival form. Two of our datasets represent rival constructions, namely the “goal-object” and “theme-object” constructions (defined below) with Russian *грузить* ‘load’ and the use of auxiliary vs. no auxiliary in past tense constructions in medieval Czech.

2.1 *O*- vs. *oб*-

The objective of this study is to address the controversy concerning the status of *о*- vs. *об*- as either a single morpheme or two separate ones. The etymologically related variants *о*- 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 *о*- 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 *о*- vs. об- covers a large portion of the space depicted in Figure 1. We saw already in the use of *о(б)стричь* ‘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 *о*- vs. *об*- are classic allomorphs expressing the same meaning in phonologically complementary (non-sonorant root onset vs. sonorant root onset) environments. However, *о*- 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 factitive uses built from adjectives meaning ‘make something Y’ (where Y is the meaning of the base adjective or noun), as in *осложнить* ‘make complicated’ (from *сложный* ‘complicated’) and *обрусить* ‘russify’ (from *русский* ‘Russian’); 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 *o*- 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 (with both a spatial sense of ‘around’ and the 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. The analysis in 4.3 addresses differences in meaning and differences in environment, as well as individual preferences of subjects and stems.

2.2 *Пере*- vs. *пре*-

The objective of this case study is quite parallel to the previous one, addressing the question of whether the variants represent one morpheme or two.Like *о*- vs. *об*-, *пере*- vs. *пре*- are etymologically related prefixes, but their history and behavior is quite different.[[2]](#footnote-2) 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 as with *о*- vs. *об*-, 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’), and there are also examples where both *пере*- and *пре*- serve the usual role of perfectivizers, as in *перетерпеть* ‘overcome’ and *претерпеть* ‘undergo, endure’ which are both perfective verbs from the imperfective *терпеть* ‘suffer’. The analysis in 4.2 reveals the various strengths of the semantic and environmental factors associated with *пере*- vs. *пре*- in Russian verbs.

2.3 -*Ну* vs. Ø

The objective of this case study is to chart an ongoing language change that serves to support a distinction between inchoative verbs that are undergoing the change and 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 1 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.

|  |  |  |
| --- | --- | --- |
|  | Forms preferring -*ну* | Forms preferring Ø |
| strongest preference | unprefixed participle:  *сохнувший* > *сохший* | non-masculine finite past:  *(об)сохнула, -о, -и* < *(об)сохла, -о, -и* |
|  | gerund:  *обсохнув* > *обсохши* | prefixed masculine finite past:  *обсохнул* < *обсох* |
|  |  | prefixed participle:  *обсохнувший* < *обсохший* |
| weakest preference |  | unprefixed masculine finite past:  *сохнул* < *сох* |

Table 1: Overall preferences for -*ну* vs. Ø among inchoative verbs

Since all of the data in this case study involves inchoative verbs, there is no variation along the meaning dimension in Figure 1, but Table 1 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 in 4.4 accounts for these and additional factors along the additional diachronic dimension of change.

2.4 *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 anaysis in 4.1 shows that this is a robust finding, even when we take into account relevant additional environmental variation, namely the use of the prefixes and 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 is not mentioned).

2.5 Auxiliated vs. unauxiliated past tense forms in Old and Middle Czech

The objective of this study is to establish whether the use vs. absence of the auxiliary in the third person with the *l*-participle and past participle to mark a semantic distinction between 1365 and 1615 in Czech. The hypothesis is that during this period the auxiliated form, as in *Posel jest přišel* coded a perfect and/or emphatic reading ‘The agent has come’, whereas the unauxiliated form as in *Posel přišel* ‘The agent came’ was the neutral means to express the past tense in narrative contexts. Subsequent to this period the auxiliary was lost across the board in the third person past tense in Czech. Our data pertains to active and passive forms attested in personal correspondence. The hypothesis thus combines a proposed semantic variation (perfect/emphatic vs. neutral) that is related to environmental variation, where the two most prominent factors are the type of clause that the past tense form appears in and the anchoring of the even to a temporal location. Forms appearing in relative clauses are more likely to be auxiliated than unauxiliated, as in this example: *kteréžto peniezie splnila mi jest úplně a docela* ‘which money she has paid to me completely and in full’ (Albrecht 1408). However, unauxiliated forms are also attested in relative clauses, as in *který byl na Telči* ‘who was in Telč’ (Bohunka z Pernštejna 1533). A temporal location can be specified either by an adverb or by the presence of a sequence of events and in this environment we are more likely to find unauxiliated forms, especially in the early part of the period, like this example: *včera přišlo mi psaní od V.M. z Prahy* ‘yesterday a letter arrived to me from Your Grace in Prague’ (Anna z Rožmberka 1556). Conversely, auxiliation is more prominent when a temporal location is missing, as in this example: *že jsou ve zdraví domů přijeli všickni* ‘that everyone arrived home in good health’ (Anna z Rožmberka 1556). However, both factors are tendencies rather than rules and our analysis in 4.5 shows how they perform along a diachronic axis.

3. Options for statistical analysis

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”, and 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 “X” vs. which combinations of those values predict “Y”? 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 outcomes “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 consumers (“X” = name brand vs. “Y” = generic) will select given possible predictors such as price, volume, advertising, packaging, etc.

The traditional method statisticians apply to such situations is logistic regression The strategy of a regression model is to discover a line that provides an approximation for the actual distribution of the data, showing the relationship between the values that are predicted and the predictors. Logisitic regression uses the natural logarithm of the odds to model the prediction and convert the predicted scores along the regression line into a categorical variable with values “X” vs. “Y”. The researcher builds a logistic regression model a given dataset by repeatedly entering various combinations of predictors (as both main effects of predictors by themselves and interactions between predictors) and gradually through a series of trial-and-error iterations, whittling the formula down to arrive at an optimal model. The output of a logistic regression model gives us information that addresses all three questions stated above: 1) We can discover which values of the predictors predict “X” vs. “Y” by tracking positive vs. negative displacement of the estimated log odds from the intercept of the regression line, which represents the predicted score for a single given combination of predictor values. 2) Information about the relative strength and importance of the predictors is provided in terms of the absolute value of the estimated log odds and the probability (“p-value”). 3) It is possible to compare the accuracy of the model with the actual outcomes for “X” vs. “Y”. A second measure is called “C” (the concordance index, which measures the probability of concordance between the model and the dataset and should ideally be 0.8 or higher), and a third measure called “AIC” (Akaike information criterion, which should be as low as possible) gives a measure of the relative goodness of fit of a statistical model.

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. The primary goal of this article is to present some alternative models that are more straightforward and intuitive in terms of both their application and their interpretation. The two major alternatives we present are classification tree & random forest (Strobl et al. 2009) and naive discriminative learning (Baayen 2011). Both alternatives eliminate the step of actually designing the model by trial-and-error, since they arrive at the optimal solution on their own. The output is relatively 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. Both random forests and naive discriminative learning can yield graphs that directly compare the importance of predictor values. In addition, both models have the advantage of being non-parametric. Classical regression models make the parametric assumption that the data is “normally distributed” (following a bell curve), which means that these models perform less well when the data is very unbalanced or there are some combinations that are absent from the data (for example, the lack of unprefixed past gerunds in the -*ну* vs. Ø dataset). The success of analyses via classification trees & random forests and naive discriminative learning do not depend upon how well the data fits a normal distribution, and are thus better suited for many types of datasets involving naturalistic data on rival linguistic forms.

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. Our abstract and hypothetical examples above would be rendered by these formulas (using “Rivals” to refer to “X” vs. “Y”):

1. rival linguistic forms:

Rivals ~ A + B + C + D

2. cancer prediction:

Rivals ~ Age + BodyMassIndex + FamilyHistory + SmokingHistory + AlcoholUse + Diet + Exercise

3. voter choice prediction:

Rivals ~ Age + Race + Religion + Income + EducationLevel + Region

4. consumer choice prediction:

Rivals ~ Price + Volume + Advertising + Packaging

While both the trees & forests and naive discriminative learning are non-parametric classification models, they work on different principles and this has implications for the kinds of datasets that can be modeled and the results. The trees & forests model uses recursive partitioning to yield a classification tree that is an optimal partitioning of the data, giving the best “sorting” of observations according to the rival outcomes. It can literally be understood as an optimal algorithm for predicting an outcome given the predictor values.

Naive discrimination 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, 201976) that summarize and bring together a wide body of results on animal and human learning. 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 an accuracy on a par with the accuracy 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 strategy with the learning model is to adjust association weights according to the combinations of predictor values and outcomes. Thus each time a rival form “X” is encountered, a set of weights for the predictor values that co-occur with the given observation of that form are increased in association with “X”, whereas the non-occurring values get their weights decreased. The result is a matrix of weights that show the relationships between the predictor factors and the rival forms. 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 random forests and naive discriminative learning provide a mechanism for cross-validation. Cross-validation assesses 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 well if we ask it to predict the outcomes of the data that were the input for its design), and one also reduces the chances that our model will “overfit” 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 a theoretically infinite dataset. 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, the training dataset and the validation dataset. Statisticians have designed a variety of cross-validation techniques in order to address the gap between the ideal situation and the limitations of reality. In many cases, however, 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 as second independent dataset that would be an equivalent sample in terms of size and sources. In the random forests model, portions of the data as well as predictors are randomly removed from the dataset, creating a “random forest” of classification trees that can be compared used for making generalizations. Naive discriminative learning uses a ten-fold cross-validation that partitions the data into ten subsamples. Nine of the subsamples serve collectively as a training dataset, while the remaining subsample is used as a validation dataset. 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.

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. Rather than merely stating hypotheses and their inverses (“null hypotheses”) and evaluating the results by reporting probabilities (p-values), we use statistical models as a sensitive multi-purpose tool for ferreting out the relationships between rival forms and their predictors.

4. The analyses

The analyses are presented according to the relative complexity of the data, starting with the most straightforward dataset. Whereas regression and tree & forest models are potentially applicable to all five datasets, naive discriminative learning aims at a speaker-level experience of the co-occurrence of rival forms and predictor values and is thus appropriate for corpus but not experimental data. 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 set dataset is the one with the *грузить* ‘load’ data (“LOAD”), which is relatively simple because it has few predictors 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 URL.

4.1 *Грузить* ‘load’ and its perfectives in the theme-object vs. goal-object constructions

Names of dataset and R script: datLOAD.csv; LOAD.R

Source of dataset: Russian National Corpus (www.ruscorpora.ru)

Size of dataset: 1920 rows, 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*

Possible predictors and their values:

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)

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, trees & forests, and naive discriminative learning.

*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]](#footnote-3) The formula for this model is (the asterisk “\*” is used to indicate an interaction between two predictors in a regression model):

CONSTRUCTION ~ VERB + REDUCED + PARTICIPLE + VERB\*PARTICIPLE[[4]](#footnote-4)

The linear model yields the following results for this formula:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | Wald Z | Pr(>|z|) |
| Intercept | -0.9465 | 0.2023 | -4.68 | <0.0001 |
| VERB=po | 6.7143 | 1.0220 | 6.57 | <0.0001 |
| VERB=za | 1.0920 | 0.2451 | 4.45 | <0.0001 |
| VERB=zero | 2.3336 | 0.2446 | 9.54 | <0.0001 |
| REDUCED=yes | -0.8891 | 0.1748 | -5.09 | <0.0001 |
| PARTICIPLE=yes | -4.1862 | 1.0220 | -4.10 | <0.0001 |
| VERB=po \* PARTICIPLE=yes | 3.8953 | 1.5978 | 2.44 | 0.0148 |
| VERB=za \* PARTICIPLE=yes | 1.4087 | 1.0774 | 1.31 | 0.1910 |
| VERB=zero \* PARTICIPLE=yes | -1.7717 | 1.4415 | -1.23 | 0.2190 |

Table Y: Coefficients for logistic regression model of LOAD data

This table may seem rather mystifying; our strategy is to focus only on the most informative numbers (Estimate and Pr(>|z|)) for the logistic regression analysis and then move on to the tree & forest and naive discriminative learning models, which are easier to interpret.

The table of coefficients begins at the Intercept, which is literally the intercept point of the line that describes the relationship between the predicted forms (*theme*, *goal*) and the predictors given in the formula. The R software package chooses as values at the Intercept those that come first alphabetically. Thus the Intercept here involves these values for the predictors: VERB=na, PARTICIPLE=no, REDUCED=no, VERBna:PARTICIPLEno (this latter item is an interaction). The term for the predicted form is also chosen alphabetically at the intercept and is thus *goal* here. The Estimate values are positive for predicting the value not at the intercept (*theme*) and negative for the value at the intercept (*goal*). Here we have the Estimate value -0.9465, which because it is negative suggests that we predict *goal* for this combination of predictor values. This makes sense because if we go back to our dataset, we find that for active forms of *нагрузить* in nonreduced constructions (i.e., examples of the type defined by the values at the Intercept),there are 70 examples of *goal*, but only 27 examples of *theme*.[[5]](#footnote-5) All of the remaining Estimate values show how much difference there is between the Intercept and the given predictor value. For example, with Verb=po, we add its Estimate value to that at the Intercept: -0.9465 + 6.7143 = 5.7678. This is a large positive value, which predicts *theme*, and that is correct since in our dataset active forms of *погрузить* in nonreduced constructions appear 207 times in the *theme* construction, with no attestations in the *goal* construction. Adding the value for Verb=za to the Intercept gives us: -0.9465 + 1.0920 = 0.1455, a rather small positive value. This weak preference for *theme* makes sense because in our dataset active forms of *загрузить* in nonreduced constructions appear 64 times in the *theme* construction, and 62 times in the *goal* construction. The remainder of the Estimate values can be interpreted in like manner. Overall, VERB is doing the bulk of the “work” in this model (giving the largest values), followed by PARTICIPLE and then REDUCED.

Now look at the column headed by “ Pr(>|z|)”. This value can be interpreted in the standard way we interpret probability values (“p-values”) in statistics. P-values indicate statistical significance, which increases as the p-values get smaller. The standard cutoff for recognizing statistical significance is p = 0.05, so any value larger than that is not significant. Values between 0.01 and 0.05 are considered minimally significant (often signified by a single asterisk “\*”), values between 0.01 and 0.001 are considered significant (double asterisk “\*\*”), and values below 0.001 are considered highly significant (triple asterisk “\*\*\*”). The values for the first six lines in the table are all <0.0001, indicating that the predictor values at the Intercept, as well as all the main effect values are highly significant. The interaction of VERB and PARTICIPLE, which is also part of the highly significant Intercept, gets lower marks, since 0.0148 earns only one asterisk, and neither of the remaining values (0.1910 and 0.2190) is significant. The C value (concordance index) of 0.96 tells us that the fit of the model is excellent and the model is 89% [CHECK THIS NUMBER] accurate in its predictions.

*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 repeated iterations of the formula. We don’t have to worry about how many predictors we put in, nor do we have to specify interactions. Both the classification tree (ctree()) and the classification forest (cforest()) will eliminate any predictors that are not significant and all interactions are taken into account, as described below.

INSERT FIGURE 2 HERE

Figure 2: Classification tree and variable importance plot for datLOAD.csv

The classification tree and variable importance plot in Figure 2 show that the selection of the prefix on the VERB is by far the most important predictor, followed by PARTICIPLE and REDUCED. The classification tree shows verb as the top node, as this is the predictor that does the most work. The first node splits the data with examples prefixed in *на*-, *за*- or no prefix (zero) on the left as opposed to examples prefixed in *по*- to the right. The probability that VERB is a significant factor at this node is good, since p < 0.001. Next we go to the left branch of the tree and look for the strongest factor within that branch, which is PARTICIPLE (also with p < 0.001), and this splits with *yes* on the left and *no* on the right. Once again VERB appears as the next node on the left (still significant but less so with p = 0.027). This yields a split between examples with the prefix *на*- or no prefix on the left as opposed to verbs prefixed in *за*- on the right, and two terminal nodes, numbered 4 and 5. The graph below each terminal node represents the percentage of goal (light grey) vs. theme (dark grey) outcomes, and “n = ” indicates the total number of outcomes 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.

The classification tree shows us that there is in fact a very complex interaction among the three factors. In a classification tree we see an interaction any time that the left branch of the tree is different from the right branch. A classification tree makes no statement about main effects; this notion is specific to the regression model. If one were to try to put all of the effects that can be captured in a classification tree into a regression model, this would result in an enormous number of them and this would cause overfitting in that model. The classification tree gives us a very good 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.

The variable importance plot shows us the relative importance of each of the predictors (variables) based on the random forest output. Recall that the random forest model includes cross-validation that involves removing predictors and checking to see how the model performs with a predictor missing. This feature makes it possible to compare the optimal model against alternative models with each predictor missing. VERB is the strongest predictor, since a model excluding VERB is 33.6% 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 0.3%.

*Naive discriminative learning*

In comparison with the regression model, the random forest gives us comparable values for concordance, with C = 0.96, and an accuracy of 89%.

Since the predictor variables all describe properties of the language, a naive discrimination learning model is appropriate. Once again our formula is simply:

CONSTRUCTION ~ VERB + REDUCED + PARTICIPLE

The naive discriminative learning model yields a matrix of the various weights that are assigned to the predictor values in association with the rival forms *goal* and *theme*, presented here in Table X:

|  |  |  |
| --- | --- | --- |
|  | goal | theme |
| PARTICIPLEno | 0.07935739 | 0.32064261 |
| PARTICIPLEyes | 0.35899493 | 0.04100507 |
| REDUCEDno | 0.17569209 | 0.22430791 |
| REDUCEDyes | 0.26266024 | 0.13733976 |
| VERBna | 0.44976692 | -0.24976692 |
| VERBpo | -0.43792012 | 0.63792012 |
| VERBza | 0.31892899 | -0.11892899 |
| VERBzero | 0.10757652 | 0.09242348 |

Table X: Matrix of weights assigned by naive discriminative learning for LOAD

Let’s see how these values compare with those in our dataset. In the first row we see that active forms (“PARTICIPLEno”) receive a weight of about 0.08 for *goal* and 0.32 for *theme*. This corresponds to 306 examples of active forms with *goal* as opposed to 589 active forms with *theme*. The balance is reversed in the next row (PARTICIPLEyes), where *goal* gets a larger weight than *theme*, and again this corresponds to the distribution of examples, since we have 565 examples of participle forms in the *goal* construction but 460 examples in the *theme* construction. All of the weights in the table are likewise related to the relative distribution of examples in our database. If we add together all the absolute values of weights for each predictor, we can easily asses the relative importance that the naive discriminative learning model assigns to the three predictors, visualized here in Figure 3:

INSERT THIS PLOT AS FIGURE 3 HERE:

[# PART REDU VERB

#0.5592751 0.1739363 2.2283851

names(summedDiffs)=c("Participle", "Reduction", "Verb")

dotplot(sort(summedDiffs), xlab="summed discrimination")]

Figure 3: Variable importance plot for LOAD using summed differences of weights

This plot is basically identical to the variable importance plot returned by the classification forest in Figure 2. In both plots we see that 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 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.

4.2 *Пере*- vs. *пре*-

Name of dataset and Rscript: 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*

Possible predictors and their values:

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

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* (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, overcom-duration, overdo, redo, seriatim, superiority, thorough, transfer, transfer metaphorical, turn over* (Note: These are the full names as listed under SemanticGroupFullName; in SemanticGroup they are abbreviated)

The aim of a statistical model for this dataset is to predict the Prefix based on the possible predictors. This prediction can be modeled using all three kinds of models considered here: logistic regression, trees & forests, and naive discriminative learning. 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 numerical (ratio scale) data, as opposed to the categorical (nominal scale) data that characterize most of our predictors (where the values of the predictors are various labels such as *yes* vs. *no* or *not stacked* vs. *stacked*).

*Logistic regression*

The optimal model for this dataset is captured by this formula:

Prefix ~ ShiftTrans + PrefixStacking + ShiftAspect + PerfectiveType + SemanticGroup + LogFreqPrefVerb

This formula yields a very large table of coefficients; we present only the Intercept and the predictor values that are most relevant in Table 4 (the reader may find the entire table in the R script):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Coef | S.E. | Wald Z | Pr(>|Z|) |
| Intercept | -13.3521 | 206.3505 | -0.06 | 0.9484 |
| ShiftTrans=tr-tr | -0.7114 | 0.3516 | -2.02 | 0.0430 |
| PrefixStacking=stacked | 2.7981 | 0.4944 | 5.66 | <0.0001 |
| SemanticGroup=intrch | -1.8945 | 0.8023 | -2.36 | 0.0182 |
| SemanticGroup=overdo | -3.1482 | 0.7311 | -4.31 | <0.0001 |
| SemanticGroup=transf | -2.4456 | 0.6330 | -3.86 | 0.0001 |
| LogFreqPrefVerb | 0.3725 | 0.0645 | 5.77 | <0.0001 |

Table 4: Selected Coefficients for logistic regression model of PERE data

The Intercept has the predictor values that come first alphabetically (ShiftTrans=intr-intr, PrefixStacking=not stacked, ShiftAspect=imp-imp, PerfectiveType=natural, SemanticGroup=bridge). The estimate at the Intercept is a large negative number (-13.4), strongly predicting *пере*-, and indeed a strong prediction of *пере*- is appropriate given the strong overall bias toward *пере*- in the dataset. SemanticGroup is a viable factor with three values that are significant, and both PrefixStacking and LogFreqPrefVerb perform well. ShiftTrans is significant, but weakly so. C = 0.95 and accuracy = 96%.

*Tree & forest*

The formula for this analysis is nearly the same as the one for the logistic regression, but it can use the raw frequency data for both the base verb and the prefixed verb:

Prefix ~ ShiftTrans + PrefixStacking + ShiftAspect + PerfectiveType + SemanticGroup + FreqBase + FreqPrefVerb

The ctree() function yields the complex classification tree and the cforest() function works out the variable importance.

INSERT FIGURE 4 HERE

Figure 4: Classification tree and variable importance plot for datPERE.csv

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. The latter two predictors are also the ones that score close to zero on the variable importance plot. 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. Three of the top-scoring predictors (those with p < 0.001) are the same as the predictors that were most important in the logistic regression model: SemanticGroup, PrefixStacking, and (Log)FreqPrefVerb. ShiftTrans barely qualified as significant in the regression model, but is eliminated in the classification tree. Interestingly ShiftAspect gets more support as a predictor in the classification tree, and this is probably due to the fact that this predictor interacts strongly with other predictors, something that is handled better in the tree & forest analysis. Overall the classification forest outperforms the logistic regression model: C = 0.98 and accuracy = 97%.

*Naive discriminative learning*

The observations in this dataset mimic 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 discrimination 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 PERE 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 discrimination 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 weight are calculated from the co-occurrence frequencies of the word’s properties. The frequencies of the derived words co-determine 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 from a word’s properties (such as aspect or transitivity shifting). The model’s classification performance, as estimated by the index of concordance C, is 0.97, and its accuracy is at 94%. This compares reasonably well with the logistic regression model and the random forest.

[HARALD -- SEE LaurasPERE.R script -- I need help with the ndl part of this, since the code is tangled up with some old, no longer necessary code for datapreparation and when I run the ndl, my computer gets stuck. Also, I am not so sure that 94% accuracy is so great for ndl since a default model would do about as well -- 94% of all the outcomes were pere-, only 6% pre-]

4.3 *О*- vs. *об*-

Names of dataset and R script: datOB.csv; OB.R

Source of dataset: Psycholinguistic experiment reported in Baydimirova 2010

Size of dataset: 2630 rows, each corresponding to a response from one of sixty subjects

Rival forms: *o*- vs. *oб*-, represented as FirstResponse[[6]](#footnote-6) with values: *O*, *OB*

Possible predictors and their values:

Subject = a code name for each subject, such as *A1*, *A2*, *A3*, etc.

Stem = the nonce stem tested, such as *bukl*, *chup*, *dukt*, *lus*, etc.

StimulusType = noun 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 onset: *TRUE*, *FALSE*

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 (of subject): Higher, IncompleteHigher, Secondary

EducationField (of subject): Humanities, Science

SubjectGroup = subjects were grouped according to stimulus type: *A* (root-stressed verb), *B* (suffix-stressed verb), *C* (root-stressed adjective)

Note that this dataset has some special features. In addition to including both numerical (Age, which is ratio scale) and categorical data (the remaining predictors), we are dealing with the responses of human subjects. It is well known in psychology that different subjects have different overall preferences, and there is evidence that this extends also to linguistic preferences (Dąbrowska 2008, 2010; Street & Dąbrowska 2010). If we fail to take these preferences into account, we risk arriving at a model that reflects the preferences instead of the phenomenon that we are interested in. Word stems are also known to have individual preferences (see Nesset et al. 2010), and although we are dealing with nonce words, they 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. Variation due to individual preferences is called “random effects” because these effects are caused by individuals that have been sampled from a population (of potential subjects or stems).

Tree & forest analyses are not designed to deal with this type of variation, so we will not include them here.

*Logistic regression*

The best overall model for this dataset is captured by this formula:

FirstResponse ~ StimulusType + ClusterOnset + Manner + (1|Stem) + (1|Subject)

The last two terms in the formula, (1|Stem) and (1|Subject), indicate that Stem and Subject are to be treated as random effects. The remaining predictors are treated as “fixed effects” (because they are repeatable rather than involving selected individuals), and because this model includes both types of effects, it is called a “mixed effects model”. The random effects are taken into account, allowing us to see the fixed effects separately, as in this table of coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| (Intercept) | -1.1771 | 0.3195 | -3.684 | 0.00023 \*\*\* |
| StimulusTypeverb | 1.3807 | 0.2830 | 4.879 | 1.07e-06 \*\*\*[[7]](#footnote-7) |
| ClusterOnsetyes | -0.6723 | 0.2085 | -3.225 | 0.00126 \*\* |
| Mannerfricative | -0.1066 | 0.2743 | -0.388 | 0.69771 |
| Mannersonorant | 1.2565 | 0.2975 | 4.223 | 2.41e-05 \*\*\* |
| Mannerstop | -0.4178 | 0.2829 | -1.477 | 0.13970 |

Table Z: Coefficients for fixed effects for logistic regression model of OB data

The Intercept reflects: StimulusType=adjective, ClusterOnset=no, and Manner=affricate, and the negative value tells us that the model predicts *O* here, which is reasonable, because with this combination of predictor values we have in our dataset 62 examples of *O*, but only 24 examples of *OB*. If we look at the p-values, we see that both StimulusType and Manner are in the highest range of significance, while ClusterOnset is in the midrange. In short, verbs and adjectives behave differently in choosing *o*- vs. *об*-, and the phonological featuresof the stem onset, namely manner of articulation and whether there is a consonant cluster, are also relevant.

There is one phonological 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. In order to discover whether stress has an effect, we need to focus only on the verb data, so we create a dataset that contains only that data (“datVerb”) and run a new logistic regression analysis. For the verb data only, we find this as the optimal formula:

FirstResponse ~ ClusterOnset + StressStimulus \* Age + 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. The table of coefficients for the fixed effects is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| (Intercept) | -1.02931 | 0.72403 | -1.422 | 0.15513 |
| ClusterOnsetyes | -0.59640 | 0.23551 | -2.532 | 0.01133 \* |
| StressStimulussuffix | -5.12788 | 2.07949 | -2.466 | 0.01367 \* |
| Age | 0.02366 | 0.02222 | 1.065 | 0.28693 |
| Mannerfricative | 0.14928 | 0.31595 | 0.472 | 0.63659 |
| Mannersonorant | 1.07911 | 0.34768 | 3.104 | 0.00191 \*\* |
| Mannerstop | -0.12425 | 0.32493 | -0.382 | 0.70216 |
| StressStimulussuffix:Age | 0.25547 | 0.08568 | 2.982 | 0.00287 \*\* |

Table A: Coefficients for fixed effects for logistic regression model of OB data (verbs only)

Here we see that both Manner and ClusterOnset are retained as significant effects for verbs alone, and in addition we have a significant interaction between StressStimulus and Age.

*Naive discriminative learning*

This experiment addressing the question of *o*- vs. *об*- allomorphy is interesting from a learning perspective because the younger subjects appear to be loosing a distinction that the older subjects make. This raises the question how learnable the *o*- vs. *об*-distinction is. To address this question, we split the data into younger (Age <= 25) and older (Age > 25) subjects. We then run the naive discriminative learning analysis twice, once for each age group:

FirstResponse ~ ClusterOnset + StressStimulus + Age + Manner + Stem

The naive discriminative learning model for the older subjects has C = 0.89, but the model for the younger subjects has C = 0.69, supporting the loss of the distinction by the younger subjects (in other words, the older group yields a better fit for the data with this formula than the younger one, which gets a low C value, below 0.8). But how learnable is the distinction given how the older subjects use the allomorphs? Other things being equal, the prediction is that, because even for the older subjects the choice between the two allomorphs comes with considerable uncertainty, the next generation will use the minority choice less often. If we take the allomorph predicted by the naive discriminative learning model for the older subjects, and use these predictions as the norm for a (hypothetical) next generation of speakers, a model fitted to these first generation speakers has an index of concordance of 0.84, which is reduced compared to the index for the older subjects, 0.89. For subsequent generations, the learning problem becomes even more aggravated. In short, naive discriminative learning is able to model the gradual loss of stress as a predictor.

Interestingly, the connection weights of the younger speakers are somewhat more similar to those of the (hypothetical) first generation speakers than to those of the older speakers (p < 0.01). It should be kept in mind, however, that for actual language data frequency of use will protect many forms with the minority allomorph against leveling. The nonce words in the present experiment have no such protection, and hence are more vulnerable to change, which may explain the sharp drop in the C value between the older and younger subjects.

The explanation for the age effect in the *o*- vs. *об*- experiment assumes that other things are equal. This need not be the case, and as a consequence, patterns of language change may go in different directions. This point is illustrated by

the data on auxiliation in Czech, as we see in 4.5 below.

4.4 -*Ну* vs. Ø

Name of dataset: datNU.csv

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*

Possible predictors and their values:

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 (of attestation): *1800-1849*, *1850-1899*, *1900-1949*, *1950-1999*, *2000-*

Genre (of attestation, as defined in the Russian National Corpus): *church*, *fiction*, *massmedia*, *mix*, *nonfiction*, *privat*

Rootfinal = type of root-final consonant: *dentalfricative*, *dentalplosive*, *labialplosive*, *none*, *velarfricative*, *velarplosive*

SemClass = designation according stative vs. inchoative and transitive vs. intransitive: *InchIntr* (inchoative intransitive), *StatIntrans* (stative intransitive), *Transitive*

SJA = presence vs. absence of -*ся/сь* reflexive marker: *Sja*, *NoSja*

Like the PERE dataset, NU presents us with very unbalanced data, since there are 31790 rows of data for examples with Ø, as opposed to only 2289 with -*ну*. Although there are numerals in the Period values, they perform as categories, so all of this data is categorical. 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.

*Logistic regression*

All seven predictors turn out to be significant, and several of them interact, yielding this formula:

NU ~ Form \* Prefix + Genre + Rootfinal + SemClass + SJA + Period + Period \* SemClass \* Genre

Because there are so many significant predictors and reactions among them, the table of coefficients is very large, so we list only the significant ones here:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| (Intercept) | -7.02317 | 1.00595 | -6.982 | 2.92e-12 \*\*\* |
| Formgerund | 8.03982 | 0.15323 | 52.469 | < 2e-16 \*\*\* |
| Formmascsg | 1.86264 | 0.12826 | 14.522 | < 2e-16 \*\*\* |
| Formpart | 3.42406 | 0.12621 | 27.129 | < 2e-16 \*\*\* |
| PrefixUnprefixed | 1.37322 | 0.28978 | 4.739 | 2.15e-06 \*\*\* |
| Genrefiction | 3.38754 | 1.01846 | 3.326 | 0.000881 \*\*\* |
| Genremassmedia | 2.33430 | 1.15921 | 2.014 | 0.044042 \* |
| Genremix | 2.72966 | 1.34745 | 2.026 | 0.042785 \* |
| Rootfinallabialplosive | -1.77747 | 0.13277 | -13.388 | < 2e-16 \*\*\* |
| Rootfinalnone | -1.59195 | 0.34983 | -4.551 | 5.35e-06 \*\*\* |
| Rootfinalvelarfricative | -1.55420 | 0.12944 | -12.007 | < 2e-16 \*\*\* |
| Rootfinalvelarplosive | -1.13334 | 0.09356 | -12.113 | < 2e-16 \*\*\* |
| SJASja | -0.56690 | 0.12296 | -4.611 | 4.02e-06 \*\*\* |
| Formmascsg:PrefixUnprefixed | 1.23596 | 0.32933 | 3.753 | 0.000175 \*\*\* |
| Formpart:PrefixUnprefixed | 5.56044 | 0.48701 | 11.417 | < 2e-16 \*\*\* |
| Genrefiction:Period1900-1949 | -3.51745 | 1.35263 | -2.600 | 0.009310 \*\* |
| Genrefiction:Period1950-1999 | -2.95309 | 1.18925 | -2.483 | 0.013023 \* |
| Genrenonfiction:SemClassTransitive:Period1950-1999 | -4.09604 | 1.83876 | -2.228 | 0.025907 \* |
| Genrenonfiction:SemClassTransitive:Period2000- | -5.65891 | 2.24472 | -2.521 | 0.011703 \* |

Table B: Selected Coefficients for logistic regression model of NU data

The Intercept reflects all of the predictor values that come first in the alphabet (and in the lists above) and negative values predict *NoNu* vs. positive values predict *nu*. The large negative value at the intercept is as expected since the vast majority of our examples have Ø. Note, however that the next line under the Intercept is Formgerund with a positive value (-7 + 8 = 1), which corresponds to the fact that the gerund actually prefers -*ну*. Form and RootFinal are the most significant factors overall. The interactions indicated by Formmascsg:PrefixUnprefixed and Formpart:PrefixUnprefixed correspond to the preferences noted in Table 1. The phonological shape of the root and the presence vs. absence of the reflexive marker -*ся/сь* also show up as significant predictors as expected. In addition we see that Genre and Period and SemClassappear in various combinations as main effects and interactions. This model performs well: C = 0.95, accuracy = 97%.

*Tree & forest*

We put all of the predictors into this formula:

NU~Form+Prefix + Genre + Rootfinal + SemClass + SJA + Period

This yields a massive tree with 9 levels of branching and 85 nodes, too large to reproduce here (readers are welcome to view it on our site: URL). This suggests that the degree of interaction is actually higher than captured in the logistic regression model. The top node of the classification tree is Form, which also heads 6 further branches, and always receives a significance of p < 0.001. An abbreviated version of the classification tree, restricted to only 4 levels of branching is presented in Figure 5. The random forest performs slightly better than the logistic regression model: C = 0.96, accuracy = 97%.

INSERT FIGURE 5 HERE

Figure 5: Abbreviated classification tree for NU data

*Naive discriminative learning*

We use this formula for modeling naive discriminative learning on this dataset:

NU~Form\*Prefix + Genre + Rootfinal + SemClass + SJA + Period

This model performs almost as well as the other two: C = 0.95, accuracy = 96%.

[HARALD: I can't run this ndl on my computer either and we should have a sentence or two about the results.]

4.5 Auxiliated vs. unauxiliated past tense forms in medieval Czech

Name of dataset: datAUX.csv

Source of dataset: Dvorský 1869

Size of dataset: 702 rows, each representing an example sentence containing an *l*-participle or a past participle with a third person singular subject

Rival forms: presence vs. absence of an auxiliary verb, represented as Aux with values *Aux* and *NoAux*

Possible predictors and their values:

Voice: *Active*, *Passive*

Number: *Sg*, *Pl*

Subject = presence vs. absence of overt subject in clause: *Subj*, *NoSubj*

Reflexive: *Reflexive*, *Nonreflexive*

TempLoc = a factor that combines TempAdv and Sequence, and is true when either of these items is present: *TRUE*, *FALSE*

Aspect (of main verb): *Impf* (imperfective) vs. *Pf* (perfective)

SpatAdv = presence vs. absence of spatial adverbial: *SpatAdv*, *NoSpatAdv*

Epoch = *early* (before 1425), *middle* (1425-1525), *late* (after 1525)

RelClause = whether or not the example appears in a relative clause: *RelCl*, *NoRelCl*

With nine possible predictors, one representing a diachronic dimension, this is a very complex dataset.

*Logistic regression model*

The formula for this model is:

Aux ~ Subject + TempLoc + SpatAdv \* Reflexive + Subject \* TempLoc + RelClause + Epoch

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.5305 0.3422 -1.550 0.121135

SubjectSubj 0.3930 0.2520 1.560 0.118819

TempLocTRUE 1.6567 0.2965 5.587 2.31e-08 \*\*\*

SpatAdvSpatAdv 0.4125 0.2136 1.931 0.053469 .

ReflexiveReflexive 0.5650 0.3436 1.644 0.100083

RelClauseRelCl -0.7116 0.1942 -3.664 0.000248 \*\*\*

Epochlate 0.3020 0.3048 0.991 0.321829

Epochmiddle -1.1366 0.3548 -3.203 0.001360 \*\*

SpatAdvSpatAdv:ReflexiveReflexive -2.5188 0.7121 -3.537 0.000404 \*\*\*

SubjectSubj:TempLocTRUE -1.6940 0.3597 -4.709 2.49e-06 \*\*\*

Table of parametric coefficients for gam() model of AUX data

[I DON'T KNOW ENOUGH ABOUT THE GAM MODEL TO COMMENT HERE -- HARALD?]

C = ??, accuracy = 67%.

*Tree & Forest*

We put all nine of the predictors into the formula:

Aux ~ Voice + Number + Subject + Reflexive + TempLoc + Aspect + SpatAdv + Epoch + RelClause

This formula gives us the classification tree and variable importance plot shown in Figure X:

INSERT FIGURE X HERE

Figure X: Classification tree and variable importance plot for AUX

The classification tree retains only six of the predictors, which are ranked in importance as follows: TempLoc > Subject > Epoch > RelClause > Voice.

The tree & forest analysis performs better than the regression model: C = 0.81, accuracy = 72%

*Naive discriminative learning*

The AUX dataset shows that patterns of language change may go in different directions, for here we see a change in time that interacts with Voice. As we have have very few observations for the early period, we focus on the middle and late periods.

The key issue of interest is whether the reduction in auxiliation over time is due in part to changes in other parts of the grammar that co-determine the use of the auxiliary. From the regression and tree & forest models, we know that RelClause, Subject, and TempLoc are predictive of the use of Aux. Over the three time periods, the distribution of these factors changed significantly (p < 0.01, chi-squared tests), as shown in Table X.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | RelClause | | Subject | | TempLoc | |
|  | NoRelCl | RelCl | NoSubj | Subj | TempLocNo | TempLocYes |
| early | 30 | 30 | 21 | 39 | 39 | 21 |
| middle | 75 | 75 | 40 | 110 | 79 | 71 |
| late | 379 | 113 | 213 | 279 | 222 | 270 |

Table X: Counts of realizations cross-classified by period for RelClause, Subject, and TempLoc.

Changing habits in these other parts of the grammar may have contributed to the increase in the use of the auxiliary from the middle to the late period. In the middle period, 120 out of 150 observations make use of the auxiliary. A naive discrimination model indicates that, given the data, performance is heavily biased towards using the auxiliary, with only 5 observations being predicted to have no auxiliary, and all remaining 145 cases being predicted to have an auxiliary. Suppose that a speaker with the connection weights from the middle period is immersed in the use of relative clauses, subjects, and temporal location of the late period. What would the rate of the use of the auxiliary be for such a speaker for these new instances of use? We can assess this question by estimating the activations of the auxiliary for the observations of the late period using the weight matrix of the middle period. Table Y summarizes the results.

middle -> middle middle -> late late -> late

Aux NoAux Aux NoAux Aux NoAux

Aux 120 0 Aux 203 2 Aux 79 126

NoAux 25 5 NoAux 227 60 NoAux 78 209

Table Y: Auxiliation as predicted by NDL for the middle and late periods.

middle->middle: weights and predictions from the middle period; middle->late:

weights from the middle period, predictions for the late period; late->late:

weights and predictions from the late period.

Importantly, there is a significant increase in the proportion of non-use of the auxiliary when predictions for the late period are derived from the weights of the middle period (from 0.03 to 0.12, X^2(1)=8.97, p = 0.0027). This suggests that the process of deauxiliation may indeed be due, at least in part, to changing use of relative clauses, temporal location, and overt subjects. If other things had been equal, with no changes in the frequency of use of other factors, the prediction would have been across the board usage of the auxiliary, as shown by a naive discrimination learning model predicting its own preferred choice from the instances from the middle period.

5. Conclusions

[TBA]

QUESTIONS:

References

SPP = František Dvorský, ed. 1869. *Staré písemné památky žen a dcer českých*. Prague: Náklad vlastní — v komisí kněhkupectví F. Řívnáče.

1. There are, however, some notable works in which scholars have confronted rival affixes in English word-formation, such as *-ity* and *-ness*, cf. Riddle 1985 and Aronoff 1976. [↑](#footnote-ref-1)
2. Note that although these prefixes can be added to adjectives and adverbs, this case study focuses exclusively on their use with verbs. [↑](#footnote-ref-2)
3. This logistic regression model is also presented in Sokolova et al. forthcoming. [↑](#footnote-ref-3)
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. [↑](#footnote-ref-4)
5. The reader is invited to follow along with these comparisons of the model and the dataset. This can be done by opening datLOAD.csv in Microsoft xl and using the “Filter” feature to select for the various predictor values. So if you set the Filter to “na” under VERB and to “no” under both REDUCED and PARTICIPLE, you find 70 examples for “goal” and 27 for “theme” under CONSTRUCTION. [↑](#footnote-ref-5)
6. Subjects were allowed to also make an additional reponse (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. [↑](#footnote-ref-6)
7. The p-values are reported here in scientific notation, where “e-X” means “move the decimal X places to the left”. So, for example, 1.07e-06 = 0.00000107. [↑](#footnote-ref-7)