PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 118 APRIL 2022 5–23

Interpreting Statistical Models for Denominal Adjective Formation in Russian

Natalia Bobkova

CLLE, Université de Toulouse 2 Jean Jaurès, France

Abstract

This study focuses on cases of suffixal rivalry in denominal adjective formations in Russian, namely on two adjectival suffixes: *-n-* and *-sk-*. We use statistical modelling (multivariate logistic regression) to shed light on properties of base nouns that contribute to the choice of one of the competing suffixes. In the first part, we provide model interpretation through traditional metrics (accuracy, confusion matrix and model coefficients with their respective *p*-values). However, model accuracy may not be uniform if we compare different samples of the data set and may take a wide range of values. In the second part of this study, we complete our interpretation of model results by performing error analysis in order to get a better understanding of the underlying properties of base nouns that cause model failure. We explore Responsible AI Toolbox widgets for this purpose. One main result of this study is that the same semantic base noun properties are related to both high model performances and model errors.

1. Introduction

The derivation of adjectives from nouns is a complex process in Russian morphology, as there is a lot of variation in the range of suffixes employed. Hence, they constitute a good testing ground for the study of the competition between rival derivational strategies for the same syntactic and semantic function (Lindsay and Aronoff, 2013; Aronoff, 2016).

The use of quantitative methods to investigate the situations of affix rivalry has increased recently. The studies rely heavily on statistical and computational methods as opposed to traditional qualitative research. Quantitative methods are exploited to evaluate the influence of different factors on the selection of rival affixes. Inferential statistics can be based on a variety of models (Baayen et al., 2013), including ana-

 Corresponding author: natalia.bobkova@univ-tlse2.fr
Cite as: Natalia Bobkova. Interpreting Statistical Models for Denominal Adjective Formation in Russian. The Prague Bulletin of Mathematical Linguistics No. 118, 2022, pp. 5–23. doi: 10.14712/00326585.017.

logical models (Chapman and Skousen, 2005; Arndt-Lappe, 2014), logistic regression (Bonami and Thuilier, 2018), word vectors (Wauquier, 2020; Guzmán Naranjo and Bonami, 2021; Huyghe and Wauquier, 2021), neural networks (Guzmán Naranjo, 2019; King et al., 2020).

The competition between adjectival suffixes is determined by a complex combination of phonological, morphological and semantic factors. In this paper we aim at modeling suffixal rivalry in the construction of denominal adjectives in Russian. The approach adopted in this paper consists in studying non-ambiguous cases for each suffix in the data set and in highlighting the emerging properties of base nouns that allow to tease apart competing suffixes. For illustration purposes we will use *-n-* and *-sk-* suffixes data, however, the approach can be applied for both binary and multiclass classification problems (i.e. to include more than two suffixes in the study).

The goal is to understand the role of base noun properties in predicting *-n-* and *-sk-*. There is a variety of models which can be used for this purpose due to their high interpretability. For instance, logistic regression, decision trees or random forest can output variable importance scores (base noun properties) in explaining the outcome (suffix). In this study we use multivariate logistic regression, a well-established statistical modelling framework. The choice of logistic regression over other models is driven by several factors: it is a tool based on statistical formulae, the direction of coefficients (positive or negative) can be associated with two classes of binary classification and, finally, the coefficients are accompanied by statistical significance tests (with p-values). Even if this model has all the advantages listed above, we will not limit our investigation to the classical tools in order to understand it (such as its table of coefficients). In this paper we will explore various quantitative methods for error analysis aiming to highlight patterns or combination of patterns which are not captured by our model, and the reasons behind them.

The error analysis was proposed by King et al. (2020) as approach to understand the output of sequence-to-sequence models, which are generally hard to interpret, for inflectional task in Russian. This paper goes further and uses quantitative and qualitative approaches for error analysis and model interpretation. Based on error analysis, our study provides a new perspective on the nature of suffix rivalry in Russian derivation and sheds light on previously unseen phenomena.

The data on which our study is performed were extracted from the Russian National Corpus. The data set is composed of highly frequent adjectives. Section 2 discusses different problems which emerge when studying adjectives in Russian. Section 3 presents the overview of the Russian National Corpus, data set constitution and base noun properties annotation. Section 4 focuses on building a logistic regression classifier, it provides data on its performance as well as model summary highlighting the base noun properties which are statistically significant for classification task. Section 5 focuses on error analysis and diagnostics, sheds light on base noun properties which may be misleading for the model and discusses the underlying reasons for errors. The error analysis here is complementary to the logistic regression task.

2. Adjectives in Russian

There are various strategies to derive adjectives from nouns in Russian. Classical grammars such as Townsend (1975) or Švedova (1980), for instance, enumerate more than 25 suffixes, which have different degrees of productivity. Three suffixes are identified as being productive in synchrony (Zemskaya, 2015; Hénault and Sakhno, 2015; Kustova, 2018): *-n-*, *-sk-* and *-Ov-* (capital *O* in both cases represents a vowel that may correspond, phonologically, to different surface forms, and orthographically to <o> or <e>). The suffixes in question can be considered as the three main adjectival suffixes (abstract entities, denoted in capital letters), while others may be interpreted as their extended variants, denoted in small letters (Bobkova and Montermini, 2019):

- -N-: *-n-*, *-Ovn-*, *-ičn-*, *-ivn-*, *-on(n)-*, *-en(n)-*, *-(e)stven(n)-*, *-ozn-*, *-al'n-*, *-onal'n-*, *-arn-*, *-in-*;
- -SK-: *-sk-*, *-esk-*, *-česk-*, *-ičesk-*, *-ističesk-*, *-ijsk-*, *-ansk-*, *-ensk-*, *-insk-*, *-istsk-*, *-Ovsk-*;
- -OV-: *-Ov-*.

Recent developments in derivational morphology, cf. Hathout (2011); Plénat (2011); Roché (2011) among others, consider that various types of constraints (phonological, morphological, semantic, pragmatic, etc.) display a complex interaction, resulting in the choice of one of the rival suffixes. However, in the existing literature on Russian language the choice of one or the other suffix is often studied theoretically, through extended data, but not necessarily by means of quantitative analysis. For instance, in Townsend (1975); Švedova (1980); Zemskaya (2015) we can encounter extensive indications on phonological, semantic or lexico-morphological factors that allow the combination with each suffix in question. Graščenkov (2019) references Švedova for the properties of base nouns discussed above and studies syntactic properties of suffixes *-n-* and *-sk-*.[1] Graudina et al. (2001); Hénault and Sakhno (2015), for instance, focus on the semantics of derived adjectives and provide evidence on distinction between *-n-* and *-sk-* based on context the adjectives appear in. However, all the indications are not supported with quantitative and/or statistical evidence.

For the purposes of the present study we will focus on phonological, morphological and semantic properties of the base nouns.

The examples of nouns combining with *-n-* are given in Table 1.[2]. In Švedova (1980), for instance, the following non-extensive indications on *-n-* can be encountered. Semantically, this suffix mainly combines with non-animate common nouns, either abstract (1) or concrete (2), although animate nouns are also possible bases (3). Phonologically, it is stress-neutral, as it combines both with bases with stress on

[1]The analysis is based on the ability of *-n-* and *-sk-* adjectives to form adverbs, to have short forms and comparative forms in their paradigms, to derive abstract nouns, to combine with evaluative suffixes.

[2]For illustration purposes we provide stress position information for the base nouns in Tables 1 and 2 In the rest of the paper these indications will be excluded, except if relevant.

the stem (4) or on inflection (5), and it selects stems displaying consonant mutation (6, 7). Etymologically, it combines both with native (8) and foreign (9) bases.

	noun	**adjective**	**gloss**
1	*gnev*	*gnevn(yj)*	'anger'
2	*kiparís*	*kiparisn(yj)*	'cypress'
3	*inženér*	*inženern(yj)*	'engineer'
4	*kómnat(a)*	*kómnatn(yj)*	'room'
5	*zim(á)*	*zímn(ij)*	'winter'
6	*jazýk*	*jazyčn(yj)*	'tongue / language'
7	*drug*	*družn(yj)*	'friend'
8	*dym*	*dymn(yj)*	'smoke'
9	*arxitektúr(a)*	*arxitekturn(yj)*	'architecture'

Table 1. Sample with -n- suffixation

Table 2 provides examples of nouns combining with *-sk-*. This suffix does not seem to be selective semantically, since it may combine with inanimate (1) and animate (2) nouns, including nouns denoting humans (3), and may also combine with proper nouns (4). Phonologically, it privileges stems ending in alveolar (5) or dental (6) consonants, and, like *-n-*, it selects nouns with stress on the stem, and mutated stems (7,8).

	noun	**adjective**	**gloss**
1	*université*	*universitetsk(ij)*	'university'
2	*kon'*	*konsk(ij)*	'horse'
3	*bandít*	*banditsk(ij)*	'bandit'
4	*Irán*	*iransk(ij)*	'Iran'
5	*soséd*	*sosedsk(ij)*	'neighbour'
6	*šef*	*šefsk(ij)*	'boss'
7	*Vólg(a)*	*volžsk(ij)*	'Volga (river)'
8	*Čéxi(ja)*	*češsk(ij)*	'Czechia'

Table 2. Sample with -sk- suffixation

The literature revision proves that the indications on these properties often lack precision: the same base noun property can be listed as favourable for different suffixes. It remains unclear which properties are statistically significant for the suffix

choice. The goal of this study is twofold: first, we will provide statistic evidence on the base noun properties that allow to discriminate between *-n-* and *-sk-* for highly frequent adjectives through logistic regression model. Second, we will identify and diagnose in depth the error patterns; this investigation will shed light on the distribution of base noun properties across different subsets of data which are prone to model failure.

3. Data

To perform our analysis, we proceeded with web scraping adjectives from the Russian National Corpus (Plungjan et al., 2005),[3] a corpus of modern Russian containing over 600 million words. This corpus is divided in several subcorpora. For the purpose of this study we are interested in standard Russian, both written and spoken. Consequently, the adjectives were extracted from five subcorpora: main (texts representing standard Russian: modern written texts from the 1950s to the present day, real-life Russian speech recordings from the same period, and early texts from the middle of the 18th to the middle of the 20th centuries), media (articles from mass media between 1990 and the 2000s), multimedia (Russian movies between 1930 and 2000), spoken (recordings of public and spontaneous spoken Russian and the transcripts of the Russian movies) and poetic (covers the time frame between 1750 and the 1890s, but also includes some poets of the 20th century).

Having established the types of subcorpora we are interested in, we automatically extracted adjectives by searching lemmas with a final sequence corresponding to *-n-* or *-sk-* immediately preceding inflectional suffixes typical of citation forms of Russian adjectives.[4] 78113 lemmas were extracted, we automatically filtered extended variants (almost 1/3 of the data set). Semi-automatic and manual cleaning further allowed to discard >70% false positives, e.g. forms corresponding to adverbs derived with *-n-* (*vnezapno* 'suddenly'), possessive adjectives in *-in* (*mamin* 'mother$_{\text{POS}}$'), proper nouns (surnames) ending in *-sk-* (*Stanislavsk(ij)* 'Stanislavsky'). This first list was additionally filtered in order to keep only adjectives clearly derived from nouns. The vast majority of remaining adjectives are denominal, other cases were removed: noun to adjective conversions (*zdorov'(e)* - *zdorov(yj)* 'health'; *tajn(a)* - *tajn(yj)* 'secret'), adverb to adjective conversions (*děševo* - *dešëv(yj)* 'cheap', *rano* - *rann(ij)* 'early'), as well as the adjectives without any motivating base. Furthermore, we only took into account adjectives having token frequency >100, excluding non frequent formations along with hapaxes from the present study.

[3] Available at https://ruscorpora.ru/. The choice of web scraping method is driven by the absence of an official API for data access in Ruscorpora.

[4] The citation form of adjectives corresponds to nominative masculine singular. Three orthographic forms are possible: <yj>, <ij>, <oj>.

Base nouns were also automatically reconstructed for each adjective. In case of multiple base candidates (*zritel'*/*zreni*(*e*) - *zritel'n*(*yj*) 'viewer/vision') and polysemy (*kamer*(*a*)$_1$/*kamer*(*a*)$_2$ - *kamern*(*yj*) 'cell/chamber'), these potential base nouns, as well as nouns with different semantics, were included as separate entries and annotated accordingly. Manual assessment at this stage led to verification of the exact shape of the reconstructed base nouns. The final data set was composed of 1048 types (620 for *-n-* and 428 for *-sk-*).

The competition between affixes is driven by a complex combination of factors. In order to examine different dimensions of rivalry, we annotated several properties of base nouns that have been highlighted in previous linguistic works as potential predictors of the suffix, as discussed in Section 2. In what follows we will present these properties in details and give a brief overview of the studies of rivalry mainly in English and French that use the same properties as predictors in modelling.

Etymological property include one binary predictor:

- `Source`: whether the base noun is of Slavic (`0`) or foreign (`1`) origin.

Phonological properties include information about the following features:

- `LastP`: the last phoneme of the stem (`Lab`: labial, `Den`: dental, `Alv`: alveolar, `Vel`: velar or `Vow`: vowel);
- `SyllB`: the length of the base noun in syllables - the only continuous property in the dataset;
- Stress position is also taken into consideration:
  - `AccSyl`: from the phonological point of view: which syllable is stressed – `D`: ultimate, `Ad`: penultimate, `Aad`: antepenultimate (*zim*(*á*) 'winter', *víšn*(*ja*) 'cherry', *rádug*(*a*) 'rainbow');
  - `AccPos`: from the morphological point of view: if the stress is positioned on `R`: the root of the base noun, or – if any – `S`: derivational or `F`: inflectional suffix (*son* 'dream', *marksízm* 'marxism', *galav*(*á*) 'head').

Both the last phoneme of the stem and the length of base noun in syllables are highlighted as important in prediction of the suffix by Lignon (2010) and Bonami and Thuilier (2018) in French, by Lindsay and Aronoff (2013) in English. We complete the list of phonological properties with information on stress position since it is not fixed in Russian and may influence the choice of the suffix.

Morphological properties include only one predictor :

- `InflCl`: the inflectional class of base nouns which is represented by the I, II or III inflectional class (*pap*(*a*)$_{\text{I.M}}$ 'dad', *pesn*(*ja*)$_{\text{I.F}}$ 'song'; *stol*$_{\text{II.M}}$ 'table', *del*(*o*)$_{\text{II.N}}$ 'business'; *ten'*$_{\text{III.F}}$ 'shadow').

We follow a canonical distinction between 3 inflectional classes, although Russian nouns may be divided into larger sets of classes and subclasses (Zaliznjak, 2003; Parker and Sims, 2019; Guzmán Naranjo, 2020). We only include inflectional class as morphological property, however, morphological structure of base nouns may be in-

teresting as well to study suffix rivalry further (Missud and Villoing, 2020; Varvara, 2020).

Morpho-phonological allomorphies typical of Russian inflection and derivation were annotated as well. They include such properties as:

- *Vowel0*: vowel / Ø alternation, binary property (*dvorec - dvorcov(yj)* 'palace');
- *ConsM*: consonant mutation, binary property (*tvorog - tvorožn(yj)* 'cottage cheese').

Both vowel alternation and consonant mutation reflect diachronic processes in Russian and do not correspond to synchronically productive phonological phenomena (Kapatsinski, 2010; Sims, 2017; Timberlake, 2004).

Possible differences in the semantics of derivatives may be considered as well, with respect to descriptive properties (Baeskow, 2012; Fradin, 2016). We include the following semantic properties of base nouns in this study:

- Binary distinct properties of [±proper], [±human], [±animate], [±concrete], [±countable];
- `A`: animacy, or the combination of the properties listed above into five groups (Thuilier, 2012):
  - `PropHum`: proper human (*Pifagor* 'Pithagoras');
  - `ComHum`: common human/animate (*sobak(a)* 'dog');
  - `ComConc`: common concrete (*dom* 'house');
  - `PropNHum`: proper non-human (*Al'p(y)* 'Alps');
  - `ComAbst`: common abstract (*sojuz* 'alliance').

After performing descriptive statistics analysis and test for multicolinearity,[5] some data were removed before modeling. For instance, the nouns with samples of properties that are not large enough to be statistically representative were dropped out (nouns with six-syllabic structure, nouns where the forth syllable from the end is stressed). Highly correlated base noun properties were also removed. This concerns binary semantic features since they strongly correlate to animacy subclasses, as well as consonant mutation which strongly correlates to velar ending stems.

The data set for modelling is composed of 1020 examples, 612 for *-n-* and 408 for *-sk-*.

4. Model

All the base noun properties listed in previous section virtually combine to form a complete picture of situations of rivalry. In what follows we will examine their predictive power for the suffix choice when they are put all together.

We use logistic regression, a multifactorial statistical tool which allows to examine the relationship between a binary dependent categorical variables and predictor

[5] For more details on methodological aspects cf. Bobkova (2022).

variables. The implementation is made with statsmodels module in Python (Seabold and Perktold, 2010).

The data were randomly divided into training and test (with test size of 20%, so the model was trained on 816 examples and tested on 204). We ran 500 simulations of train-test split with a different random state.[6] The goal of this manipulation is twofold. First, we aimed at assessing overall model AUC score when trained and tested of different subsets of original data (`mean AUC`: 0.8957, `min AUC`: 0.8345; `max AUC`: 0.9502, `std`: 0.020)).[7] Second, since overall model performance is high it does not make a lot of mistakes, and given the relatively small test set, we searched for the worst performing model in order to maximize error rates and have enough material for further analysis.

We will now focus on the model with the lowest AUC (0.8345) and investigate its performance and properties. We will use logistic regression table of coefficients (Table 3) to evaluate statistical significance of predictors.

First, we use p-values in order to understand if a particular base noun property is useful for suffix prediction. The p-value less than 0.05 suggests that the property has a significant effect on the suffix choice. The model summary states that [+common, +human] and [-common, +human] semantic properties, as well as [+dental]-ending stems are statistically significant for predicting suffix ($p<0.000$). The following parameters are also significant, but to a lesser extent: [+labial]- ($p<0.012$), [+alveolar]- ($p<0.032$) and [+velar]-ending stems ($p<0.042$), inflectional class 2 ($p<0.021$) and 1 ($p<0.031$). Source, the length of base noun in syllables, vowel-∅ alternation, [+common, +concrete] semantic property, morphological and phonological stress positions are not statistically significant for *-n-* and *-sk-* classification problem.

Second, we can interpret coefficients which compare the outcome for each level of a base noun property with the reference level (the reference levels for each categorical predictor correspond to Slavic origin, absence of Ø vowel, common abstract, stressed root, stressed antepenultimate syllable, inflectional class 3, vowel-ending stem). Positive coefficients increase the chances for the model to predict *-sk-* ([+common, +human], [-common, +human], inflectional class 1 and 2), negative coefficients, in turn, decrease odds for *-sk-* and increase the probability for predicting *-n-* ([+dental], [+labial]-, [+alveolar]- and [+velar]-ending stems).

Table 4 provides confusion matrix. 31 nouns out of 204 were misclassified, the error rate is 14.7%. This table also suggests that more classification errors were made for *-sk-* (25.3% of misclassified data) rather than for *-n-* (8.5% of errors). We will proceed with an in-depth investigation of these errors as well as underlying possible reasons for them in the following section.

[6] 500 is an arbitrary choice in order to have a large number of simulations.

[7] Compared to AUC score, overall accuracy score is higher: `mean accuracy`: 0.9079, `min accuracy`: 0.8534; `max accuracy`: 0.9559, `std`: 0.018.

	coef	std err	z	P>|z|
Intercept	-4.0000	1.124	-3.559	0.000
Source	0.0909	0.306	0.297	0.766
BaseLen	0.2158	0.184	1.173	0.241
Vowel0	-1.0549	0.680	-1.550	0.121
A_ComConc	0.2601	0.342	0.754	0.451
A_ComHum	4.2509	0.394	10.785	0.000
A_PropNHum	11.4359	1.177	6.461	0.000
StressMo_DerS	0.5390	0.584	0.924	0.356
StressMo_InfS	-0.1964	0.732	-0.268	0.789
StressPho_ad	-0.7678	0.524	-1.464	0.143
StressPho_d	-0.4269	0.604	-0.684	0.494
InflCl_1	2.7431	1.274	2.153	0.031
InflCl_2	2.8106	1.219	2.306	0.021
LastPh_cAlv	-0.7735	0.361	-2.143	0.032
LastPh_cDent	-1.3741	0.379	-3.627	0.000
LastPh_cLab	-1.0472	0.416	-2.518	0.012
LastPh_cVel	-0.8042	0.396	-2.031	0.042

Table 3. Model summary

	predicted *-n-*	**predicted *-sk-***
true *-n-*	118	11
true *-sk-*	19	56

Table 4. Confusion matrix

Classification report is shown in Table 5. Even if the chosen model has the lowest accuracy, it still performs quite well: with accuracy of 85.3% and AUC of 83.5%, good precision and descent recall. However, these metrics, especially accuracy, may not be uniform across different subsets of data. Moreover, these metrics do not allow to identify important conditions of inaccuracies. The model may perform better for some initial base noun properties and worse for others. Therefore, an in-depth analysis is needed to convey a detailed interpretation of model behavior.

5. Error analysis

In this section we will further investigate the performance of the model, namely through data exploration and interpretability techniques as well as through an anal-

metric	value
Accuracy	0.853
AUC	0.835
Precision	0.836
Recall	0.747
False Positive Rates	0.085
False Negative Rates	0.253

Table 5. Classification report

ysis of how failure is distributed for a model. We will use visualisation methods provided by Responsible AI.[8]

The Error analysis[9] and Interpretability[10] dashboards are integrated within the Responsible AI Widgets. They enable a better understanding of overall and local predictions of the model as well as of model errors (Nushi et al., 2018; Amershi et al., 2019; Bansal et al., 2019; Srivastava et al., 2020). These tools allow to work with regression and classification problems, both binary and multiclass. Responsible AI tools can be used to assess any kind of models (statistical or machine learning), even the models which are not easily interpretable (for instance, deep learning models). In what follows we will complete the assessment of the logistic regression classifier used in this study.

Error analysis dashboard enables the visualization of data subsets with higher error rates than the overall error score. These errors may occur when the model faces specific set of properties among independent variables, i.e. the properties of base nouns for which the model underperforms.

As assessed in the previous section through confusion matrix, the overall error rate is 14.71% since 31 out of 204 base nouns were associated with the wrong suffix.

However error patterns may be complex and involve several properties of base nouns. The Figure 1 groups all misclassified data into subsets which can be easily interpreted in a tree-like structure. This tree uses the mutual information between each property and the error on the true labels to best separate error instances from success instances hierarchically in the data. This allows to visualize common patterns in model failure. The following information is available for this binary tree: error rate (portion of instances in the node for which the model is incorrect, shown through the

[8]https://github.com/microsoft/responsible-ai-toolbox

[9]https://github.com/microsoft/responsible-ai-toolbox/blob/main/docs/erroranalysis-dashboard-README.md

[10]https://github.com/microsoft/responsible-ai-toolbox/blob/main/docs/explanation-dashboard-README.md

intensity of color); error coverage (portion of all errors that fall into the node, shown through the fill rate of the node) and data representation (number of instances in the node, shown through the thickness of the incoming edge to the node along with the actual total number of instances in the node).

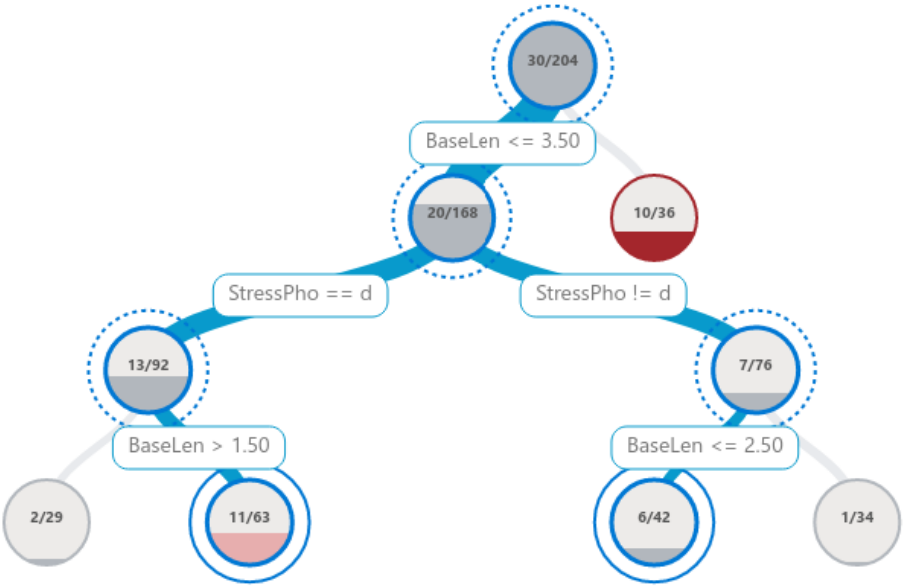

Figure 1. Error tree for logistic regression model

This decision tree represents combined data on two branches. The root node contains the information about the length of base noun in syllables. It allows for further partitioning data into two groups, based on the following condition: if the number of syllable is less than or greater than 3.5.

While the overall error rate is 14.71% for the whole dataset, the error rate can be as high as 27.78%, which corresponds to the extreme right branch with only one node, 10 out of 36 cases of wrong classification (for base nouns of 4 or 5 syllables). Six nouns are 5-syllabic (*gumanitarij* 'humanitarian', *bogoslovi*(*e*) 'theology', *artillerij*(*a*) 'infantry', *universitet* 'university', *žurnalistik*(*a*) 'journalism', *professional* 'professional'), the other four are 4-syllabic (*veterinar* 'vet', *vselennaj*(*a*) 'universe', *čudoviščˊ*(*e*) 'monster', *distribˊjutor* 'distributor').

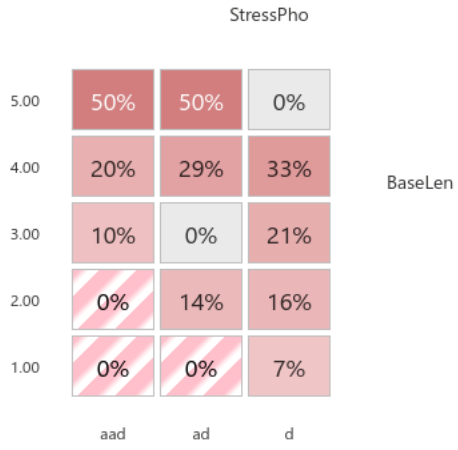

Figure 2. Error rate for the length of the base noun in syllables and phonological stress position

More information can be found on the left branch. It can be divided into two subbranches and concerns errors for the nouns based on their phonological stress position.

The out-most left subbranch concerns errors that occur in case when the length of base noun in syllables is less than 3.5 and more than 1.5 (i.e. 2 and 3 syllabic nouns), combined with the last stressed syllable property. The hierarchical error pattern here shows that the error rate for this particular combination of properties is higher than the average: 17.46%, 11 out of 63 nouns were misclassified. Among the misclassified nouns we encounter six 2-syllabic nouns (*glav(a)* 'leader', *dekabr'* 'December', *latyn'* 'Latin', *sentjabr'* 'September', *senat* 'senate', *raspad* 'disintegration') and five 3-syllabic nouns (*kardinal* 'cardinal', *xoxlom(a)* 'khokhloma (painting)', *seminar* 'seminar', *komitet* 'committee', *monastyr'* 'monastery').

The right subbranch of the tree is less interesting, since less errors can be found here. The error rate is 14.29% which is slightly lower than the overall error rate, only 6 out of 42 selected nouns were incorrectly classified (monosyllabic and 2-syllabic nouns where any syllable is stressed except for the last one). We will not focus on these error subset and analyze two previous subsets in more details.

The error heat map shown on Figure 2 allows to further investigate how the phonological properties in question impact the error rate across data subsets. Indeed, the highest error rates (up to 50%) are encountered for 5-syllabic base nouns, regardless phonological stress position. This heat map reveals that the error rates are also visibly higher for the nouns where the last syllable is stressed.

In previous section we assessed properties of base nouns which are statistically significant for suffix choice. Both the length of base noun in syllables and phonological stress position were not listed among these properties. Hence error analysis suggests that based one these properties we can isolate subsets of data with the highest error rates. But does this mean that these features are correlated to model errors?

Interpretability dashboard allows the exploration of the top important features that impact the overall model predictions. In previous section we saw that animacy, the last phoneme of the stem and the inflectional class are statistically significant in predicting if the suffix is *-n-* or *-sk-*. Not surprisingly, the visualizations available within Responsible AI toolbox prove the same, as shown on Figure 3 (`All data`). Moreover, it is possible to compare feature importance values for different selected subgroups of data side by side, for instance, the subgroups with the highest error rates (`BaseLenStressPho`: 2 and 3 syllabic nouns where the last syllable is stressed; `BaseLen`: 4 and 5 syllabic nouns).

Based on the information on feature importance and the ordering we can conclude that, in general, the model behaves in the same way on the whole data set and the two subgroups with highest errors (the only difference concerns 4 and 5 syllabic nouns: inflectional class appears to be slightly more important than the last phoneme for this data subset). This means that the same base nouns features are leveraged for predicting suffix across the three sets and that phonological stress position as well as the length of base noun in syllables are useful to isolate the majority of model errors, but they are not necessary correlated to these errors.

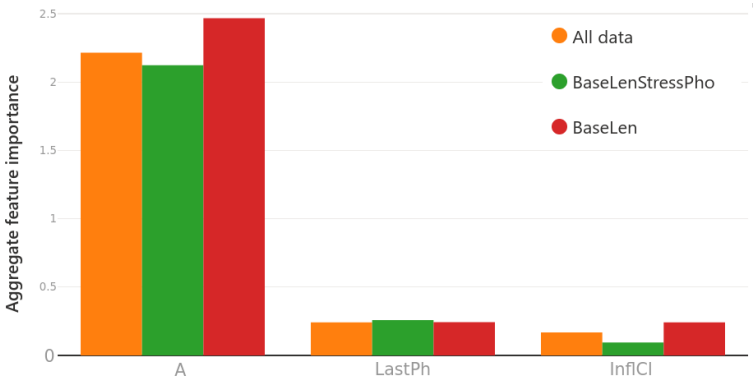

Figure 3. Top 3 features by their importance

In order to understand the reasons behind the erroneous predictions in test set we will contrast them to train data and to correctly classified data. For consistency, we

will isolate the same subgroups in train subset and correct predictions as for incorrect predictions: 2 and 3 syllabic nouns with the last stressed syllable; 4 and 5 syllabic nouns.

subset	**CH**	**CC**	**PNH**	**CA**	**CH**	**CC**	**PNH**	**CA**
train: *-n-*	9	42	10	113	2	2	0	43
test: correct *-n-*	0	15	0	25	0	1	0	13
test: incorrect *-n-*	2	0	0	0	4	0	0	0
train: *-sk-*	55	6	43	16	20	2	47	10
test: correct *-sk-*	7	0	5	0	3	0	9	0
test: incorrect *-sk-*	0	1	0	8	0	2	0	4

Table 6. Distribution of animacy across subsets:
2- and 3-syllabic nouns, the last stressed syllable | 4- and 5-syllabic nouns
(CH: ComHum, CC: ComConc, PNH: PropNHum, CA: ComAbst)

Table 6 presents the distribution of animacy across two subset: 2- and 3-syllabic nouns with the last stressed syllable - in the left part of the table; 4- and 5-syllabic nouns - in the right part. Three main trends are observed here. First, the distributions of the most important base noun property to the suffix choice - animacy - are similar between 2-3-syllabic nouns with the last stressed syllable and 4-5-syllabic nouns. For instance, common abstract nouns are more numerous in both subsets for *-n-* training data (113 and 43 cases respectively). We observe the same tendencies in training set for *-sk-*: common human and proper non human nouns are the most represented (55 and 43 cases for the first subset and 20 and 47 - for the second). Second, train data distributions and correctly predicted data distributions follow the same patterns as well (common abstract nouns are the ones that are most numerous for *-n-* classification - 25 and 13 respectively in both subsets; similarly to train set, common human and proper non human nouns correctly predicted are the most numerous for *-sk-* (7 and 5; 3 and 9)). The third observation concerns test set where animacy has distinct distributions between correctly and incorrectly predicted data. For instance, if we take into consideration *-n-* distribution, we can see that common concrete nouns and common abstract nouns were correctly predicted with *-n-* suffix, whereas common human nouns (2 and 4 in both subsets) were mistakenly associated with *-sk-*. Similarly, with *-sk-* distribution, common human and proper non-human nouns are correctly identified with *-sk-*, but some common concrete (1 and 2) and common abstract nouns (8 and 4) were mistakenly classified with *-n-*.

The error cases are the following:

1. 2- and 3-syllabic nouns, the last stressed syllable
    - actual *-n-* suffix

    - ComHum: *glav*(*a*) 'leader', *kardinal* 'cardinal'
  - actual *-sk-* suffix
    - ComConc: *monastyr'* 'monastery'
    - ComAbst: *dekabr'* 'December', *komitet* committee', *latyn'* 'Latin', *raspad* 'disintegration', *seminar* 'seminar', *senat* 'senate', *sentjabr'* 'September', *xoxlom*(*a*) 'khokhloma (painting)'
2. 4- and 5-syllabic nouns
  - actual *-n-* suffix
    - ComHum: *gumanitarij* 'humanitarian', *professional* 'professional', *veterinar* 'vet', *čudoviščе*(*e*) 'monster'
  - actual *-sk-* suffix
    - ComConc: *distrib'jutor* 'distributor', *universitet* 'university'
    - ComAbst: *artillerij*(*a*) 'infantry', *bogoslovi*(*e*) 'theology', *vselennaj*(*a*) 'universe', *žurnalistik*(*a*) 'journalism'

Even if the examples of common error patterns are not numerous, the conclusion is that misclassified data follows in general the distribution which is the opposite to the true suffix label. This can explain model errors: the model fails to discriminate correctly between two rival suffixes if the distribution of base noun properties is unusual (compared to the training data) for a specific suffix.

6. Conclusion

A brief literature overview given in Section 2 suggests that the topic of affix rivalry in denominal adjective formation in Russian is mostly approached with descriptive methods, statistical studies performed on a big corpus are missing. The modelization performed in Section 4 confirms the conclusions encountered in literature; in addition, it provides evidence on statistical significance of the properties of base nouns that allow to discriminate between the rival suffixes. Moreover, the error analysis performed in Section 5 sheds light on specific combination of properties that may behave differently and have a specific preference for the suffix which can't be drawn from the model.

This study was made using a logistic regression classifier in order to discriminate between *-n-* and *-sk-* adjectival suffixes in Russian. Overall, the model performs very well, with AUC ranging from 0.83 to 0.95, depending on train-test split. The choice of a simple logistic regression classifier is driven by its high transparency, since it allows an easy access to model parameters with feature importance and relevant statistics. For instance, the following base noun properties are statistically significant to predict *-n-* or *-sk-*: [+common,+human], [-common,+human] [+dental]-ending stems; to a lesser extent - [+labial]-, [+alveolar]-, and [+velar]-ending stems, inflectional class 2 and 1.

Compared to logistic regression, other classification models may not be interpretable that easily. Therefore, Responsible AI tools contribute to a better understanding of the

output of ”black box” models. Even if logistic regression is transparent, it is nevertheless possible to get extra insights for this model through error analysis, and Responsible AI provides dashboards for relevant visual explorations which are easily interpretable as well.

The main tool used for the present study is binary tree which allows to isolate subsets of test data with the highest error rates. This complements the information about the most relevant features for the classification task with information on features that group data into subsets where model fails more often than on average. The overall error rate of the model is 14.71%, however, *-n-* and *-sk-* data may be grouped into subsets where error rates are even higher based on the length of the base noun in syllables and phonological stress position: 2 and 3 syllabic nouns with the last stressed syllable (11 nouns misclassified, error rate 17.46%), 4 and 5 syllabic nouns (10 nouns misclassified, error rate: 27.78%). These two subsets group data with more than two thirds of all misclassified nouns (11 false positives for *-sk-* and 19 false positives for *-n-*).

However, if it is possible to isolate error cases by certain phonological patterns, it does not necessarily implies that these exact patterns cause model failure. A closer look on aggregate feature importance suggests that the same properties are important for subclasses with the highest error rate and the whole data set. For instance, the most statistically significant property of the base noun that contribute to the suffix choice is animacy, and it remains significant across all the studied data sets (all data and two data sets with highest errors). The model failure can be explained by some cases of base noun properties distributions which do not follow the same patterns as in training set.

One possible extension of this approach would be including the combination of properties which lead to higher error rates as interaction terms in our model and to test weather it improves overall accuracies of the model and decreases the error rate. The approach used in this study should also be extended to additional binary classification problems (*-n-*/*-Ov-* and *-sk-*/*-Ov-*) and it may be applied to a multiclass classification involving all the three suffixes. This could provide a finer-grained quantitative evidence and potentially complete the discussion on suffix rivalry for denominal adjectives in Russian.

Bibliography

Amershi, Saleema, Andrew Begel, Christian Bird, Robert DeLIne, Harald Gall, Ece Kamar, Nachi Nagappan, Besmira Nushi, and Tom Zimmermann. Software Engineering for Machine Learning: A Case Study. In *International Conference on Software Engineering (ICSE 2019) - Software Engineering in Practice track*. IEEE Computer Society, 2019. doi: 10.1109/ICSE-SEIP.2019.00042.

Arndt-Lappe, Sabine. Analogy in suffix rivalry: The case of English-ity and-ness. *English Language & Linguistics*, 18(3):497–548, 2014. doi: 10.1017/S136067431400015X.

Aronoff, Mark. Competition and the lexicon. In Elia, Annibale, Iacobini, Claudio, and Voghera, Miriam, editors, *Livelli di Analisi e fenomeni di interfaccia. Atti del XLVII congresso internazionale della Società Linguistica Italiana*, pages 39–52. Bulzoni, 2016.

Baayen, Harald, Anna Endresen, Laura A Janda, Anastasia Makarova, and Tore Nesset. Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian linguistics*, 37(3):253–291, 2013. doi: 10.1007/s11185-013-9118-6.

Baeskow, Heike. -ness and-ity: Phonological exponents of n or meaningful nominalizers of different adjectival domains? *Journal of English Linguistics*, 40(1):6–40, 2012. doi: 10.1177/0075424211405156.

Bansal, Gagan, Besmira Nushi, Ece Kamar, Dan Weld, Walter Lasecki, and Eric Horvitz. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. In *AAAI Conference on Artificial Intelligence*. AAAI, 2019. doi: 10.1609/aaai.v33i01.33012429.

Bobkova, Natalia. Statistical modelization of suffixal rivalry in Russian: adjectival formations in -sk- and -n-. *Corpus*, (23), 2022. doi: 10.4000/corpus.6580.

Bobkova, Natalia and Fabio Montermini. Suffix rivalry in Russian: what low frequency words tell us. In *Mediterranean Morphology Meetings*, volume 12, pages 1–17, 2019.

Bonami, Olivier and Juliette Thuilier. A statistical approach to rivalry in lexeme formation: French *-iser* and *-ifier*. *Word Structure*, 11(2), 2018.

Chapman, Don and Royal Skousen. Analogical modeling and morphological change: the case of the adjectival negative prefix in English. *English Language & Linguistics*, 9(2):333–357, 2005. doi: 10.1017/S136067430500167X.

Fradin, Bernard. L'interprétation des nominalisations en N-age et N-ment en français. In Rainer, Franz, Russo, Michela, and Sanchez Miret, Fernando, editors, *Actes du XXVIIe congrès international de linguistique et philologie romanes* (*Nancy, 15-20 juillet 2013*). Société de linguistique romane/Eliphi, 2016.

Graudina, Ljudmila, Viktor Ickovič, and Lija Katlinskaja. *Grammatičeskaja pravil'nost' russkoj reči: stilističeskij slovar' varintov*. Nauka, 2001.

Graščenkov, Pavel. *Grammatika prilagatel'nogo. Tipologija ad"ektivnosti i atributivnosti*. Litres, 2019.

Guzmán Naranjo, Matías. *Analogical classification in formal grammar*. Language Science Press, 2019.

Guzmán Naranjo, Matías. Analogy, complexity and predictability in the Russian nominal inflection system. *Morphology*, 30(3):219–262, 2020. doi: 10.1007/s11525-020-09367-1.

Guzmán Naranjo, Matías and Olivier Bonami. Comparing derivational processes with distributional semantics. *ParadigMo II*, page 25, 2021.

Hathout, Nabil. Une approche topologique de la construction des mots: propositions théoriques et application à la préfixation en anti. *Des unités morphologiques au lexique*, pages 251–318, 2011.

Huyghe, Richard and Marine Wauquier. Distributional semantics insights on agentive suffix rivalry in French. *Word Structure*, 14(3):354–391, 2021. doi: 10.3366/word.2021.0194.

Hénault, Christine and Sergueï Sakhno. Çem supermarket-n-yj luçşe supermarket-sk-ogo? Slovoobrazovatel'naja sinonimija v russkix ad"ektivnyj neologizmax po dannym interneta. *B. Tošovic, A. Wonisch. Wortbildung und Internet*, 2015.

Kapatsinski, Vsevolod. Velar palatalization in Russian and artificial grammar: Constraints on models of morphophonology. *Laboratory phonology*, 1(2):361–393, 2010. doi: 10.1515/labphon.2010.019.

King, David, Andrea Sims, and Micha Elsner. Interpreting sequence-to-sequence models for Russian inflectional morphology. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 409–418, 2020.

Kustova. Prilagatel'nye. *Materialy k korpusnoj grammatike russkogo jazyka. Vyp.3. Časti reči i leksiko-grammatičeskie klassy*, pages 40–107, 2018.

Lignon, Stéphanie. –iser and –ifier suffixations in French: Verify data to verize hypotheses? In *Décembrettes 7*, 2010.

Lindsay, Mark and Mark Aronoff. Natural selection in self-organizing morphological systems. In *Morphology in Toulouse. Selected Proceedings of Décembrettes 7 (Toulouse 2-3 December 2010)*, pages 133–153. Lincom Europa, 2013.

Missud, Alice and Florence Villoing. The morphology of rival-ion,-age and-ment selected verbal bases. *Dany Amiot Delphine Tribout*, page 29, 2020.

Nushi, Besmira, Ece Kamar, and Eric Horvitz. Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure. In *HCOMP 2018*. AAAI, 2018.

Parker, Jeff and Andrea Sims. Irregularity, paradigmatic layers, and the complexity of inflection class systems: A study of Russian nouns. *InP. Arkadiev & F. Gardani Eds. The Complexities of Morphology*, 2019. doi: 10.1093/oso/9780198861287.003.0002.

Plénat, Marc. Enquête sur divers effets des contraintes dissimilatives en français. In Roché, Michel, Boyé, Gilles, Hathout, Nabil, Lignon, Stéphanie, and Plénat, Mark, editors, *Des unités morphologiques au lexique. Paris: Hermès-Lavoisier*, pages 145–190, 2011.

Plungjan, Vladimir, Tat'jana Reznikova, and Dmitrij Sičinava. Nacional'nyj korpus russogo jazyka: obščaja xarakteristika. *Naučno-texničeskaja informacija. Serija 2: Informacionnye processy i sistemy*, (3):9–13, 2005.

Roché, Michel. Quel traitement unifié pour les dérivations en-isme et en-iste? In Roché, Michel, Boyé, Gilles, Hathout, Nabil, Lignon, Stéphanie, and Plénat, Mark, editors, *Des unités morphologiques au lexique. Paris: Hermès-Lavoisier*, pages 69–143. Hermès-Lavoisier, 2011.

Seabold, Skipper and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010. doi: 10.25080/Majora-92bf1922-011.

Sims, Andrea. Slavic morphology: Recent approaches to classic problems, illustrated with Russian. *Journal of Slavic Linguistics*, 25(2):489–524, 2017. doi: 10.1353/jsl.2017.0019.

Srivastava, Megha, Besmira Nushi, Ece Kamar, Shital Shah, and Eric Horvitz. An Empirical Analysis of Backward Compatibility in Machine Learning Systems. In *KDD*, 2020. doi: 10.1145/3394486.3403379.

Švedova, Natal'ja. *Russkaja grammatika*, volume 1. Moskva: Nauka, 1980.

Thuilier, Juliette. *Contraintes préférentielles et ordre des mots en français*. PhD thesis, Université Paris-Diderot-Paris VII, 2012.

Timberlake, Alan. *A reference grammar of Russian*. Cambridge University Press, 2004.

Townsend, Charles Edward. *Russian word-formation*. Slavica Publishers, 1975.

Varvara, Rossella. Constraints on nominalizations: Investigating the productivity domain of Italian-mento and-zione. *Zeitschrift für Wortbildung/Journal of Word Formation*, 4(2):78–99, 2020. doi: 10.3726/zwjw.2020.02.05.

Wauquier, Marine. *Confrontation des procédés dérivationnels et des catégories sémantiques dans les modèles distributionnels*. PhD thesis, Université Toulouse II Jean Jaurès, 2020.

Zaliznjak, Andrej. *Grammatičeskij slovar' russkogo jazyka*. Russkie slovari, 2003.

Zemskaya, Elena. *Jazyk kak dejatel'nost'. Morfema, slovo, reč*. Moskva: Flinta, 2015.

Address for correspondence:
Natalia Bobkova
natalia.bobkova@univ-tlse2.fr
Université de Toulouse 2 Jean Jaurès, Maison de la recherche, B503
5, allée Antonio Machado 31058 Toulouse cedex 9, France