

# Statistical modelling of doublets in denominal adjective formation in Russian

**Natalia Bobkova**  
CLLE, CNRS  
University of Toulouse  
natalia.bobkova@univ-tlse2.fr

## Abstract

This paper presents a quantitative study of doublets in denominal adjective formation in Russian and aims at identifying the underlying phonological, morphological and semantic properties of base nouns which allow the choice of more than one suffix to form adjectives. First, we extracted doublets from National Corpus of Russian language, then we annotated the properties of base nouns, trained logistic regression models to learn patterns and, finally, analyzed characteristics of nouns which allow the combination with both rival affixes.

## 1 Introduction

The derivation of adjectives from nouns is a complex process in Russian morphology, as these lexemes display a great deal of variation in the range of suffixes employed. Consequently, 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; Bonami and Thuilier, 2018). As various strategies are employed to form adjectives from nouns, doublets (and even triplets) of adjectives formed on the same base with distinct suffixes exist.

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. In general, three approaches may be applied to address the problems of rivalry. The first one consists in studying non-ambiguous cases for each suffix in the data set and highlighting the emerging properties of base nouns that allow to tease apart the suffixes, making them mutually exclusive. The second approach aims at studying ambiguous cases, e.g. cases where the base noun allows more than one suffix to form adjectives. The third approach is hybrid and consists in combining the previous ones and in investigating the properties of base nouns that allow for multiple adjective-forming affixation as opposed to nouns that do not. This approach would allow to establish a comparison between nouns that do and do not allow for adjectival doublets. In the present paper we focus on the second approach and leave the others for distinct studies. The goal of this paper is thus to shed light on the properties on base nouns that are less restrictive for the choice of the suffixes.

The data on which our study is performed were extracted from the National corpus of Russian language. The data set is composed of doublets: cases where two adjectives are attached to one stem. As various suffixes are employed to form adjectives, we first explore to which extent each of the suffix can be statistically predicted. We use the following statistical tools: correlation coefficients (Cramer's V for categorical variables) and one-to-rest logistic regression. We then focus on the properties of base nouns that allow for the formation of adjectival doublets with a given pair of suffixes as opposed to nouns that allow for doublets with all the other pairs. Binomial logistic regression is used in this case.

## 2 Rivalry in denominal adjectival formations

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-*.

In this paper we are interested in the rivalry between the following suffixes: *-n-*, *-sk-*, *-Ov-*, *-Ovsk-*, *-ičesk-*, *-ičn-*, *-esk-*, and, in particular, in cases where two different suffixes can result in the coexistence of two adjectives. The choice of this particular set of suffixes and their extended variants is motivated by the fact that they constitute the most frequent cases of rivalry in our data set (cf. Section 3).

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, or in the emergence of doublets. As far as the Russian language is concerned, the doublets are commonly encountered in denominal adjective formation, along with triplets, however less numerous, as shown in Table 1.

Base noun	Adj_1	Adj_2	Adj_3	Suffixes
ZIMA 'winter'	ZIMOVOJ	ZIMNIJ		<i>-Ov-/n-</i>
MUZEJ 'museum'	MUZEJNYJ	MUZEJSKIJ		<i>-n-/sk-</i>
LONDON 'London'	LONDONOVSKIJ	LONDONSKIJ		<i>-Ovsk-/sk-</i>
ANEMIJA 'anemia'	ANEMIČESKIJ	ANEMIČNYJ		<i>-ičesk-/ičn-</i>
DRUID 'druid'	DRUIDIČESKIJ	DRUIDSKIJ		<i>-ičesk-/sk-</i>
LOGIKA 'logic'	LOGIČESKIJ	LOGIČNYJ		<i>-esk-/n-</i>
BOEC 'fighter'	BOJCOVYJ	BOJCOVSKIJ		<i>-Ov-/Ovsk-</i>
KON 'horse'	KONEVOJ	KONNYJ	KONSKIJ	<i>-n-/sk-/Ov-</i>

Table 1: Doublets and triplets in Russian adjectival formation

The choice of one or the other of the suffixes is accounted for by scholars (Townsend, 1975; Švedova, 1980; Hénault and Sakhno, 2015) by purely phonological factors, semantic or lexico-morphological ones:

- *-n-* tends to form more qualitative adjectives, whereas *-sk-* is used to form more relational ones;
- *-Ov-* appears with inanimate base nouns, *-Ovsk-* chooses to combine with animate ones;
- *-esk-* privileges nouns with stems ending with velars;
- *-ičesk-* appears in particular in lexemes of foreign origin, and consequently also with lexemes containing specific suffixes / combining forms (e.g. *-ija-*, *-izm-*, *-ik-*, etc.).

However, little studies are devoted to the existence of doublets or triplets (Antipina, 2012), namely to the properties of base nouns which do not restrict the choice of one affix. The goal of this paper is to use statistical approaches to reveal the main properties of base nouns (constraints) which may allow the choice of more than one affix (for instance, exactly two suffixes).

### 3 Data

#### 3.1 Corpus

To perform our analysis, we extracted adjectives from the National corpus of Russian language (<https://ruscorpora.ru/>), a corpus of modern Russian containing over 600 million words. This corpus is divided in several subcorpora:

- Main subcorpus: texts representing standard Russian. It can be subdivided into 3 parts, each of which has its distinguishing features: modern written texts (from the 1950s to the present day), a subcorpus of real-life Russian speech (recordings of oral speech from the same period), and early texts (from the middle of the 18th to the middle of the 20th centuries);
- Media subcorpus: articles from mass media between 1990 and the 2000s;
- Multimedia subcorpus: Russian movies between 1930 and 2000;
- Corpus of Spoken Russian: recordings of public and spontaneous spoken Russian and the transcripts of the Russian movies;
- Poetry subcorpus: covers the time frame between 1750 and the 1890s, but also includes some poets of the 20th century;
- Dialectal subcorpus: recordings of dialectal speech (presented in loosely standardized orthography) from different regions of Russia;
- Educational subcorpus: small disambiguated corpus adapted for the Russian educational program;
- Parallel text subcorpus: texts in Russian are complemented by their translations into different languages, and vice versa.

For the purpose of this study we are interested in standard Russian, written or spoken. Dialectal, as well as educational and parallel subcorpora were therefore ruled out. The adjectives thus come from five subcorpora: main, media, multimedia, oral and poetic.

#### 3.2 Data collection

Having established the types of subcorpora we are interested in, we automatically extracted adjectives based on their derivational suffixes. The raw extraction resulted in more than 75 thousands of adjectives. We then automatically grouped adjectives derived from the same base noun; base nouns were automatically reconstructed as well. This operation generated a list of 1968 raw base nouns with at least two adjectives derived from each noun.

Manual verification followed and concerned the verification of the exact shape of the base noun and the correct assignment of all the adjectives which might be potentially formed on it. Manual cleaning resulted in suppression of false positives as well. Among false positives we encounter:

- proper nouns formed mainly with *-Ovsk-*, *-Ov* and *-sk-* suffixes: STANISLAVSKIJ, MENDELEEV, AJVAZOVSKIJ;
- forms of nouns corresponding to genitive plural with *-ov* as an inflectional suffix: *dvor<sub>NOM</sub>* 'yard' - *dvorov<sub>GEN</sub>*;
- possessive adjectives with the suffix *-ov-*: *DED* 'grandpa' - *DEDOV*. Despite of the fact that they are denominal, these adjectives were also excluded from this study due to their morphological and semantic peculiarities.

Manual verification led us to a data set composed of 773 base nouns with 1593 derived adjectives (729 cases of doublets, 41 cases of triplets, 3 cases of quartets). The individual suffixes distribution is presented in Table 2, showing the most frequent suffixes used to form adjectives in this data set.

	<i>-n-</i>	<i>-Ov-</i>	<i>-sk-</i>	<i>-ičesk-</i>	<i>-Ovsk-</i>	<i>-ičn-</i>	<i>-esk-</i>
Frequencies	450	322	320	180	169	81	71

Table 2: Distribution of individual suffixes

The data in Table 2 suggest that the three main suffixes are the most frequent among doublets, while their extended variants are less numerous. Table 2 also shows some trends concerning suffixes that form doublets: extended variants of *-sk-* are more frequent than those of *-n-* (*-ičn-* is the only extended variant of *-n-* encountered here).

Since doublets represent 95% of the data, we kept only them for the present study. Triplets and quartets were excluded as statistical models can perform poorly due to the little data. Our final data set is thus composed of 773 base nouns and 1458 adjectives. The most common couples of suffixes that can combine with the same base and form doublets are displayed in Table 3.

Suffixes	N of bases
<i>-Ov-/-n-</i>	226
<i>-n-/-sk-</i>	100
<i>-Ovsk-/sk-</i>	75
<i>-ičesk-/ičn-</i>	71
<i>-ičesk-/sk-</i>	63
<i>-esk-/-n-</i>	51
<i>-Ov-/-Ovsk-</i>	41

Table 3: Distribution of doublets

As shown in Table 3, the three main suffixes enter in competition not only with other suffixes but with each other as well - these are the most frequent cases of rivalry. The exception is the rivalry between *-n-* and *-Ov-* which seem to privilege distinct nominal bases. Similarly, *-Ovsk-* competes with both *-Ov-* and *-sk-* but not with *-n-*: the doublets with *-Ovsk-/-n-* are not frequent.

### 3.3 Annotation

Since the competition between affixes is driven by a complex combination of factors, base nouns were annotated according to some of their properties.

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, etc (`\zim'a \winter'`, `\viʃnʲa \cherry'`, `\raduga \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'`, `\mark'sizm \marxism'`, `\galav'a \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.

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

- **VowAlt**: vowel /  $\emptyset$  alternation, binary property (DVOREC 'palace' - DVORCOVYJ);
- **ConsM**: consonant mutation, binary property (TVOROG 'cottage cheese' - TVOROŽNYJ).

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

Morphological properties include only one predictor :

- **InfCl**: the inflectional class of base nouns. We follow a canonical distinction between 3 inflectional classes (PAPA<sub>I,M</sub> 'dad', PESNJA<sub>I,F</sub> 'song'; STOL<sub>II,M</sub> 'table', DELO<sub>II,N</sub> 'business'; TEN<sub>III,F</sub> 'shadow').

Semantic properties include the following features:

- Binary distinct properties of [ $\pm$ proper], [ $\pm$ human], [ $\pm$ animate], [ $\pm$ concrete], [ $\pm$ countable];
- **Anim**: animacy, or the combination of the properties listed above into five groups ([Thuilier, 2012](#)):
  - **AnimA**: proper human (PIFAGOR 'Pithagoras');
  - **AnimB**: common human/animate (SOBAKA 'dog');
  - **AnimC**: common concrete (DOM 'house');
  - **AnimD**: proper non-human (AL'PY 'Alps');
  - **AnimE**: common abstract (SOJUZ 'alliance').

The choice of these properties was motivated by their presence in the literature as potential factors to distinguish between two rival affixes. We hypothesize that the same properties could be less restrictive for some couples of rival affixes and allow the combination of the base noun with both of them.

## 4 Results

### 4.1 Exploration

Before diving into the statistical analysis we investigate if the data contain strongly correlated features among the predictors. A multicollinearity problem arises when there are two or more features heavily correlated to each other. Multicollinearity does not really affect the quality of the logistic regression but can have an impact on the reliability of effects of individual predictors in the model. If some of the predictors overlap in their measures, their effects become indistinguishable.

To detect if the predictors in the data set are affected by multicollinearity we create dummy variables for non binary categorical data and use a Pearson correlation test. The results are shown in [Table 4](#).

[Table 4](#) keeps the features whose correlation to other features is  $\geq 0.3$ . As it could have been anticipated, all the classes of animacy are highly correlated to their constituents ([ $\pm$ proper], [ $\pm$ human], [ $\pm$ animate], [ $\pm$ concrete], [ $\pm$ countable]). The multicollinearity analysis revealed other correlations. It is possible, to a certain extent, to derive some stress position values from the constituents of animacy. Similarly, the shift in semantic properties of the base noun can be associated with changes in values of inflectional class.

To address the multicollinearity issue we proceed with a straightforward method of dropping highly correlated features: we only used animacy for further investigations, and remove its constituents. Phonological and morphological stress positions and inflectional class are kept as well. The final set contains thus quite independent features.

	Propre	Concr	Compt	Anim	Human
AnimA	<b>0.66</b>	0.13	-0.24	0.24	0.25
AnimB	-0.17	<b>0.51</b>	<b>0.46</b>	<b>0.92</b>	<b>0.87</b>
AnimC	-0.16	<b>0.46</b>	0.05	<b>-0.44</b>	<b>-0.41</b>
AnimD	<b>0.73</b>	0.13	-0.26	-0.11	-0.13
AnimE	-0.19	<b>-0.99</b>	<b>-0.31</b>	<b>-0.54</b>	<b>-0.51</b>
AccSylAad	-0.05	<b>-0.38</b>	<b>-0.30</b>	-0.26	-0.25
AccPosR	0.12	-0.13	-0.17	<b>-0.30</b>	<b>-0.33</b>
AccPosS	-0.10	0.21	0.19	<b>0.41</b>	<b>0.44</b>
InflCl1	-0.07	<b>-0.36</b>	-0.28	<b>-0.35</b>	<b>-0.33</b>
InflCl2	0.05	<b>0.37</b>	<b>0.31</b>	<b>0.37</b>	<b>0.35</b>

Table 4: Correlation coefficient for predictors, where  $\rho \geq 0.3$

Next, we examine how strong the correlation between each predictor (independently) and the suffix is. We use Cramer’s V test, the measure of correlation between two categorical variables based on Pearson’s Chi squared statistics. Feature importance for every suffix choice is displayed in Table 5.

<i>-Ov-/-n-</i>	<i>-n-/-sk-</i>	<i>-Ovsk-/-sk-</i>	<i>-ičesk-/-ičn-</i>	<i>-ičesk-/-sk-</i>	<i>-esk-/-n-</i>	<i>-Ov-/-Ovsk-</i>
Anim (0.59)	Anim (0.24)	Anim (0.51)	SyllB (0.36)	AccPos (0.41)	AccSyl (0.49)	SyllB (0.19)
SyllB (0.35)	ConsM (0.24)	LastP (0.16)	LastP (0.28)	Anim (0.38)	LastP (0.49)	InflCl (0.13)
AccPos (0.26)	AccSyl (0.19)	InflCl (0.15)	Anim (0.27)	LastP (0.25)	InflCl (0.35)	LastP (0.10)
VowAlt (0.23)	LastP (0.16)	AccPos (0.14)	AccSyl (0.25)	AccSyl (0.14)	SyllB (0.24)	Anim (0.09)
InflCl (0.17)	AccPos (0.15)	SyllB (0.13)	InflCl (0.23)	InflCl (0.13)	Anim (0.22)	ConsM (0.08)
AccSyl (0.16)	SyllB (0.14)	ConsM (0.07)	ConsM (0.21)	SyllB (0.12)	AccPos (0.11)	AccSyl (0.08)
LastP (0.006)	InflCl (0.12)	AccSyl (0.05)	AccPos (0.20)	ConsM (0.10)	ConsM (0.11)	AccPos (0.06)
ConsM (0.04)	VowAlt (0.08)	VowAlt (0.04)	VowAlt (0.06)	VowAlt (0.05)	VowAlt (0.00)	VowAlt (0.00)

Table 5: Cramer’s V for each suffix, from  $\rho$  max to  $\rho$  min

Animacy appears to be one of the features that are most correlated to the affix choice. It seems to be strongly correlated to the choice of both *-Ov-* and *-n-*, as well as of *-Ovsk-* and *-sk-*; to a lesser extent - of *-n-* and *-sk-*. The last phoneme of the stem appears to be related to the emergence of *-esk-* and *-n-*, *-ičesk-* and *-ičn-*, *-Ovsk-* and *-sk-*. Vowel-zero alternations, as well as consonant mutations, seem to be the least correlated to the choice of suffix. The strongest correlations between the properties of base noun and suffixes are observed for *-Ov-/-n-*, *-Ovsk-/-sk-* and *-esk-/-n-* rivalries. For both *-n-/-sk-* and *-Ov-/-Ovsk-* the correlations seem to be weak. Lastly, a significant gap in correlation coefficient values is observed for the *-Ovsk-/-sk-* rivalry: after  $\rho = 0.51$  for animacy it drops to 0.16 for the last phoneme of the stem.

While Cramer’s V can provide some insights about the data, it only indicates to which extent each predictor correlates independently to the suffix. Cramer’s V does not allow the visualisation of correlation coefficients when all the predictors act simultaneously. Moreover, the predictors in our data set are categorical (all but the length of base noun in syllables), and some of the predictors are non binary. For instance, animacy appears to be the most highly correlated feature to the suffix choice. However, at this stage it is not clear whether all the five constituents of animacy are equally relevant.

To address these issues and to go deeper into the investigation we proceed with a logistic regression. The choice of logistic regression is driven by an easy interpretability and visualization of its results. It also provides a fine grained analysis of predictors (entering into the constituents of categorical variables), its coefficients allow to establish a ranking of the most important predictors.

In what follows we use two types of logistic regression: first we assess to which extent the individual suffixes may be predicted, given that we face a multilabel classification problem. We use one-to-all

logistic regression for this purpose. Next, we transform multilabel classification problem into a binary classification and analyse to which extent a pair of competing suffixes may be predicted, as opposed to all other suffixes. Binomial logistic regression is used for this task.

## 4.2 Multilabel classification

To perform a statistical modelling of suffixal rivalry we use a multi-label logistic regression, in particular one-to-rest approach. This heuristic method allows the decomposition of one multilabel classification problem into multiple binary classification tasks. A set of binary classifiers is thus leveraged for a multiclass classification. Our data set suggests building 7 binary classifiers, since we take 7 suffixes for this study. Instead of being mutually inclusive the labels become mutually exclusive, since each classifier solves such problems as "- *n-* vs. all the rest", "-*sk-* vs. all the rest", etc.

To evaluate the performance of these models we used the exact match ratio. This strict metric indicates the percentage of samples that have all their labels classified correctly. The data set was randomly split into train and test sets, the results of our models for test set are shown in Table 6.

	<i>-esk-</i>	<i>-ičn-</i>	<i>-Ovsk-</i>	<i>-Ov-</i>	<i>-sk-</i>	<i>-ičn</i>	<i>-n-</i>
Exact match ratio	90	86	82	81	75	70	53

Table 6: Results of logistic regression model (Multilabel approach)

These results show that the suffix *-esk-* is highly predictable, which allows us to think that there might be distinct properties of base nouns that allow the combination with this suffix. As far as other suffixes are concerned, the accuracies of the models are also quite high, except for *-n-*, for which the prediction is just slightly better than pure chance. This may mean that suffix *-n-* can potentially combine with all types of base nouns regardless of their properties, and the restrictions here can be less specific. Another explanation is that the data set contains exactly two labels for the outcome of the classification, it therefore may be confusing for the classifier to correctly predict *-n-* suffix. To assess fully the classification of *-n-* and to verify the numbers given by one-to-rest approach, a data set with only one dependent variable is necessary. This study however lies beyond the scope of this paper.

## 4.3 Binary classification

The logistic regression model allows us to access the parameters of the model and to visualize their weights. We are particularly interested in properties of base nouns allowing the derivation of two adjectives with distinct suffixes. Binary classification of the pair of competing suffixes as opposed to all other suffixes provides some insights, in particular due to the *p*-value, a measure of statistical significance of independent variables (typically  $\leq 0.05$ ).

As far as the doublets with *-Ov-/n-* suffixes are concerned, the model shows a number of properties that can allow both of these rival suffixes:

- nouns designating common concrete entities,  $p < 0.001$  (OGUREC 'cucumber' - OGURCOVYJ / OGUREČNYJ);
- nouns with stressed radical,  $p < 0.027$ , not a derivational or inflectional affix (/smet'ana/ 'cream' - SMETANNYJ / SMETANOVYJ);
- nouns affected by vowel alternation seem to prefer these two suffixes to the others as well,  $p < 0.049$  (PEPEL 'ashes' - PEPLOVYJ / PEPEL'NYJ).

Both *-ičesk-* and *-ičn-* combine with base nouns possessing certain morphological and phonological properties:

- inflectional class II,  $p < 0.034$  (SINONIM 'synonym' - SINONIMIČNYJ / SINONIMIČESKIJ);

- stress on derivational affix of the base noun,  $p < 0.050$ , namely *-izm* or *-ist*, as in CINIZM 'cynicism' - CINIČESKIJ / CINIČNYJ.

The acceptance of both *-esk-* and *-n-* suffixes seem to rely on the semantic and phonological properties of base nouns:

- these nouns are mostly common abstract,  $p < 0.003$  (MISTIKA 'mysticism' - MISTIČESKIJ / MISTIČNYJ);
- their stems end with velars,  $p < 0.010$ , which, in turn, provokes a consonant mutation in derived adjectives, though this property is not considered to be a statistically significant factor ( $p < 0.483$ ); cf. ISTERIKA 'hysterics' - ISTERIČESKIJ / ISTERIČNYJ.

Morphological and phonological properties determine the choice of both *-n-* and *-sk-*:

- inflectional class II,  $p < 0.028$  (INVALID 'disabled person' - INVALIDNYJ / INVALIDSKIJ);
- the length of the stem in syllables,  $p < 0.012$  (ZRITEL' 'viewer' - ZRITEL'NYJ / ZRITEL'SKIJ)

The same properties are also significant for the choice of both *-ičesk-/sk-*:

- inflectional class II,  $p < 0.050$  (VAMPIR 'vampire' - VAMPIRIČESKIJ / VAMPIRSKIJ);
- the length of the stem in syllables,  $p < 0.043$  (monosyllabic nouns and nouns with two syllables).

The combination of a base noun with both *-Ov-* and *-Ovsk-* seems to be driven by phonological and semantic properties:

- the length of the stem in syllables,  $p < 0.000$ . Both of these affixes privilege short stems (of 1-2 syllables);
- two classes of animacy - common concrete nouns,  $p < 0.019$  (BOJEC 'fighter' - BOJCOVYJ / BOJCOVSKIJ) and common abstract nouns,  $p < 0.001$  (ONLAJN 'online' - ONLAJNOVYJ / ONLAJNOVSKIJ).

Lastly, morphological and semantic properties seem to allow both *-Ovsk-* and *-sk-*:

- inflectional class II,  $p < 0.018$  (BANKIR 'banker' - BANKIROVSKIJ / BANKIRSKIJ);
- two subsets of animacy: common human or common animate,  $p < 0.001$  (SULTAN 'sultan' - SULTANOVSKIJ / SULTANSKIJ), common abstract,  $p < 0.000$  (INTERNET 'internet' - INTERNETOVSKIJ / INTERNETSKIJ).

To evaluate the predictive power of all the binary classifiers we use accuracy metric, the results are shown in Table 7.

	<i>-esk-/n-</i>	<i>-ičesk-/sk-</i>	<i>-Ovsk-/sk-</i>	<i>-Ov-/Ovsk-</i>	<i>-ičesk-/ičn-</i>	<i>-n-/sk-</i>	<i>-Ov-/n-</i>
Accuracy	95.79	94.52	93.84	92.47	91.78	87.67	81.50

Table 7: Results of logistic regression model (Binary classification)

The results of binary classification are globally superior to the results of one-to-rest classification, given in Table 6. Despite the fact that *-esk-* and *-n-* are individually predicted with the best and the worst results respectively, their combination has the best accuracy. As for other combinations with *-n-* (*-n-/sk-* and *-Ov-/n-*), they can be found in the end of the ranking, however, the accuracy of predictions remains high. The results of logistic regression are not always congruent with the observations based of Cramer's V analysis: each property of the base noun does not have the same weight for the suffix choice when taken separately from other parameters, or jointly - when all the predictors are taken into consideration.

## 5 Discussion

Our study based on statistical models gave us some insights in order to identify the properties of base nouns which may allow the choice of two rival affixes. The most recurrent are the semantic factor of animacy and the morphological factor of inflectional class. Animacy is relevant for *-Ov-/-n-*, *-esk-/-n-*, *-Ov-/-Ovsk-* and *-Ovsk-/-sk-* doublets. Inflectional class plays a role in the acceptance of both *-ičesk-/-ičn-*, *-n-/-sk-*, *-ičesk-/-sk-* and *-Ovsk-/-sk-* suffixes. Phonological factors are determinant to a lesser extent: the length of stem in syllables (namely for monosyllabic nouns) allows both *-n-/-sk-*, *-ičesk-/-sk-* and *-Ov-/-Ovsk-* suffixes.

The results, however, are based only on properties of base nouns, the discussion on doublets would be incomplete without a deeper investigation on the nature of these doublets. The data extracted from National Corpus of Russian Language allow us to include for further studies such properties of adjectives as their frequency and the type of subcorpus they appear in.

The frequency of doublets needs further investigation because of two factors. First, one of the doublets may have undergone phenomena of lexicalization and be formally or semantically opaque, whereas another one is more transparent, as in TRUDNYJ / TRUDOVOJ, both derived from TRUD 'labor', however the first adjective means 'difficult', and the second one - 'labor, or work related'. Moreover, different adjectivizing affixes can be used to derive adjectives which correspond to two distinct senses of the underlying noun; the semantic of the whole adjective in a couple formed with two rival suffixes needs to be assessed. Second, even if both doublets are semantically and formally transparent, one may be frequently and commonly used, whereas another one may be an hapax, reflecting the result of the creative use of morphological constructions by speakers (Dal and Namer, 2012), as in PRIZRAČNYJ / PRIZRAKOVYJ, both derived from PRIZRAK 'ghost' and both transparent, however the first one is attested with frequency 2724, the second one appears in the corpus only once.

The type of subcorpus the doublets appear in can shed a light on their linguistic specialization. For instance, there might be a difference between suffixes chosen in general and newspaper subcorpora and the poetic and oral subcorpora: ALMAZNYJ 'diamond' and NOVOSTNOJ 'news' are both attested in main subcorpus, whereas ALMAZOVYJ and NOVOSTEVOJ - in poetic and oral subcorpora respectively. Furthermore, the two adjectives attested in the general subcorpus are very frequent ones, their doublets in oral and poetic subcorpora are hapaxes. The correlation between the frequency and the type of subcorpora the adjectives appear in also needs further investigation.

## References

- Ol'ga Pavlovna Antipina. 2012. *Sopostavitel'nyj analiz paronimov russkogo i anglijskogo jazykov*. Ph.D. thesis, Bashkir State University.
- Mark Aronoff. 2016. 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*. Bulzoni, pages 39–52.
- Natalia Bobkova and Fabio Montermini. 2019. Suffix rivalry in russian: what low frequency words tell us. In *Mediterranean Morphology Meetings*. volume 12, pages 1–17.
- Olivier Bonami and Juliette Thuilier. 2018. A statistical approach to rivalry in lexeme formation: French *-iser* and *-ifier*. *Word Structure* 11(2).
- Georgette Dal and Fiammetta Namer. 2012. Faut-il brûler les dictionnaires? ou comment les ressources numériques ont révolutionné les recherches en morphologie. In *SHS Web of Conferences*. EDP Sciences, volume 1, pages 1261–1276.
- Nabil Hathout. 2011. 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.
- Christine Hénault and Sergueï Sakhno. 2015. Çem supermarket-n-yj luçse supermarket-sk-ogo? slovoobrazovatel'naja sinonimija v russkix ad"ektivnyj neologizmax po dannym interneta. *B. Tošovic, A. Wonisch. Wortbildung und Internet* .

- Vsevolod Kapatsinski. 2010. Velar palatalization in Russian and artificial grammar: Constraints on models of morphophonology. *Laboratory phonology* 1(2):361–393.
- Galina Ivanovna Kustova. 2018. Prilagatel'nye. *Materialy k korpusnoj grammatike russkogo jazyka. Vyp.3. Časti reči i leksiko-grammatičeskie klassy* pages 40–107.
- Stéphanie Lignon. 2010. –iser and –ifier suffixations in French: Verify data to verize hypotheses? In *Décembrettes 7*.
- Mark Lindsay and Mark Aronoff. 2013. Natural selection in self-organizing morphological systems. In *Morphology in Toulouse. Selected Proceedings of Décembrettes 7 (Toulouse 2-3 December 2010)*. Lincom Europa, pages 133–153.
- Marc Plénat. 2011. Enquête sur divers effets des contraintes dissimilatives en français. *M. Roché, G. Boyé, N. Hathout, S. Lignon & M. Plénat, Des unités morphologiques au lexique. Paris: Hermès-Lavoisier* pages 145–190.
- Michel Roché. 2011. Quel traitement unifié pour les dérivations en-isme et en-iste?
- Andrea D Sims. 2017. Slavic morphology: Recent approaches to classic problems, illustrated with Russian. *Journal of Slavic Linguistics* 25(2):489–524.
- Juliette Thuilier. 2012. *Contraintes préférentielles et ordre des mots en français*. Ph.D. thesis, Université Paris-Diderot-Paris VII.
- Alan Timberlake. 2004. *A reference grammar of Russian*. Cambridge University Press.
- Charles Edward Townsend. 1975. *Russian word-formation*. Slavica Publishers.
- Natal'ja Švedova. 1980. *Russkaja grammatika*, volume 1. Moskva: Nauka.
- Elena Andreevna Zemskaya. 2015. *Jazyk kak dejatel'nost'. Morfema, slovo, reč*. Moskva: Flinta.