A quantitative approach to doublets in Russian denominal adjective construction

Natalia Bobkova, Fabio Montermini

A quantitative approach to doublets in Russian denominal adjective construction

Natalia Bobkova – CLLE, CNRS and Université de Toulouse Jean Jaurès
Fabio Montermini – CLLE, CNRS and Université de Toulouse Jean Jaurès

Abstract

This article is devoted to the rivalry between suffixes in the derivation of denominal adjectives in Russian. In particular, it proposes a large-scale quantitative analysis based on the Russian National Corpus. Its main goal is to contribute to identify the properties that determine the choice of the suffix in these derivatives. Denominal adjectival derivation in Russian makes use of a wide variety of exponents. Most of them are phonological variants (extensions) of three main suffixes, *-n-*, *-sk-* and *-Ov-*. The latter, that can be considered as basic, constitute the focus of our analysis. Two datasets were built for this research, a general one containing one of the suffixes above, and a more specific one containing doublets, i.e. adjectives constructed on the same base with different suffixes. Data from the two sets are analyzed by means of various statistical models. Our results globally provide a quantitatively robust confirmation of observations previously made in the literature. In particular, we show that -n- occupies a specific position in the derivational system of Russian, as it is less productive, and its derivatives are tendentially less transparent and more prone to display lexicalized meanings that point towards the qualitative pole of the qualitative-relational semantic spectrum. Moreover, *-sk-* and *-Ov-* are more likely to form doublets (be attached to the same bases), a further argument in favor of a greater homogeneity between them as opposed to *-n-*.

Keywords: Russian, derivational morphology, denominal adjectives, quantitative linguistics, corpus linguistics, statistical approaches to word-formation

1. Introduction

The derivation of adjectives from nouns in Russian constitutes an interesting ground for the observation and the analysis of affix rivalry. Denominal adjectives (which we may globally characterize as having a relational value) may in fact be derived in this language mainly by means of three different suffixes, *-n-*, *-sk-* and *-Ov-*, or by a number of variants, basically extended variants of the latter. Several attempts have been made in the literature to isolate the factors that determine the choice of one or the other suffix / variant, categorically or tendentially (cf. Townsend 1975; Švedova 1980; Zemskaja 2015; Hénault & Sakhno 2016, among others). The factors identified include phonological, morphological, semantic and etymological properties of base nouns and of derivatives, or properties connected with the relationship between the two. However, apart from some studies on small sets of lexemes, observations of this phenomenon in a quantitative perspective are still lacking. The research we present in this paper is a first step in the direction of a large-scale analysis of Russian denominal adjectives. Our main goal is to build robust statistical models in order to predict the choice of a suffix over the other in the derivation in question. In particular, we built two distinct datasets from the Russian National Corpus: a general dataset of denominal adjectives containing one of the main suffixes listed above, and a dataset of all doublets encountered in the corpus, i.e. adjectives constructed with different suffixes on the same base. As we will show, in fact, the study of doublets may shed light on the global dynamics of the system, in particular when such

properties as productivity or the transparency of the derivative with respect to the base (parsability) are taken into account.

Our article is organized as follows: in Section 2 we present the general characteristics of denominal adjectival derivation in contemporary Russian, along with a brief discussion of the observations previously made in the literature. Section 3 presents the datasets on which our analysis is based. Section 4 proposes a quantitative exploration of the two datasets, in particular concerning the three parameters of frequency, productivity and morphological complexity. In Section 5 we present the results of a statistical modelization specifically realized on the three series of doublets (one for each pair of affixes) contained in our dataset. Finally, Section 6 contains some concluding remarks and perspectives for future work.

2. Denominal adjectives in Russian

As in several other European languages, there are various exponents (suffixes) that are used in Russian for the derivation of adjectives from nouns. It is currently acknowledged that three suffixes are mainly productive in synchrony, *-n-*, *-sk-* and *-Ov-*[1] (Zemskaja 2015; Hénault & Sakhno 2016; Kustova 2018).[2] An example for each suffix is provided in (1).[3]

(1) *kanal* → *kanal'**n**(yj)* 'canal'
universitet → *universitet**sk**(ij)* 'university'
begemot → *begemot**ov**(yj)* 'hippopotamus'

In addition, *-n-* and *-sk-* also possess several variants in which extra phonological material is added to the basic form. The variants in question may correspond to a combination of two of the suffixes above (2a), to a sequence of Slavic origin only found in the derivation in question (2b), to a sequence used to adapt a foreign suffix[4] (2c) with these categories sometimes overlapping (2d).

(2)	a.	*bank*	→	*bank**ovsk**(ij)*	'bank'	*-Ov-* + *-sk-*
	b.	*Budd(a)*	→	*budd**ijsk**(ij)*	'Buddha'	*-ij* + *-sk-*
	c.	*cikl*	→	*cikl**ičesk**(ij)*	'cycle'	*-ič-* + *-(e)sk-*
		respublik(a)	→	*respublik**ansk**(ij)*	'republic'	*-an-* + *-sk-*
		Satan(a)	→	*satan**insk**(ij)*	'Satan'	*-in-* + *-sk-*
	d.	*sestr(a)*	→	*sestr**insk**(ij)*	'sister'	*-in-* + *-sk-*

[1] The notation *-Ov-* is intended to cover the various realizations of this suffix, that may correspond phonologically to different outputs, and orthographically to <ov> or <ev>, depending on the value for the [±palatalized] feature of the preceding segment.

[2] A confirmation of this state of affairs comes from our observation of corpora. For instance, about 96% of the hapaxes of suffixed adjectives contained in RusCorpora (cf. Section 2.1 for details) contain one of these three suffixes or their variants.

[3] The following conventions apply to Russian examples: denominal adjectival affixes are indicated in bold; inflectional affixes (of the citation form, i.e. nominative singular for nouns and nominative masculine singular for adjectives) are given in brackets, both for bases and derivatives. Note that the nominative masculine singular suffix of adjectives may correspond to three orthographic and phonological forms (<yj>=/ɨj/, <ij>=/ij/, <oj>=/oj/), depending on segmental and prosodic properties that we do not detail here. Concerning phonological forms, we use a broad IPA transcription in the lines of Yanushevskaya & Bunčić (2015). By default, translations are only provided for base nouns, the meaning of the adjective systematically corresponding to 'related to X', when not otherwise specified.

[4] The sequences *-ič-*, *-an-* and *-in-* in particular are used in Russian to adapt foreign adjectives containing Latin suffixes, like English *-ic*, *-an* or *-ine*.

Overall, grammars and handbooks of Russian (e.g. Townsend 1975 or Švedova 1980) list up to 25 different variants for the derivation in question. Note that extended variants may sometimes partially overlap with other derivational suffixes (3a-b), although the presence of the latter in the base is not a necessary condition (3c), as in the following series involving the suffix *-estv-* used to forms property nouns.

(3)	Noun	*-estv-* noun	Adjective	
a.	*Ø*	*xudožestv(o)*	*xudožestv**enn**(yj)*	'art'
b.	*bog*	*božestv(o)*	*božestv**enn**(yj)*	'god'/'deity'
c.	*um*	*Ø*	*um**stvenn**(yj)*	'mind'

As in other languages, semantically Russian denominal adjectives have mainly a relational meaning, which we globally translate as 'related to X'. As it is common for derivatives, however, these adjectives may undergo lexicalization phenomena reducing their transparency, and eventually acquire the status of qualitative adjectives, as in *um**n**(yj)* ('smart', vs. *um**stvenn**(yj)* in (3c), simply meaning 'mental').

Formally, as the examples above show, the default case consists in adjoining the suffix to the stem of the base, although some lexemes display allomorphies that imply consonant mutation or vowel / Ø alternation. Specifially, some derivatives display a systematic alternation between a velar or dental consonant (/ts/, /k/, /t/, /x/, /g/) and a palatal (/tʃ/, /ʃ/, /ʒ/) and/or an alternation between a fully articulated vowel and Ø. Denominal adjectives may display either the first phenomenon (4a), the second (4b), or both (4c) (the transcription of stems is provided in brackets).[5]

(4)	a.	*knig(a)* (/knʲig/)	→	*knіž**n**(yj)* (/knʲiʒn/)	'book'
		mal'čik (/malʲtʃik/)	→	*mal'čіš**esk**(ij)* (/malʲtʃiʃisk/)	'boy'
	b.	*ugol* (/ygal/)	→	*ugl**ov**(oj)* (/uglav/)	'angle'
	c.	*komponovk(a)* (/kampanovk/)	→	*komponovoč**n**(yj)* (/kampanovatʃn/)	'assembly'

All the above shows that suffixal variation in the construction of denominal adjectives in Russian is a complex phenomenon that interacts with phonological, morphological and possibly semantic factors. It should be noted, moreover, that this rivalry may manifest itself at each level, i.e. both in the choice of one of the 'main' suffixes (5a) or of one of their variants (5b). The examples in (5) correspond to triplets we extracted from RusCorpora.

(5)	a.	*ieroglif* →	*ieroglif**n**(yj)* / *ieroglif**ičesk**(ij)* / *ieroglif**ov**(yj)*	'hieroglyph'
	b.	*okean* →	*okean**sk**(ij)* / *okean**ovsk**(ij)* / *okean**ičesk**(ij)*	'ocean'

For the purposes of this study, however, we will focus on cases where the adjectives are formed exclusively with the main suffixes (*-n-*, *-sk-*, *-Ov-*) and exclude extended variants from our analysis.

Establishing the criteria according to which speakers choose one affix / variant over another, either in a categorical way or tendentially, appears thus a major issue in the study of this particular case of affixal rivalry. Authors having dealt with denominal adjectivization in the past have identified some properties of these derivatives and/or of their bases that seem to drive the choice of the exponent. Below, we provide a non-exhaustive list of base noun

[5] We do not discuss in detail consonant mutation and vowel/Ø alternation, and we limit ourselves to consider that they both correspond to lexical rather than to phonological phenomena in synchrony (cf. Timberlake 2004; Kapatsinski 2010; Sims 2017 for detailed discussions).

properties that are listed in the literature (cf. in particular Townsend 1975, Švedova 1980, Zemskaja 2015, Hénault & Sakhno 2016).

-n-:
Semantically, this suffix mainly combines with non-animate common nouns, either abstract (6a) or concrete (6b), although animate nouns are also possible bases (6c).

(6)	a. *gnev*	→	*gnev**n**(yj)*	'anger'
	b. *kiparis*	→	*kiparis**n**(yj)*	'cypress'
	c. *inžener*	→	*inžener**n**(yj)*	'engineer'

Phonologically, it is stress-neutral, as it combines both with bases with stress on the stem (7a) or on inflection (7b), and it can select stems displaying consonant mutation (8).

(7)	a. *kómnat(a)*	→	*kómnat**n**(yj)*	'room'
	b. *zim(á)*	→	*zím**n**(yj)*	'winter'

(8)	a. *jazyk*	→	*jazyč**n**(yj)*	'tongue / language'
	b. *drug*	→	*druž**n**(yj)*	'friend'

Etymologically, it combines both with native (9a) and foreign (9b) bases.

(9)	a. *dym*	→	*dym**n**(yj)*	'smoke'
	b. *arxitektur(a)*	→	*arxitektur**n**(yj)*	'architecture'

-sk-:
This suffix does not seem to be selective semantically, since it may combine with inanimate (10a) and animate (10b) nouns, including nouns denoting humans (10c), and may also combine with proper nouns (10d).

(10)	a. *universitet*	→	*universitet**sk**(ij)*	'university'
	b. *kon'*	→	*kon**sk**(ij)*	'horse'
	c. *bandit*	→	*bandit**sk**(ij)*	'bandit'
	d. *Iran*	→	*iran**sk**(ij)*	'Iran'

Phonologically, it privileges stems ending in alveolar (11a) or dental (11b) consonants, and, like *-n-*, it also selects nouns with stress on the stem, and mutated stems (12).

(11)	a. *sosed*	→	*sosed**sk**(ij)*	'neighbor'
	b. *šef*	→	*šef**sk**(ij)*	'boss'

(12)	a. *Volg(a)*	→	*volž**sk**(ij)*	'Volga (river)'
	b. *Čex(ija)*	→	*češ**sk**(ij)*	'Czechia'

-Ov-:
Semantically, this suffix combines mainly with concrete nouns, both inanimate (13a) and non-human animate (13b).

(13)	a. *dom*	→	*dom**ov**(yj)*	'house'
	b. *tigr*	→	*tigr**ov**(yj)*	'tiger'

Moreover, lexemes containing the (mostly evaluative) suffixes *-ik*, *-ok*, *-nik*, *-čik* (possibly displaying vowel/*Ø* alternation) also tend to be suffixed by *-Ov-*:

(14) a. *šarik* → *šarik**ov**(yj)* 'ball'
b. *gribok* → *grib**k**ov**(yj)* 'yeast'

In addition to the properties listed above in the examples (6)-(16), it has been observed that adjectives in *-n-* manifest a stronger tendency to acquire lexicalized meanings (as in *um**n**(yj)*, mentioned above or in *zvuč**n**(yj)* 'resounding' vs. *zvuk**ov**(oj)* 'related to sounds', both from *zvuk* 'sound').[6]

To sum up, the properties identified in the literature as relevant for the choice of a particular suffix in denominal adjective derivation include at least the following:

- the semantic reading of the base noun, in particular concerning the feature [±animate];
- the morphological constitution of the base noun, some derivational suffixes favoring the choice of a particular adjectival suffix;
- the origin of the base noun (Slavic vs. foreign);
- the tendency for a derived adjective to have a clearer relational meaning (vs. a qualitative one).

Although some of these properties have been tested via the observation of quantitative data (cf. the Internet-based analysis by Hénault & Sakhno 2016 of adjectives derived from the noun *supermarket* 'supermarket'), a large-scale quantitative analysis of corpus data is still lacking. In what follows, we propose some observations based on extensive corpus data using statistical tools.

3. The data

3.1. Database constitution

Our study is based on a dataset extracted from the Russian National Corpus (RusCorpora), a general corpus of contemporary Russian.[7] The dataset was automatically extracted from the corpus by systematically searching lemmas containing a sequence corresponding to one of the above-mentioned suffixes immediately preceding inflectional suffixes typical of citation forms of adjectives (see footnote 3). For this purpose we used a set of regular expressions summed up in the following formula: *{ev,n,ov,sk}{ij,oj,yj}.[8] A manual cleaning of the data was performed, which led to discard >70% false positives, e.g. forms corresponding to masculine or neuter genitive plurals in *-Ov* (e.g. *dvorov* 'yard$_{GEN.PL}$'), to possessive adjectives in *-Ov* (e.g. *dedov* 'grandpa's' from *ded* 'grandpa') or to proper nouns (surnames) ending in *-Ov* or *-sk(ij)*.

[6] Some authors (e.g. Hénault & Sakhno 2016) make a difference between what they call "lexical" derivation (like *-n-*) and "syntactic" derivation, and consider that *-Ov-* belongs to the latter category, since this derivational suffix is homophonous with the inflectional suffix for the genitive plural of masculine and neuter nouns, and with the suffix forming possessive adjectives (cf. Section 3.1). This issue is however marginal with respect to the focus of our article.

[7] The corpus is available at the URL https://ruscorpora.ru/new/. Our dataset was extracted in spring 2017 from an old version of RusCorpora containing ~600M tokens. A new version of the corpus, containing >900M tokens has been made available since.

[8] One of the conceivable combinations, **skyj*, corresponds to a phonologically forbidden sequence.

This first list was further filtered in order to keep only adjectives clearly derived from nouns. For example, many adjectives derived with *-n-* may be deverbal formations (*sdelat'* 'make' - *sdel**ann**(yj)*; *osmotret'* 'observe' - *osmotr**enn**(yj)*); they were automatically filtered based on their endings *-annyj*/*-ennyj*. For other adjectives we analyzed their morphological family. If there is a noun which may formally or semantically serve as a base for a given adjective, the entry with a base noun is kept. For instance, *pokupatel'**n**(yj)* is formally closer to agent noun *pokupatel'* 'buyer'; however, semantically it may refer to the verb *pokupat'* 'buy'. In this case, only the noun was considered the base of the adjective. In other cases, several nouns may be candidates to form an adjective: *zreni(e)* 'vision' / *zritel'* 'viewer' - *zritel'**n**(yj)* (cf. *zritel'nyj zal* 'auditorium' and *zritel'nyj nerv* 'optic nerve'). Consequently, two entries were kept for this adjective, which resulted in two distinct annotations on semantic, morphological and phonological levels. Finally, polysemy was treated in the same way: *kamer**n**(yj)* may be related to *kamer(a)*$_1$ 'cell' or to *kamer(a)*$_2$ 'chamber'. In this case, only the semantic annotation for the two entries is different.

For this study we only took into account adjectives containing one of the three main suffixes (*-n-*, *-sk-*, *-Ov-*), and, for further annotation and modelling, only very highly frequent adjectives (with a token frequency >100) were kept.[9] Our final dataset contains thus 1,768 adjective types, whose distribution is given in Table 1, whereas Table 2 presents a sample of the entries of our dataset with the indication of their frequency.

Table 1: Type frequencies per suffix in the main dataset.

Suffix	Type frequency
-n-	638
-sk	450
-Ov-	680

Table 2: Sample of entries in the main dataset.

Base	Adjective	Token frequency	Gloss
sneg	snež**n**(yj)	14,944	'snow'
Kiev	kiev**sk**(ij)	17,982	'Kiev'
trud	trud**ov**(oj)	32,482	'labour'

In addition to the main dataset, we created a subset containing all doublets / triplets (adjectives that are formed from the same base with different suffixes) contained in the corpus, regardless of their token frequency. This second dataset was constructed automatically by using the algorithm in (15):

(15) Given two adjectives A and A' whose structure is X+der+inf, they are doublets if $X_{(A)}=X_{(A')}$
(ex.: *mir**n**(yj)* (*–n* (*–yj*)) = *mir**ov**(oj)* (*–ov* (*–oj*))).

[9] We arbitrarily chose frequency 100 as a cutoff for highly frequent adjectives in order to have the best balance between the two datasets (highly frequent lexemes vs. hapaxes).

The algorithm states that if two adjectives (A and A') are orthographically identical once their derivational (der) and inflectional suffix (inf) are stripped, they are doublets. This algorithm based on the orthographic form of adjectives does not allow taking into account stem allomorphies such as vowel/Ø alternation and consonant mutation (see Section 2). Fleeting vowels were dealt for by automatically dropping all vowels; stems were thus treated as consonant strings, which generated some noise in the dataset, as shown in (16), where the potential ambiguities of the sequence <grz> are presented.

(16) *gruz* → *gruz**n**(yj)* / *gruz**ov**(oj)* 'weight'
groz(a) → *groz**n**(yj)*[10] / *groz**ov**(oj)* 'storm'

As far as consonant mutation is concerned (see Section 2), it was dealt for by introducing an extra condition to the algorithm presented above.

(17) Given two adjectives A and A' whose structure is X+der+inf, they are doublets if $X_{(A)}$=_<c, k, t, x, g># and $X_{(A')}$=_<č, š, ž>#

The combination of this algorithm with vowel dropping generated additional noise in the data. For instance, the string <vs{kč}> may correspond to the base of either of the following adjectives:

(18) *visok* → *visk**ov**(oj)* / *visoč**n**(yj)* 'temple'
vosk → *vosk**ov**(oj)* 'wax'

After manual verification, almost 60% of false positives were discarded (i.e. triplets, doublets formed with extended variants of *-n-* or *-sk-*, couples of adjectives formed on distinct noun bases). The final dataset contains 375 doublets showing different combinations of the three suffixes *-n-*, *-sk-* and *-Ov-*. Their distribution is given in Table 3 whereas Table 4 presents a sample of the dataset.

Table 3. Type frequencies per competing suffixes in the doublet dataset.

Suffixes	Type frequency of doublets
-n-/-Ov-	227
-n-/-sk-	108
-sk-/-Ov-	40

Table 4. Sample of entries in the doublet dataset.

Base	A1	Token frequency	A2	Token frequency	Gloss
diagonal	diagonal'**n**(yj)	205	diagonal**ev**(yj)	64	'diagonal'
izobretatel'	izobretatel'**n**(yj)	829	izobretatel'**sk**(ij)	281	'inventor'
ad	ad**ov**(yj)	102	ad**sk**(ij)	2,961	'hell'

[10] The adjective *groz**n**(yj)* has the lexicalized meaning 'terrible'.

The last step for the constitution of the datasets consisted in reconstructing base nouns for adjectives included in both of them. This step was performed automatically with subsequent manual verification. Token frequencies of all base nouns were also extracted automatically from RusCorpora.

3.2 Data annotation

The second step consisted in the manual annotation of our datasets. As noted in Section 2, several authors (mainly Švedova 1980; Sorokina 1984; Bottineau 2012; Zemskaja 2015) provide lists of the phonological, morphological and semantic properties of base nouns that can be considered relevant in the choice of the suffix. More recently, the relevance of these properties was tested on adjectival neologisms by Alekseeva (2011). In the present study we focus in particular on phonological, morphological, semantic and etymological properties of base nouns in relation with the three main suffixes *-n-*, *-sk-* or *-Ov-*. The detailed list of properties according to which adjectives in our dataset were annotated is the following:

- Phonology
    - phonological stress position (antepenultimate / penultimate / final stress);
    - morphological stress position (stress on the stem / on the derivational suffix / on the inflectional suffix);
    - last phoneme of the stem (dental, labial, velar, alveolar, vowel);
    - length of the base noun in syllables;

- Morphophonological
    - fleeting vowel;
    - consonant mutation;

- Morphological
    - inflectional class of the base noun (class 1, 2, 3);[11]

- Semantic
    - animacy of the base noun ([+proper; –human] / [+proper; +human] / [–proper; +human] / [–proper; +concrete] / [–proper; –concrete]).[12]

The models presented in Section 5 are based on the formal and semantic properties of base nouns, and do not take into consideration token frequency neither of base nouns, nor of the corresponding adjectives. However, frequency was considered in the exploratory data analysis presented in Section 4.1.

4. Data exploration

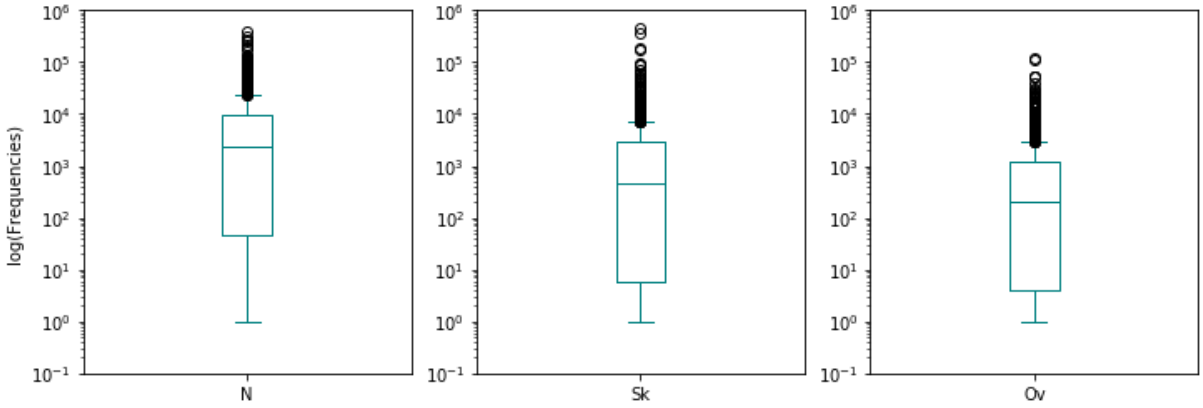
4.1 Absolute frequency

[11] In our annotation we used a simplified model based on three inflectional classes (Fraser & Corbett 1995:132-137), although Russian nouns may be divided into larger sets of classes and subclasses (Zaliznjak 2003; Parker & Sims 2019; Guzmán Naranjo 2020).

[12] We took inspiration from Thuilier (2012) for the identification of these categories. For details about the relevance of base noun features in the rivalry between *-n-* and *-sk-* cf. Bobkova & Montermini (2019), Bobkova (to appear).

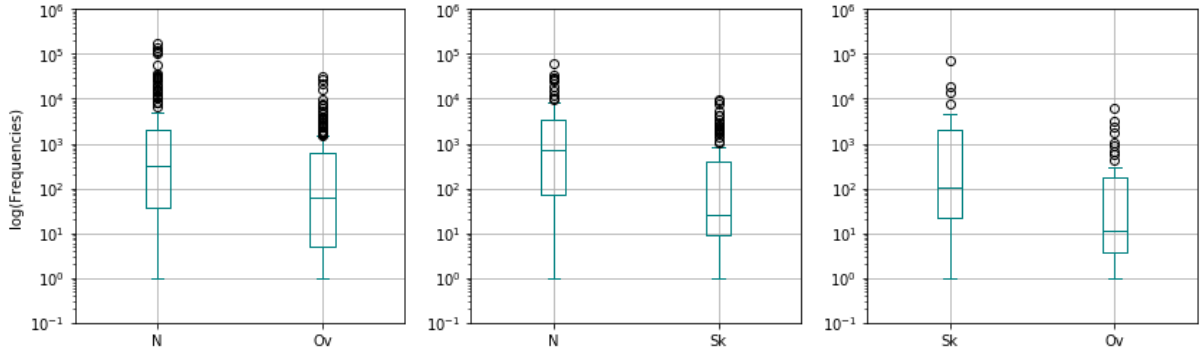
A first observation of our doublets dataset concerns the absolute frequency of the adjectives they contain. As far as the global dataset is concerned, adjectives in *-n-* are the most frequent in terms of tokens, whereas adjectives in *-Ov-* are the least frequent. The plots in Figure 1 show overall token frequency of adjectives for the three suffixes in question in the dataset[13]. Since all frequencies follow a Zipfian distribution, they are visualized on a log scale, with base 10. The mean frequencies are significantly different for all the suffixes (*-n-*/*-sk-*: t=5.11, p<0.05; *-n-*/*-Ov-*: t=9.84, p<0.05; *-sk-*/*-Ov-*: t=3.79, p<0.05)

Figure 1: Token frequency of adjectives in *-n-*, *-sk-* and *-Ov-* in the global dataset, logarithmic scale.



The same tendency is observed for doublets: when a couple includes adjectives in *-n-* they are on average significantly more frequent than the corresponding adjectives in *-Ov-* or *-sk-* (significance tests respectively: t=3.52, p<0.05 and t=3.62, p<0.05). In *-sk-*/*-Ov-* doublets, the adjectives in *-sk-* are on average more frequent, however, the difference is not significant (t=1.61, p=0.11) (see Figure 2).

Figure 2: Token frequency of adjectives in *-n-*, *-sk-* and *-Ov-* in the doublets dataset, logarithmic scale.



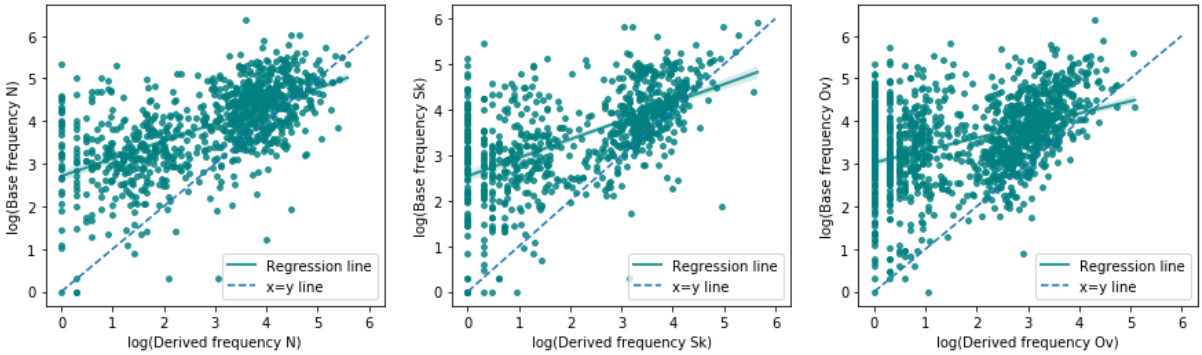
Frequency may be a good indicator of the semantic properties of a doublet's member. We may expect, for instance, to observe a large difference in frequency in a pair in which an adjective is lexicalized, while the other has been created 'on the spot' by speakers. Absolute frequency has traditionally been associated with semantic transparency (Baayen, 1993; Bybee, 1995): the higher the absolute frequency, the lower semantic transparency, and the less productive the suffix is. Hay (2001), however, provides evidence that not absolute but relative frequency is related to morphological decomposition.

[13] The position of the whiskers is set to 1.5*IQR (IQR=Q3-Q1) from the edges of the box. Outlier points are those past the end of the whiskers.

4.2 Relative frequency

By relative frequency we mean the difference between the frequency of a derivative and the frequency of its base. Figure 3 shows positive and significant correlations between base frequencies and derived frequencies for each suffix: the Pearson correlation coefficient for *-n-* is 0.21 (p<0.05), for *-sk-* it is 0.51 (p<0.05) and for *-Ov-* it is 0.25 (p<0.05). The x=y line indicates equal base and derived (adjective) frequency. Most of the data fall above this line, which provides evidence that adjective frequency is on average lower than the frequency of base nouns. Previous studies have pointed out that derived forms tend to be less frequent than their bases (Harwood & Wright, 1956; Hay, 2001), and the same tendencies are observed in Russian both for prefixed (Antić, 2012) and suffixed derivatives (Sims & Parker, 2015). Points below the x=y line correspond to data where the frequency of an adjective is higher than the frequency of its nominal base.

Figure 3. Base versus derivative frequencies in the global data set, logarithmic scale.



Relative frequency has a direct impact on morphological decomposition (Hay, 2003): it corresponds to the likelihood that a word is accessed as a whole or by referring to its morphological structure. In this case, the majority of words which fall above the x=y line may be considered as decomposable and therefore semantically transparent. Words falling below the x=y line are on the contrary accessed as independent lexical entries, are less semantically transparent and may display semantic shifts. The first line of Table 5 suggests that decomposable adjectives represent between 83% and 97% of our data. However, *-n-* tends to form adjectives with a lower degree of decomposability, whereas adjectives in *-Ov-* present the highest decomposability rate. The same trend holds both for the general and for the doublet datasets.

Table 5: Base versus derivative frequencies ratio.

	-n-	*-sk-*	*-Ov-*	*-n-/-Ov-*		*-n-/-sk-*		*-sk-/-Ov-*	
				-n-	*-Ov-*	*-n-*	*-sk-*	*-sk-*	*-Ov-*
base > derivative frequency	0.831	0.882	0.937	0.899	0.969	0.880	0.963	0.925	0.95
complexity (median)	0.125	0.105	0.043	0.053	0.016	0.103	0.012	0.032	0.006

Our intuition, however, is that the x=y line is not the best indicator in order to make a clear distinction between decomposable and non-decomposable adjectives, since most of them are less frequent than their base nouns. We need to measure their degree of decomposability. To do that, we rely on the notion of complexity, which we define as the ratio between the token frequency of a derived word and the token frequency of its base in a corpus. The higher the complexity index is, the more the derivative tends to be interpreted as a whole; lower complexity corresponds to the situation in which a word is accessed on the basis of its constituents. The second line of Table 5 represents the median complexity index for our dataset: tendencies are reversed with respect to base > derivative frequency. According to their complexity, we can order the three suffixes as follows: *-n-* > *-sk-* > *-Ov-*.

This measure gives more information on the nature of doublets when their frequencies are lower than those of their base nouns. A hapax like *ijun**ev**(yj)* (freq.=1), for instance, is much more likely to be analyzed according to its structure than *ijun'**sk**(ij)* (freq.=2,120), although both are derived from the much more frequent noun *ijun'* ('June', freq. 77,069). Complexity may be linked to the fact that a derived word displays a lexicalized meaning. Take for instance the couple of adjectives derived from *vkus* ('taste', freq. 34,598), *vkus**n**(yj)* (freq. 13,462) vs. *vkus**ov**(oj)* (freq. 484). The first, meaning 'tasty', is partially lexicalized in meaning, while the latter simply means 'related to taste', a fact that can be correlated with their respective complexity indexes (0.39 vs. 0.014), in spite of the fact that *vkus**n**(yj)* falls above the x=y line.

Adjectives in *-n-*, which, as shown above, are the most frequent on average are also those that tend to display lexicalization more frequently. Some more examples are given in Table 6 (glosses are only provided for A1, whereas A2 corresponds systematically to the transparent reading 'related to X').

Table 6: Relative frequencies and semantic reading of doublet adjectives

Base	Base Freq.	A1	A1 Freq.	A2	A2 Freq
bol'	44,925	bol'**n**(oj)	106,056	bol**ev**(oj)	1,774
'pain'		'sick'			
grjaz'	23,516	grjaz**n**(yj)	27,095	grjaz**ev**(oj)	460
'mud'/'dirt'		'dirty'			
nabljudatel'	16,401	nabljudatel'**n**(yj)	6,297	nabljudatel'**sk**(ij)	23
'observer'		'observant'			

4.3. Productivity

At the end of section 4.1 we pointed out that several studies suggested that the lower the semantic transparency of the derived words is, the less productive the rule that formed them. In this section we focus on productivity measures, namely on the capacity of the three suffixes in question to form hapaxes. The traditional productivity measure proposed by Baayen (1991) consists in dividing the number of hapaxes containing a particular suffix by the total number of tokens containing that same affix in a corpus. It has been observed, however, that this measure boosts productivity for suffixes with a low number of types. It has also been pointed

out that the measure does not take into account the token frequency of each derived word. To tackle this issue, Gaeta & Ricca (2006) adjusted the productivity measure proposed by Baayen by calculating it at equal token numbers for different affixes. One approach to get equal token numbers is to perform calculations for different affixes on different corpus sizes (Gaeta & Ricca, 2006). Another method, which we adopt for the present study, is to use binomial interpolation, as proposed by Baayen (2001). This method uses the whole frequency spectrum for derivatives with a given affix. It is therefore possible to calculate the productivity of an affix for a fixed value of tokens.[14] The implementation of this productivity measure was performed in R using ziprR library, in particular with the finite Zipf-Mandelbrot (fZM) LNRE model (Evert & Baroni, 2007). The first column of Table 7 presents the results of calculating the productivity $P(*10^3)$ as proposed by Baayen (1991). The other three columns represent the productivity calculated at an equal number of tokens $P(V_n)$, $P(V_{sk})$ and $P(V_{Ov})$ for *-n-*, number of tokens for *-sk-* and number of tokens for *-Ov-*.

Table 7: Productivity of *-n-*, *-sk-* and *-Ov-*.

	*P(*10³)*	$P(N_n)$	$P(N_{sk})$	$P(N_{Ov})$
-n-	0,00334	0,09367	0,08795	0,08138
-sk-	0,02589	0,11836	0,11327	0,10741
-ov-	0,08142	0,13382	0,12550	0,11150

Productivity measured by interpolation does not contradict the traditional productivity measure[15]. *-Ov-* appears to be more productive than *-sk-*, which, in turn, is more productive than *-n-*. On the basis of these measures, we can establish a scale of productivity for the three suffixes: *-Ov-* > *-sk-* > *-n-*.

These results confirm the idea that a lower degree of semantic transparency (or higher complexity, as sketched in section 4.2) is associated with lower productivity, as the order of productivity we obtain is reversed with respect to the ordering based on complexity. Our measure on doublet data is also in line with other observations, such as the quantitative and qualitative analysis of neological adjectives proposed by Alekseeva (2011) who cites *-n-*, *-sk-* and *-Ov-* among the most productive suffixes in Russian and establishes the same order of productivity *-Ov-* > *-sk-* > *-n-* by counting the number of neological formations attested in dictionaries between 2001 and 2009.

4.4 Proportions in doublet couples

The observation of frequencies for doublets may result in two different problems. Consider first the examples provided in Table 8.

Table 8: Sample of doublets with various token frequencies.

[14] See Baayen (2001) for a more detailed description of the mathematical intuition behind binomial interpolation.
[15] Even if in our study both measures of productivity are in line with each other, it is not always the case: measuring productivity for a fixed number of tokens may results in different rankings by their degree of productivity, as compared to traditional productivity measure (cf. the discussion on suffixes in Italian in Gaeta & Ricca (2006) and Varvara (2019)).

Base	A1	Freq1	A2	Freq2	Gloss
master	master**sk**(ij)	18,410	master**ov**(oj)	2,434	'master'
zritel'	zritel'**n**(yj)	6,224	zritel'**sk**(ij)	2,890	'observer'
pokupatel'	pokupatel'**n**(yj)	1,295	pokupatel'**sk**(ij)	1,032	'buyer'
gžel'	gžel'**sk**(ij)	86	gžel**ev**(yj)	4	'gzhel pottery'
forum	forum**n**(yj)	16	forum**sk**(ij)	6	'forum'
Evromajdan	evromajdan**n**(yj)	1	evromajdan**ov**(yj)	1	'Euromaidan'[16]

First, difference in token frequency may be extremely large within a couple, (cf. *master**sk**(ij)* vs. *master**ov**(oj)*). To better deal with these differences, we calculated the proportions of the two adjectives for every couple of doublets. We define the proportion of the two members of a pair constituting a doublet as the frequency of a given adjective (Freq1) divided by the total frequency of the two adjectives in a couple (Freq1+Freq2). The proportions of two adjectives in a couple sum up to 1. However, although the examples in Table 8 display similar proportions, the absolute frequencies of the adjectives vary significantly. The first three examples correspond to highly frequent adjectives, whereas the last three correspond to low-frequency adjectives or hapaxes. This leads to the second consideration, i.e. that proportions alone are insufficient to capture differences between doublets. In order to deal with this issue, we introduce the notion of volume, which we define as the absolute difference between the frequency of two adjectives in a doublet couple. Volumes are high for the first three examples in Table 8, and are lower for the last three. Details about proportions and volumes are given in Table 9.

Table 9: Proportions and volume for a sample of doublets

Couple	Proportion (A1-A2)	Volume
master**sk**(ij) ~ master**ov**(oj)	0.88-0.12	15,976
zritel'**n**(yj) ~ zritel'**sk**(ij)	0.68-0.32	3,334
pokupatel'**n**(yj) ~ pokupatel'**sk**(ij)	0.56-0.44	263
gžel'**sk**(ij) ~ gžel**ev**(yj)	0.96-0.04	82

[16] *Euromaidan* is the name given to anti-government demonstrations in Ukraine in 2013-2014.

forum**n**(yj) ~ forums**k**(ij)	0.73-0.27	10
evromajdan**n**(yj) ~ evromajdan**ov**(yj)	0.50-0.50	0

Proportions were further converted to discrete intervals of proportion ranges. By convenience, an equal number of bins corresponding to fractions of 20% were chosen, resulting in one categorical variable with 5 levels: 0-20, 20-40, 40-60, 60-80, 80-100. These intervals will be further referenced by means of quantiles for the sake of readability (1-5; 2-4; 3-3; 4-2; 5-1). Table 10 shows the number of doublets in every interval (the first interval corresponds systematically to the proportion of the first suffix, and vice-versa; note the absence of data for the 2-4 interval for the *-sk-*/*-Ov-* couples).

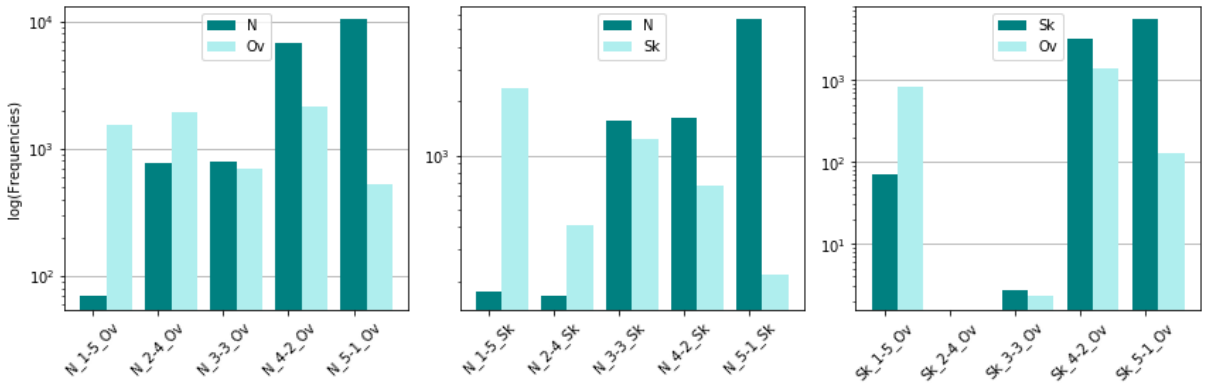
Table 10: Distribution of doublets by proportion intervals in the doublet dataset

	-n-/-Ov-	-n-/-sk-	-sk-/-Ov-
1-5	54	20	10
2-4	19	5	0
3-3	15	5	3
4-2	27	12	5
5-1	112	66	22

The tendencies observed in the three doublet groups are similar: the distribution of doublet couples is convex, with more cases on the extreme edges, where there is a big difference in frequency between the two suffixes. In particular, *-n-* is clearly dominating in both doublet couples where it appears, whereas *-sk-* prevails in *-sk-/-Ov-* doublets.

Figure 4 shows not only the proportions of doublets but also their frequencies, and implicitly their volumes. The frequency dominance of *-n-* as compared to *-Ov-* and *-sk-* is evident in this case too, as well as the dominance of *-sk-* compared to *-Ov-*. Moreover, a further tendency emerges: the higher the proportion of the dominant suffix, the higher its frequency in a couple, while the distribution of frequencies of the competing suffix is smoother. However, even for the non-dominant suffix the same tendencies are observed in extreme proportion intervals: frequencies are the lowest within the 1-5 interval and higher within the 5-1 interval. As for volume (frequency difference), the highest values are also found on the extreme edges, where the dominant suffix is within the 5-1 proportional interval. This means that there is a significant gap in token frequency between two doublets, one of them being highly frequent and the other rare or close to hapaxic. On the opposite edges, where the dominant suffix is within the 1-5 interval, the tendencies are reversed, the volume of the non-dominant suffix is higher, although volume values are way lower.

Figure 4: Token frequencies of doublet suffixes by proportion intervals, logarithmic scale



In this section we discussed absolute and relative frequency of denominal adjectives, their degree of complexity as well as their productivity. The main finding was an inverse correlation between complexity and productivity: suffixes that form more complex adjectives (more semantically and morphologically opaque) are less productive. This concerns in particular *-n-*. Conversely, *-Ov-* tends to derive less complex (morphologically and semantically transparent) adjectives, and it is also the most productive suffix. In what follows we propose an assessment of, which properties of base nouns privilege one of the competing suffixes through statistical modelling on high frequency data. We also discuss to which extent these models can be applied to doublet data.

5. Statistical modelization of doublets

Three different binary classifiers were trained by means of the annotated datasets, each distinguishing a pair of suffixes. This task can be performed using logistic regression, decision trees or random forests, since all these methods are easily interpretable in terms of variable importance scores, and, therefore, are widely used for linguistic data. In what follows we will implement a random forest classifier (RFC)[17] on highly frequent adjectives (>100) and assess which base noun properties allow to distinguish between competing suffixes. The RFC is implemented with Scikit-learn library in Python (Pedregosa et al, 2011). RFC is a model used for classification purposes (e.g. to predict that the suffix will be *-n-* or *-sk-* given the properties of base nouns). The intuition behind random forests algorithm is the following: it builds several decision trees on different subsets of data; the class which is predicted by the majority of trees would be the final classification result.

To assess the performances of RFC classifiers we used a 10-fold cross-validation[18] and we compared the performances of the three models to the baseline (a model which always assigns the larger class). The results are presented in Table 11; these are the best accuracies we received since every model was optimized with the best hyperparameters[19]. The accuracies of classification are very high for *-n-*/*-sk-* and *-sk-*/*-Ov-*; the performances on the classification of

[17] Initially this paper used a decision tree classifier. However, we decided to follow the suggestion of one of the anonymous reviewers, who recommended to use random forests in order to minimize potential overfitting problems.

[18] This is a statistical method used to assess the performance of machine learning models. The 10-fold cross-validation method randomly divides the data on which a model is trained into ten parts; the first nine parts are used to train the models, while the tenth one is used to assess its performance. This process is repeated ten times, and the overall accuracy is calculated as the average of the results for each testing group.

[19] The following hyperparameters were optimized by means of grid search, a fine-tuned technique that computes the optimum values of hyperparameters: max number of levels in each decision tree, min number of data points allowed in a leaf node, min number of data points placed in a node before the node is split, number of trees in the forest

-n-/*-Ov-* are slightly worse. Nevertheless, all the models perform way better than their respective baselines.

Table11: Model accuracies in high frequencies (cross-validation)

Classification	FRC Accuracy	Baseline accuracy
-n- vs *-Ov-*	0.78	0.52
-n- vs *-sk-*	0.92	0.58
-sk- vs *-Ov-*	0.91	0.60

We can interpret the results of RFC by assessing the properties of base nouns that are the most important for each classification. Variable importance, shown in Table 12, is based on mean decrease in impurity.[20] The relevance of each predictor is computed according to value distribution after each split.

Table 12. Ranking of the most important properties of the base nouns

-n-/*-Ov-*	*-n-*/*-sk-*	*-sk-*/*-Ov-*
base length	[–proper];[–concrete]	[–proper];[+concrete]
[-proper];[-concrete]	[+proper];[–human]	[+proper]; [-human]
[-proper];[+concrete]	[–proper];[+human]	base length
last phoneme: [+velar]	last phoneme: [+dental]	[–proper];[+human]
infl. class 2		last phoneme: [+velar]
last phoneme: [+labial]		

The results concerning feature relevance show that not all the properties of base nouns play a role in determining suffix choice. Nevertheless, some classes of properties, such as animacy or the last phoneme of the stem, are recurrent. Despite the fact that there are three distinct models, some predictors are shared among them for the same suffix: [–proper]; [–concrete] for doublets with *-n-*, [–proper; +human] for *-sk-*, [-proper; +concrete] and [+velar]-ending stem for *-Ov-*.

In what follows we propose an assessment of the performance of the constructed models and test them on doublets. Since there is no 'right' or 'wrong' answer for doublets – both suffixes are possible –, we cannot assess the performance of our models through canonical metrics (for example, accuracy or confusion matrix). Instead, we will analyze tendencies in the predicted probabilities of models[21]. What are the predicted probabilities for each of the competing suffixes, given the properties of base nouns? Table 13 shows the mean predicted probabilities for each suffix in doublet couples. In general, all the models make little distinction between two competing suffixes, while *-sk-*/*-Ov-* predictions differ in 18 points, the preference for *-Ov-* over *-n-* is 12 points higher; the preference for *-sk-* over *-n-* - 10 points higher.

Table13: Mean predicted probabilities for doublet data.

Doublets	Predicted probabilities

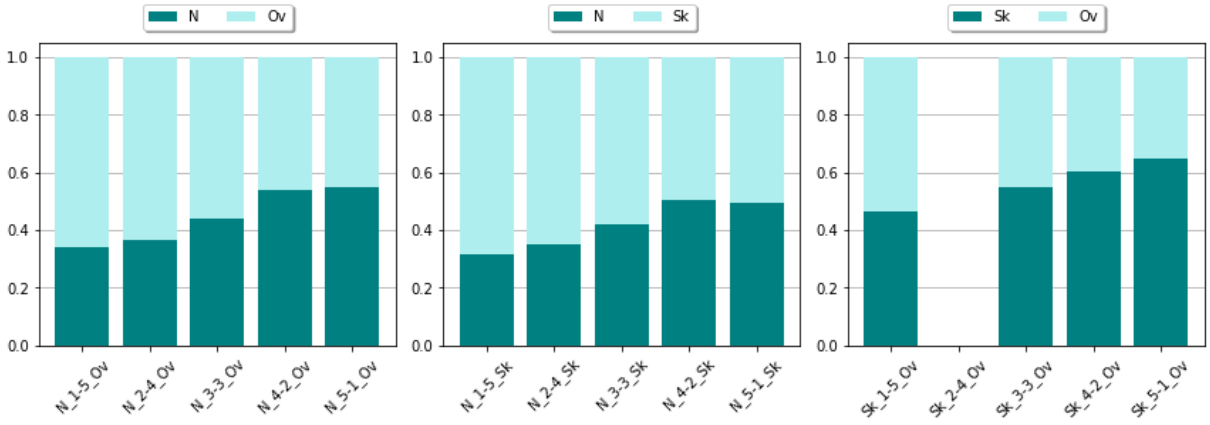
[20]Feature importance based on mean decrease in impurity is computed as the mean and standard deviation of accumulation of the impurity decrease within each tree.

[21] The scikit-learn *predict_proba()* function outputs probability of each class instead of class labels. It assumes that the probability is positive and uses the logistic function.

	suffix 1	suffix 2
-n- / -Ov-	0.44	0.56
-n- / -sk-	0.45	0.55
-sk- / -Ov-	0.59	0.41

Figure 5 shows details of predicted probabilities using proportion intervals. Here a general tendency emerges: the higher the proportion of an adjective in a doublet couple, the more chances its suffix has to be predicted. As frequency had not been included as a variable in the training, we consider this is an interesting finding. This means that in a doublet couple an adjective with one suffix which is more frequent than the adjective with another suffix may display specific base noun properties which would lead the model towards the preference of the first suffix over the second.

Figure 5: Predicted probabilities for doublet suffixes by proportion intervals.



In the case of doublets containing an adjective in *-n-*, the frequency analysis performed above showed that this is a dominant suffix (i.e. it displays a higher global token frequency). We have also shown that frequency is positively correlated to morphological complexity: adjectives with *-n-* are generally accessed as a whole compared to adjectives with *-Ov-* which are accessed via their constituents. Finally, we have seen that *-n-* is the least productive suffix in doublet data. Even if both models seem to capture frequency distribution tendencies through proportional intervals, the possible correlation to frequency is no longer supported and challenges our hypothesis according to which high frequency lexemes may influence morphological competence and, consequently, the formation of new words. Models predict *-Ov-* and *-sk-* in doublets data with higher probability than *-n-*, a fact that is more aligned with the low parsability of *-n-* adjectives and its low productivity. Apparently, the token frequencies of one or another suffix are insufficient to determine the tendencies at work in the formation of new words.

In what follows we take a closer look at the properties of base nouns that are the most relevant according to their p-value in order to determine the preference for a suffix, and we analyze doublet data consequently. Table 14 shows base noun distributions in doublets data according to the most relevant properties in suffix prediction.

Table 14: Nominal base property distribution for *-n-/-Ov-* doublets

Animacy	Base length	Inflectional Class	Last Phoneme
[–proper];[–concrete]: 86 [–proper];[+concrete]: 139 [–proper];[+human]: 2	1 syll: 48 2 syll: 118 3 syll: 49 4 syll: 10 5 syll: 2	1: 46 2: 156 3: 15	[+alveolar]: 93 [+dental]: 69 [+labial]: 17 [+velar]: 48

The most important predictors for *-n-* ([–proper];[–concrete], monosyllabic stem, 1st inflectional class) are in minority. On the other hand, there are more bases with properties that were observed to be relevant in the prediction of *-Ov-* (2nd inflectional class, [+alveolar]-ending stems). This disparity can bias the model towards *-Ov-* in spite of *-n-*. In its turn, Table 15 shows the distribution of base nouns according to the most relevant properties for *-n-/-sk-* prediction.

Table 15: Nominal base property distribution for *-n-/-sk-* doublets

Animacy	Last Phoneme
[–proper];[–concrete]: 24 [–proper];[+concrete]: 15 [–proper];[+human]: 68 [+proper];[–human]: 1	[+alveolar]: 66 [+dental]: 28 [+labial]: 8 [+velar]: 6

As in the case of *-n-/-Ov-* doublets, properties relevant for *-n-* ([–proper];[–concrete]) are in minority, whereas features favoring *-sk-* ([–proper];[+human], [+alveolar]-ending stems) are in majority, except for [+proper];[–human] base nouns. The same tendency is thus observed: despite the fact that adjectives in *-n-* display the highest token frequencies in both *-n-/-Ov-* and *-n-/-sk-* couples both models predict *-Ov-* and *-sk-* respectively, since the formal and semantic properties of base nouns disfavor *-n-*. This may be due to the lowest productivity of *-n-* in doublet data, where one of the adjectives displays high frequency, and the other a very low one. We suggest that this fact can be linked to diachronic changes in progress that have been observed independently. For instance, while Nemčenko (1973) identified *-sk-* as the most productive suffix in synchrony, Alekseeva (2011) observed a shift in tendencies and argues that *-Ov-* has become the most productive.

6. Conclusion and perspectives

In this study we performed a comparative analysis of highly frequent lexemes and of doublets of denominal adjectives in Russian in terms of frequency. The same tendencies hold for both sets: overall, *-n-* is more frequent than *-sk-* and the latter is more frequent than *-Ov-*. We have shown that relative frequency may be associated with the degree of complexity of adjectives, making *-n-* the least parsable (more semantically and morphologically complex) and *-Ov-* the most parsable suffix (more transparent from the morphological and semantic point of view). This measure is in line with observations previously made in the literature according to which adjectives in *-n-* tend to display more qualitative readings, while adjectives in *-Ov-* tend generally more towards the relational pole. In this respect, adjectives in *-n-* are more inclined to lexicalization, and are more prone to be accessed as a whole instead of being parsed

according to their morphological structure. We have also shown that *-n-* has the lowest productivity: its token frequencies are quite high compared to the frequencies of *-sk* and *-Ov-*, the latter giving rise to hapaxes more often than *-n-*.

In the second part, we presented the results of training a random forest classifier on highly frequent data in order to assess the extent to which each suffix can be distinguished from another on the basis of the formal and semantic properties of base nouns. The reason to build a RFC is the interpretability of this model which can provide insights on base noun properties that are the most relevant to discriminate between two competing suffixes. As a result, we provided evidence that some properties of base nouns are recurrent for the choice of the suffix (e.g. animacy or the last phoneme of the stem). We tested our models on doublets data and explored the predicted probabilities for each of the competing suffixes. What we wanted to assess is whether the most frequent suffix is also the one with higher predicted probability. Even if we observed such tendencies in data distribution, overall results show that the most frequent suffix (for instance *-n-*) is predicted with lower probability than the other. We showed that the most common properties of base nouns in the doublet dataset correspond to the most relevant properties in the prediction of suffixes *-sk-* and *-Ov-*. This fact explains why the most frequent suffix is not also the most probable.

At this stage we performed our analysis mainly on the formal and semantic properties of base nouns. This choice leaves us with some open questions. All the numerical representations for frequencies and complexity provided in the data exploration section were performed at the stage of statistical data analysis. However, none of these properties was included as a predictor for our model. We plan to build other classifiers taking into consideration both frequency and complexity. In their turn, semantic properties were limited to a set of categorical subclasses of animacy and were coded by hand. Their reliability depends on the robustness of the annotation protocol as well as on the agreement rate between annotators (although complex cases were resolved by discussion between them). In order to avoid any bias in data, it could be useful to explore numerical semantic representations of base nouns instead of categorical ones, for instance by means of a distributional semantic analysis. Finally, the etymological origin of the base noun, in particular concerning the Slavic vs. foreign distinction, is generally considered to play a role in affix selection. A first analysis of this parameter, performed by building a scale of 'nativeness' based on bigram frequency, showed its relevance, in particular when it is considered as a continuous rather than categorical property (Bobkova & Montermini 2021).

Acknowledgments

We wish to thank the participants in the on-line workshop *Affixal Rivalry* which took place on March 19, 2021 for their valuable comments. We also thank the editors of this issue, Richard Huyghe and Rossella Varvara, as well as the anonymous reviewers for helping us to improve greatly the content of this article.

Supplementary materials

The datasets and scripts used in this research are available at the following URL: https://zenodo.org/record/7396347#.Y4zX5ezMJR0.

Bibliography

Alekseeva, Evgenija V. 2011. *Адъективные новообразования в современном русском языке*. Sankt-Peterburg: Sankt-Peterburg Gosudarstvennyj Universitet. (Doctoral dissertation).

Antić, Eugenia. 2012. Relative frequency effects in Russian morphology. *Frequency effects in language learning and processing*, 1, 83-107.
Baayen, R. Harald. 1992. Quantitative aspects of morphological productivity. In *Yearbook of Morphology 1991*. Springer, Dordrecht. 109-149.
Baayen, R. Harald. 1993. On frequency, transparency and productivity. In *Yearbook of Morphology 1992*. Springer, Dordrecht. 181-208.
Baayen, R. Harald. 2001. *Word Frequency Distributions*. Springer Science & Business Media.
Bobkova, Natalia. 2022. Statistical modelization of suffixal rivalry in Russian: Adjectival formations in *-sk-* and *-n-*. *Corpus* 23.
Bobkova, Natalia & Montermini, Fabio. 2019. Suffix rivalry in Russian: What low frequency words tell us. In Audring, J., Kotsoukos, N., Manoulidou, K. (eds.), *Rules, Patterns, Schemas and Analogy. MMM12 online Proceedings*. 1–17. Patras: University of Patras.
Bobkova, Natalia & Montermini, Fabio. 2021. Suffixal variation in Russian denominal adjectives of Slavic and foreign origin. Paper presented at the workshop *Internationalisms in Slavic as a window into the architecture of grammar*. Interslavic 2020/2021. February 24-26, 2021 (online).
Bottineau, Tatiana. 2012. Les variations sémantiques du suffixe russe *-ovat-*. *Slavica Occitanica* 84:211–227.
Bybee, Joan. 1995. Diachronic and typological properties of morphology and their implications for representation. *Morphological aspects of language processing*, 225-246.
Evert, Stefan & Baroni, Marco. 2007. zipfR: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Session,* Prague, 29-32.
Fraser, Norman M. & Corbett Greville G. 1995. Gender, animacy, and declensional class assignment: A unified account for Russian*. Yearbook of Morphology* 1994. 123–150.
Guzmán Naranjo, Matías. 2020. Analogy, complexity and predictability in the Russian nominal inflection system. *Morphology*. 30(3). 219–262.
Gaeta, Livio & Davide Ricca. 2006. Productivity in Italian word formation: A variable- corpus approach. *Linguistics* 44(1). 57–89.
Hay, Jennifer. 2003. *Causes and Consequences of Word Structure*. New York: Routledge.
Hay, Jennifer. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics,* 1041-1070.
Harwood, Frank. & Wright, Alison. 1956. Statistical study of English word formation. *Language*. 260-273.
Hénault, Christine. & Sakhno, Serguei. 2016. *Чем* супермаркет-н-ый *лучше* супермаркет-ск-ого*? Словообразовательная синонимия в русских адъективных неологизмах по данным Интернета*. In Tošovic, B. & Wonisch A. (eds.). *Wortbildung und Internet*. 107-124, Graz: Institut für Slawistik.
Kapatsinski, Vsevolod. 2010. Velar palatalization in Russian and artificial grammar: Constraints on models of morphophonology. *Laboratory Phonology* 1(2). 361–393.
Kustova, Galina I. 2018. Прилагательные. In: *Материалы к корпусной грамматике русского языка*. III. *Части речи и лексико-грамматические классы*. 40–107. Sankt-Peterburg: Nestor-Istorija.
Nemčenko, Vasilij N. 1974. *Словообразовательная структура отсубстантивных суффиксальных имён прилагательных в современном русском языке.* Gor'kij. (Doctoral dissertation).
Parker, Jeff & Sims, Andrea. 2019. Irregularity, paradigmatic layers, and the complexity of inflection class systems: A study of Russian nouns. In: Arkadiev, P., Gardani, F. (eds.). *The Complexities of Morphology*. 23–51. Oxford: Oxford University Press.

Pedregosa, Fabian & Varoquaux, Gaël & Gramfort, Alexandre, et all. 2011. Scikit-learn: Machine learning in Python. In *the Journal of machine Learning research 12*. 2825-2830.
Sims, Andrea. 2017. Slavic morphology: Recent approaches to classic problems, illustrated with Russian. *Journal of Slavic Linguistics*. 25(2). 489–542.
Sims, Andrea & Parker, Jeff. 2015. Lexical processing and affix ordering: Cross-linguistic predictions. *Morphology* 25(2). 143–182.
Sorokina El'vira A. 1984. *Прилагательные-неологизмы современного русского языка*. Moskva: Gosudarstvennyj Universitet im. M.V. Lomonosova. (Doctoral dissertation).
Švedova, Natal'ja Ju. 1980. *Русская грамматика*. Moskva: Nauka
Thuilier, Juliette. 2012. *Contraintes préférentielles et ordre des mots en français*. Paris: Université Paris Diderot. (Doctoral dissertation).
Timberlake, Alan. 2004. *A Reference Grammar of Russian*. Cambridge: Cambridge University Press.
Townsend, Charles E. 1975. *Russian Word-Formation*. Columbus, OH: Slavica Publishers.
Varvara, Rossella. 2019. Misurare la produttività morfologica: i nomi d'azione nell'italiano del ventunesimo secolo. In *Le tendenze dell'italiano contemporaneo rivisitate. Atti del LII Congresso Internazionale di Studi della Società di Linguistica Italiana*. Roma: Società di Linguistica Italiana, 187-201.
Yanushevskaya, Irena & Bunčić, Daniel. 2015. Russian. *Journal of the International Phonetic Association* 45(2): 221–228.
Zaliznjak, Andrej A. 2003. *Грамматический словарь русского языка. Словоизменение*. Moskva: Russkij jazyk.
Zemskaja Elena A. 2015. *Язык как деятельность. Морфема, слово, речь.* Moskva: Flinta.