Quantitative approaches to suffixal rivalry in denominal adjective formation in Russian

Natalia Bobkova, Fabio Montermini

CLLE-ERSS, CNRS & Université de Toulouse Jean Jaurès

1 Introduction

The derivation of adjectives from nouns is a complex issue 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, among others).

The strategies used to derive adjectives from nouns in Russian are varied. Švedova (1980), for instance, enumerates more than 25 suffixes, which have various degrees of productivity. The main three adjectival suffixes are *-n-*, *-sk-* and *-Ov-*[1], all other suffixes may be considered as their extended variants, for instance, *-esk-*, *-ičesk-* are variants of *-sk-*, whereas *-ičn-* is a variant on *-n-* (Bobkova and Montermini 2019).

Recent developments in derivational morphology (cf. 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:

(1)	*slesar'*	'locksmith'	↔	*slesar**n**(yj) / slesar**sk**(ij) / slesar**ev**(yj)*
	dinamika	'dynamics'	↔	*dinamič**esk**(ij) / dinamič**n**(yj)*
	simmetrija	'symmetry'	↔	*simmetr**ičesk**(ij) / simmetr**ičn**(yj)*
	bojec	'fighter'	↔	*bojc**ovsk**(ij) / bojc**ov**(yj)*[2]

In this paper we are particularly interested in the rivalry between the following productive suffixes, as they frequently appear with the same nominal bases:

(2) *-n- / -sk- / -Ov-*
-esk- / -n-
-ičesk- / -ičn-
-Ovsk- / -Ov-

We aim to establish properties of nominal bases which allow a distinction between these suffixes, regardless of doublets. The choice of one or the other of the suffixes is accounted for by scholars (Švedova 1980, Hénault 2016) by either purely phonological factors, semantic or lexico-morphological ones:

[1] The notation *-Ov-* indicates the variation of the vowel of this suffix, capital *O* may correspond to orthographically different surface forms, <o> or <e>.

[2] Note that the examples of *dinamika* and *bojec* display 2 cases of stem allomorphy: the first one concerns a mutation of the last phoneme of the stem ('dinamik'-'dinamič'); the second – vowel-zero alternation ('bojec'-'bojc'). Both phenomena are typical of Russian language, however, they can not be described by productive morphophonological processes in synchrony.

(3) *-n-* tends to form more qualitative adjectives, whereas *-sk-* is used to form more relational ones;

-Ov- appears with inanimate base nouns, *-Ovsk-* choses 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.).

Our goal is to use quantitative approaches to reveal the main predictors (constraints) which result in the choice of a particular suffix. We show the results of 2 models based on a multifactorial analysis: logistic regression and decision trees. Since both allow an easy visualisation of the list of predictors (the most important features for the choice of the suffix, and – in case of decision trees – the visualization of the procedure of classification), we expect to highlight the properties of the base nouns which can motivate the choice of a particular affix.

2 Data and methodology

To perform our analysis, we extracted the adjectives from the National corpus of Russian language (https://ruscorpora.ru/), proceeded to manual cleaning and automatically reconstructed the bases for each adjective. Our final data set is composed of 4351 entries. Since the competition between the affixes listed above is driven by a complex combination of factors, the base nouns were annotated according to some of their properties:

(4) phonological: last phoneme of the stem, length of the base noun in syllables, stress position;

morphological: inflexional class;

semantic: animacy (Thuilier 2012), which combines in different ways such properties as [±common], [±human], [±concrete];

etymological: native or loanword.

The properties listed in (5) form the list of predictors for both models.

The data were further divided into two subcorpora: the highest frequent lexemes (>100; 2275 entries) and hapaxes (frequency 1; 2076 entries) lexemes. Dal & Namer (2012) show for instance that very low-frequency lexemes, if observed on a large scale, are likely to be good indicators of the creative use speakers do of morphological constructions, since they are less likely to have undergone phenomena of lexicalization and thus to be formally and/or semantically opaque.

The main goal of dividing the data into two subcorpora is to build a statistical model which learns the adjectival formation from the high frequency subcorpus (training set) and to evaluate how well the same model can apply its knowledge to predict the suffix in the low frequency subcorpus (test set). The training set was randomly divided into a proper training set and a dev set – for evaluation of the model on high frequency data as well.

Since a large number of predictors can introduce noise in models, we perform feature selection for every suffix as well.

3 Results

First, we observe descriptive statistics and data distribution between high and low frequency subcorpora for emerging tendencies.

The distribution of *-n-/-sk-/-Ov-* is even between both subcorpora, there are more lexemes in high frequency data set, comparing to low frequency one. The same tendencies are observed for *-esk-/-n-* distribution.

As for *-ičesk-/-ičn*, they are equally distributed between the subcorpora, although the lexemes are more numerous in the low frequency data set, especially with *-ičesk-*; we can hypothesize that this suffix is productive in synchrony and is used by speakers more often than other ones to form new adjectives.

-Ovsk-/-Ov- represents another interesting case for study: the proportions of both suffixes are inversed in 2 subcorpora. If in high frequency lexemes *-Ov-* is attested more often than *-Ovsk-*, in low frequency lexemes it is the opposite: the adjectives formed with *-Ovsk-* are more numerous than with *-Ov-*. We would expect to find proper nouns, of Russian and foreign origin, among noun bases – since their inventory can be potentially unlimited, this can explain the high productivity of *-Ovsk-* among low frequency lexemes.

3.1Rivalry of *-n-/-sk-/-Ov-*

Since the choice here is made between 3 suffixes, we used a multinomial logistic regression to evaluate the results. Both the logistic regression and decision trees can apply quite well the tendencies learned from the training set on dev and test sets, with an accuracy of 72 and 61, respectively. The main constraints which interact here and determine the choice of the suffix are animacy (*-n-* choses common abstract nouns, *-sk-* combines with common human and proper non-human, and *-Ov-* privileges common concrete nouns); to a lesser extent - the length of stem in syllables (whereas *-n-* and *-sk-* choses polysyllabic bases, *-Ov-* has a clear preference for monosyllabic ones) and the last phoneme of the stem (*-n-* privileges dental consonants, *-sk-* and *-Ov-* both combine more often with alveolars, *-Ov-*, in its turn, has also a preference for velars).

3.2Rivalry of *-esk-/-n-*

Both binomial logistic regression and decision trees provide excellent results in classification for both dev and test set with accuracy of 97 and 92. This proves that the same tendencies are preserved between the lexicalized adjectives formed with these suffixes and new emerging adjectives. Phonological constraints are the strongest: stress position is crucial to determine the choice of the suffix: *-esk-* is chosen more often for nouns where the antepenultimate syllable is stresses; *-n-* combines with nouns where the ultimate and penultimate syllables are stressed. Another phonological factor involved to determine the choice of the suffix is the last phoneme of the stem: *-n-* combines with dentals whereas *-esk-* with velars. Etymological factor is also important, however to a lesser extent: *-n-* privileges native stems more often than foreign, whereas the tendency is the opposite for *-esk-*.

3.3Rivalry of *-ičesk-/-ičn-*

Comparing to the previous competing suffixes, both statistical models perform more poorly to solve the rivalry between *-ičesk-* and *-ičn-*: the accuracy on the dev set is 95, and the generalization to the test set is 82. According to the models, semantic constraints prevail (animacy), the phonological factor is present again (last phoneme of the stem). However, if we take a closer look on misclassified data, we can see that *-ičesk-* was classified correctly in almost all the cases in dev and test sets, the large majority of misclassified data concerns -*ičn*-. Unfortunately, the tendencies learned by the models cannot shed light on the rivalry between these two suffixes.

3.4 Rivalry of -*Ovsk*-/-*Ov*-

As mentioned above, there is more data for testing the models than for training them. Unsurprisingly, the models can learn tendencies for high frequency lexemes and apply the knowledge quite well for the data coming from the same distribution: the accuracy on dev set is 94; as for classification on test set, the performance of both models drop, the accuracy is only 75. However, the conclusions about the main predictors can be made: semantic factors constitute the main constraint (-*Ov*- combines mostly with common abstract and common concrete nouns, -*Ovsk*- choses common human and proper human nouns); the phonological factor can also play a role (-*Ov*- privileges stems ending with velars whereas the is no clear preference for -*Ovsk*-).

4 Discussion

Our study based on statistical models allowed us to identify the constraints determining the choice of different rival suffixes forming denominal adjectives in Russian. The strongest constraints concern phonology, semantics and etymology of the base noun. Morphological factors, such as inflectional class, play a less significant role. The results of our study show that the factors often cited in the literature are good predictors for the choice of the suffix. Our results may contribute to improve the list of the best predictors for the choice of the affix, and to order these predictors according to their force in imposing the choice of an affix. The use of statistical models needs some precautions: the issues of inequalities in distributions as well as the lack of data should be addressed. The models used for the study capture only formal properties of base nouns and do not allow to take into account details concerning the semantics of derived adjectives. For this study we did not include in the list of parameters the type of corpus adjectives appear in (literature, newspapers, oral texts, poetry, etc.). We keep for further investigations the inclusion of the type of corpus among the predictors, the study of semantics of the adjectives using distributional methods and a more detailed study of existing doublets as well.

References

Aronoff, M. 2016. Competition and the lexicon. In A. Elia, C. Iacobini & M. Voghera (Eds), *Livelli di Analisi e fenomeni di interfaccia. Atti del XLVII congresso internazionale della Società di linguistica Italiana*. Roma: Bulzoni, 39-52.

Bobkova, N. & F. Montermini (2019). Suffix rivalry in Russian: what low frequency words tell us, *Mediterranean Morphology Meetings* 12:1-17.

Dal, G. & F. Namer. Faut-il brûler les dictionnaires? ou comment les ressources numériques ont révolutionné les recherches en morphologie. In *SHS Web of Conferences*, volume 1, pages 1261–1276. EDP Sciences, 2012.

Lindsay, M. & M. Aronoff. 2013. Natural selection in self-organizing morphological systems. In N. Hathout, F. Montermini & J. Tseng (Eds.), *Morphology in Toulouse. Selected prodeedings of Décembrettes 7*. Munich: Lincom Europa, 133-153.

Bonami, O. & J. Thuilier. 2018. A statistical approach to rivalry in lexeme formation: French -iser and -ifier. *Word Structure* 11(2): 4-41.

Hénault, C. & S. Sakhno. 2016. Čem supermarket-n-yj lučše supermarket-sk-ogo? Slovoobrazovatel'naja sinonimija v russkix ad"ektivnyx neologizmax po dannym Interneta. In Branko Tošović & Arno Wonisch (eds.), *Wortbildung und Internet, xxx–xxx*. Graz: Institut für Slawistik.

Plénat M. 2011. Enquête sur divers effets des contraintes dissimilatives en français. In M. Roché, G. Boyé, N. Hathout, S. Lignon et M. Plénat, *Des unités morphologiques au lexique*, Hermes-Lavoisier, Paris, pp.145-190.

Roché M. 2011. Quel traitement unifié pour les dérivations en -isme et en -iste ?. In M. Roché, G. Boyé, N. Hathout, S. Lignon et M. Plénat, *Des unités morphologiques au lexique*, Hermes-Lavoisier, Paris, 69-143.

Švedova, N. 1980. *Russkaja grammatika*. Moskva: Nauka.

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