Quantitative Analysis of Suffix Variability of Comparative Adjectives in Russian

Timur I. Galeev
Kazan Federal University, Kazan, Russia
Justus-Liebig-Universität Giessen, Germany
https://kpfu.ru/timur.galeev
tigaleev@kpfu.ru

Vladimir V. Bochkarev
Kazan Federal University, Kazan, Russia
https://kpfu.ru/vladimir.bochkarev
vbochkarev@mail.ru

Abstract

There are two variants of the productive suffix of comparative adjectives used in modern Russian. They are a full two-syllable form and a reduced one-syllable suffix. Both variants are normative. However, they are slightly different in terms of stylistics. The suffix -ee makes the word sound neutral and the word with the suffix -ei sounds more colloquial. The article presents a quantitative study of variability of the suffixes of comparative adjectives and analyzes linguistic and extralinguistic factors that influence the frequency of the variants. The authors concluded that there is no previously anticipated influence of phonetic and morphological factors on the choice of the suffix of an adjective in a bookish speech.

2012 ACM Subject Classification Computing methodologies → Phonology / morphology

Keywords and phrases Adjectives, language change, variability, Google Books Ngram, Russian language

Digital Object Identifier 10.4230/OASIcs.SLATE.2019.21

Funding This research was financially supported by the Russian Government Program of Competitive Growth of Kazan Federal University, state assignment of Ministry of Education and Science, grant agreement № 34.5517.2017/6.7 and by RFBR, grant № 17-29-09163.

1 Introduction and Literature Review

The presence of extra-large diachronic corpora led to significant changes in linguistics. It became possible to conduct a microanalysis of linguistic phenomena, such as morphosyntactic variability. Such studies have changed significantly in both theoretical and methodological terms over the past ten years. Many theoretical works on syntactic variability were inscribed in the paradigm of generative linguistics and reduced themselves to identifying semantic differences between variants and determining the primacy of one of the variants. In terms of methodology, a lot of works reduced to mono-factorial studies based on relatively simple text fragments (resolving collocations).

Nowadays variability studies have become much more functional and methodologically more complex: language facts are now interpreted and motivated by psycholinguistic factors, and the study material is analysed in terms of the cognitive or usage-based approach [3, 6, 7, 10]. Regression analysis of linguistic phenomena has gained popularity and been used in many recent works [12, 13, 14].

The common feature of the above-mentioned works is that they use statistical analysis of corpus or experimental (survey) data and try to identify new indicators that influence the


© Timur I. Galeev and Vladimir V. Bochkarev;
licensed under Creative Commons License CC-BY

choice of a competing form. Our study, in contrast to the studies of foreign colleagues, is carried out on the material of the standardized literary Russian language.

Using regression analysis that considers many factors and can be widely used for comparative studies, European and American scientists illustrate that, on the one hand, grammatical variability is found within a particular regional and cultural tradition (indigenization, or "privatization"). But on the other hand, different regional dialects can show the same cases of variability as in the standard language (literary language) [14]. Within the same ethnic group that has mastered a particular language and, according to A.A. Shakhmatov [16], "treated the language as his own property", such variability arises after some time despite the initial absence of negative language experience.

In this case, some factors (such as different cultures of speakers of different languages) prove the universal nature of grammatical variability and its potential as a marker of certain cognitive processes typical of human brain. Within the framework of a psycholinguistic-based grammar model, we also interpret these results from different explanatory points of view, including the phenomena of language contact, assimilation of the second language (albeit to a lesser extent), semantic variations and changes.

2 Methods

We searched for examples of variability in the Google Books Ngram, unprecedently extra-large corpus [11]. For example, the pattern of morphological variability is "N-ee \N-ei", where N is a combination of letters same for the first and second variants (umn-ee\umn-ei (smarter)). Henceforth, we used the ALA-LC transliteration system.

These patterns are tested after the discussion. Data are collected from the Google Books Ngram corpus. To be more objective, the search is performed (1) for the whole Russian corpus (the period 1607-2009) and (2) for the period 1920-2009 (after the spelling reform). The "purity" of the obtained results indicates the quality of the pattern. If the obtained data are incorrect the pattern is refined. Part-of-speech tagging of the obtained data array was checked using the OpenCorpora dictionary [2]. Not only Russian comparative adjectives but also some forms of nouns and pronouns can end in -ei. Some of them can be homonymous to the studied part of speech. For example, the forms umnei and starei can both correspond to verbs (singular imperative form) and comparative adjectives.

Such isomorphism of dynamic verb forms and forms of comparative adjectives in slavic languages [1], as well as statistical probability of homonymy in inflectional languages in general [15] were studied by J.D Bobaljik and his coauthors. Homonyms belonging to different parts of speech were excluded. For example, verbs in the imperative form like umnei (be clever!) (V) were excluded from the list, which allowed us to analyse only the occurrences of the adjectives umnei (smarter) (Adj), which are variants of the word umnei (smarter) (Adj).

At the second stage of the research, we manually excluded the homonymous forms. The examples that were not variants were excluded (for example, vechernee// vechernei (zarei) (evening/evening (dawn)). The selected list of words contains quantitative data: the number of each variant occurrence in the corpus (for example, bolnee - 49060, bolnei - 22691 (more sick)) and the ratio of the variants (bolnee - 68%, bolnei - 32%). The ratio of the frequencies was summed, and the arithmetic average of the whole group was calculated. For example, the percentage of adjectives with the suffixes -ee// -ei is 71.54% // 28.46% within the entire period, and 71.41% // 28.59% over the last 100 years.

Graphs of changes in the frequency use of both variants were constructed for each of the pairs and for the sum of the variants of the whole group in order to demonstrate the

language dynamics and compare the trends and total quantitative data. The obtained data on the number and dynamics were interpreted and described to assess the degree of relevance and expediency of certain standard prescriptions.

To improve the accuracy of the results, we performed preprocessing of the corpus raw data. After that, we selected only the vocabulary 1-grams from the obtained ones. By 1-grams we understand words composed only of the letters of the Russian alphabet. Word forms that differ only by the presence or absence of the letter "er" at the end of them were regarded as the same words. To normalize and calculate the relative frequencies, the number of vocabulary 1-grams was calculated for each year (unlike Google Books Ngram Viewer, where normalization is performed for the total number of 1-grams). Another difference is that we calculate frequencies without taking into account differnces in capitalisation.

3 Results and Discussion

Graphs of changes in the frequency use of both variants were constructed for each of the pairs and for the sum of the variants of the whole group in order to demonstrate the language dynamics and compare the trends and total quantitative data.

In total, 1571 pairs of comparative adjective were analyzed. The suffix -ee prevails in 1395 cases, the suffix -ei prevails in 174 cases. The same number of word usage with the suffixes -ee/-ei was found in 2 cases. The average ratio of the variants for all pairs of comparative adjectives is as follows: 71.54% (-ee) // 28.46% (-ei).

The obtained quantitative and diachronic data were interpreted and described to assess the degree of relevance and expediency of certain standard prescriptions taking into account linguistic factors (phonetics, morphology and style).

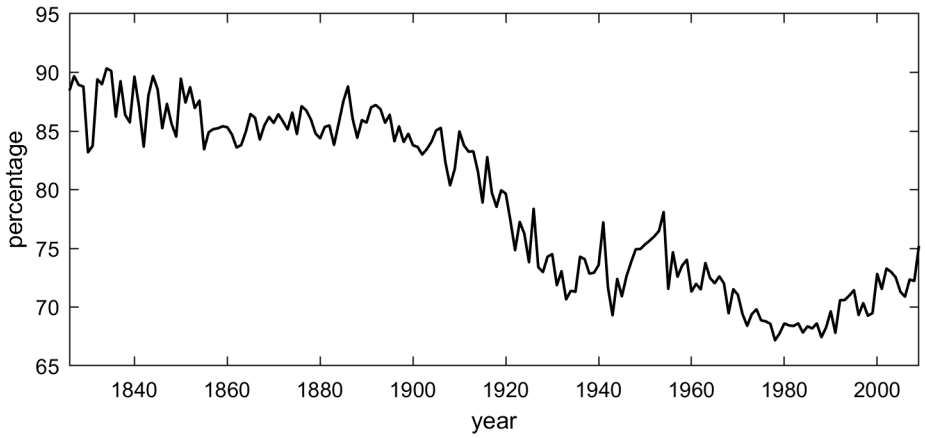

Figure 1 The mean value of the word share (for the total number of words with -ee/-ei).

The following tendency was detected. Low-frequency words are characterized by approximately the same number of occurrences of the variants (for example, nepriglyadnee (60) // nepriglyadnei (ugly) (41) (59.41% vs. 40.59%); dramatichnee (51) // dramatichnei (dramatic) (56) (47.66% vs 52.34%). High-frequency words show the opposite tendency, the number of their use significantly differs (sern'esnee (1900960) // ser'eznei (more serious) (197850) (90.57% vs 9.43%). High-frequency words with the suffix -ee prevail over the words with the suffix -ei (vazhnee, trogatelnee, effektnee (more important, more touching, more effective)).

The only exception is the word masshtabnei. Low-frequency words with the suffix -ei slightly prevail over the words with the suffix -ee (dramatichnei, conservativnei).

Analysis of changes in the frequency dynamics of the whole group of comparative adjectives showed the following overall tendency: the number of words with the suffix -ei gradually increased within the period from the end of the 20th century to the beginning of the 20th century and rapidly increased in the 20th century. Figure 1 shows the ratio of the variants in different years. The variants ratio with the suffixes -ee/-ei was approximately 90%/10% in the middle of the 20th century. The share of the adjectives had gradually decreased over the century and a half and reached a minimum by 1980 (below 70%).

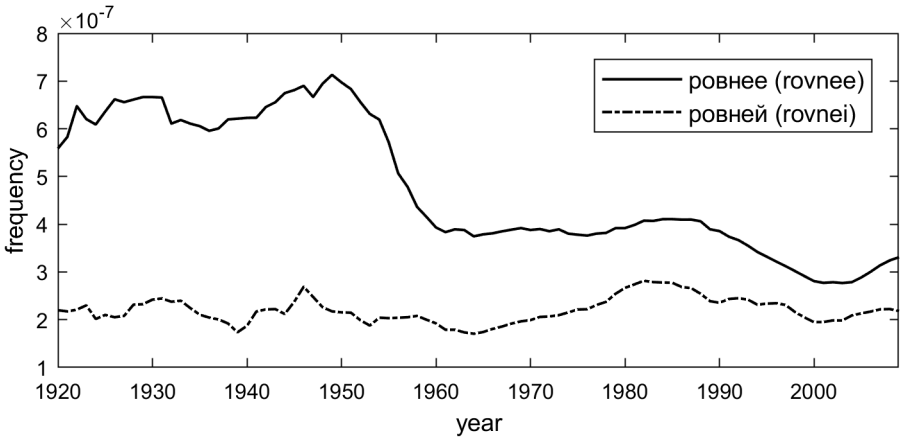

Figure 2 Frequency graph of the adjectives rovnee/rovnei (more even).

This tendency can be traced while analyzing the pair of words rovnee/rovnei (more even). The frequency of these variants will probably be equal in some time. An example of this case is shown in figure 2. Like in the Google Books Ngram Viewer service, a moving average with a window +- 3 years (the window length is 7 years) is used.

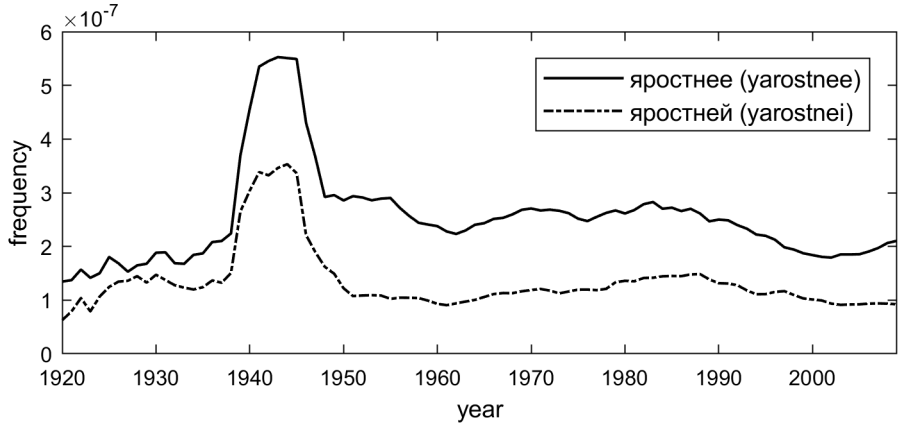

Figure 3 Frequency graph of the adjectives yarostnee/yarostnei (more violent/fierce).

Some pairs show frequency peaks due to historical events. For example, during the Great Patriotic War, there was a sharp increase in adjectives yarostnee/yarostnei (more violently/ fiercely) in Russian press and fiction (see figure 3). It was caused by appeal to defend the motherland.

The predominance of the variant with the suffix -ei was detected in some rare cases at the beginning of the 20th century. As a rule, in such cases, the initially less frequent variant with the suffix -ee has been used more frequently in recent time (see figure 4).

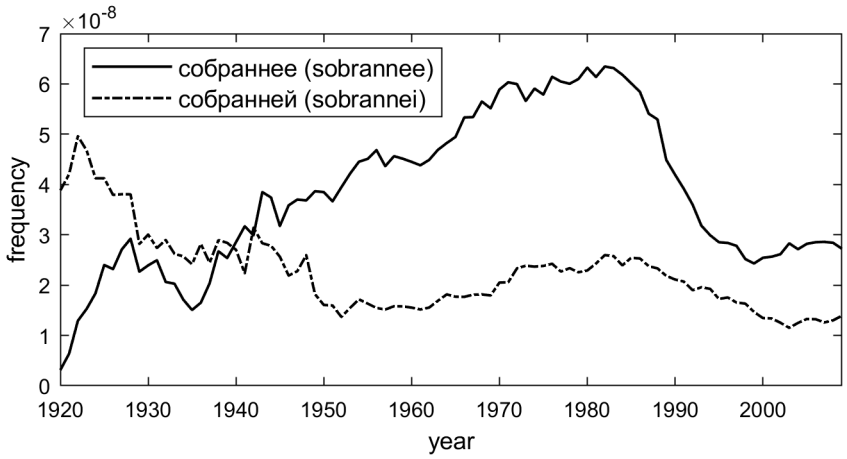

Figure 4 Frequency graph of the adjectives sobrannee/sobrannei (self-collected).

4 Conclusion

Russian morphology has been an extensively studied area of linguistics. Lots of synchronic studies on Russian words variability were performed in the pre-corpus era when the study material was analyzed manually.

It was previously believed that the choice of a variant form depends not only on the style of the book and an author‘s idiolect but also on the complexity of the word root and place of stress in the word and prefix. Some researches state [4, 5] that if a root is monosyllabic, the suffix -ee is preferable and if a root is multi-syllabic, the suffix -ei is preferable.

The main feature of our work is that we performed diachronic analysis of the given type of variability for the first time. Google Books Ngram was used as a study material and allowed for obtaining more objective data than that obtained manually from a relatively small number of sources.

The results were the following. Only ten pairs of adjectives (out of 102 pairs) with a monosyllabic root show predominance of the variant with the suffix -ei (10% versus 12% for the whole group). At that, in 10 cases, where the suffix -ei prevails over the competing suffix -ee, the maximum ratio is 75%, while the share of -ee is more than 90% in many cases. Thus, the cases of the predominance of the suffix -ei in the monosyllabic words are less than in the group on average. Moreover, there is no meaningful predominance in these cases.

In other works [8, 9], the variants with the prefix po-, which has the meaning of “a small increase or change in the quality of an adjective” are often mentioned: pobistrei, poslozhnei and posil‘nei (faster, more complicated, stronger). In a bookish speech, this additional meaning is conveyed by the word nemnogo (bit): nemnogo bistree, nemnogo slozhnee, nemnogo sil‘nee (a bit faster, a bit more complicated, a bit stronger).

The adjective suffix -ee was more frequent in 15 cases out of 190 (less than 8%). Comparing this result with the average value for the whole group (12%), we conclude that the prefix has no influence on the choice of the suffix -ei as a preferable one.

We detected a decrease in the proportion of forms with the suffix -ee in the book speech in the 80s years of the 20th century in comparison with the language situation in the middle of the 19th century by more than 20%. After the 80s of the 20th century, the proportion of the -ee forms fixed at the level of 75%. This indicates changes in morphology of the bookish speech which can be due to language liberalization and economy.

Thus, the Google Books Ngram corpus, being a source of "hidden" knowledge of bookish speech, allows one to identify language anomalies, as well as qualitatively and quantitatively improves the existing scientific description and interpretation of language features by verifying the previously obtained empirical data.

References

1 Jonathan D. Bobaljik. *Universals in Comparative Morphology: Suppletion, superlatives, and the structure of words*. MIT Press, Cambridge, USA, 2012.

2 V. Bocharov, S. Bichineva, D. Granovsky, N. Ostapuk, and M. Stepanova. Quality assurance tools in the OpenCorpora project. In *Komp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog»*, pages 101–109. Russian State University for the Humanities, Moskow, 2011.

3 Claire Childs, Christopher Harvey, Karen P. Corrigan, and Sali A. Tagliamonte. Transatlantic perspectives on variation in negative expressions. *English Language and Linguistics*, pages 1–25, 2018. doi:10.1017/S1360674318000199.

4 I.A. Es'kova. Obrazovanie sinteticheskikh form stepenei sravneniia v sovremennom russkom literaturnom iazyke. In *Razvitie grammatiki i leksiki sovremennogo russkogo iazyka*. Nauka, Moskow, 1964.

5 L. K. Graudina, V. A. Itskovich, and L. P. Katlinskaia. *Grammaticheskaia pravil'nost' russkoi rechi. Stilisticheskii slovar' variantov*. Nauka, Moskow, 2004.

6 Stefan Th. Gries. Syntactic alternation research: Taking stock and some suggestions for the future. *Belgian Journal of Linguistics*, 31(1):8–29, 2017. doi:10.1075/bjl.00001.gri.

7 Benedikt Heller, Benedikt Szmrecsanyi, and Jason Grafmiller. Stability and Fluidity in Syntactic Variation World-Wide: The Genitive Alternation Across Varieties of English. *Journal of English Linguistics*, 45(1):3–27, 2017. doi:10.1177/0075424216685405.

8 O.K. Kochineva. Stepeni kachestva bezlichno-predikativnykh slov v sovremennom russkom literaturnom iazyke. *Uch. zap. LGPI im. Gertsena*, 402:59–67, 1968.

9 V.Iu. Koprov. *Variantnye formy v russkom iazyke*. Voronezh state university, Voronezh, 2001.

10 Aet Lees. 4 Synchronic Corpus Study of Object Case Alternation. In *Case Alternations in Five Finnic Languages*, pages 52–92. Brill, Leiden, The Netherlands, 2015. doi:10.1163/9789004296367_005.

11 Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

12 Terttu Nevalainen, Elizabeth Closs Traugott, and Phillip Wallage. Quantitative evidence for a feature-based account of grammaticalization in English: Jespersen's Cycle, November 2012. doi:10.1093/oxfordhb/9780199922765.013.0060.

13 Dirk Pijpops and Freek Van de Velde. Constructional contamination: How does it work and how do we measure it? *Folia Linguistica*, 50(2):543–581, 2016. doi:10.1515/flin-2016-0020.

14 Melanie Röthlisberger, Jason Grafmiller, and Benedikt Szmrecsanyi. Cognitive indigenization effects in the English dative alternation. *Cognitive Linguistics*, 28(4):673—-710, 2017. doi:10.1515/cog-2016-0051.

15 Uli Sauerland and Jonathan Bobaljik. Syncretism Distribution Modeling: Accidental Homophony as a Random Event. In *Proceedings of GLOW in Asia IX 2012*, pages 31–53. Mie University, Mie, Japan, 2013.

16 A. A. Shakhmatov. *Ocherk sovremennogo russkogo literaturnogo iazyka*. Urait, Moskow, 2018.