The Behaviour of Russian Competing Verbs:
a Computer-Assisted Approach

Timur Galeev[1][0000-0002-7322-6225], Anna Shevlyakova[1][0000-0002-2659-1887],
Vladimir Bochkarev[1][0000-0001-8792-1491]

[1] Kazan Federal University, Kremlyovskaya str. 18, 420008 Kazan, Russia
tigaleev@kpfu.ru

Abstract. The article studies morphological variability of Russian verbs. The distributional and quantitative analyses of these verbs were performed based on the extra-large diachronic corpus Google Books Ngram. The obtained frequency data were interpreted in terms of language norm and evolution. The accurate time of the norm change was identified for each pair of verbs. It was found that distribution of the competing verbs can both coincide and be markedly different. Four main trends in the frequency behaviour of the competing pairs of verbs were revealed. The analysis of the variability type and frequency of the variants showed that usage of a particular variant form is largely context dependent and is not determined only by a speaker`s individual preferences. Each of the variant forms has its niche in the language. It was also revealed that the observed active return of unproductive forms of verbs to the Russian language indicates the general stability of the verb system of the Russian language and tendency to unification by productive type.

Keywords: variability, Russian verb, linguistic norm, language evolution, Google Books.

1   Introduction

The problem of linguistic variation has been extensively studied in the past half-century, and it has now become a highly productive subfield of research in corpus and computer linguistics. Variability is an inherent feature of any human language caused by the asymmetric dualism of a linguistic sign when any content can be expressed by different means [1].

Variations can occur at all levels within language and can be due to different factors. At the beginning of the 20th century, the founder of the Kazan linguistic school Baudoin de Courtenay [2] wrote that variability is the driving force of language evolution, which involves constant oscillations and fluctuations in the structure of any language. Variation analysis is one of the most fruitful areas in studies of language change and evolution because change involves competition.

Changes of language variants may not be obvious. However, thorough diachronic analysis can trace their behavior within time and indicate how language evolves. Several models of change have been proposed that describe the way linguistic changes start and their stages [3, 4].

Frequency of variant occurrence is a critical issue in such studies since analysis of fluctuation between variants can show tendency in which variants have a greater or lesser likelihood of occurring under certain conditions and help to observe language change in progress.

The beginning of the 21th century has witnessed remarkable growth in the quantitative study of linguistic variation due to new research opportunities associated with creation of extra-large text corpora. One of such text databases is the electronic library Google Books. It contains a great number of texts that can be used to estimate frequency of certain language phenomena and objectively deduce regularities of their use.

Different types of language variation were studied using the Google Books corpus. The dependence of rate of change of semantic meaning and connotative characteristics of a word on its frequency usage was studied by Hamilton, Leskovec & Jurafsky (2016). Two quantitative patterns of semantic changes were identified for a period of 200 years based on data from six historical text corpora written in four languages (including the Google Books corpus): 1) the law of correspondence – the rate of se-mantic changes is inversely proportional to the word frequency; 2) the law of innovation – regardless of frequency, the more meanings the polysemantic word has, the higher the rate at which it acquires new semantic meanings is [5].

Google Books Ngram data on morphological variability are increasingly interpreted using native speakers` behavioral models when they choose a linguistic form [6] and text style (in which the variant is used) [7]. Such data are also used to study how context influence the use of one or another variant, as well as cognitive factors in general [8, 9].

In our work, we study morphological variability of Russian verbs, which was first studied by Smirnitsky [10] and Vinogradov [11]. Variability of grammatical forms is especially relevant for the Russian language due to its inflectional nature.

As it is known, the Russian language has shown the following tendency for at least several centuries. Some verbs forming a relatively small and (almost) unproductive class, such as iskat` ('look for') transfer to the superproductive class of verbs, such as igrat` ('play').

The work objective was to study pairs of verbs and verbal forms that partially differ in shape but have close meaning. The first task was to analyse the distribution of these words and find some regularities of their use. The second task was, provided that the distribution of some of the words are almost identical (muchit / muchaet – 'torture', muchaetsia / muchitsia – 'anguish'), to analyze frequency of their use over time in terms of language norm and evolution.

2   Materials and Methods

We studied pairs of verbs, which had the same semantics and similar form. They were selected in the following way. Some of the verbs were taken from the book by Graudina [12], the rest were obtained using the Google Books Ngram corpus. More than 1000 verbs with alternating consonants at the end of the stem before the ending were extracted automatically from the Google Books Ngram corpus (*miau ...ch-et/...ka-et; bryz ...zh-et/...ga-et* ('meows', 'splashes'). Then, pairs of verbs that met the requirements were selected manually from the list. The list of the studied verb pairs included 122 verbs (each verb has two verb paradigms).

Then, the Ngram Viewer service was used to conduct a distributive analysis and study the contexts of use of the verb pairs. Ngram Viewer allows one to see the words that are most often used with a given word. Word distribution makes it possible to draw conclusions about differences in semantics of words, whether these words are completely interchangeable or not, and what patterns of usage they have.

Having performed the distributional analysis, we studied frequency of words used in the most similar contexts in the course time to see if one verb is supplanted by another and which of the forms dominates.

3   Results

The most common contexts of use of each pair of verbs and verbal forms were analyzed. It was found that these contexts coincide in some cases and are markedly different in other cases. For example, the distribution of words slomannyi/slomlennyi ('broken') is significantly different.

Table 1. Co-occurrence of the verbs *slomanyi* and *slomlennyi*

сломанный 'broken$_1$'	*nos, zabor, stul, zub, mech, zamok, nozh, klinok* 'nose, fence, chair, tooth, sword, lock, knife, blade'
сломленный 'broken$_2$'	*chelovek, dukhom, gorem, pytkami, zhizn'u, ustalost'u, bolezn'u* 'person, in spirit, by grief, by torture, by life, by fatigue, by disease'

The word slomannyi often collocates with concrete nouns in the meaning of 'object' or 'body part' in the form of the nominative case. The word slomlennyi collocates with the word chelovek or abstract nouns dukh, gore, zhizn', pytki, ustalost', bolezn' ('spirit', 'grief', 'life', 'torture', 'fatigue', 'illness') in the form of the ablative case (in the meaning of cause, method, means or stimulus).

Lexical and syntactic variability is also observed for many forms of verb pair paradigms, for example:

Table 2. Co-occurrence of the verbs *blistaet* and *bleshchet*

blistaet 'shines₁'	*krasotoi* 'by beauty'
bleshchet 'shines₂'	*noviznoi, umom, original'nost'iu* 'by novelty, by intelligence, by originality'

The participles *blistaiushchii* ('shining') in combination with the words *mir, svet, mech, zolotom, ogniami, chistotoi* ('world', 'light', 'sword', 'gold', 'lights', 'clean') and *bleshchushchii* ('glittering, shining') *sneg, almaz, zdorov'em, umom, ostroumiem* ('snow', 'diamond', 'health', 'intelligence', 'wit') differ in the same way.

Table 3. Co-occurrence of the verbs *blistaiushchii* and *bleshchushchii*

blistaiushchii 'shininig₁'	*zolotom, mech, mir, ogniami, chistotoi* 'gold, sword, peace, lights, purity'
bleshchushchii 'shininig ₂'	*zdorov'em, sneg, ostroumiem, almaz* 'with health, snow, with wit, diamond'

The participle *dvizhushchii* (guiding) acts as an agreed definition in combination with abstract nouns such as *faktor, motiv, printsip, moment, impul's* ('factor', 'motive', 'principle', 'moment', 'impulse'), as well as *nerv* ('nerve') and *mekhanizm* ('mechanism'), possible used in figurative meaning.

Table 4. Co-occurrence of the verbs *dvigaiushchii* and *dvizhushchii*

dvigaiushchii 'moving₁'	*vpered, nogami, dushi* 'forward, on feet, souls'
dvizhushchii 'moving₂'	*faktor, motiv, mekhanizm, printsip, moment, impul's, nerv, stimul, napor* 'factor, motive, mechanism, principle, moment, impulse, neur, stimulus, pressure'

In this case, the difference in the contexts does not arise any questions as it is known that the language tends to linguistic economy and two words which forms are different but the meanings are the same can rarely be used in the texts of the same style for a long time because one word often displaces the other. Though sometimes the process of phraseologisation can take place and historical and archaic words and grammatical phenomena can be saved and used in phraseological units due to linguistic memory (*zhit' pripevaiuchi* – 'live happily ever after', *sidet' slozha ruki* – 'sit back', *pritcha vo iazytsekh* – 'a byword'). However, it should be noted that words with an excess paradigm exist in the language due to various distribution. This reflects the cognitive mechanisms of the language functioning.

Besides, the study of distribution of some words (*muchaet/muchit* – 'tortures / tortures', *lazaet/lazit* – 'climb/climbs') showed that the contexts of their use are identical. Therefore, the semantics of these words has no differences. The use of words with the

same meanings but slightly different in form can be explained by difference in style, the difference between bookish and colloquial speech.

After the distributive analysis, the frequency analysis was performed to find how the words under study behave and whether one word displaces another one with time and becomes normative. The Google books data were used to build graphs of frequency change of the finite and infinite forms of verbs with an excessive paradigm.

We obtained 232 graphs described diachronic changes of verb pairs in 3Sg and 3Pl. The verb pairs were classified according to the frequency dynamic of their use.

The following results were obtained.

1. In 50% of cases, the unproductive form dominates the productive one, the norm change is not expected. For example, *kolyshet > kolykhaet* ('to wave'), *zhazhdu > zhazhdaiu* ('to yearn'), *mechus', > metaius'* ('flounce'), *khnych' > khnykai* ('whimper'), *pashut > pakhaiut* ('plough').
2. In 37% of cases, the productive form dominates the unproductive one. For example, *fyrkaesh' > fyrchish'* ('to snort'), *mykajutsja > mychutsja* ('to torment'), *muchish'>muchaesh'* ('to torture'), *mykaet> mychet* ('to wander').
3. In 10% of cases, an unsuccessful attempt to change the norm was observed. For example, the forms of the verbs *dvigat'* ('to move') and *sekat'sia* ('to split').
4. In 13% of cases, there was a long-term competition between the two forms. For example, between the verbs *zametat'sia* ('to sweep'), *klikat'sia* ('to shriek')
5. In almost 3% of cases, only a form with the stem of one type was found. For example, the verb *nianchit'* ('to nurse').
6. Significantly lower number of graphs (8%) showed that productive forms are displaced by unproductive forms (the form *kaplet* 'to trickle' is displaced by the form *kapaet*). For example, the verb *slomat'sia* ('to break').

The productive declination class is much more common in the other forms. In approximately 64% of cases, the declination type of the 3Sg and 3Pl forms (*kudakhchet, kudakhchut* – 'to cackle') does not coincide with the type of declination in other forms (*kudakhtaiu, kudakhtaia, kudakhtaiushchii, kudakhtai*).

Four tendencies were identified during the frequency analysis of competing forms of words:

1. Absence of competition between the verb forms (only one form is found or dominate another one throughout the target period).
2. Both forms have almost the same frequency.
3. Norm change (frequency of the less widespread form is increasing, and frequency of the more widespread form is decreasing (the frequency curves tend to become closer).
4. One form displaces another one (X-shaped chart).

Accurate time of the norm change concerning each pair of verbs was also revealed. It was found that the norm change most often occurs within two twenty-year periods: 1860-1880, 1910-1930 (see the Tables 5 and 6).

Table 5. Average annual relative frequency of the members of the verbal paradigm and the year of prevalence of one variant over another: productive form → unproductive form (of the type *khlestaet → khleshchet – 'to slash'*)

Verbal paradigm	Year	Annual average relative frequency
pleskat'sia	1830	1.600E-08
goniat'sia	1855	5.932E-09
poloskat'sia	1870	1,754E-08
blestet'	1875	1.584E-08
murlykat'	1910	3.202E-08
kudakhtat'	1915	3.202E-08
ryskat'	1915	5.045E-08
khlestat'	1915	1.265E-07
tykat'sia	1920	4.592E-08
tykat'	1920	5.630E-08
sypat'	1940	4.903E-08
zashchipat'	1940	1.238E-07
kurlykat'	1950	9.401E-08
pometat'sia	1950	7.144E-08

Table 6. Average annual relative frequency of the members of the verbal paradigm and the year of prevalence of one variant over another: productive form → unproductive form (of the type *muchit → muchaet – 'to torture'*)

Verbal paradigm	Year	Annual average relative frequency
merit'sia	1840	8.139E-08
meriat'sia	1840	1.522E-07
fyrkat'	1860	1.087E-07
fyrchat'	1860	1.083E-07
mykat'sia	1865	1.088E-07
miaukat'	1915	2.127E-07
pryskat'	1915	3.130E-07
muchit'sia	1920	2.561E-07
muchat'sia	1920	3.031E-07
mykat'	1920	5.203E-07
slomat'sia	1920	1.253E-06
muchit'	1960	1.549E-06
natolkat'	1975	4.125E-06

It was also found that more frequent verbs require more time for changing by this or that type. Verbal paradigms changed in the middle of the 20th century are 4-6 times more frequent than those that changed in the middle of the 19th century (Fig. 1 and 2).
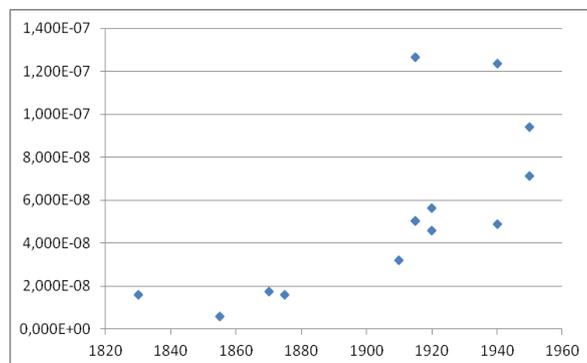
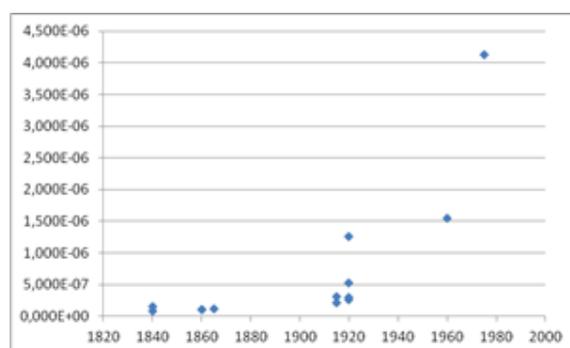**Fig. 1.** Chronology of the norm change (of the type *khlestaet→khleshchet* – 'to slash')

Fig. 2. Chronology of the norm change (of the type *muchit→muchaet* – 'to torture')

The large corpus data allowed us to determine which verb forms are often used in the bookish texts and which forms are less frequent.

As it was expected, the most frequent verbs were verbs in the forms 3Sg and 3Pl. Present passive participle and present active participle were the least frequent.

Table 7. Regularity of the forms

1Sg	*83 % (100 out of 122)*
2Sg	*75 % (92 out of 122)*
3Sg	*100% (122 out of 122)*
1Pl	*75% (91 out of 122)*
2Pl	*64% (78 out of 122)*
3Pl	*90% (110 out of 122)*
Active participle (present)	55 % (67 out of 122)
Active participle (past)	*57 % (70 out of 122)*
Passive participle (present)	*13% (16 out of 122)*
Gerund (Imperfect)	*80% (97 out of 122)*
Imperative Sg.	*75% (92 out of 122)*
Imperative Pl.	*54 % (66 out of 122)*

It should be noted that the most "conservative" form is 2Pl (for example, mashete, ma-khaete – 'to wave').

Table 8. "Conservativeness" of the forms

	I non-productive class (...-et), 'Old' form	I productive class (...a-et), 'New' form
1Sg	*53% (54 out of 101)*	*47% (47 out of 101)*
2Sg	*64% (59 out of 92)*	*36% (33 out of 92)*
3Sg	*58% (71 out of 122)*	*42% (51 out of 122)*
1Pl	*63% (57 out of 91)*	*37% (34 out of 91)*
2Pl	*67% (52 out of 78)*	*33 % (26 out of 78)*
3Pl	*51% (56 out of 110)*	*49% (54 out of 110)*
Active participle (present)	*52% (35 out of 67)*	*48% (32 out of 67)*
Active participle (past)	*11 % (8 out of 70)*	*89% (62 out of 70)*
Passive participle (present)	*12% (2 out of 16)*	*88% (14 out of 16)*
Gerund (Imperfect)	*42% (41 out of 97)*	*58% (56 out of 97)*
Imperative Sg.	*52 % (48 out of 92)*	*48% (44 out of 92)*
Imperative Pl.	*53% (35 out of 66)*	*47% (31 out of 66)*

The "conservativeness" (tendency to save the original form) of the verbs turned out to be directly proportional to their frequency: the more often a verb is used, the more often it saves its original form.

4   Discussion

Differences in historically changing variant forms are often difficult to determine and appropriately characterize. They relate to the sphere of native speakers` communicative habits.

However, modern corpus-based studies serve as a valuable tool for distributive and statistical studies of word usage which can reveal regularities of their use [13, 14]. In this work, the frequency characteristics of the studied excess verbs were first obtained.

Besides, their usage was studied using the distributive semantics approach. These results can be further used to deeply research cognitive processes relating to Russian morphology and inflexion types. Study of such processes can allow one to solve some problems of Russian grammar: mechanisms of development of grammatical semantics, factors that influence appearance of irregular and non-standard inflection models of words referring to different parts of speech in Russian.

The obtained results are of great value for descriptive morphology of the modern Russian language and can be useful for teaching practice of Russian as a foreign language [15].

The methods used to describe frequency dynamics of the variant verb forms can be used in other cases. Moreover, these methods can allow one not only to describe and explain linguistic phenomena, but also to make reasonable quantitative predictions of the development of linguistic forms.

The applied approach and obtained results can be used to compile dictionaries of collocations and cognitive dictionaries.

5   Conclusion

The data on characteristic properties of use of the excessive verbs were first obtained in this work.

The distributive analysis of the studied pairs of words showed that the distribution of the competing forms can be the same or have significant differences which are due to different factors.

The frequency analysis of the competing forms revealed 4 main trends in the frequency behaviour of the competing pairs of verbs. The accurate time of the norm change was identified for each pair of verbs.

The active return of unproductive forms of verbs to the Russian language was revealed which indicates the general stability of the verb system of the Russian language and tendency to unification by productive type. 3Sg. and 3Pl. forms are at the top of the frequency rating of the verb paradigm. They are more resistant to changes and unification.

Thus, the importance of considering the quantitative data of competing verb forms in combination with the dynamics of their frequency is substantiated.

Acknowledgment

References

1. Elizarenkova, T.: O fakul'tativnosti i ee osobennostiakh v drevneindiiskom iazyke [About optionality and its features in the ancient Indian language]. In: Vostochnoe iazykoznanie: fakul'tativnost'. Moscow, Russia (1982).
2. Boduen de Kurtene, I.: Izbrannye trudy po obshchemu iazykoznaniiu [Selected works on general linguistics]. Saint-Petersburg, Russia (1912).
3. Bailey, C-J.: Variation and linguistic theory. Washington, DC, USA (1973).
4. Pintzuk, S.: Variationist approaches to syntactic variation. In Joseph, B., Janda, R. (eds.) The handbook of historical linguistics, 509-528. Blackwell, Malden/Oxford (2003).
5. Hamilton, W., Leskovec, J., Jurafsky D.: Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1489–1501 (2016).

6. Amato, R.: Human collective behavior: language, cooperation and social conventions. Universitat de Barcelona, Spain (2018).
7. Klaussner, C.: Elements of Style Change. University of Dublin, Ireland (2017).
8. Brand, J.: The role of cognitive factors on the development of the vocabulary. Lancaster University, United Kingdom (2018).
9. Nesset, T., Kuznetsova, J.: Stability and Complexity: Russian Suffix Shift over Time. Scando-Slavica, 57(2), 268–289 (2011).
10. Smirnitskii, A.: K voprosu o slove (problema tozhdestva slova) [To the question of the word (the problem of the word identity).]. Trudy Instituta iazykoznaniia AN SSSR, 4, 3–9 (1954).
11. Vinogradov, V.: Russkii iazyk. Grammaticheskoe uchenie o slove [The Russian Language. Grammatical studies of the word]. Moscow, Russia, (1972).
12. Graudina, L.: Grammaticheskaja pravil'nost' russkoj rechi. Stilisticheskij slovar' variantov [Grammatical correctness of Russian speech. Stylistic dictionary of variants]. Astrel', Moscow (2004).
13. Zakharov, V., Azarova, I.: Semantic Structure of Russian Prepositional Constructions. Lecture Notes in Computer Science, 11697, 224–235 (2019).
14. Mukhamedshin, D., Suleymanov, D., Nevzorova, O.: Choosing the Right Storage Solution for the Corpus Management System. Smart Innovation, Systems and Technologies, 146, 105–114 (2020).
15. Gracheva, O., Kopylova, P.: Konsonantnye cheredovaniia pri slovoizmenenii sovremennykh russkikh glagolov [Consonant alternations in the inflection of modern Russian verbs]. Vestnik Volgogradskogo gosudarstvennogo universiteta, Seriia 2: Iazykoznanie, 5(29), 153–158 (2015).