

A FREQUENCY DICTIONARY OF RUSSIAN VERBS BASED ON THE SYNTAGRUS TREEBANK

Gudzina Viktoriya Anatolyevna

PhD, Associate Professor, Department of Russian Language and
Literature, Bukhara State University, Uzbekistan, Bukhara

Abstract

The article presents a frequency dictionary of Russian verbs compiled from the SynTagRus treebank of the Russian National Corpus (1,234,915 words, 733 documents) and describes the quantitative structure of the verbal lexicon. After the exclusion of 27 predicative words that the treebank tags as verbs, the dictionary contains 9,482 verb lemmas with 171,651 occurrences, which make up 13.90% of all words. For every lemma it records frequency, frequency per million, range, the dispersion coefficient DP and aspect. The distribution follows Zipf's law with a slope of -0.83 for the first thousand ranks. 251 lemmas cover half of all verb occurrences, 1,290 cover 80% and 2,650 cover 90%, while 30.5% of lemmas occur only once. Imperfective verbs account for 53.8% of tokens but only 48.4% of lemmas. Dispersion analysis shows that only 19 verbs are evenly distributed across the corpus. The results agree with an independent multi-source frequency list ($p = 0.816$ for the top 200 verbs) and can be used in learner's lexicography and in teaching Russian as a foreign language.

Keywords: Frequency dictionary, Russian verbs, SynTagRus, corpus lexicography, dispersion, aspect, Zipf's law, vocabulary coverage.

Introduction

Frequency dictionaries are one of the oldest products of quantitative linguistics. For Russian, the dictionary edited by L.N. Zazorina, built on a corpus of about one million words, served for decades as the main source of frequency data [4]. The frequency dictionary of O.N. Lyashevskaya and S.A. Sharov, based on the Russian National Corpus, replaced it with a much larger and more balanced data set [3], and S. Sharoff and his colleagues prepared a frequency dictionary of Russian core vocabulary for learners [10]. These works list all parts of speech

together. The verb, the central element of the Russian sentence and the most complex part of speech in terms of aspect and form, deserves a separate frequency description.

Two methodological developments make such a description more informative than a simple list. The first is the measurement of dispersion. S.Th. Gries proposed the DP coefficient, which shows how evenly a word is distributed across the parts of a corpus and allows frequent but thematically bound words to be separated from words of general use [6]. The second is the availability of corpora with manual morphological and syntactic annotation, such as SynTagRus, the deeply annotated part of the Russian National Corpus [1], distributed in the Universal Dependencies format [9]. In such a corpus lemmas, aspect and verb forms are disambiguated by hand rather than guessed by a tagger.

The aim of the study is to compile a frequency dictionary of Russian verbs on the basis of SynTagRus and to describe the quantitative structure of the verbal lexicon.

MATERIALS AND METHODS

The material is the current Universal Dependencies version of SynTagRus as distributed in May 2026. After the exclusion of punctuation it contains 1,234,915 words in 733 documents of fiction, popular science, journalism and online news. All tokens tagged as VERB were extracted, which in the UD scheme includes finite forms, infinitives, participles and converbs. The auxiliary uses of *быть* are tagged as AUX (11,695 tokens) and were not included. SynTagRus also tags predicative words as VERB. These 27 lemmas, among them *можно* (1,810 tokens), *нет* (1,060), *надо* (958) and *нельзя* (408), were excluded from the dictionary, since they are not verbs in the Russian grammatical tradition.

For every lemma the dictionary records absolute frequency, frequency per million words (ipm), range, that is the number of documents in which the lemma occurs, the dispersion coefficient DP calculated over the 733 documents [6], and the aspect according to the corpus annotation [2]. DP ranges from 0 for perfectly even distribution to 1 for concentration in a single document. The rank-frequency distribution was compared with Zipf's law [12] by linear regression in logarithmic coordinates. Coverage was calculated as the cumulative share of verb tokens covered by the most frequent lemmas [8]. To test reliability, the frequencies of the 200 most frequent verbs were compared with the wordfreq database [11], whose

word forms were lemmatised with the pymorphy3 analyser, a continuation of the pymorphy2 system [7].

RESULTS

3.1 General parameters

Verbs account for 13.90% of all words of the corpus. The lexicon is highly uneven. The single most frequent verb, *мочь*, covers 2.36% of all verb tokens, the ten most frequent verbs cover 11.39%, the hundred most frequent 34.47% and the thousand most frequent 75.79%. At the other end, 2,896 lemmas (30.5%) occur only once and 1,368 occur twice.

Figure 1 shows the rank-frequency distribution. Between ranks 30 and 1,000 the data follow a straight line in logarithmic coordinates with a slope of -0.83 ($R^2 = 0.978$). The head of the distribution lies below the line, since the most frequent verbs are less dominant than Zipf's law predicts, and the tail falls more steeply because of the limited size of the corpus.

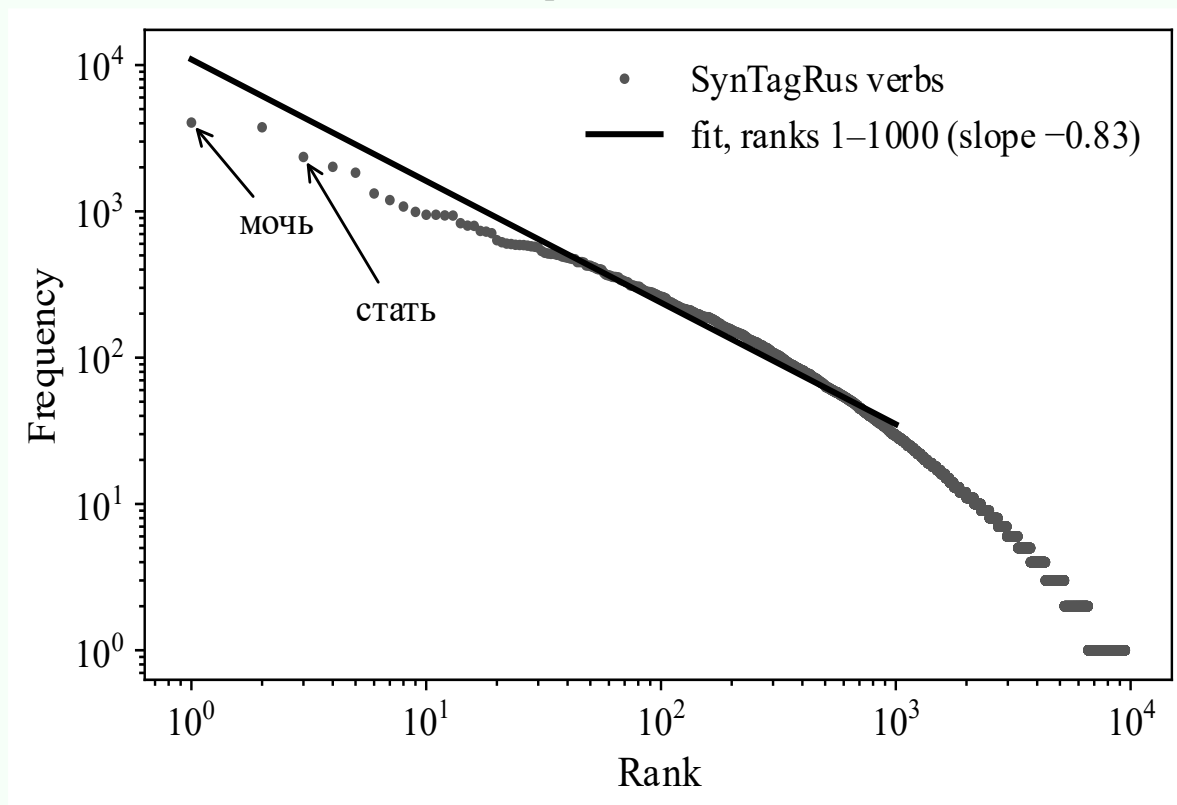


Figure 1. Rank-frequency distribution of Russian verbs in SynTagRus (logarithmic scales).

The coverage curve in Figure 2 has direct practical meaning. A learner who knows the 251 most frequent verbs understands half of all verb occurrences, 1,290 verbs give 80% and 2,650 verbs give 90%. Beyond this point each new percentage point of coverage requires hundreds of additional lemmas.

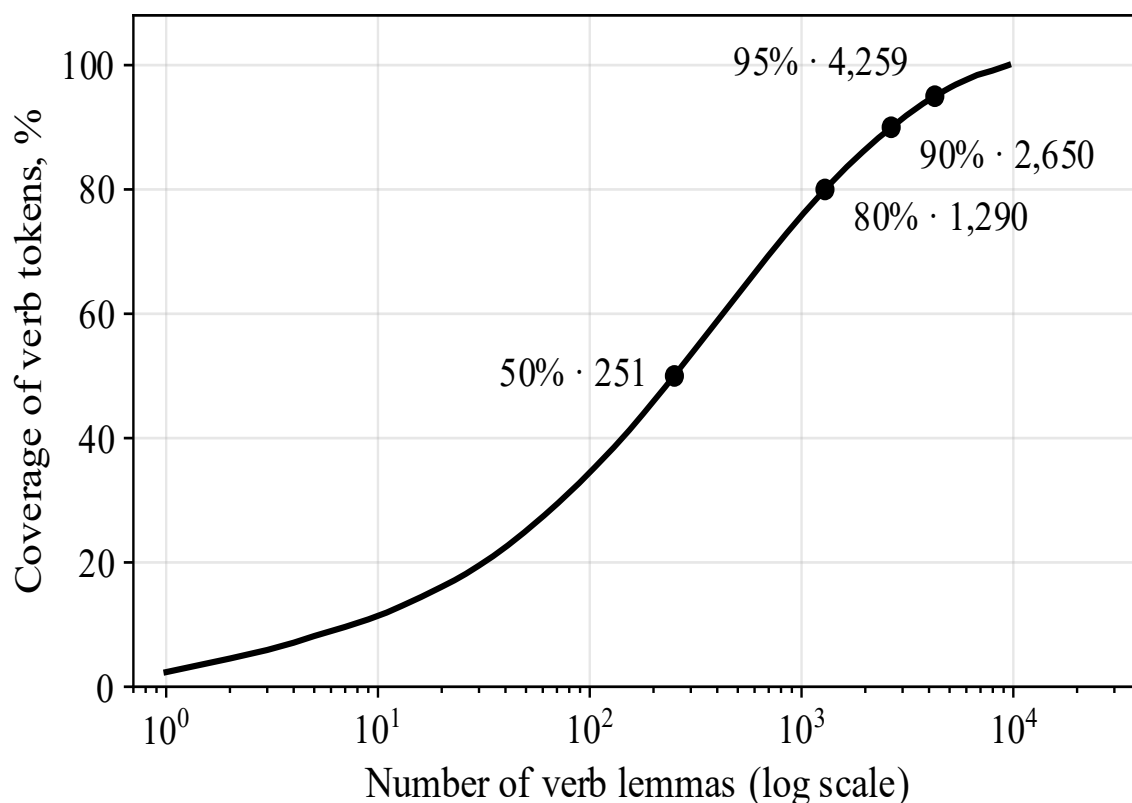


Figure 2. Cumulative coverage of verb tokens by the most frequent lemmas.

3.2 The most frequent verbs

Table 1 lists the thirty most frequent verbs. The list is headed by verbs of modality and existence (мочь, быть, стать, иметь, хотеть), verbs of speech and thought (говорить, сказать, знать, считать, думать), and verbs of action and motion (делать, сделать, идти, работать). Several verbs typical of written and academic style appear in the top thirty, namely являться, использовать, существовать and находиться, which reflects the share of popular science and journalism in the corpus.

Table 1. The thirty most frequent Russian verbs in SynTagRus

Rank	Verb	Aspect	f	ipm	Range	DP
1	мочь	ipf	4,045	3,275.5	662	0.242
2	быть	—	3,756	3,041.5	596	0.301
3	стать	pf	2,354	1,906.2	570	0.291
4	говорить	ipf	2,016	1,632.5	535	0.358
5	сказать	pf	1,838	1,488.4	428	0.456
6	иметь	ipf	1,325	1,072.9	475	0.365
7	знать	ipf	1,194	966.9	350	0.452
8	идти	ipf	1,079	873.7	427	0.374
9	делать	ipf	992	803.3	396	0.412
10	считать	ipf	948	767.7	422	0.445
11	сделать	pf	948	767.7	418	0.417
12	получить	pf	938	759.6	414	0.456
13	работать	ipf	936	757.9	395	0.465
14	оказаться	pf	830	672.1	369	0.435
15	являться	ipf	799	647.0	349	0.536
16	хотеть	ipf	796	644.6	301	0.501
17	жить	ipf	735	595.2	262	0.520
18	видеть	ipf	727	588.7	283	0.496
19	использовать	ipf	709	574.1	285	0.588
20	начать	pf	635	514.2	327	0.475
21	давать	ipf	615	498.0	298	0.513
22	называть	ipf	602	487.5	326	0.480
23	казаться	ipf	598	484.2	264	0.509
24	находиться	ipf	591	478.6	303	0.533
25	стоять	ipf	589	477.0	245	0.538
26	понимать	ipf	588	476.1	239	0.543
27	существовать	ipf	583	472.1	279	0.562
28	думать	ipf	577	467.2	235	0.556
29	принять	pf	573	464.0	292	0.584
30	дать	pf	565	457.5	320	0.455

3.3 Dispersion

Figure 3 relates frequency to dispersion. DP decreases steadily with frequency, but the spread is wide. Among the 656 verbs with a frequency of at least 50, the median DP is 0.840, and only 19 verbs have DP below 0.5, that is, are distributed relatively evenly across the documents. The most even are мочь (0.242), стать (0.291), быть (0.301) and говорить (0.358). At the same time some verbs with more than 100 occurrences are concentrated in a few texts, such as подписать (DP 0.887), влиять (0.868), применять (0.854) and спасти (0.854). Their frequency depends on the topics of particular news and popular science texts rather than on general usage.

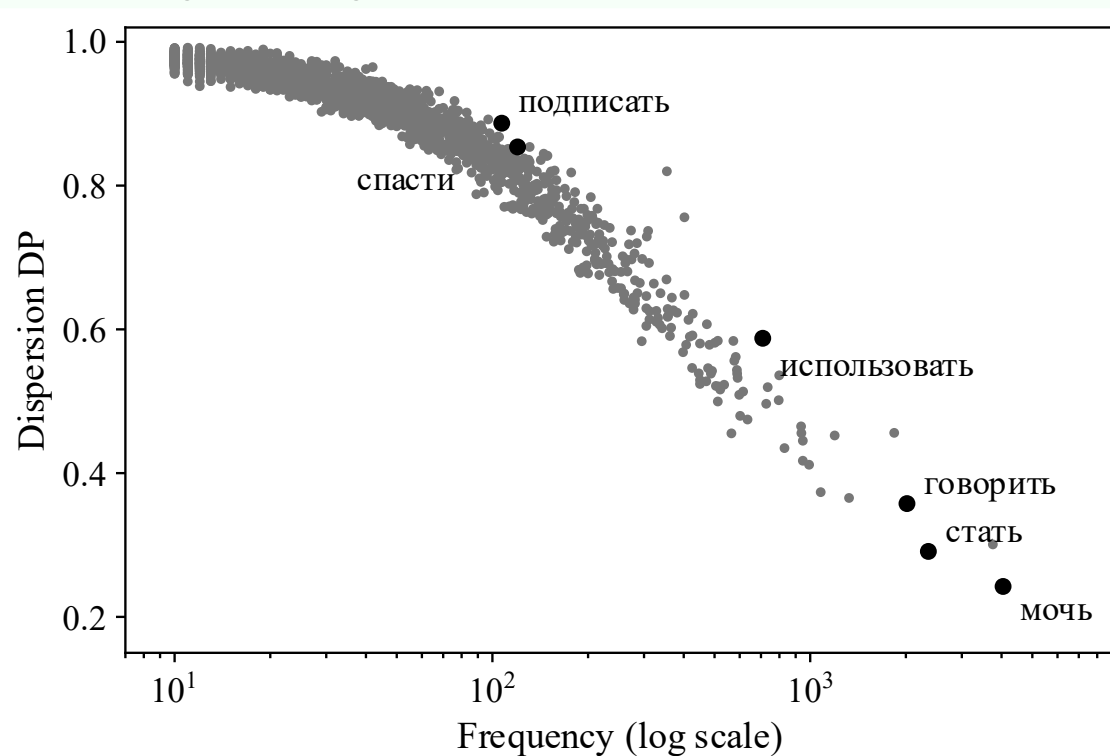


Figure 3. Frequency and dispersion (DP) of verbs with at least 10 occurrences.

3.4 Aspect and verb forms

Imperfective verbs account for 53.8% of verb tokens and perfective verbs for 44.0%, the rest being быть, which the corpus does not mark for aspect. Among lemmas the proportion is reversed, 51.6% perfective and 48.4% imperfective. Perfective verbs are therefore more numerous but less frequent on average, which corresponds to the derivation of many perfectives by prefixation from a smaller

set of imperfective bases [2]. Among the hundred most frequent verbs, 57 are imperfective and 42 perfective.

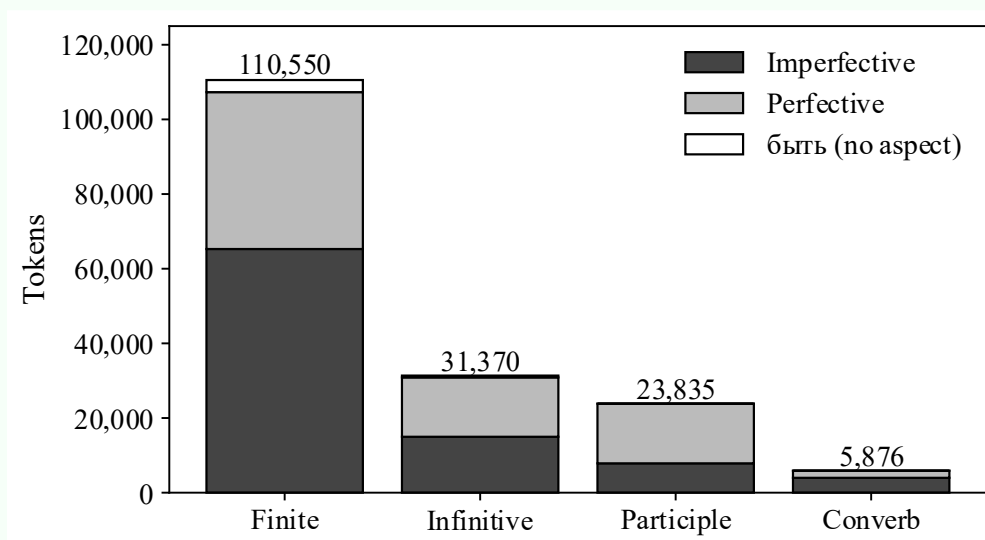


Figure 4. Verb forms in SynTagRus by aspect.

Figure 4 shows the distribution of verb forms. Finite forms make up 64.4% of verb tokens, infinitives 18.3%, participles 13.9% and converbs 3.4%. The aspect profile differs sharply between forms. Finite forms are mostly imperfective (60.8% of aspect-marked forms), infinitives are almost balanced with a slight perfective majority (51.5%), participles are predominantly perfective (67.0%), and converbs predominantly imperfective (67.8%). The perfective preference of participles is explained by the frequency of past passive participles such as *сделанный* or *принятый* in written texts.

3.5 Reliability

The frequencies of the 200 most frequent SynTagRus verbs correlate with the lemma frequencies derived from the wordfreq database ($\rho = 0.816$, $p < 0.001$). For the top 100 the coefficient is 0.778. The agreement is strong, given that the two sources differ greatly in size and composition.

DISCUSSION

The dictionary confirms that the Russian verbal lexicon is organised around a small nucleus. About 250 verbs account for half of all verb usage, and these verbs are also the most evenly dispersed. For learners this means that a well-chosen list

of several hundred verbs gives access to most verb occurrences in texts, in line with the coverage approach of I.S.P. Nation [8]. The coverage figures obtained here are of the same order as the core vocabulary proposed by S. Sharoff and colleagues for all parts of speech [10].

The dispersion data show that frequency alone is not a sufficient criterion for selecting vocabulary. Verbs such as подписать or применять reach high ranks because of the topics of particular documents. A ranking by adjusted frequency, that is frequency multiplied by $1 - DP$, places verbs of general use first and is more suitable for learner's dictionaries. In SynTagRus the first fifteen verbs by adjusted frequency coincide with the first fifteen by raw frequency, except that хотеть replaces являться, which shows that the very core of the lexicon is stable under both criteria.

The predominance of perfective participles and imperfective converbs reflects the written registers of SynTagRus and should be taken into account in teaching. The treatment of predicatives as verbs in the UD annotation shows that corpus-based lexicography must check annotation categories against the grammatical tradition. The study has limitations. SynTagRus is smaller than the main corpus of the Russian National Corpus and contains written texts only, so the frequencies of rare verbs are unstable.

CONCLUSION

The study leads to four conclusions. First, the frequency dictionary of Russian verbs compiled from SynTagRus contains 9,482 lemmas with 171,651 occurrences, which make up 13.90% of the words of the corpus. Second, the distribution follows Zipf's law with a slope of -0.83 , and 251 verbs cover half of all verb usage, while 2,650 verbs cover 90%. Third, only 19 frequent verbs are evenly dispersed, so dispersion must complement frequency in the selection of vocabulary. Fourth, imperfective verbs predominate in tokens and perfective verbs in lemmas, participles are mostly perfective and converbs mostly imperfective. The dictionary can be used in learner's lexicography, in teaching Russian as a foreign language, including to Uzbek students, and as a reference list for further corpus studies of the Russian verb.

REFERENCES

1. Дяченко П.В., Иомдин Л.Л., Лазурский А.В., Митюшин Л.Г., Подлесская О.Ю., Сизов В.Г., Фролова Т.И., Цинман Л.Л. Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус) // Труды Института русского языка им. В.В. Виноградова. 2015. Вып. 6. С. 272-299.
2. Зализняк Анна А., Шмелёв А.Д. Введение в русскую аспектологию. М.: Языки русской культуры, 2000.
3. Ляшевская О.Н., Шаров С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.
4. Частотный словарь русского языка / под ред. Л.Н. Засориной. М.: Русский язык, 1977.
5. Baayen R.H. Word Frequency Distributions. Dordrecht: Kluwer Academic Publishers, 2001.
6. Gries S.Th. Dispersions and adjusted frequencies in corpora // International Journal of Corpus Linguistics. 2008. Vol. 13, No. 4. P. 403-437.
7. Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages // Analysis of Images, Social Networks and Texts. AIST 2015. Communications in Computer and Information Science. Vol. 542. Cham: Springer, 2015. P. 320-332.
8. Nation I.S.P. Learning Vocabulary in Another Language. Cambridge: Cambridge University Press, 2001.
9. Nivre J., de Marneffe M.-C., Ginter F., Hajič J., Manning C.D., Pyysalo S., Schuster S., Tyers F., Zeman D. Universal Dependencies v2: An evergrowing multilingual treebank collection // Proceedings of the 12th Language Resources and Evaluation Conference. Marseille: European Language Resources Association, 2020. P. 4034-4043.
10. Sharoff S., Umanskaya E., Wilson J. A Frequency Dictionary of Russian. Core Vocabulary for Learners. London; New York: Routledge, 2013.
11. Speer R. rspeer/wordfreq: v3.0 (v3.0.2) [Software]. Zenodo, 2022. DOI: 10.5281/zenodo.7199437.
12. Zipf G.K. Human Behavior and the Principle of Least Effort. Cambridge, MA: Addison-Wesley, 1949.