Verbs of Motion in Russian: A Corpus Study of Quantitative Features

Maria Khokhlova

St Petersburg State University
Universitetskaya emb. 7-9-11,
199034 St Petersburg, Russia
m.khokhlova@spbu.ru

Abstract. The distribution of lexical units in texts is characterized by skewness and asymmetry due to the fact that they occur with different frequencies, so traditional descriptive statistics may not always be indicative when presenting vocabulary. The paper provides a brief overview of metrics that are used to describe high- and low-frequency units in different areas, with special attention paid to the rank average. The lexical and semantic group of verbs of motion, which is well described in the literature, was chosen as the material for the study; however, the question of ordering units within it remains open. The author analyzed the frequencies of about 900 lexical units from the dictionary "Lexical-semantic groups of Russian verbs" according to the Russian National Corpus, which were then presented in the form of a frequency dictionary. Using the rank average, the core and peripheral parts of the list were identified. The core vocabulary makes up about 12.74% of the total number of verbs described in the dictionary, while they account for about 70% of all word usages. The meanings of high-frequency verbs imply the absence of direction of movement or, in contrast, the presence of a final destination point. The results of the study may be in demand in language teaching, for example, in compiling lexical minimums.

Keywords: Verbs of motion; Russian language; Core; Periphery; Rank distribution; Frequency dictionary; Rank average

1 Introduction

The term "field" is regularly used to describe phenomena that share certain similarities. In relation to vocabulary, researchers distinguish lexical-semantic fields (or, more narrowly, groups) that include units connected by paradigmatic relations (such as synonymy or hyponymy) [1], as well as by thematic relatedness [2]. At the same time, it is very important to divide the set of lexical items into those that differ by their frequency characteristics and reproducibility, and those that, although numerous, are characterized by low occurrence in speech.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2025, pp. 145–152, 2025.

2 Related Work

Vocabulary can be supplied with quantitative characteristics, one of which is the number of occurrences of a given unit in a text (or corpus). This makes it possible to organize the lexical data in the form of a frequency dictionary. The distribution of vocabulary in such a dictionary is asymmetric and follows Zipf's empirical law, which relates the frequency of a word to its rank (quoted from [3]):

$$k_r = \frac{k_{max}}{r^{\gamma}} \tag{1}$$

where r is the rank of a word, k_r is the frequency of the word with rank r, k_{max} is the frequency of the most frequent word, and γ is a coefficient that characterizes the unevenness of the frequency distribution. The values of k_{max} and γ vary greatly depending on the language and genre.

The frequency of the n-th word in a text is approximately inversely proportional to its rank: it occurs n times less often than the most frequent word (the first item in the ranked list). For research purposes, therefore, it is important to distinguish between high- and low-frequency segments, which would allow one to classify the vocabulary accordingly.

Since a frequency dictionary can be represented as a rank distribution, this allows the use of appropriate statistical metrics for quantitative data. Some of these metrics are applied to organize linguistic data, while others originate from different domains. For example, the Pareto index [4] is related to the distribution of income among populations, for which the Gini index [5] was also proposed. Bradford's law [6] describes the distribution of journals by the number of articles, dividing them into several zones (a core and subsequent zones) according to the ratio 1 : n : n^2, where $n > 1$. The widely known Hirsch index, introduced by [7], serves to measure researchers' productivity based on the number of citations of their publications. The representation of linguistic data is discussed in [8], which provides an overview of concentration indices for frequency dictionaries.

As a metric applied for the division of a frequency dictionary into a head and a tail, [9] introduces the concept of a rank average, calculated using the following formula:

$$\bar{r} = \sum \frac{r * f_r}{N} \tag{2}$$

where r is the rank of a lexeme in the frequency list, f_r is the frequency corresponding to that rank, and N is the sample size.

The rank average can be regarded as an indicator of frequency concentration, representing the upper zone of the frequency distribution [3, p. 118]. This makes it possible to use the obtained value as a boundary dividing the rank distribution into a head and a tail, or, in the case of lexical units, into a core and a periphery.

3 Research Material

For the purposes of this study, we selected a group of verbs of motion, since the lexical units it includes are characterized by concrete meanings and considerable variety, while the question of their exact number remains open.

In the Academic Grammar of the Russian Language, 18 pairs of motion verbs are listed, differing by the aspectual opposition of single vs. repeated action and unidirectional vs. multidirectional motion [10]: *bezhat'* — *begat'* 'to run', *bresti* — *brodit'* 'to trudge, to wander', *vezti* — *vozit'* 'to carry', *vesti* — *vodit'* 'to lead, to drive', *gnat'* — *gonjat'* 'to drive', *gnat'sja* — *gonjat'sja* 'to chase', *ehat'* — *ezdit'* 'to go, to drive', *idti* — *hodit'* 'to go, to walk', *katit'* — *katat'* 'to roll, to wheel', *katit'sja* — *katat'sja* 'to roll, to wheel', *lezt'* — *lazit'* 'to climb', *letet'* — *letat'* 'to fly', *nesti* — *nosit'* 'to carry', *nestis'* — *nosit'sja* 'to rush', *plyt'* — *plavat'* 'to swim', *polzti* — *polzat'* 'to crawl', *tashhit'* — *taskat'* 'to carry, to drag', *tashhit'sja* — *taskat'sja* 'to drag, to tail'.

Verbs of motion are also represented in other sources — notably in reference dictionaries, which serve both descriptive and didactic purposes. The most comprehensive treatment is found in the ideographic dictionary [11], which reflects a semantic classification of Russian verbal vocabulary comprising about 2,500 units. At the highest level of the hierarchy there are three semantic fields — actions, states, and relations — which are further divided into subfields. The minimal unit of description is a lexical-semantic group (LSG), whose verbs share a common archiseme and exhibit typical collocability. In total, the dictionary describes 108 LSGs. The verbs are presented in their basic meanings [ibid., p. 5], i.e., the primary senses listed in explanatory dictionaries. For each LSG, the authors identify basic lexemes using a stepwise identification procedure, which gradually reduces the meaning of the verb under analysis to a more general one [ibid.]. In the case of aspectual pairs, only the form that has a direct definition in the Dictionary of the Russian Language [12] is included.

In the aforementioned dictionary [11], verbs of motion are described in the subfield of movement, which includes several groups [1]:

1. verbs of spatial movement (*peredvigat'*, *peredvigat'sja*, *peremeshhat'*, *peremeshhat'sja* 'to move');
2. verbs of vertical movement (*opuskat'*, *opuskat'sja* 'to lower, to descend', *podnimat'*, *podnimat'sja* 'to raise, to rise', *ronjat'* 'to drop');
3. verbs of movement oriented towards the destination point (*dostavit'*, *dostignut'* 'to deliver, to reach', *priblizit'sja* 'to approach', *pridvinut'* 'to move');
4. verbs of movement oriented away from the starting point (*otdalit'*, *otdalit'sja* 'to move away', *udalit'* 'to remove', *udalit'sja* 'to move away');
5. verbs of movement oriented relative to an intermediate point (*minovat'* 'to pass', *obvesti* 'to walk around', *obhodit'* 'to go around', *perevesti* 'to take across', *perehodit'* 'to cross', *provesti* 'to guide', *prohodit'* 'to pass').

[1] The basic verbs are given hereafter in brackets according to [11] for different groups, the proposed numbering of groups will be further used in the article.

Some verbs are listed in multiple groups — for example, pereletet' / pereletat', which belong to both the first and fifth groups. In total, this subfield includes 895 unique verbs, with aspectual pairs presented together (e.g., *skol'zit' / skol'znut'* 'to slide', *razjehat'sja / razjezzhat'sja* 'to slide apart', *uplyt' / uplyvat'* 'to float away, to swim away'), whereas the pairs differing by iterativity or directionality (e.g., *ehat' / ezdit'* 'to go, to drive', *idti / hodit'* 'to go, to walk', *nesti / nosit'* 'to carry') are listed separately. In the subsequent analysis, frequency counts were calculated for all occurrences of both perfective and imperfective forms of each verb pair.

4 Research Objectives and Methodology

The list provided in the Lexical-Semantic Groups dictionary [11] is exhaustive; however, it requires further structuring and systematization. In this study, we applied the rank average metric described above to divide the list of lexemes into core and peripheral parts. The frequencies of the verbs were obtained from the Russian National Corpus (RNC) [13], expressed in instances per million words (ipm). The subcorpus analyzed was manually disambiguated for grammatical ambiguity, and participles were excluded from the search results.

5 Results

First, it should be noted that about 4% of lexemes do not have any occurrences in the corpus. These include the following verbs: *vzmetyvat'sja* 'to soar away', *vyskol'zat'* 'to slip', *dobrasyvat'*, *dobrosit'* 'to toss', *dokidyvat'* 'to throw', *doskakivat'* 'to skip ahead', *immigrirovat'* 'to immigrate', *obvalivat'* 'to collapse', *obvezti* 'to carry', *obkatit'* 'to roll', *obskakivat'* 'to skip', *obsýpat'sja*, *obsypát'sja* 'to crumble away', *otgresti* 'to row away', *otkatyvat'* 'to roll back', *otkochevat'* 'to migrate away', *otsazhivat'sja* 'to settle away; to be removed', *otsest'* 'to move away; to step aside', *perevolakivat'* 'to drag across', *perekachivat'* 'to pump over; to transfer (liquid)', *perelech'* 'to fly over', *pereplevyvat'* 'to swim across', *peretjanut'sja* 'to stretch over', *podstupat'sja* 'to approach, to advance on', *podchalivat'* 'to tow up (to a shore or ship)', *podchalit'* 'to tow up, to warp (a ship)', *prigresti* 'to row toward, to shovel toward', *prizemljat'* 'to bring down, to land (an aircraft)', *pripolzat'* 'to crawl up, to creep toward', *priskakivat'* 'to gallop up; to run up (on horseback)', *promel'kivat'* 'to flash by, to flicker past', *propalzyvat'* 'to crawl through, to slip past', *prosypat'* 'to spill, to scatter through', *razveivat'* 'to scatter, to disperse', *strjasat'* 'to shake, to rattle', *strjasti* 'to shake off, to jolt', *uparhivat'* 'to soar away, to fly off'. Among these examples, most verbs imply the presence of both a subject and an object, and many contain the semantic component of vertical movement. All verbs with zero frequency were excluded from further analysis.

Table 1 presents the descriptive statistical values for verbs in each subgroup of LSGs corresponding to subject, object, and subject-object movement. The highest metric values were observed for the subgroup of subject movement oriented toward the endpoint, which, overall, includes the most frequent lexical units. However, it should be noted that the values show considerable dispersion, and therefore the descriptive statistics in question are not sufficiently representative.

Table 1: Mean values and standard deviations for five groups of verbs

	Group 1 (movement in space)		**Group 2** (vertical movement)		**Group 3** (movement relative to the destination point)		**Group 4** (movement relative to the starting point)		**Group 5** (movement relative to an intermediate point)	
	subject	object and subject	subject	object and subject	subject	object and subject	subject	object and subject	subject	object and subject
mean (ipm)	45.05	31.16	25.80	28.96	71.38	52.54	42.10	16.45	27.27	35.31
sd (ipm)	131.88	54.21	45.44	51.10	146.30	54.07	121.05	20.21	85.90	69.64

For each LSG listed in the dictionary, the core and peripheral parts were identified separately by calculating the rank average. The higher the value of this metric, the more evenly distributed the frequencies are. Conversely, lower values indicate frequency concentration, i.e. the presence of a zone containing lexical units characterized by high quantitative indicators. In the following sections, the LSGs represented in the dictionary, as well as the overall list of verbs, will be examined.

Table 2 presents the rank average values, the proportion of verbs forming the core, and their frequency for each subgroup of LSGs corresponding to subject, object, and subject–object movement.

The largest share of cumulative frequency (74.74%) is shown by the core of verbs belonging to the first group, associated with movement in space. For this LSG, 10 core lexemes were identified, including the following ones: *bezhat′* 'run', *dvigat′sja / dvinut′sja* 'move', *ezdit′* 'to go', *ehat′* 'to go', *idti* 'to go', *letet′* 'to fly', *sledovat′* 'to follow', *uehat′ / uezzhat′* 'to leave', *ujti / uhodit′* 'to leave', and *hodit′* 'to walk'. They make up 10.3% of all verbs in this LSG, accounting for 74.74% of the total frequency, which represents the maximum value compared with the units of other groups.

There is no correlation between the rank average and the number of units that can be classified as core. The core of the second LSG subgroup (subject movement — vertical movement) has the highest rank average (11.14) and, at the same time, is formed by the largest number of units due to the inclusion of aspectual pairs (24 lexemes): *vskochit′ / vskakivat′* 'to jump up', *lezt′* 'to climb', *podnjat′sja / podnimat′sja* (*podymat′sja*) 'to rise / go up', *spustit′sja / spuskat′sja* 'to

Table 2: Rank average values

	Group 1 (movement in space)		**Group 2** (vertical movement)		**Group 3** (movement relative to the destination point)		**Group 4** (movement relative to the starting point)		**Group 5** (movement relative to an intermediate point)	
	subject	object and subject	subject	object and subject	subject	object and subject	subject	object and subject	subject	object and subject
rank average	10.58	10.21	11.14	6.10	7.80	9.49	6.22	8.29	5.14	2.21
core (%)	10.22	15.15	21.62	22.58	12.17	15.13	10.81	23.29	12.66	15.38
freq. (%)	74.74	66.77	66.64	65.37	70.98	67.27	72.17	63.09	71.54	71.31

descend', and others. This allows us to assume that the frequency distribution of lexical units in this group is more uniform and shows a less distinct separation between the core and the tail part of the distribution. At the same time, there is a strong negative correlation between core size and total frequency (Spearman's correlation coefficient = –0.9, $p << 0.05$): the smaller the core, the greater the number of occurrences in the corpus for the verbs it contains.

The greatest difference in frequency concentration is observed in the fifth LSG subgroup (object and subject–object movement), which involves verbs containing the semantic feature of an intermediate point and is also characterized by small size: its rank average equals 2.21, which allowed the inclusion of only two aspectual pairs (*provesti* / *provodit'* 'to conduct, to lead' and *perevesti* / *perevodit'* 'to take across') in the core.

For the overall frequency list of verbs, the rank average value was 63.69, resulting in a list of 114 units (mostly aspectual pairs). The core vocabulary thus constitutes 12.74% of all verbs described in the dictionary. This set includes verbs such as *idti* 'to go', *hodit'* 'to walk', *prijti* / *prihodit'* 'to come', *vyjti* / *vyhodit'* 'to exit', *projti* / *prohodit'* 'to pass', *letet'* 'to fly', *vstupit'* / *vstupat'* 'to enter', etc. The average frequency of verbs within the core is 193.04 ipm, and together they account for 69.38% of all word occurrences. More than one-third (34.21%) of the lexical units contain the semantic component of a destination point (*vernut'sja* 'to return', *dostignut'* / *dostigat'* 'to reach', *priblizit'sja* / *priblizhat'sja* 'to approach', *prinesti* / *prinosit'* 'to bring', *javit'sja* / *javljat'sja* 'to appear'), while about one-fifth (22.81%) denote movement in space without reference to a specific location (*bezhat'* 'to run', *guljat'* 'to walk', *ezdit'* 'to go', *ehat'* 'to go', *uehat'* / *uezzhat'* 'to leave, to depart').

6 Conclusion

The distribution of verbs belonging to the lexical-semantic group of motion is characterized by a wide frequency range and heterogeneity, which complicates the identification of core and peripheral parts within this lexical group. The rank average metric used in the study allowed one to determine the part of the distribution in which lexemes with high frequencies are concentrated; these constitute over 12% of all verbs represented in the examined dictionary. The meanings of these high-frequency verbs typically imply either the absence of a specific direction of movement or, conversely, the presence of a defined endpoint. The results of this study may be useful for language teaching, particularly in the development of lexical minima and the selection of core vocabulary for educational purposes.

Acknowledgements. The author acknowledges St. Petersburg State University for the research project "Modeling the communicative behavior of Russian citizens in social, spoken, and pragmatic aspects using AI methods" #124032900006-1 (Pure ID 131371614).

References

1. Trier, J.: Der Deutsche Wortschatz im Sinnbezirk des Verstandes. Die Geschichte eines Sprachlichen Feldes. Bd. 1. Heidelberg: Winter (1931)
2. Porzig, W.: Wesenhafte Bedeutungsbeziehungen. In: Beiträge zur Geschichte der Deutschen Sprache und Literatur, 58. pp. 70–97 (1934)
3. Martynenko, G. Ia.: Metody matematicheskoi lingvistiki v stilisticheskikh issledovaniiakh [Methods of mathematical linguistics applied to research in stylistics]. St. Petersburg: Nestor-Istoriia. 296 p. (2019)
4. Pareto, W.: Cours d'economie politique. Vol. 1–2. Lausanne (1896–1897)
5. Gini, C.: Srednie velichiny [The averages]. Moscow: Statistika. 448 p. (1970)
6. Bradford, S. C.: Sources of Information on Specific Subjects. In: Engineering: An Illustrated Weekly Journal, London. 26 January 1934, pp. 85–86 (1934)
7. Hirsch, J. E.: An index to quantify an individual's scientific research output. In: PNAS, 102 (46). pp. 16569–16572 (2005) http://www.pnas.org/content/102/46/16569.full Last accessed 1 Nov 2025
8. Martynenko, G. Ia., Grigor'ev, Iu. D.: Indeksy kontsentratsii chastotnykh slovarei [Concentration indices of frequency dictionaries]. In: Vestn. Novosib. gos. un-ta. Seriia: Lingvistika i mezhkul'turnaia kommunikatsiia. pp. 41–53 (2017)
9. Martynenko, G. Ia.: Osnovy stilemetrii [Foundations of stylometry]. Leningrad: Izd-vo Leningradskogo universiteta. 173 p. (1988)
10. Shvedova, N. Yu. (ed.): Russkaya grammatika [Russian Grammar]. Vol. 1: Fonetika. Fonologiya. Udarenie. Intonatsiya. Slovoobrazovanie. Morfologiya. Moscow: Nauka. 789 p. (1980)
11. Matveeva, T. V. (ed.): Leksiko-semanticheskie gruppy russkikh glagolov [Lexico-semantic groups of Russian verbs]: Ucheb. slovar'-spravochnik [Educational dictionary-reference]. Sverdlovsk: Izdatel'stvo Ural'skogo Universiteta. 153 p. (1988)

12. Academy of Sciences of the USSR, Institute of Russian Language (Evgenyeva, A. P., ed.): Slovar' russkogo yazyka [Dictionary of the Russian language]. 2nd ed., revised and expanded. Moscow: Russkii Yazyk (1981–1984)
13. Russian National Corpus (RNC): `http://ruscorpora.ru`. Last accessed 1 Nov 2025