

**A CORPUS-BASED ANALYSIS OF NUCLEUS AND
PERIPHERY IN THE RUSSIAN LEXICO-SEMANTIC GROUP OF
VERBS OF LOCOMOTION**

Asal Tagaeva Rustamovna

Teacher in the Department of Practical English

Navoi State University

Tog'ayeva@nspi.uz

<https://orcid.org/0009-0003-1150-8408>

Abstract. The article identifies the nucleus and periphery of the Russian lexico-semantic group of verbs of locomotion on corpus grounds. The material is the SynTagRus treebank of the Russian National Corpus (1,234,915 words, 733 documents), with the wordfreq database used for external validation. Thirty verbs sharing the archiseme of self-propelled movement are described by frequency, dispersion, subject combinability, semantic extension to inanimate subjects and derivational activity, and a nucleus index combining these criteria divides the group into four zones. The nucleus contains seven verbs (*идти, ходить, бежать, двигаться, бегать, гулять, нестись*) that account for 91.2% of all occurrences, while *идти* alone covers 56.0%. Frequency, dispersion and combinability correlate strongly (ρ from 0.88 to 0.93), derivational activity only moderately. The corpus nucleus does not coincide with the semantic hyperonym, and derivational families preserve traces of former centrality.

Keywords. linguistic nucleology, nucleus, periphery, lexico-semantic group, verbs of motion, corpus linguistics, SynTagRus, dispersion.

INTRODUCTION

Field theory describes vocabulary as a set of groupings whose members differ in the degree of their membership. F. Daneš regarded the opposition of a compact centre and a diffuse periphery as a universal property of language systems [7, p. 9-21]. G.S. Shchur traced this idea through the main field conceptions of European and Soviet linguistics [6], and Yu.N. Karaulov made it a principle of the ideographic description of Russian vocabulary [4]. Within a lexico-semantic group the nucleus is usually understood as the unit that expresses the archiseme of the group most directly and most neutrally, while peripheral units add differential semes of manner, intensity or evaluation [2, p. 105-113].

In most studies the nucleus is still identified by introspection and by the analysis of dictionary definitions. Such procedures establish semantic generality well, yet they say little about how units actually function in texts. Ju.D. Apresjan showed that members of a synonymic series differ not only in meaning but also in combinability [1]. Corpora provide exactly this kind of evidence.

The frequency dictionary of O.N. Lyashevskaya and S.A. Sharov made the frequency of Russian lexemes measurable on the basis of the Russian National Corpus [5]. The SynTagRus treebank, a fully annotated part of the same corpus, provides morphological and syntactic information for every sentence [3, p. 272-299; 10]. Measures of dispersion such as the DP coefficient proposed by S.Th. Gries distinguish words that are frequent everywhere from words that are frequent in a few texts only [8, p. 403-437].

Verbs of locomotion offer a convenient test case. Russian belongs to the languages that express the path of motion mainly by prefixes and the manner

of motion by verb roots [12, p. 57-149]. It therefore possesses a large set of manner verbs, from neutral *идти* to highly specific *семенить* and *ковылять*.

The aim of the study is to identify the nucleus and periphery of the Russian lexico-semantic group of verbs of locomotion on corpus grounds. Three research questions are addressed. Which verbs form the nucleus when frequency, dispersion, combinability, semantic extension and derivational activity are taken into account? Do these criteria agree with each other? Does the semantically most general verb of the group coincide with its corpus nucleus?

MATERIALS AND METHODS

2.1 Corpus data

The main source is SynTagRus, the deeply annotated subcorpus of the Russian National Corpus developed at the Institute for Information Transmission Problems of the Russian Academy of Sciences, whose morphological and syntactic annotation is corrected manually [3]. The study used the Universal Dependencies version of the treebank [10] as it stood in the public repository in May 2026. After the exclusion of punctuation the corpus contains 1,234,915 words in 733 documents of fiction, popular science, journalism and online news.

External validation relied on the wordfreq database, version 3.1.1, which combines Russian frequencies from Wikipedia, film subtitles, news, books and Twitter [11]. Its word forms were lemmatised with the pymorphy3 analyser, a continuation of the pymorphy2 system of M. Korobov [9, p. 320-332]. The frequency of each form was distributed among its possible lemmas in proportion to the probabilities of the parses, and the results were summed by lemma.

2.2 Inventory of the group

The group includes thirty verbs whose dictionary meanings contain the archiseme «to move in space by one's own force». It covers the pairs of unidirectional and multidirectional verbs *идти* and *ходить*, *бежать* and *бегать*, *ползти* and *ползать*, twelve verbs of manner (*шагать*, *стунать*, *брести*, *плестись*, *тащиться*, *ковылять*, *семенить*, *красться*, *пробираться*, *топать*, *шестьвовать*, *маршировать*), seven verbs of aimless movement (*бродить*, *гулять*, *слоняться*, *шататься*, *прогуливаться*, *прохаживаться*, *расхаживать*), two verbs of speed (*мчаться*, *нестись*) and three hyperonyms (*двигаться*, *передвигаться*, *перемещаться*). Verbs of movement by vehicle such as *ехать* were excluded. Only tokens tagged as VERB were counted, which in the UD scheme includes participles and converbs. Concordances of ambiguous verbs such as *нестись*, *тащиться* and *шататься* were checked manually for homonymy.

2.3 Criteria and the nucleus index

Five criteria were operationalised. Frequency was measured in instances per million words (ipm) in both sources, and the two percentile ranks were averaged. Dispersion was measured by the DP coefficient across all documents, where a value close to 0 indicates even distribution and a value close to 1 indicates concentration in a few texts [8]. Combinability was measured as the number of different lemmas filling the subject position (the *nsubj* relation) of the verb. Semantic extension was measured as the number of different inanimate nominal subjects, since inanimate subjects signal figurative or abstract uses of verbs of locomotion. Derivational activity was measured as the number of different prefixed derivatives attested in the corpus. For *идти* the suppletive stem *-йти* was taken into account, and

formally similar verbs with a different root, such as *обрести* or *заслоняться*, were excluded.

Each criterion was converted into percentile ranks within the group, so that all values lie between 0 and 1. The nucleus index (NI) is the arithmetic mean of the five ranks. The group was then divided into zones by the Jenks natural breaks method. A four-zone model was chosen because the goodness of variance fit grows from 0.897 with three classes to 0.936 with four, while a fifth class adds only 0.025. Agreement between criteria was tested with Spearman's rank correlation. All calculations were made with Python scripts.

RESULTS

3.1 Frequency and dispersion

The thirty verbs occur 1,928 times in SynTagRus, and their distribution is extremely uneven. *Идти* alone accounts for 1,079 occurrences (873.7 ipm), that is 56.0% of the group, and together with *ходить* it covers 69.6%. Two verbs, *ковылять* and *семенить*, do not occur in SynTagRus at all, although wordfreq registers them at 0.4 ipm. Frequencies in the two sources are strongly correlated ($\rho = 0.864$, $p < 0.001$), so the frequency ranking does not depend on the composition of a single corpus.

Dispersion separates the nucleus even more sharply. *Идти* occurs in 427 of 733 documents (58.3%) with $DP = 0.374$, and *ходить* has $DP = 0.692$. All other verbs have DP above 0.80, and most manner verbs exceed 0.95, which means that they appear in a handful of texts. The case of *нестись* is instructive. Ten of its 23 occurrences come from one document of war prose, where bullets, shells and a blizzard rush over the battlefield. Table 1 gives all measurements.

Table 1. Corpus profile of the verbs of locomotion

Verb	f	ipm	ipm wf	Range	DP	Subj.	Inan.	Der.	NI	Zone
<i>идти</i>	1079	873.7	670.4	427	0.374	313	230	24	0.993	Nucleus
<i>ходить</i>	262	212.2	187.1	132	0.692	71	18	27	0.966	Nucleus
<i>бежать</i>	159	128.8	91.9	67	0.808	40	8	16	0.897	Nucleus
<i>двигаться</i>	138	111.7	64.3	79	0.803	53	34	7	0.893	Nucleus
<i>бегать</i>	53	42.9	35.7	39	0.883	27	5	11	0.848	Nucleus
<i>гулять</i>	45	36.4	33.0	31	0.901	16	5	7	0.797	Nucleus
<i>нестись</i>	23	18.6	4.6	13	0.956	10	9	8	0.759	Nucleus
<i>бродить</i>	28	22.7	12.2	24	0.928	19	5	1	0.714	Near
<i>ползти</i>	11	8.9	5.4	7	0.979	8	4	9	0.657	Near
<i>передвигаться</i>	20	16.2	8.9	18	0.952	11	1	0	0.572	Near
<i>стунать</i>	6	4.9	5.7	6	0.968	1	1	18	0.555	Near
<i>перемещаться</i>	13	10.5	9.8	11	0.970	5	3	0	0.541	Near

ilmiy –amaliy anjuman

<i>брести</i>	12	9.7	2.1	10	0.966	6	0	5	0.536	Near
<i>мчатся</i>	10	8.1	4.9	7	0.979	6	0	3	0.486	Far
<i>ползать</i>	8	6.5	4.1	7	0.975	3	0	6	0.460	Far
<i>шагать</i>	6	4.9	6.5	6	0.980	3	1	1	0.445	Far
<i>топать</i>	11	8.9	2.1	9	0.970	4	0	1	0.431	Far
<i>шестьковать</i>	5	4.0	0.6	4	0.991	4	2	2	0.417	Far
<i>шататься</i>	6	4.9	2.5	6	0.971	2	0	1	0.359	Far
<i>тащиться</i>	3	2.4	3.2	2	0.994	2	2	1	0.345	Far
<i>пробираться</i>	5	4.0	3.2	5	0.984	3	1	0	0.343	Far
<i>расхаживать</i>	8	6.5	1.0	8	0.974	4	0	0	0.340	Far
<i>слоняться</i>	5	4.0	0.9	5	0.983	2	0	1	0.284	Marginal
<i>маршировать</i>	2	1.6	1.9	1	0.999	1	1	1	0.257	Marginal
<i>плестись</i>	3	2.4	0.9	3	0.987	2	0	1	0.255	Marginal
<i>красться</i>	1	0.8	1.2	1	0.995	0	0	4	0.224	Marginal

<i>прогуливаться</i>	4	3.2	3.3	4	0.987	1	0	0	0.214	Marginal
<i>прохаживаться</i>	2	1.6	0.2	2	0.991	2	0	0	0.167	Marginal
<i>ковылять</i>	0	0.0	0.4	0	1.000	0	0	2	0.166	Marginal
<i>семенить</i>	0	0.0	0.4	0	1.000	0	0	0	0.079	Marginal

Note. *f* is the absolute frequency in SynTagRus, *ipm* wf the lemma frequency in wordfreq, Range the number of documents containing the verb, Subj. the number of subject lemmas, Inan. the number of inanimate subject lemmas, Der. the number of prefixed derivatives, NI the nucleus index. Zones are named as in Table 3.

3.2 Combinability and semantic extension

Subject combinability follows the same pattern. *Идти* takes 313 different subject lemmas, *ходить* 71, *двигаться* 53 and *бежать* 40, whereas most manner verbs have no more than six. The nuclear verb also has the widest range of figurative uses. Of its 650 nominal subjects 584 are inanimate, and they belong to 230 different lemmas. The most frequent of them is *речь* (204 cases), which forms the construction «речь идёт о». Even without it the share of inanimate subjects remains 85.2%. Other frequent subjects such as *процесс* (23), *разговор* (11) and *борьба* (11) also name processes. *Ходить* shows a smaller but distinct extension, as in «ходят слухи» (10 cases) or with subjects denoting transport. The hyperonym *двигаться* has 34 inanimate subject types, mostly physical bodies such as *электрон* or *планета*, which reflects its general physical meaning rather than figurative use.

3.3 Derivational activity

Derivational activity is the only criterion that does not place *uđmu* first. The largest families belong to *ходить* (27 prefixed derivatives) and *uđmu* (24), followed by *ступать* (18), *бежать* (16) and *бегать* (11). The hyperonyms *передвигаться* and *перемещаться*, themselves prefixed, have no derivatives. The behaviour of *ступать* is remarkable. With only six occurrences it ranks among the peripheral verbs, yet its family includes *выступать*, *поступать*, *наступать*, *отступать*, *уступать* and other frequent verbs whose meanings have moved far from physical stepping. Table 2 shows how far the five criteria agree.

Table 2. Spearman correlations between the criteria

Criterion	Frequency	Dispersion	Combinability	Extension
Dispersion	0.930			
Combinability	0.898	0.879		
Extension	0.774	0.656	0.748	
Derivation	0.567	0.550	0.493	0.551

Frequency, dispersion and combinability form a tight cluster with coefficients from 0.879 to 0.930. Semantic extension correlates with them moderately to strongly (0.656 to 0.774). Derivational activity shows the weakest links (0.493 to 0.567), although all coefficients are significant at $p < 0.01$.

3.4 Nucleus index and zonal structure

The nucleus index ranges from 0.993 for *идти* to 0.079 for *семенить*. Natural breaks divide the group into four zones, summarised in Table 3. Seven verbs form the nucleus and account for 91.2% of all occurrences of the group. The near periphery includes six verbs (4.7%), the far periphery nine verbs (3.2%), and the marginal zone eight verbs that together make up only 0.9% of the occurrences.

Table 3. Zones of the lexico-semantic group

Zone	NI range	Verbs	Share of occurrences
Nucleus	0.759– 0.993	<i>идти, ходить, бежать, двигаться, бегать, гулять, нестись</i>	91.2%
Near periphery	0.536– 0.714	<i>бродить, ползти, передвигаться, стунать, перемещаться, брести</i>	4.7%
Far periphery	0.340– 0.486	<i>мчаться, ползать, шагать, топтать, шестьвовать, шататься, тащиться, пробираться, расхаживать</i>	3.2%
Marginal zone	0.079– 0.284	<i>слоняться, маршировать, плестись, красться, прогуливаться, прохаживаться, ковылять, семенить</i>	0.9%

DISCUSSION

The corpus data give a quantitative form to the model of a compact centre and a diffuse periphery [7]. Less than a quarter of the group, seven verbs out

of thirty, carries more than nine tenths of its occurrences, while the eight marginal verbs are almost invisible in a corpus of more than a million words.

Идти behaves as the dominant of the group in every respect but one. It is first by frequency, dispersion, combinability and semantic extension, and second by derivation. Its nuclear status also shows itself in delexicalised constructions such as «речь идёт о», where the verb has lost the meaning of physical movement altogether.

The answer to the third research question is negative. The semantically most general verbs, *передвигаться* and *перемещаться*, which describe any change of location without restriction of manner, fall into the near periphery. Their bookish character, limited combinability and lack of derivatives outweigh their semantic generality. *Двигаться* reaches the nucleus, but only in fourth position, and its subjects show that it functions largely as a verb of the physical motion of bodies. The corpus nucleus is thus formed not by the most abstract unit but by everyday, stylistically neutral verbs. This qualifies a purely componential definition of the nucleus [2] and agrees with the view of Ju.D. Apresjan that combinability is part of the semantic description of a unit [1].

The zonal structure also reflects the type of information encoded in the verb root. In a language where path is expressed by prefixes [12], the nuclear verbs *идти*, *ходить*, *бежать* and *бегать* specify only the basic mode of motion and serve as bases for numerous path derivatives such as *выйти*, *войти*, *прийти* or *убежать*. Verbs with a specific manner component, among them *плестись*, *семенить*, *красться* and *ковылять*, occupy the marginal zone. The more semes a verb adds to the archiseme, the further it lies from the nucleus.

The weak correlation of derivational activity with the other criteria needs interpretation. The family of *cmynamь* suggests that derivatives can preserve the centrality a verb had at an earlier stage of the language, while the base itself has moved to the periphery. Derivational activity therefore reflects historical depth as much as present-day salience, and future models may give it a lower weight. The borderline position of *нестись* points to another limitation. Its index depends on concentrated use in a few texts and contrasts with its low frequency in wordfreq (4.6 ipm), so its place in the nucleus is probably an effect of corpus composition. A larger corpus, unequal weights of criteria and a finer indicator of figurative use than animacy are tasks for further research.

CONCLUSION

The study leads to four conclusions. First, the nucleus of the Russian lexico-semantic group of verbs of locomotion can be identified on corpus grounds by an index combining frequency, dispersion, combinability, semantic extension and derivational activity. Second, the nucleus includes seven verbs with *uɔmu* as the dominant, and these verbs account for 91.2% of the occurrences of the group, while the marginal zone of eight manner verbs accounts for 0.9%. Third, frequency, dispersion and combinability agree closely, whereas derivational activity partly reflects historical rather than current centrality. Fourth, the semantically most general verbs of the group do not form its corpus nucleus, so semantic generality is a necessary but not a sufficient condition of nuclear status. The procedure is reproducible and applicable to other groups of Russian vocabulary and to teaching Russian as a foreign language.

REFERENCES

1. Апресян Ю.Д. Лексическая семантика. Синонимические средства языка. М.: Наука, 1974. 367 с.
2. Васильев Л.М. Теория семантических полей // Вопросы языкознания. 1971. № 5. С. 105-113.
3. 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.
4. Караулов Ю.Н. Общая и русская идеография. М.: Наука, 1976. 338 с.
5. Ляшевская О.Н., Шаров С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.
6. Щур Г.С. Теории поля в лингвистике. М.: Наука, 1974.
7. Daneš F. The relation of centre and periphery as a language universal // Travaux linguistiques de Prague. Vol. 2. Les problèmes du centre et de la périphérie du système de la langue. Prague: Academia, 1966. P. 9-21.
8. Gries S.Th. Dispersions and adjusted frequencies in corpora // International Journal of Corpus Linguistics. 2008. Vol. 13, No. 4. P. 403-437.
9. 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.

10. 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.
11. Speer R. rspeer/wordfreq: v3.0 (v3.0.2) [Software]. Zenodo, 2022. DOI: 10.5281/zenodo.7199437.
12. Talmy L. Lexicalization patterns: semantic structure in lexical forms // Language Typology and Syntactic Description. Vol. 3. Grammatical Categories and the Lexicon / ed. by T. Shopen. Cambridge: Cambridge University Press, 1985. P. 57-149.