# STATISTICAL ANALYSIS OF RUSSIAN MULTIWORD PREPOSITIONS

**Assoc. Prof. Dr. Victor Zakharov**[1]
**Mgr. Anastasia Golovina**[2]
**Assoc. Prof. Dr. Irina Azarova**[3]
[1, 2, 3] St Petersburg University, Russia

## ABSTRACT

This paper is part of a larger study that aims to create the first quantitative grammar of the Russian prepositional system. The present study deals with Russian secondary multiword prepositions. Prepositions are a heterogeneous class consisting of a small group of about 25 primary prepositions and hundreds of secondary ones, the latter being motivated by content words (nouns, adverbs, verbs), which may be combined with primary prepositions to form multiword prepositions (MWPs). A strict division between secondary multiword prepositions and equivalent free word combinations is not specified. This is a task for a special corpus-based research.

Prepositions are characterized as function words used to express various relationships between main and dependent members of a phrase. The difficulty is that relations expressed by prepositions are multi-sided, grammatical and lexical. Primary prepositions are said to have no real lexical meaning. It is not quite true as regards primary prepositions and even more so for secondary ones. Prepositions express semantic relations between words, and their meanings directly correspond to these relations.

Multiword prepositions perform the grammatical function of a preposition in a certain position of a syntactic structure in some contexts and can be a free combination in others. This paper is devoted to the statistical analysis of the use of multiword prepositions in corpora. The features of multiword prepositions in the function of a preposition are described. Statistical data on the ratio of the use of individual multiword expressions as prepositional units and as free combinations are provided.

***Keywords:*** *Russian language, secondary prepositions, multiword prepositions, corpus statistics*

## INTRODUCTION

Multiword units in Russian, like in many other languages, are an integral part of the lexicon. They are made up of various parts of speech including prepositions. Multiword prepositions (MWPs) make up a large part of secondary prepositions. Their structural models vary, but in most cases they include a simple adposition.

While much research has been dedicated to primary prepositions, the same cannot be said about secondary prepositions, many of which are multiword expressions. We have yet to obtain an exhaustive list of the elements of this set. The reason for that lies partly in the lack of close attention to these units and partly in their complex nature. The same words and word units may perform as prepositions as well as other parts of speech or free combinations (e.g. (in Rus.) *вокруг 'around', накануне 'on the eve of', относительно 'relative to', снаружи 'outside', в форме 'in the form of', исключая 'excluding', что до 'as for')*. As a rule, the distinction between these ambiguous entities is outlined neither in grammars nor in dictionaries.

Perhaps the most well-defined inventory of Russian secondary prepositions can be found in the "Russian Grammar" [1]. The description of the subclass of secondary prepositions is given according to the part of speech they derive from (nominal, adverbial, verbal) and their structural type (simple and complex); a list of secondary prepositions along with their case government is presented after the description of each subtype. No common list of secondary prepositions is provided, however; neither is it implied that the preposition lists are exhaustive. Rather the opposite: it is noted in the "Russian Grammar" that a lot of the units listed are entities of uncertain part-of-speech status due to their preserved ability to include determiners and combine selectively with the other parts of the potential prepositional phrase [1]. As a result, the lists provided in the source cited are to be regarded as mere examples of different subtypes of secondary prepositions. The "Explanatory Dictionary of Functional Parts of Speech of the Russian Language" [2], another work touching on the subject, contains only 199 MWPs. Even fewer - just 157 - are found in the more relevant to the topic of multi-word units "Explanatory Dictionary of Combinations Equivalent to a Word" [3].

## MATERIALS AND METHODS

Based on a number of linguistic sources we have compiled a table of Russian prepositions totalling 740 entries (including variations). Naturally, the degree of "prepositionality" of these entities varies. The contents of this table were used as the base material of our study.

Analysis of our preposition index allows us to discern the core and the periphery of the Russian prepositional system. The core is made up of primary prepositions and some secondary ones that have been identified as prepositions in most of the sources used. For instance, word combinations *в соответствии с 'in accordance with', независимо от 'independent of', судя по 'judging by'* are listed as prepositions in all of the studied sources. Some word combinations are recognised as prepositional units only in a few sources, e.g. *в защиту 'in defence of' (*see [4]), *без разрешения 'without permission of'* ([3]), *за рамки 'beyond the scope of'* ([2]), *вверх по 'up'* ([4], [5]), *во исполнение 'in pursuance of' (*[2], [6]), etc.

However, the core and the periphery of the Russian prepositional system can be identified on other grounds as well, namely through statistical analysis. It has already been mentioned that the same word combinations may be free or perform the functions of MWPs. The question is, how often either of these states occur.

In this article we propose the methodology of a study the overarching aim of which is to provide a systematic statistical description of Russian multiword prepositions, which, as we hope, will broaden the traditionally narrow definition of the preposition.

## THEORY: CHARACTERISTIC FEATURES OF MULTIWORD PREPOSITIONS

As we have already mentioned, prepositional multiword entities do not *always* function unambiguously as MWPs, but rather do so *sometimes*. Those units whose prepositional function is their dominant one can be regarded as the core elements of the multiword preposition subclass. In order to define its borders we have formulated the main characteristic features of multiword prepositions.

- MWP performs the grammatical function of a preposition in a certain syntactic position as part of a prepositional phrase; that is, it governs a noun or a nominalised word (sometimes an infinitive).
- MWP inherits the semantics of the notion word (noun, verb); it derives from as well as its valency (*на основе 'on the grounds of' – основа чего? 'the grounds of what?'; в зависимости от 'depending on' – зависеть от чего? 'to depend on what?'; с целью 'with the aim to' – цель что сделать? 'aim to do what?'*).
- As a rule, it contains one or two primary prepositions.
- Its nominal components tend to have abstract semantics.
- It has a relatively high frequency among multiword units of the same structural type.
- It is idiomatised, i.e. its nominal component loses its lexical meaning to an extent (which is why MWPs are sometimes called "prepositional idioms").
- The grammatical number of the noun cannot be changed (it is either singular or plural).
- It has a primary preposition as a synonym.
- In most cases, it does not allow for insertion or separation (as a rule, the noun cannot have a possessive or adjectival determiner).
- All of these features are characterised by significant statistical regularity.

This paper deals with the features that lend themselves to statistical description and analysis on corpus material. The results presented in this article have been acquired on the Araneum Russicum III Maius corpus (1.25 billion tokens) created by Vladimír Benko of Comenius University in Bratislava,

Slovakia (www.unesco.uniba.sk). Those features that are more lexically or semantically inclined are subject to future qualitative and quantitative studies.

## RESULTS AND DISCUSSIONS

### Structure of Multiword Prepositions

Something of interest is the structure of MWPs and their components. These are primary prepositions *(в 'in', с 'with')*, nouns *(помощь 'help', лицо 'face')*, adverbs *(отдельно 'separate', неподалёку 'not far away')*, gerunds *(смотря 'looking', считая 'counting')*, numerals *(один 'one')*, loan words *(а-ля 'à la', пандан 'pendant')*, pronouns *(то 'that', что 'what')*, particles *(не 'not')*, conjunctions *(и 'and')*.

Usually MWPs come in the form of bigrams and trigrams whose structure can be denoted as Prep+N, Prep+N+Prep and Adv+Prep, where Prep stands for preposition, N for noun, Adv for adverb. On some occasions these units do allow for separation or insertion of a determiner.

The most common MWP components are the primary prepositions *в* (156 entries), *на* (58), *по* (58), *с* (52), *от* (38), *под* (33) and some others. The most commonly occurring nouns are *начало 'beginning'* (in 8 MWPs), *путь 'way', предел 'limit', граница 'border'* (7), *сторона 'side', рамка 'frame', направление 'direction', лицо 'face'* (6), *имя 'name'* (5). Most of these words carry the semantics of spatiality or abstraction. The abstract semantics of the notional components of MWPs warrants their broad combinability with various semantic classes of governing and dependent members of a prepositional phrase. We might even go so far as to suggest that the prevalence of words with spatial semantics in MWPs reflects the general extralinguistic reality of the semantic structure of the language. Partial proof of that is the prevalence of the preposition *в* (*'in'*) among MWP components.

We have also observed the frequencies of MWPs and their components (Table 1). We have compared the frequency counts of nouns found in two or more MWPs to the total frequency counts of all MWPs containing these nouns. The table shows those items whose MWP use amounts to or exceeds a quarter of all usage cases.

*Table 1. Frequency counts of nouns found in MWPs, ipm*

| Base word | MWPs | Base word ipm | As MWP component | % of MWP use |
|---|---|---|---|---|
| довершение | в довершение, к довершению | 0.63 | 0.61 | 97 |
| неимение | за неимением, по неимению | 1.49 | 1.36 | 91 |
| рамка | в рамках, в рамки, вне рамок, за рамками, за рамки, из рамок | 204.30 | 175.84 | 86 |
| зависимость | в зависимости от(о), вне зависимости от(о) | 171.10 | 126.68 | 74 |
| сравнение | в сравнении с(о), в сравненье с(о), по сравнении с(о), по сравнению с(о) | 118.10 | 82.50 | 70 |
| предел | в пределах, в пределы, вне пределов, за пределами, за пределы, из пределов, из-за пределов, на пределе | 113.3 | 75.08 | 66 |
| рубеж | за рубеж, за рубежом | 43,00 | 25.95 | 60 |
| счет | в счет, за счет, на счету | 303,00 | 175.05 | 58 |
| ход | в ходе, по ходу | 186.70 | 108.05 | 58 |
| помощь | без помощи, при помощи, с помощью | 611.40 | 332.12 | 54 |
| середина | в середине, к середине | 64.6 | 33.5 | 52 |
| честь | в честь, к чести | 55.8 | 29.18 | 52 |
| конец | в конце, к концу, под конец | 355,00 | 161.3 | 45 |
| начало | в начале, к началу, на начало, перед началом, под начало, под началом, с начала, с началом | 363.30 | 142.90 | 39 |
| сфера | в сфере, вне сферы | 209.50 | 75.98 | 36 |
| адрес | в адрес, по адресу | 157.90 | 51.35 | 33 |
| сторона | в стороне, в стороне от(о), в сторону, в сторону от(о), на стороне, со стороны | 550.60 | 158.14 | 29 |

As indicated in the table, for many of the nouns (e.g. *зависимость, счет, сравнение, помощь)* over a half of all usage cases are as part of a prepositional unit. Some words (*довершение, неимение, рамка*) are used almost exclusively as MWP components.

**Statistics of Multiword Prepositions in Russian Vocabulary**

Another point of interest is the general statistics of all MWPs included in our list (Table 2). As the entire frequency list is too long, only the top part is provided below.

*Table 2. Top-20 most common MWPs*

| Rank | MWP | Frequency, ipm |
|---|---|---|
| 1 | во время *('during')* | 268.24 |
| 2 | с помощью *('with the help of')* | 240.18 |
| 3 | в качестве *('as')* | 215.43 |
| 4 | в течение *('over the course of')* | 207.66 |
| 5 | в результате *('as a result')* | 173.78 |
| 6 | в случае *('in the event of')* | 172.00 |
| 7 | вместе с *('along with')* | 168.60 |
| 8 | в рамках *('as part of')* | 166.10 |
| 9 | в виде *('in the form of')* | 143.83 |
| 10 | несмотря на *('despite')* | 133.40 |
| 11 | в области *('in the area of')* | 132.66 |
| 12 | в конце *('at the end of')* | 129.62 |
| 13 | в соответствии с *('in accordance with')* | 127.55 |
| 14 | в связи с *('due to')* | 114.63 |
| 15 | в зависимости от *('depending on')* | 111.62 |
| 16 | за счет *('by virtue of')* | 111.12 |
| 17 | на основе *('based on')* | 105.38 |
| 18 | в ходе *('in the course of')* | 97.83 |
| 19 | в мире *('in the realm of')* | 97.71 |
| 20 | в процессе *('in the act of')* | 94.63 |

The relative frequency counts of about 90 MWPs range from 20 to 270, which is comparable to the frequencies of many Russian notion words; 21 have frequency counts exceeding 90. For instance, we find such common vocabulary words as *положение 'position'* (ipm 268.2, ranked 390 out of 20000), *красный 'red'* (ipm 240.5, ranked 442), *принимать 'to accept'* (ipm 207.7, 521), *врач 'doctor'* (ipm 173.1, 653), *уметь 'to be able to'* (ipm 172.5, 657) and some others in the same frequency range in the Frequency Dictionary of the Modern Russian Language [7].

As we can see, a lot of MWPs share frequencies with common vocabulary words, the most commonly used MWPs being found in the top frequency stratum of the Russian lexicon. In addition to that, frequency counts of some MWPs are comparable to those of some primary prepositions (*над, между, из-за, перед, про*) or even exceed them (*из-под, по-над, по-за*). This fact signifies that MWPs are regular and essential elements of the Russian syntactic sentence structure.

**Place of Multiword Preposition in Russian Ngram Vocabulary**

If we take frequency counts of units made up of 2 and 3 words following the structural patterns of Prep+N, Prep+N+Prep and Adv+Prep and mark those entries that are found on our MWP list, we will see that the top part of these frequency lists (25 most common ngrams) is comprised mostly of MWPs (Table 3).

*Table 3. Top-25 of the most frequently observed constructions, ipm (non-MWP entities are greyed out)*

| Rank | Prep + N | Freq. | Prep + N + Prep | Freq. | Adv + Prep | Freq. |
|---|---|---|---|---|---|---|
| 1 | во время | 268.24 | в соответствии с | 127.55 | вместе с | 168.6 |
| 2 | с помощью | 240.18 | в связи с | 114.63 | уже в | 67.00 |
| 3 | в качестве | 215.43 | в зависимости от | 111.62 | рядом с | 61.10 |
| 4 | в течение | 207.66 | по сравнению с | 64.90 | еще в | 68.9 |
| 5 | в России | 194.70 | в отличие от | 57.96 | совместно с | 40.35 |
| 6 | в результате | 173.78 | по отношению к | 39.44 | независимо от | 37.24 |
| 7 | в случае | 172.15 | в сочетании с | 18.80 | наряду с | 29.30 |
| 8 | в рамках | 166.10 | вне зависимости от | 15.06 | вплоть до | 28.99 |
| 9 | в соответствии | 144.10 | в соответствии с/со | 14.34 | сегодня в | 27.30 |
| 10 | в виде | 143.83 | со ссылкой на | 13.59 | особенно в | 27.20 |
| 11 | в области | 132.66 | в сравнении с | 13.01 | уже на | 26.90 |
| 12 | в конце | 129.62 | во главе с | 12.76 | сразу после | 25.30 |
| 13 | на сайте | 125.50 | в период с | 12.60 | можно в | 24.20 |
| 14 | на территории | 121.50 | в случае с | 12,00 | специально для | 22.80 |
| 15 | в связи | 121.10 | в ответ на | 11.35 | недалеко от | 18.57 |
| 16 | в Москве | 121,00 | по уходу за | 11.10 | вместе со | 17.67 |
| 17 | в целом | 114.50 | по состоянию на | 11,00 | можно с | 16.80 |
| 18 | в зависимости | 112.80 | для участия в | 10.80 | еще на | 19.30 |
| 19 | за счет | 111.12 | в борьбе с | 10,00 | уже через | 15.50 |
| 20 | в частности | 104.40 | для работы с | 9.90 | еще до | 18.60 |
| 21 | в мире | 97.71 | по работе с | 9.60 | примерно в | 15.20 |
| 22 | со стороны | 93.84 | в возрасте от | 8.60 | прямо на | 18.10 |
| 23 | в день | 89.10 | при работе с | 8.10 | уже с | 14.50 |
| 24 | в начале | 87.13 | на участие в | 7.90 | вслед за | 14.32 |
| 25 | к сожалению | 83.80 | на пути к | 7.66 | одновременно с | 14.06 |

The table shows that MWPs are found in the upper zone of frequency lists of constructions with the same part-of-speech structure. Namely, out of 25 of the

most commonly used word combinations following the Prep+N and Prep+N+Prep patterns 14 units are found in our MWP list, as well as 10 out of 25 word combinations of the Adv+Prep type.

The figures demonstrate not only a high degree of use of MWPs but also the strength of the bond between their elements, which allows us to regard these word combinations as actual units with common features stemming from the unique aspects of their structure and function.

**Ambivalent Nature of Multiword Prepositions**

In order to detect the most regularly occurring combinations in the function of a preposition we have conducted a frequency analysis of a number of potential MWPs on corpus material. For this purpose a random sample of 100 contexts (or fewer in the case of low frequency units) was selected from the corpus data for each of the prepositional units under study. The samples were then examined by hand to determine the proportion of free versus prepositional uses of these entities among the selected contexts.

A few of the units in question were found to be used exclusively as MWPs. For instance, the prepositional combinations *в виде 'in the form/shape of', по сравнению с 'compared to', через посредство 'by dint of'* were used in the function of a preposition in all observed contexts (*«фигурки **в виде** кораблей и якорей» 'figurines **in the shape of** ships'; «меньше **по сравнению с** 2009 годом» 'less **compared to** 2009'; «защитить **через посредство** насилия» 'to defend **by dint of** force')*. In this case we can make the claim that these units belong to the core of the MWP subclass with a reasonable degree of certainty.

The proportion of prepositional and free uses of a sizeable number of entities proved to be less straightforward. As such, *смотря по 'judging by'* was used as a preposition in over 90% of the contexts (91 out of 100); *по части 'in terms of'* in almost two thirds (68 из 100), and *что до 'as for'* in only one quarter (18 out of 100) of its sample.

In some cases qualitative analysis by hand led to the discovery of prepositions that had not been listed in any other source. For one, the sample of 20 contexts containing the low-frequency combination *в пандан 'as a complement'* contained 3 instances of non-prepositional usage *(«**в пандан** заметил» 'noted **as a complement**')* and 17 cases of MWP occurrences, out of which only 6 were in the original form of the preposition (*«**в пандан** этим работам <...> создала» 'created <...> **as a complement to** those works'*), 1 was located in the postposition *(«ему **в пандан** вторил» 'was echoing him **as a complement**'),* 8 of the contexts included the variation *в пандан к/ко 'as a complement to'* (*«Второе предисловие **в пандан к** первому» 'second foreword **as a complement to** the first one')* and 2 contexts revealed a yet to be catalogued variation *в пандан с/со 'as a complement with'* (*«Ресницы **в пандан с** ежами и хип-хопом» 'eyelashes*

***as a complement with*** *the sea urchins and hip-hop').* Such examples demonstrate the value of corpus studies in tasks dealing with nomenclature and status definition of composite entities.

The results acquired in the current study will aid in the refinement and improvement of theoretical descriptions of the Russian prepositional system by supplying them with statistical data. At the same time, the data itself can be used in stochastic syntax models.

## CONCLUSION

The current paper deals with the complex aspects of the multiword preposition (MWP) subclass. It is demonstrated that MWPs as a class are widely represented in Russian. Characteristic features of MWPs as multiword units are described. The ambivalent nature of their use is discussed. These units may perform as prepositional entities with particular grammatical semantics or manifest as free combinations in which each word has its own meaning and syntactic function. Statistical data on the proportion of prepositional and free use of multiword units in a corpus are provided.

The data presented signify that MWPs are a large and diverse subclass that is nonetheless characterised by a number of common features and, therefore, lends itself to description, definition and measurement. The experiments described in this paper are exploratory in nature but will be calibrated and conducted further with the purpose of acquiring the first comprehensive description of the subclass in question.

Further stages of our research include acquisition of statistical data on the entirety of the MWP subclass. Clusters of conditional synonymy between primary and secondary prepositions will be defined. Preposition stranding and the separability of the components of a multiword preposition are to be described. The latter is of significant importance to sentence structure representation in automated syntactic and semantic analysis. An additional task is the description of prepositional use in fixed phrases and idioms.

## ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation for Basic Research [grant No. 17-29-09159 "Quantitative grammar of Russian prepositional constructions"].

## REFERENCES

[1] Shvedova N.Ju, Russian Grammar. Vol. 1: Phonetics. Phonology. Word Stress. Intonation. Word Formation. Morphology, Moscow, 1980.

[2] Efremova T.F., Explanatory Dictionary of Functional Parts of Speech of the Russian Language, Moscow, 2004.

[3] Rogozhnikova R.P., Explanatory Dictionary of Combinations Equivalent to a Word: Approx. 1500 fixed phrases of the Russian language, Moscow, 2003.

[4] Corpus Dictionary of Multiword Lexical Units. URL: https://ruscorpora.ru/new/obgrams.html (last access 21.08.2020).

[5] Dictionary of the Russian Language (4 volumes), 3rd edition, Moscow, 1985.

[6] Hagen M., Full Paradigm of the Russian Language. Morphology. URL: http://www.speakrus.ru/dict/hagen-morph.rar (last access 21.08.2020).

[7] Lyashevskaya O.N., Sharov S.A., Frequency Dictionary of the Modern Russian Language (on the Materials of the National Corpus of the Russian Language), Moscow, 2009