Research complex for the corpus-based study of Russian prepositional constructions

Victor Zakharov

Saint-Petersburg State University
E-mail: v.zakharov@spbu.ru

Abstract

The creation of a complex for the study of Russian prepositional constructions is a part of a research project, which is aimed at the development of corpus-driven semantic-grammatical description of Russian prepositional constructions. One can hardly mention any corpus-based works dedicated to the Russian prepositions. Our objective is to create a corpus-based semantic and grammatical description of Russian prepositional constructions using empiric corpus data. We are going to develop a toolkit, which will include corpus linguistics tools and procedures for calculating conjunction of prepositions with their “governors” and “governees”. This toolkit implements the corpus-driven approach combined with distributional, statistical and other techniques.

Keywords: Russian prepositional constructions; corpus linguistics; Aranea corpora; methodology; technology

1 Introduction

The creation of the complex is a part of a research project, which is aimed at the development of corpus-driven semantic-grammatical description of Russian prepositional constructions. One can hardly mention any corpus-based works dedicated to the Russian prepositions. Our aim is to create a corpus-based semantic and grammatical description of Russian prepositional constructions using empiric corpus data.

In contrast to the classical linguistic methodology focusing on the simplest units of different language levels, modern studies practice synthetic methods trying to catch and describe language structures, which integrate different language units: words, collocations, etc. Constructions – the combinations of lexical, semantic, morphological, syntactical and other units realized in phrases – are of peculiar interest for modern linguists. Complex description and systematization of constructions seek for elaboration of identification methods using manual and automatic techniques as well as for analysis of their paradigmatic and syntagmatic features and quantitative analysis of their frequency and strength. In classical linguistic papers, prepositional constructions used to be described from the grammatical point of view and their semantics used to be neglected. Integrated description and classifying of the constructions are directed to elaboration of identification methods using manual and automatic techniques as well as for analysis of their paradigmatic and syntagmatic features and quantitative analysis of their frequency and strength. In classical linguistic papers, prepositional constructions used to be described from the grammatical point of view only and their semantic features were neglected.

2 Methodology

Prepositional constructions have high frequency of occurrence and are important for the generation and understanding of Russian texts. Here we understand construction as combination of a preposition with a main word (governor) and a dependent word (governee) in terms of dependency grammar. To meet the need, we are going to develop a technology to extract and analyze prepositional constructions.

This technology includes corpora, corpus tools, other software instruments, and manual procedures. Our final aim is creation of a corpus-based semantic and grammatical description of Russian prepositional constructions using empiric data and also formalization of basic ontological semantic patterns. The semantic-grammatical analysis of relations between lexical items certainly cannot be performed entirely automatically and it needs participation of linguists.

The elaborated technology is based on the corpus-driven approach combined with distributional and statistical techniques. One can hardly mention any corpus-based works dedicated to the Russian prepositions. Primarily, this paper deals with developing a methodology of using corpus tools for solving tasks of the description of prepositions and prepositional constructions. The nearest tasks are as follows:

- to get sets of prepositional constructions from corpora of different types and different functional styles;
- to get a number of statistical characteristics for each preposition from corpora of different types and functional styles, namely:
- ipm in a corpus (precisely, in various corpora);
- percentage of each meaning of appropriate preposition;
- a list of most frequent semantic classes and/or lexemes acting as a “governor” for each prepositional meaning;
- a list of most frequent semantic classes and/or lexemes acting as a “governee” for each prepositional meaning;
- etc.

The meanings of prepositions are defined according ontological description of simple non-derivative prepositions based on the Syntactic Dictionary by G. Zolotova (2011). These meanings could be named «semantic rubrics». We use the term «semantic rubric» as a general term for the group of meanings of different prepositions.

The prepositions in Russian are heterogeneous and diverse: there is a small group of primary prepositions and a large number of secondary ones, the latter being motivated by the content parts of speech (nouns, adverbs and verbal forms), which may be combined with the primary prepositions forming multiword expressions (combinations of several words). This fact shows that corpus frequencies of the primary prepositions are regularly overrated because they may be used as parts of secondary prepositions. One of the project tasks is to clarify real frequencies of primary prepositions, excluding their use as part of secondary prepositions. This fact shows that frequency values of the primary prepositions in corpus texts are often wrongly overrated since they may be used as parts of secondary prepositions. One of the project tasks is to exclude from results cases of their usage as parts of secondary prepositions

3 Research Tools

To solve the task mentioned above we need appropriate corpora. These corpora have to meet next requirements. They should be representative, balanced, annotated and highly functional. Pilot research shows that we do not have a corpus of Russian with semantic annotation, which would satisfy our requirements. The other point of our research is that we are going to extract constructions "governor – preposition – governee" that suppose syntactic connection between elements. Within the Russian National Corpus exists Deeply Annotated Corpus (treebank) but test shows that it is practically useless for our tasks. Thus, we carried out our work on the basis of morphologically annotated corpora.

At first stage we've chosen 2 corpora of Russian texts: Russian National Corpus (RNC) (http://www.ruscorpora.ru/en/index.html), Russian corpora of the Aranea corpus family (http://unesco.uniba.sk). They are different in size and in balance of textual genres. Russian National Corpus, among the balanced Main corpus (283 mln. token), includes subcorpora of other types such as Corpus of Spoken Russian, Newspaper corpus, Dialectal corpus, Poetry corpus and others.

The Aranea family consists of web corpora created by the wacky technology [Benko 2014]. For Russian there are three region-specific variants with volume of 120 and 1200 mln. token plus Araneum Russicum Maximum that increases the total number of Russian corpora to seven. We used mainly Araneum Russicum Minus (120 mln. tokens, 91 mln. lexical words). This volume is enough for investigation of such units as prepositions. It is worthy also to mention that Araneum Russicum Externum corpus permits to create domain-specific subcorpora such as .ua, .by, .il, etc. All this gives a possibility to study Russian prepositional constructions in regional variations.

4 Corpus-Based Analysis

We have done first stage experiments to receive scientific results concerning prepositions and to tune and improve the methodology of long-term research. Prepositional constructions have high frequency of occurrence in corpora and there is a task to develop technology (and methodology) in which way we can ensure the reliability of results and minimize labor costs to receive appropriate data.

Since the aim of our project is to compile the quantitative grammar for prepositional constructions, we use the term «semantic rubric» as a generic name of the group of meanings of prepositions. It is often the case that these rubrics correspond to the names of semantic actants (or cases, roles), but due to the variety of interpretations for these concepts in functional grammars [Mustajoki 2006] and the absence of clear boundaries between the items of the list, we will start our description from «the bottom», from the most frequent basic prepositions. The different variants of meanings, more or less explicitly formulated in the explanatory dictionaries, may be associated with usually secondary prepositions, creating chains of synonymous or quasi-synonymous constructions.

Synonymous constructions are selected on the base on information partly provided by dictionaries. In addition, the synonymy is verified by the possibility of replacement one preposition with other. In case of the discrepancy between prepositional co-occurrence and semantic classes of nouns or some other restrictions on the implementation of semantic rubric meanings they are regarded as partial synonyms, differences in usage being measured in accordance with corpus-based statistics.

We started our description from «the bottom», i.e. from the most frequent basic prepositions. The different variants of meaning more or less explicitly formulated in the explanatory dictionaries, may be associated with secondary prepositions, creating chains of synonymous or quasi-synonymous constructions. That is why we use semantic rubrics as a common denominator of the meanings of different prepositions. Let's begin with three semantic rubrics.

Mediative as semantic rubric has a narrow and a wide interpretation. Generally it is considered as a particular semantic role in a predicate structure of a verb. In the narrow sense mediative is understood as a means, that is a substance or an object are being used during the performance of an action or a process. In a broader sense mediative is a tool (instrumentative) and includes its material and abstract implementations [Mustajoki 2006]. In Russian language both mediative and instrumentative are regularly expressed by the instrumental case form (*красить стены валиком, рисовать картину красками*), however, we can observe more complex syncretic instances in the form of prepositional constructions.

Transitive is one of the possible ways of the proposition localization. Unlike the characteristics of location, which are applicable for diverse set of actions, states and processes, this specification is often associated with a "framework" structure of a prefix *пере-* for the verbs of motion and their derivatives: *перейти через дорогу, перевозки нефти через Атлантику*, etc.

Temporative is as complex a semantic rubric as a locative, therefore it's rather difficult to represent all variants of its realization immediately. The preposition *через* ('through') in this case means after a certain period of time.

Let's show a fragment of our research on the example of a preposition *через* and its synonyms, referring to three semantic rubrics, mediative, transitive and temporative (see table 1).

preposition	semantic rubric	RNC (balanced), the disambiguated subset (a sample of 200 examples)	*RNC,* newspapers subcorpus (a sample of 200 examples)	Web-corpus Araneum Russicum Minus (a sample of 200 examples)
через	**mediative**	**ipm 173,53** **20,5%** **(41 occurrences)**	**ipm 221,30** **34,5%** **(69 occurrences)**	**ipm 185,21** **32,5%** **(65 occurrences)**
с помощью	mediative	ipm 76,63 460 occurrences	ipm 113,21 25870 occurrences	ipm 228,64 27437 occurrences
при помощи	mediative	ipm 22,32 134 occurrences	ipm 27,66 6322 occurrences	ipm 82,60 9916 occurrences
посредством	mediative	ipm 16,49 111occurences	ipm 12,31 2813 occurrences	ipm 42,90 5154 occurrences

через посредство	mediative	ipm 0,99 6 occurrences	ipm 0,09 22 occurrences	ipm 0,42 50 occurrences
через	**transitive**	**ipm 245,39 29,5% (57 occurrences)**	**ipm 118,65 18,5% (37 occurrences)**	**ipm 157,71 27,5% (55 occurrences)**
сквозь	transitive	ipm 119,83 719 occurrences	ipm 19,25 4398 occurrences	ipm 23,90 2869 occurrences
поперек	transitive	ipm 12,00 72 occurrences	ipm 3,15 720 occurrences	ipm3,40 408 occurrences
через	**temporative**	**ipm 390,96 47% (94 occurrences)**	**ipm 319,95 50% (100 occurrences)**	**ipm 222,25 27,5% (55 occurrences)**
Спустя	temporative	ipm 54,67 328 occurrences	ipm 82,28 18800 occurrences	ipm 52,50 6300 occurrences
по истечении	temporative	ipm 2,17 13 occurrences	ipm 5,81 1328 occurrences	ipm 6,88 826occurences

Table 1: Frequencies of preposition *через* (through) and synonyms.

The experiments demonstrate the capabilities of corpus tools to obtain data on the representation of individual meanings of prepositions in Russian texts still missing in the scientific literature. Also, a preliminary analysis of the data in the table shows why different corpora should be used to receive reliable data. We see that both values of separate meanings and IPM values varies noticeably from one corpus to another corpus both for primary prepositions and for secondary ones (see table 1).

5 Technological Issues

The second task of this research was to develop a methodology of using corpus tools for the description of prepositional constructions.

The experiments show that corpus manager of the Russian National Corpus (Yandex-Server) provides a sufficiently flexible query language and selection of required constructions. Much worse is the situation with the processing of results. The output is possible only in small portions without automatically obtaining statistical data. Therefore, the main instrument in our project is the NoSketch Engine system (Rychlý 2007), which supports the Aranea corpora. The main features of this system are a powerful query language (CQL) and built-in tools for handling obtained data.

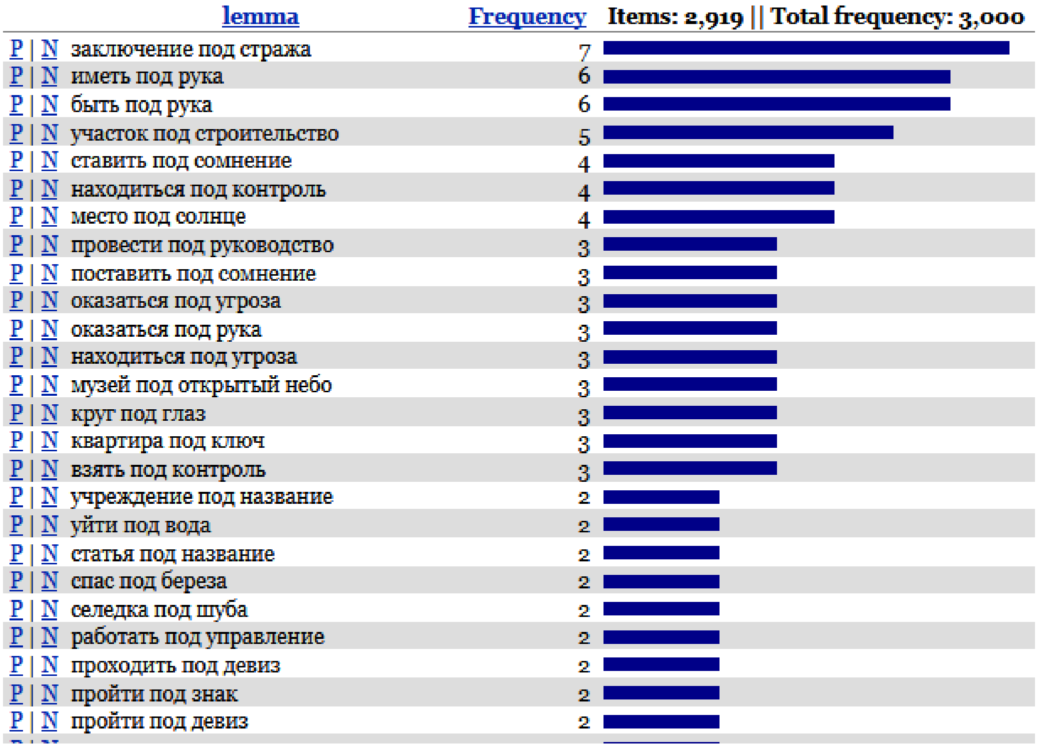
The experiments allow us define the basic scheme of research procedures. The preliminary technology could be roughly described as follows.

We create a concordance in accordance with a query of the following template “governor” (verb or noun) + "our preposition” + “governee” (noun or pronoun or abbreviation), with other words between these constituents. Of course, details of a query have to be thought out and experimentally debugged.

The other challenge is that we cannot possibly handle the whole concordance, which could be too large. To avoid it we select a random sample from it. The size of the sample is a question of further experiments. Preliminary, the following procedure seems quite reliable:

- to sort concordance in shuffling mode;
- to set a volume of random sample (number of lines) rather big, for example, 3000 lines;
- to make frequency list for this sample (see Fig. 1) ;
- to save and process first 200 lines of this list.

Such a procedure provides us with a list of prepositional constructions, which presents both randomly selected and the most frequent ones inside this sample.



	lemma	Frequency	Items: 2,919 || Total frequency: 3,000
P | N	заключение под стража	7	
P | N	иметь под рука	6	
P | N	быть под рука	6	
P | N	участок под строительство	5	
P | N	ставить под сомнение	4	
P | N	находиться под контроль	4	
P | N	место под солнце	4	
P | N	провести под руководство	3	
P | N	поставить под сомнение	3	
P | N	оказаться под угроза	3	
P | N	оказаться под рука	3	
P | N	находиться под угроза	3	
P | N	музей под открытый небо	3	
P | N	круг под глаз	3	
P | N	квартира под ключ	3	
P | N	взять под контроль	3	
P | N	учреждение под название	2	
P | N	уйти под вода	2	
P | N	статья под название	2	
P | N	спас под береза	2	
P | N	селедка под шуба	2	
P | N	работать под управление	2	
P | N	проходить под девиз	2	
P | N	пройти под знак	2	
P | N	пройти под девиз	2	

Figure 1: frequency list for construction with preposition *под* ('under').

In course of further processing of this list we define preposition meanings, calculate percentage of each meaning, and analyze governors and governees – first, define quantity of each lexeme, second, classify them across the semantic classes. We extrapolate frequency of meanings into a corpus size and express it in IPM, that is 'instances per million tokens in a corpus'. Probably, ranks and fractions of IPMs are more reliable than actual magnitudes of IPMs. For proportions of frequency it is reasonable to establish a threshold of sparseness (for example, 5%).

6 Conclusion

The project will result in the corpus-based creation of a completely new linguistic resource for the Russian language. In this investigation we use corpus-based analysis and techniques.

The NoSketch Engine system and Aranea corpora in general meet our requirements. However, Aranea corpora are created on the base of texts from the web. Studies carried out with this corpora showed that there are problems that can be divided into 3 parts: insufficient quality of linguistic annotation due to "dirty" data, lack of metadata and technical problems associated with removing duplicates, elements of hypertext markup languages, and so on. In practice, it can be said that web-corpora are unbalanced. That is, we get large-volume corpora, but the question arises of assessing the reliability of results obtained. The problem of verification and reliability of data obtained on the basis of statistical methods is a problem of balanced and correctly interpreted corpus data.

Therefore, in addition to web-corpora, we proceed to create our own corpora based on the Sketch Engine system, which has this function. This will provide us with well-balanced data and allow us to create corpora of different functional styles. The methodology and technology of processing corpus data will remain the same as it is described in this paper.

7 References

Zolotova, G. A. (2011). *Syntactical Dictionary: a Set of Elementary Units of Russian Syntax* [Sintaksicheskiy slovar': repertuar elementarnykh edinits russkogo sintaksisa], 4th edition, Russia, Moscow.

Benko, V. (2014) Aranea: Yet Another Family of (Comparable) Web Corpora. In *Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655*. Springer International Publishing Switzerland, pp. 257-264.

Mustajoki, A. (2006). *Theory of Functional Syntax* [Teorija funtsionalnogo sintaksisa]. Russia, Moscow.

Rychlý, Pavel (2007). Manatee/Bonito – A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University, pp. 65-70.

Acknowledgements

This work was implemented with financial support of the Russian Foundation for Basic Research, the project No. 17-29-09159 “Quantitative grammar of Russian prepositional constructions”.