THE OUTLINE OF THE QUANTITATIVE ONTOLOGY FOR RUSSIAN PREPOSITIONAL CONSTRUCTIONS

Assoc. Prof. Dr. Irina Azarova [1]
Assoc. Prof. Dr. Victor Zakharov [2]
[1, 2] St. Petersburg State University, Russia

ABSTRACT

The dependency grammars for such languages as Russian usually treat the prepositions in combination with subordinate nouns as major elements as if the case form in the prepositional construction had some self-contained meaning subjected to the regular transformation. This scheme may be valid for languages with restricted declensional paradigms, however, in Russian due to the fact that several case forms are combined with primary prepositions the specific system of joint interpretation of a preposition and a case form has resulted in a system of syntaxemes, minimal syntactical items designed to express semantic notions according to the lexical nature of the governor words in the texts.

The syntaxeme structure being the part of the grammatical system has a number of vague manifestations in modern Russian texts which may be acquired from the corpus statistics. The evidence of syntaxeme structure are presented in the ranks of collocations with rough semantic classes of governee nouns with frequent primary prepositions, on the one hand, and the semantico-syntactic role specification of the prepositional construction in the sentence, from the other hand.

This structure has several levels of syntactic abstraction. The syntaxeme level is the central layer showing the most frequent combinations of prepositions with noun forms of subordinate semantic classes. Several syntaxemes may be united into the group of so-called semantic rubrics, more or less equal to the semantico-syntactic arguments or roles of the construction in the sentence. The witnesses of semantic rubrics in the texts are usually expressed by secondary prepositions. The bottom level of syntaxeme units is determined by subtle sense variants of the construction produced by the marked semantic classes of governor words of prepositions and governee nouns, which resulted in the merged synonymic and quasi-synonymic usage of prepositional constructions.

Keywords: *prepositional construction, Russian language, corpus statistics, syntaxeme, prepositional meaning*

INTRODUCTION

The paper presents the preliminary stage and further guidelines for the quantitative ontology of Russian prepositional constructions. This prepositional ontology is usually investigated in terms of semantic roles assigned to prepositional phrases [1]. However, the absence of prepositional meanings in the semantic annotation leads to parallel series of prepositional frequencies along with that of semantic roles marked up in the annotated corpora (e.g. PropBank). All we can conclude from them: they are different. The use of data from annotated prepositional meanings [2] has another deficiency – the scarcity of representative bulk of data instances.

In Russian the extra factor is added: primary prepositions are combined with several case forms, which increases the combinatorial effect for prepositional expressions due to the shift of the typical syntactic usage for the nominal case forms, which may be totally different or have vague traces of the primary meaning. We use the concept of syntaxeme [4] as a minimal syntactic prepositional construction. This item can not be split into the sum of preposition's meaning and that of the noun case form. It acts as the syntactico-morphological function, which interprets the whole construction having its particular meaning dependent on semantic types of governee nouns and governor words if used. The notion of syntaxeme [3] was defined in the functional direction of traditional linguistic analysis, so we redefine it inside our own quantitative corpus approach.

We developed the procedure for describing the continuum of prepositional meanings [4] basing on the corpus data starting from the bottom – textual analysis of sense distribution for a small group of frequent primary prepositions in random context samples from different corpora. The semantic class of governee nouns in the construction with some specific prepositional meaning is typically conveyed by frequent lexemes from these classes, so-called class-representatives. Sometimes they are top-level hypernyms from the WordNet basic concepts hierarchy, though lexemes with another semantic status are also mentioned in [5]. Anyway these lexemes are used in order to fix prototypical contexts for some particular syntaxeme specification.

Other primary prepositions may appear in prototypical contexts organizing the ranked list on the basis of the corpus frequency, which depends on the functional and thematic text balance in the source corpus in question. It is necessary to stress that syntaxeme involving primary prepositions has the grammatical nature, which is proved by high corpus frequency – several decimal exponents bigger than highest lexical corpus frequencies (for example 2500 IPM (instances found per million corpus tokens) versus 300) – and intricate inconsistency of prepositional synonymy in combination with particular nouns. Moreover, a syntaxeme may be expressed in the lexical form with a help of so-called secondary prepositions. Usually lexicalized forms are less frequent, less intricate, and have a variety of textual nuances. For example, the temporal

syntaxeme of the particular period or moment of time is expressed by preposition "в" ('in') plus the locative or prepositional case form for nouns denoting months and years: *в 1999 году* ('in 1999'), *в августе* ('in August'), but for nouns denoting day of the week or the time of day the accusative noun form is used: *в пятницу* ('on Friday'), *в 10 часов* ('at 10 o'clock'), the lexicalized variants being numerous: *во время войны* ('during war'), *в период беременности* ('during pregnancy'), *в момент опасности* ('at the moment of danger'), *во времена крестовых походов* ('during the crusades'). Usually the problem of grammatical nature of secondary prepositions is treated as unresolvable: there are pros and cons, however, all of them form the list of potential grammatical units, which may "cross the frequency threshold" and become the authentic member of the syntaxeme expressors. We can see the results of grammaticalization: the ablative form of the noun *посредством* ('by means of') is used in the mediative syntaxeme with corpus frequency 19 IPM whereas the noun *посредство* ('instrumentality') does not exceed 0,01 IPM. The last stage of grammaticalization is usually fixed in orthography: *в течение года* ('during a year'); *несмотря на непогоду* ('despite the bad weather').

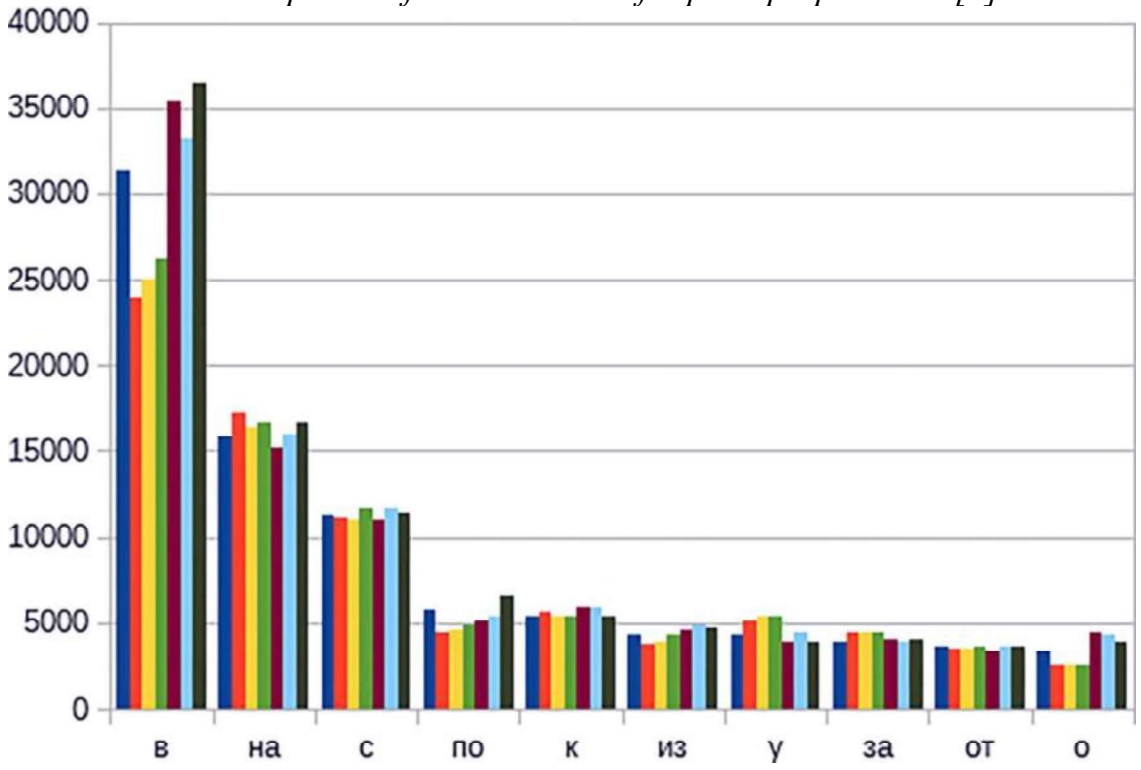
CORE GROUP OF RUSSIAN PRIMARY PREPOSITIONS

The core group of Russian primary prepositions may be set by two methods. The first one is based on the traditional linguistic approach. There are a group of primary prepositions which are proved or considered to be the heritage of the mutual Slavonic stratum. We may see parallel correspondences in various Slavonic languages. In some works it is hinted that there were several conventional ways of expressing semantic roles. For example, for the semantic role "addressee" of communication [6] we see two case forms: dative (*сказать/ отвечать/ возразить соседу* 'to say/ answer/ object to a neighbor') and accusative (*известить/ ругать соседа* 'to notify/ scold a neighbor'), and 3 prepositional: <"на" + accusative> (*ворчать/ ругаться на соседа* 'grumble on/ curse at the neighbor'), <"к" + dative> (*взывать/ приставать к соседу* 'to appeal to/ pester a neighbor'), <"до" + genitive (*дозвониться до соседа* 'to call a neighbor'). As a matter of fact there two additional prepositional phrases for communication addressee specification: <"с" + ablative> (*разговаривать/ договориться с соседом* 'to talk/ negotiate with a neighbor'), <"у" + genitive> (*спросить/ выпрашивать у соседа* 'to ask /beg a neighbor').

Primary prepositions from the ancient group have counterpart prefixes, which modify the verbal stem and with a strong probability predict the appearance of the governed prepositional phrase with the same preposition (*войти в дом* 'to enter the house'; *заползти за куст* 'to crawl over a bush') or that with the same meaning (*выйти из дома* 'to to get out of the house'; *перепрыгнуть через куст* 'to jump over a bush'). This "frame" construction on the large scale describe the appearance of many prepositional phrases which are governees of verbs, mainly travel verbs and other semantic groups.

We will follow the second method for picking up the core group which is based on the corpus statistics. It is well known that two Russian prepositions “в” (‘in’) and “на” (‘on’) are the upmost frequency leaders. However, the frequency differentiation concerning periods of literary and publicistic texts from [7] show the substantial variability (see Fig. 1).

Fig. *1. Frequency variation in literary and publicistic text subcorpora from three time periods for the ten most frequent prepositions [7].*



Prepositions: “в” (‘in’), “на” (‘on’), “с” (‘with’), “по” (‘by’), “к” (‘to’), “из” (‘from’), “у” (‘at’), “за” (‘behind’), “от” (‘from’), “о” (‘about’).

The order of columns and their colors: 1. dark blue – total,
literary: 2. red – 1950-60s, 3. yellow – 1970-80s, 4. green – 1990-2000s;
publicistic: 5. brawn – 1950-60s, 6. light blue – 1970-80s, 7.dark green – 1990-2000s.

It is seen from the Fig.1 that the most frequent preposition “в” (‘in’) has the biggest variation, however, its corpus frequency exceeds others anyway. Prepositions with the second “на” (‘on’) and the third rank “с” (‘with’) vary less and surpass frequencies of other top 7, which may change its rank in different subcorpora. For example, “по” (‘by’) gives way to “к” (‘to’) and “у” (‘at’) in literary texts of the 1950-60ties. We assign ranks to frequencies in different subcorpora in order to highlight their variations (see Table 1 below).

It is easy to see that corpus frequencies of medium frequent prepositions (from rank 4 to 10) are subjected to variation in texts with communicative goals) and topics discussed. It will be more obvious for less frequent prepositions (with rank equal and less than 11). The frequency variation of primary and secondary

prepositions will be the issue of the future research. We take this group in order to demonstrate technique developed in our project.

Table 1. *Ranks for 10 the most frequent prepositions in literary and publicistic subcorpora on the base [7].*

		Literary				Publicistic		
Preposition	Total	1950-60s	Literary 1950-60s	1970-80s	1990-2000s	1950-60s	1970-80s	1990-2000s
в	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
на	2,0	2,0	2,0	2,0	2,0	2,0	2,0	2,0
с	3,0	3,0	3,0	3,0	3,0	3,0	3,0	3,0
по	4,0	4,0	6,5	6,0	6,0	5,0	5,0	4,0
к	5,5	5,5	4,0	4,5	4,5	4,0	4,0	5,0
из	5,5	5,5	8,0	8,0	8,0	6,0	6,0	6,0
у	7,0	7,0	5,0	4,5	4,5	9,0	7,0	9,0
за	8,0	8,0	6,5	7,0	7,0	8,0	9,0	7,5
от	9,0	9,0	9,0	9,0	9,0	10,0	10,0	10,0
о	10,0	10,0	10,0	10,0	10,0	7,0	8,0	7,5

Fractional ranks (such as 6,5 and similar) denote that the frequency difference for two prepositions is less than a possible statistical error, so both have the same rank.

ANNOTATION TAGS FOR PREPOSITIONAL MEANINGS

The prepositional construction is viewed in Russian grammatical tradition after [8] as a chain of two pairs of "governor-governee": the first pair connects the full word with a preposition, the second – a preposition and a noun case form. We stress in the introduction that this point of view is inappropriate for the idea of syntaxeme. The opposite attitude which reverses the link between the preposition and a noun has another mismatch because interpretation of the same pair "preposition-noun" may vary according to the semantics of the governor word in the prepositional construction (examples follow). So we postulate that the syntaxeme is attached to its governor as a whole component and an entire unit is interpreted with semantic tags of different level of specification.

G.Zolotova [3] mentioned two types of syntaxemes: bound and free ones. We supposed that all prepositional construction are bound. In case of absence in the context a governor word, as is the case in the title of the play by M.Gorky "На дне" ('On the Bottom') which denotes something like "being in the state" where "bottom" is understood metaphorically as "the environment of the declassed descended people". Another pseudo-free prepositional construction is very natural in Russian: *в лесу много снега* ('there is a lot of snow in the forest'). The locative at the beginning describes the situation in which the proposition is fulfiled

or exists. We don't want to insert imaginary predicates such as *exist, happen*, etc, we may consider that there is one from the top of the semantic predicate's taxonomy.

G.Zolotova's taxonomy [3] of syntaxemes was designed for the special goals of the functional approach to syntactic analysis. Some syntaxemes resemble semantic arguments or roles in various paradigms [2] though they were invented in the earlier period (the first edition of Zolotova's conception appeared in 1988). It was an attempt to grasp specific syntactic fundamental quality of noun case forms as well as prepositional phrases. However, straightforward use of Zolotova's syntaxemes leads in some cases to the conflict in interpretation, which was shown in the dictionary itself. We understand that several syntaxemes tend to overlap and organize a group with mutual semantic component, so we assign this group to the ***semantic rubric***. For example, Locative (*гулять в лесу* 'to have a walk in the forest'), Directive (*пойти в лес* 'to go to the forest') and Transitive (*бродить по лесу* 'to wander through the woods') syntaxemes have a common constituent stressing the Localization of some event. At the same time we see that prepositional constructions mentioned as realizations of one syntaxeme are not interchangeable, as they have particular meanings. For example, a number of constructions put into Temporative syntaxeme have its particular meanings: the preposition "до" ('until') indicates a limit of time duration or that of an event: *до пятого марта* ('until March 5'), *до отъезда* ('until departure'). Another variant is the specification of the duration of some time interval: "*от… до*", "*с… до*": *от 12 до 15 лет/ с 12 до 15 лет* ('from 12 to 15 years'). The time lapse before the beginning of some period or event may be shown by construction "*за… до*": *за день до встречи* ('a day before meeting'). There are quite a number of other variants, and the same is valid for Locative syntaxeme. This variety may be described from the logical point of view starting from the locomotion in the three-dimensional space and after this model the moments and periods of time are structured. These variants are ***syntaxeme's subtypes***.

Some prepositional constructions denoting the same thing use different morpho-syntactic schemes, which separate the denoting objects into two or more groups. The most obvious example of this division is use of frequent prepositions "в" ('in') and "на" ('on') in Locative syntaxeme. The idea of localization of some event or state in a place is expressed more often [4000 IPM] with "в" ('in'): *сидеть/ работать в саду/ комнате* ('to sit/ work in the garden/ room') and less often [2500 IPM] with "на" ('on'): сидеть/ работать на стуле/ полу ('to sit/ work on a chair/ floor'). The opposition of meanings between "в" ('in') and "на" ('on') in this syntaxeme is usually described by the three-dimensional type of the concrete object designated by the dependent noun and the idea of possible inclusion into its inner space for the former, and the idea of contiguity and support for the latter [9]. Some objects designated in the locative syntaxeme may give the both opportunities and thus used with two prepositions: *в столе/ на столе* ('in/inside the table' / 'on [the surface of] the table'), but very often there is the only one possibility: *в городе* ('in the town'), *на полу* ('on the floor'). Sometimes

the division is fulfiled on the base of the linguistic ethno-specific classification, and we see controversial object specification in different languages: *на дереве* (‘in the tree’), *на улице* (‘in the street’, Czech ‘v ulici’) and suchlike. Anyway this division gives grounds for syntaxeme’s subtypes as well.

So we use for tagging the hierarchical system with topmost semantic rubrics, then corrected set of syntaxemes, and then detailed level of syntaxeme subtypes. This tagging groups correlate with corpus frequencies of realization in texts. As we show in the previous section that exact frequency values predispose to alter according to stylistics and thematic text specificity, then we will find out the perceptible dissimilarity between frequencies observed. We assume that on the higher levels of the syntaxeme’s system the main impact is produced by the grammatical resource in contrast to the lexicalized and more logic level of syntaxeme’s subtypes. The grammatical features are not purely conscious [10] that explain some logical discrepancies mentioned above in examples. We will use the idea of grammatical structuring proposed by R.Jackobson [11] that there is an opposition between the marked *indicative* grammatical category specifying some feature and its unmarked counterpart which may express opposite feature on a par with that of the indicative’s. For corpus statistics this idea has a by-product: unmarked categories are more frequent than indicative ones due to the partial ambivalent use. The surplus is not identical to the original amount but about its half so the proportion of realization of indicative and unmarked grammatical features roughly will be equal to 1:1,5. The difference of frequencies in this case is rather evident.

CONCLUSION AND FUTURE WORK

In this paper, we present preliminary results of our research aiming at the construction of the quantitative grammar of prepositional construction based on statistics from corpora of modern Russian texts. Its first stage requires the specification of ontological concepts used for structuring these multifaceted set. We put as a foundation a set of syntactico-morphological units – syntaxemes – which represent a union of a preposition and a noun case form with its particular meaning attached to some real or implied governor word. Syntaxemes with a mutual semantic component are grouped into semantic rubrics. Notions from both levels are interpreted as the grammatical representations of pertinent meaningful oppositions with some marked member and its ambivalent counterpart in Jakobson’s sense. Oppositions are reflected in the corpus statistics augmenting substantially the corpus frequency of the ambivalent unit, which may be used for prepositional construction alignment and figuring out their hierarchy. An example of such frame is shown in the paper “Prepositional Grammar Component for Syntactic and Lexical Disambiguation in Russian Based on Corpus Statistics” in this volume.

Syntaxemes are further subjected to the more logical classification producing syntaxemes’s subtypes which have a lexical origin, may be expressed with a help of secondary prepositions, and are infrequent usually.

It is shown in the paper that corpus frequencies are affected by stylistics and the thematic balance of the corpus processed. The more precise data on this issue will be obtained in a series of planned experiments with a variation of communication goals of the text and a number of topics presented.

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".

REFERENCES

[1] Schneider N., Hwang J.D., Srikumar V., Green M., Suresh A., Conger K., O'Gorman T., Palmer M., A Corpus of Preposition Supersenses, The 10th Linguistic Annotation Workshop, Germany, pp. 99–109, 2016.

[2] Dahlmeier D., Hwee T.Ng, Schultz T., Joint Learning of Preposition Senses and Semantic Roles of Prepositional Phrases, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, pp. 450–458, 2009.

[3] Zolotova, G.A., Syntactical Dictionary: a Set of Elementary Units of Russian Syntax, 4th edition, Moscow, 2011.

[4] Azarova I., Zakharov V., Temporal Aspects Expressed by Russian Prepositional Phrases, SGEM International Multidisciplinary Scientific Conference on Science and Arts, vol. 5, issue 3.6, 74, 2018, pp. 571–577.

[5] O'Hara T., Wiebe J. Exploiting Semantic Role Resources for Preposition Disambiguation, Computational Linguistics, vol. 35, issue 2, pp. 151–184, 2008.

[6] Skoblikova E.S., The Role of Vocabulary in Phrases with a Governed Component, Essays on the Theory of Phrases and Sentences, Kuibyshev, Russia, pp. 25-46, 1990.

[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.

[8] Tesnière L. Basics of Structured Syntax, Russian translation by V.G. Gak, Moscow, 1988.

[9] Herskovits, A., Semantics and Pragmatics of Locative Expressions, Cognitive Science, vol 9, pp. 341-378, 1985.

[10] Steblin-Kamenskiy M.I., Controversial Issues in Linguistics, Moscow, 1974.

[11] Jakobson R.O., Selected Woks, Moscow, 1985.