PREPOSITIONAL GRAMMAR COMPONENT FOR SYNTACTICAL AND LEXICAL DISAMBIGUATION IN RUSSIAN BASED ON CORPUS STATISTICS

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

ABSTRACT

Grammatical description of the sentence generation for the particular language is usually split into special morphological and Syntactical modules applied autonomously: the variance of morphological forms posed into the enumeration of constructional augmentations produces the enormous list of possible expositions of structural complexity paying no attention to the statistical plausibility of a construction in question. The usual method of reducing the complexity score of the Syntactical construction is to put morphological block inside the Syntactical one thus determining item structures strictly possible for this entity, to take into consideration the preferences of item occurrences.

The Russian prepositional constructions are the clear case of exuberant variability of the structural complexity in case we are to interpret the meaning of the govenee nouns, its syntactical semantics and the governor element – some full word in a sentence or a predicative or nominal centre. In Russian the ambiguity of interpretation of a prepositional construction is formed by several meanings of primary prepositions plus several noun forms with different senses combined with the preposition plus possible difference of semantic classes implied by the govenee nouns.

We construct an ontology for Russian prepositional constructions based on the corpus statistics and propose a sample from the grammatical module aimed at the analysis of the above mentioned structural variables.

Keywords: *grammatical generation, prepositional construction, Russian language, corpus statistics, prepositional meaning*

INTRODUCTION

We present in this paper the prototype variant of the grammar description of the Russian prepositional constructions. Text grammatical analyses is usually split into special morphological and Syntactical modules, though the autonomous grammatical interpretation produces the combination of errors on different levels. The semantic module may be helpful in disambiguation of grammatical constructions as well as lexical choice that is the consequence of lemmatization procedure, however, it is not clear in which terms to describe the semantics of the sentence and the text as a whole. We will use a AGFL formalism from the group of the affix generative grammars [1] which allows to insert hierarchy of categories and their values. They

may be syntactical, morpho-syntactical, and semantico-syntactical. In this paper we concentrate our efforts to show in what manner to introduce the interpretation of the latter type for Russian prepositional constructions.

The identification of semantic units and their relations to each other is an essential part of the automatic text analysis, though the recent indulgence to linguistic processing from neural network methods in natural language processing turned out to be a deadlock in near future. The real effectiveness will be in applying tactics of neural networks to the strategy of semantico-syntactical analysis which is indispensable if we want to extract information or content from the text.

Prepositional constructions are the crucial part of syntactical and semantico-syntactical automatic text analysis. The first problem is the prepositional phrase (PP) attachment, the second – interpretation of relations between the governor word for PP attachment and nouns or pronouns in the prepositional construction (governees).

Prepositions in the Russian language for quite a long time remained without the scrutiny of specialists in automatic text analysis. In information retrieval systems, they were included in "stop words" lists, which prevented their use in search models of information retrieval. Indeed, from the point of view of information retrieval, they, as a rule, can be neglected, because of their “low” nominativity, that is, they are not semantic identifiers of the document content. However, for semantically oriented analysis of the text, they are certainly important, since they convey certain semantic-syntactical relations between content words, clarify characteristics of a predicate, space-time specifications of propositions, etc.

CORE GROUP OF RUSSIAN PRIMARY PREPOSITIONS

In [2] in these proceedings we presented the core group of Russian primary prepositions which we use in order to illustrate our method of prepositional ontology construction and its use in the affix generative grammar [1]. They are as follows: “в” (‘in’), “на” (‘on’), “с” (‘with’), “по” (‘by’), “к” (‘to’), “из” (‘from’), “у” (‘at’), “за” (‘behind’), “от” (‘from’), “о” (‘about’). The preposition “в” (‘in’) is the most frequent. We see that according to statistics given in [3] any preposition tends to vary its frequency according to the stylistic and thematic corpus balance, though “в” (‘in’) has never moved from the highest rank. This gives us a clue that the distribution of semantic prepositional groups in corpus contexts may be the outline of grammatical oppositions presented in the semantic continuum of prepositional constructions.

We described in [2] that prepositional ontology has hierarchical structure. The most abstract concepts are semantic rubrics, which are realized by means of syntaxemes – the minimal Syntactical morphological prepostional constructions with particular meanings. Syntaxemes are further detailed into subtypes, which convey lexico-grammatical meanings and may be expressed with secondary prepositions in a variety of textual forms. Notions from the two topmost levels of ontology are of grammatical nature, that require the special approach. In [2] we posit a quantative interpretation of Jakobson’s idea of the indicative categories [4] that some

approximation to the prepositional frequency ratio “1.5 to 1” may be interpreted as a manifestation that lesser member of this pair has some grammatical markedness.

As a matter of fact we do not know the exact number and nature of semantic rubrics. The same indeterminacy exists in relation to the number and realization of syntaxemes. The syntaxemes mentioned in [5] may be looked as a basis for this list, but they were described in the frame of the functional approach without reference to the corpus statistics and generative grammar perspective. The number of prepositional senses, for example, “в” (‘in’) in the Russian explanatory dictionary [6], is enormous, and they have the same hindrance: there is no statistical assessment of all variants, and their granularity level is an issue of lexicographical principles.

We take the list of semantic roles as a suggestion for semantic rubrics: for example, relative frequencies for semantic roles from the annotated English corpus Penn TreeBank [7]: subject (.35), temporal (.113); locative (.075); direction (.026); manner (.021); purpose (.017); extent (.010).

We look for the abstract distribution of senses for the most frequent preposition “в” (‘in’) in the random sample of corpus contexts in the manner which may align presented frequencies in consent with proportions for indicative categories. We distinguish the following rubric’s frequencies on the basis of our balanced corpus: *localization* 8090 IPM (instances per million corpus tokens); *temporative* 5090 IPM; *objective* 3240 IPM; *derivative*, that is, secondary prepositions and phrasal expressions 2080 IPM; *qualificative* 1160 IPM; *partitive* 690 IPM; *quantificative* 430 IPM. As corpus frequencies may vary in correlation with stylistic and thematic corpus balance, the proportional numerals may be more informative: localization (.35); temporative (.22); objective (.14); derivative (.09); qualificative (.05); partitive (.03); quantificative (.02). The diagram of semantic rubric distribution for preposition “в” (‘in’) is shown in Fig. 1 below.

We are to clear some points. Firstly, rather small portion of contexts expressing objective concepts is explained by the fact that they are conveyed by means of case forms: the nominative renders a subject, the accusative – an object, the dative – an addressee, the ablative – an instrument, though interpretation of these case forms are not thus straightforward, but this issue is beyond the boundaries of this paper. Secondly, a small portion of propositional constructions realize syncretically several rubrics. For example, *лежать в нескольких метрах* (‘to lie a few meters away’) – localization plus quantificative; *попасть в чужие руки* (‘to fall into the wrong hands’) – localization + objective + set phrase.

Syntaxemes of the localization rubric include proper locative, expressed by “в” (‘in’) plus the locative case form [3700 IPM]: *сидеть в саду* (‘to sit in the garden’), *гулять в лесу* (‘to have a walk in the forest’), the same meaning may be expressed by the preposition “на” (‘on’) with the locative case [1800 IPM] as well: *сидеть на стуле* (‘to sit on the chair’), *дышать воздухом на веранде* (‘to breathe air on the veranda’). In [8] the difference was connected with idea of “inclusion” for the former in the contrast “support” and “contiguity” for the latter. We point out that this “classification” is purely linguistic because a veranda is three-dimensional object and

a sitting person is inside it. Moreover, both syntaxemes may be used: *сидеть в кресле на веранде* (‘to sit in a chair on the veranda’), *висеть в бильярдной на стене* (‘to hang in the pool room on the wall’). This fact is usually taken as an evidence that they have different roles, and we see that places of localization are included into each other, but which one into which? So we will consider the first variant to be the locative1, and the second – the locative2 because it concedes in frequency. We see the same parallelism in the directive syntaxeme denoting the end point of the travel trajectory: “в” (‘in’) plus the accusative case form [3700 IPM]: *прийти в сад* (‘to come to the garden’), *положить в шкаф* (‘to put in the closet’), and “на” (‘on’) with the accusative case [1570 IPM]: *поставить на стол* (‘to put on the table’), *прийти на веранду* (‘to come to the veranda’). The sequence of the directive and locative has a standard interpretation: the locative in postposition is an attribute for the directive: *приехать в город на Неве* (‘to come to the city on the Neva River’), *приехать на виллу в Мексике* (‘to come to the villa in Mexico’). The sequence of directives is as ambivalent as that of locatives: *отвезти в деревню на виллу* (‘to take someone to the village to the villa’), *отправиться на дачу в Барвиху* (‘to go to the cottage in Barvikha’).

The temporative rubric concedes in frequency to the localization rubric that is usually interpreted as “time is space” [7], that is, temporal syntaxemes are structured on the localization model. The concept of “time” may be expressed as the deictic category referring to the time of a speech, that is usual for verbal predicates, this is an “absolute” time characteristic. If the referential point differs from the moment of a speech act, it is a “relative” time characteristic [4]. Due to the time metaphor we can see “imagined movement” on the time scale presenting the continuum of our experience in which events pass from the future through the present to the past. The complete isomorphism is impossible but the opposition of locative and accusative case forms in temporative construction reminds the selectional rules for included and supported object, though they are more simple: the temporative syntaxeme with a locative case form [2780 IPM] is used for nouns denoting months, years, longer periods such as a century or an epoch: *в пятом году* (‘in the fifth year’), *в октябре* (‘in October’), *в 19 веке* (‘in the 19th century’), *в неолите* (‘in the Neolithic’). Temporative syntaxeme with an accusative case form [2300 IPM] is applied to abstract nouns denoting time or to quantified expressions of hours of day and night, days of the week: *во время зимовки* (‘during wintering’), *в период нереста* (‘during spawning ’), *в пять часов утра* (‘at five in the morning’), *в пятницу* (‘on Friday’). The sequence with a temporative with the locative case is impossible, the member in the postposition is expressed by a genitive case form: *в октябре 1995 года* (‘in October 1995’), though accusative temporatives may be concatenated: *в воскресенье в девять часов утра* (‘on Sunday at nine o'clock in the morning’).

The temporative with the preposition “на” (‘on’) used with an accusative case with a frequency [930 IPM] comparable with that of grammatical constructions, it is a quantified temporal period (it is syncretic with the quantificative syntaxeme below): *на 5 дней* (‘for 5 days’), *на 10 веков* (‘for 10 centuries’) or so-called bound constructions with main nouns denoting time periods with some attribute, usually adjectival: *на ближайшее десятилетие* (‘for the next decade’), *на данное время* (‘at this time’). Temporal constructions for the preposition “на” (‘on’) with locative

case form are quite a few, it's better to regard them as set phrases: *на будущей неделе* ('next week'), *на данном этапе* ('at this stage').

The objective prepositional rubric includes various types of objects such as an object of action, object of thought and nomination, an addressee or participant, etc., further specification is possible on the subsyntaxeme level. This rubric is "marked" according to the frequency of corpus realization in the contrast with the previous one. It's not a common view but this sense domain is structured on the model of localization [9]. There are as well the parallel syntaxemes for the prepositions "в" ('in'), "на" ('on') with a locative case form (2300 IPM and 1120 IPM correspondingly) and an accusative case form (930 IPM and 2000 IPM). The former objective syntaxeme is more active for "в" ('in') than for "на" ('on'). There are several types of objects characteristic for the first preposition: an object of perception: *видеть в старых фильмах* ('to see in old movies'), an object of application: *использовать в технике* ('to use in engineering'), the object linked to the abstract noun replacing the direct object of the verb: *изменения в анализах крови* ('changes in blood tests'). As for the second preposition "на" ('on') there is some vehicle: *ехать на велосипеде* ('to ride a bike') or device: *спуститься вниз на веревке* ('to come down on the rope'). G.Zolotova's [5] proposed for this construction an instrumentive or mediative syntaxeme, thogh there is a wide range of object types: *играть на гитаре* ('to play guitar'), *называться на иврите* ('to be called in Hebrew'), *выставить кандидатуру на выборах* ('to run for election').

The objective syntaxeme with an accusative case form is more active for the preposition "на" ('on'). There are used verbs of communication: *отвечать на вопрос* ('to answer the question'), emotional verbs: *обижаться на власть* ('to take offense at the authority'), metaphorically shifted travel verbs: *заступить на вахту* ('to stand on watch'), *пойти на ваши условия* ('to go to your terms'). This syntaxeme for the preposition "в" ('in') is one of the most infrequent, it is attached to the verbs of transfiguration: *превратиться в густую массу* ('to turn into a thick mass'), *превратить жизнь в театр* ('to turn life into theater') or sometimes social: *выбирать в органы власти* ('to elect to the authorities'), *назначить на должность руководителя* ('to appoint a manager').

The next syntaxeme rubric – the derivative – is a heterogeneous group incorporating secondary prepositions and set phrases comprising the primary prepositions as a component of their structure, secondaries with the pronominal specification transforming into adverbial constructions. In this rubric the division into constructions with locative or accusative case forms is not thus important, so we take them as a whole. The examples for the prepositions "в" ('in') [2000 IPM] are: *в виде таблетки* ('in pill form'), *в области науки* (in the field of science), *проявить себя в деле* ('to prove oneself in business'), *иметь в виду* ('to keep in mind'), *в знак благодарности* ('in gratitude'), *в пользу рекламодателя* (in favor of the advertiser),*приводить себя в порядок* ('to trim oneself up'), *бросаться в глаза* ('to strike the eye'). The preposition "на" ('on') is more active in this rubric [200 IPM] collating with its total frequency: *шевелиться на ветру* ('to stir in the wind'), *оказаться у всех на виду* ('to be in public view'), *принять на борт судна* ('to take aboard ship').

The three infrequent rubrics – the qualificative, the partitive, the quantificative – have indiscernible frequency portion comparable the statistical error. They are syncretic with other syntaxemes, so some examples are given above. In the Fig. 1 we show diagrams showing proportion of corpus realization of semantic rubrics for the topmost preposition "в" ('in') as a whole and that of described syntaxemes for "в" ('in') and "на" ('on').

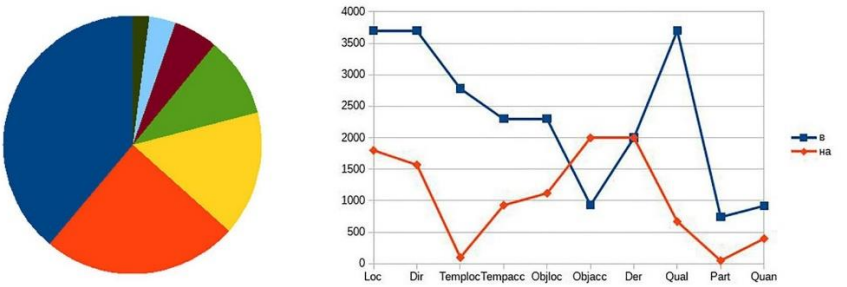

***Fig. 1**. Corpus frequency proportions of the semantic rubrics for the "в" ('in') (there are diminishing progression of proposed semantic rubrics on the left diagram, these are localization, temporative, objective, derivative, qualificative, partitive, quantificative and corpus frequency proportions of inserted syntaxemes for "в" ('in') and "на" ('on') (the right chart).*

To understand the correct model for organizing syntaxemes we are to understand what information is necessary to specify on the syntaxeme's or syntaxeme type's level. For example, we use syntaxemes in the generative grammar formalism. The first problem is the main word (or dummy predicate) to which the syntaxeme analyzed is attached. A well-known example of ambiguity of PP attachment in English "I saw a man with a teleskope" in Russian is rendered unequivocally: *Я видел человека с телескопом* versus *Я видел человека в телескоп*. Naturally, there are ambiguous cases: *взять тетрадь в клетку* ('to take a squared notebook' versus 'to take a notebook into the cage') where the latter variant is hardly come to someone's mind. There is syntactical device of so-called redistribution of Syntactical links: when the prepositional construction is lineally divided from its governor verb by an object, the Syntactical link between a verb and PP is lessened and a link between a noun and PP becomes possible. This device gives a chance to appearance of nominal constructions with PP such as *картина на стене* ('a picture on the wall'). We stated above that for the goal of systematic analysis we interpret the governors of these constructions as dummy verbs.

CONCLUSION AND FUTURE WORK

The paper provides a construction grammar perspective to identifying meaning of prepositions in Russian. In order to solve natural language processing tasks, we need to learn how to uncover semantic relations in texts, especially in Russian a great number of them are conveyed by prepositions.

We are collecting a serias of prepositional constructions and arrange them according to frequencies of specified meanings in corpora of modern Russian texts. Different semantic aspects of prepositional constructions are described with semantic rubrics which are based on a notion of syntaxeme proposed by G. Zolotova. Our final goal is to create a corpus-based quantitative ontology of Russian prepositions.

The semantic rubrics presented in our approach help to organize rather vague prepositional meanings. Their affinity and difference may be explicated through the overlap of semantic classes of governing and subordinate words. The whole structure of prepositional frequencies that has not investigated so far and arrangement of semantic units expressed in text contexts are resources for the compilation of the quantitative prepositional grammar for Russian.

We are going to compile the first version of essential semantic rubrics to proceed in the outlined direction and to grasp the sense distribution for primary prepositions. Then we will assign the secondary prepositions to these sets. Thus we will check the initial hypothesis that the granularity of prepositional meanings are restricted by the meaningful diversity of secondary prepositions.

Further stages of our project include:

to clarify the set of syntaxemes for prepositional constructions referring to governors and governees semantic types on the base of corpus data;

to compile sets of prepositional constructions from corpora of different genres in order to discover the significant variation of statistical parameters;

to describe prepositional constructions in terms of predominant semantic classes and/or lexemes used as “governors”;

to list predominant semantic classes and/or lexemes used as “governees” for different semantic rubrics and/or syntaxemes;

to create a database of Russian prepositional constructions accumulating corpus material with statistical information obtained;

to compile rules of the hybrid generative grammar showing the use of prepositional phrases for expressing the comprehensive set of syntaxemes.

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] Koster C.H.A., Affix grammars for natural languages, Attribute Grammars, Applications and Systems, Springer, Berlin, Heidelberg, 1991.

[2] Azarova I., Zakharov V., The Outline of the Quantitative Ontology for Russian Prepositional Constructions, NordSci proceedings, pp. $$–$$, 2019.

[3] 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.

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

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

[6] Dictionary of the Russian language in 4 volumes, 3rd edition, vol. 1, Moscow, 1985.

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

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

[9] 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.