Semantic Analysis of Russian Prepositional Constructions

Victor Zakharov[1], Kirill Boyarsky[2], Anastasia Golovina[1], and Anastasia Kozlova[2]

[1] Saint Petersburg University
Universitetskaya emb. 7-9
199034 Saint Petersburg, Russia
v.zakharov@spbu.ru, st070508@student.spbu.ru
[2] ITMO University
Kronverkskiy av. 49
197101 Saint Petersburg, Russia
boyarin9@yandex.ru, stasia.kozlova@gmail.com

Abstract. The paper deals with semantics of Russian prepositions. We consider prepositional meaning to be the relation found in prepositional constructions where it should be regarded as a special type of relationship between content words. There is a rather small set of common prepositional meanings that we call syntaxemes after G.A. Zolotova that encompass the larger part of the Russian prepositional semantics as a whole. In this paper we propose a methodology of prepositional phrase extraction and syntaxeme identification based on text corpora and corpus statistics. Prepositional construction meaning is detected automatically using the SemSin parser. We demonstrate our methodology on constructions with the polysemous Russian preposition *через* (through).

Keywords: Russian prepositional constructions, preposition meaning, corpus statistics, parsing, semantic classes

1 Introduction

This paper presents a study on semantics of Russian prepositions. It is part of a larger project. In this paper, we demonstrate the way to identify preposition meaning on constructions with the polysemous Russian preposition *через*.

The preposition is a part of speech found in many languages. Russian linguistics divides this class into primary and secondary prepositions by origin as well as simple (one word) and complex (multiword) units by structure. Primary prepositions in particular are highly polysemous. For instance, the Russian preposition ‘with’ has 26 meanings in the Dictionary of the Russian Language [4] (11 meanings with the genitive case, 2 with the accusative one and 13 with the ablative). The majority of them are quite rare, in some cases the preposition is a part of an idiom.

Prepositional ambiguity is manifested in the complex nature of the prepositional meaning and in selective preferences of certain prepositions, depending

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2020, pp. 103–111, 2020. © Tribun EU 2020

on context. That alone makes the systematization of the prepositional class a very complicated and tedious undertaking.

Prepositions are often regarded as having no lexical meaning. However, we have an alternative view on prepositional semantics. We consider prepositional meaning to be the relation found in prepositional constructions where it should be regarded as a special type of relationship between content words. An additional factor in the proposed view on the prepositional meaning is the case of the prepositional governee. We believe that the preposition should be studied in conjunction with the associated case. It becomes possible, then, to speak of prepositional homonymy where every "preposition+case" pair is a homonym within the paradigm of a given preposition.

We have adopted the approach suggested by G.A. Zolotova [10] for the task of describing prepositional meanings. The preposition-case unit is regarded as a syntaxeme – the minimal lexical-grammatical construction expressing a certain meaning. Syntaxemes can be relatively autonomous, but usually they form blocks which attach themselves to notion words, mostly verbs. There are about 30 syntaxemes listed in Zolotova's dictionary. Different preposition-case units may form semantically comparable syntaxemes, which is why in this study we take the syntaxeme to mean the common semantic invariant of these units. The names of some syntaxemes (temporative, directive, instrumentive, etc.) correlate with the idea of syntactic-semantic roles that have been introduced by Ch. Fillmore with the idea of syntactic-semantic sentence description [5].

We also use the concept of the syntaxeme as a node in prepositional ontology. The notion of the syntaxeme was defined in the functional direction of traditional linguistic analysis, so we redefine it inside our own quantitative corpus approach [1].

In contrast to the classical linguistics focusing on the simplest units of different language levels, modern studies practice synthetic methods attempting to capture and describe the more complex language structures which integrate different language units: words, collocations, etc. In classical linguistic papers, prepositional constructions used to be described from the grammatical point of view and their semantics used to be neglected. Complex description and systematization of prepositional constructions demand elaboration of identification methods using manual and automatic techniques as well as analysis of their paradigmatic and syntagmatic features and quantitative analysis of their frequency and strength.

2 Russian Preposition *Через*

This article presents the methodology of semantic analysis of prepositions based on constructions with the Russian preposition *через* ('through, across, in, after').

Existing approaches to prepositional semantics description differ in their methodology, both formally and in their content. Deep research on the semantics of individual prepositions, including *через*, can be found in linguistic literature. Such is, for instance, the specialized comparative analysis of the semantics of *через* and its close synonym *сквозь* ('through') by V.A. Plungyan and E.V. Rakhilina [8]. Their paper presents an investigation on the semantics of these two prepositions based on their various aspects. Full descriptions of polysemous prepositions are provided in the form of semantic nets consisting of blocks of ready-made and constructible language material, including idioms. It can be said, however, that the description is provided in the terms of the construction-based approach practiced by us. Also introduced are the notions of "stable" and "flexible" word parameters which are descriptions of situations gained through inferring semantic characteristics of words in context.

However, most sources operate with a significantly simpler system of meanings. Wiktionary, for instance, lists 4 meanings of the preposition *через*:

1. *сквозь, поперёк* ◆ *Он помог слепой женщине перейти* **через** *дорогу на другую сторону.*
2. *поверх чего-либо* ◆ *Я легко перепрыгнул* **через** *забор.*
3. *по истечении некоего отрезка времени* ◆ *Перезвони мне минут* **через** *пять.*
4. *с помощью, посредством* ◆ *Оплата производится на почте при получении заказа,* **через** *Сбербанк либо по WebMoney.*

As can be seen from this example, the meanings of prepositions in explanatory dictionaries are typically expressed descriptively or by means of other synonyms, forming a "vicious circle". However, the same set of meanings could be interpreted as the transitive (1, 2), temporative (3) and mediative (4) syntaxemes as per the Syntactic Dictionary by G.A. Zolotova [10].

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

The **temporative** rubric is quite diverse. In some cases, the time specification appears to be relative: the time interval precedes some event, follows it or is simultaneous with it. Another variation conveys a sequence of events, which can be expressed with the preposition *через* 'in, after' with the accusative case: *прийти через день* ('to come in a day'), *произойти через 2 столетия* ('to happen after 2 centuries'). However, the corresponding Wiktionary definition of this meaning may need to be reformulated or divided into two, the temporative and the locative, as there exists a very similar meaning of *через* that refers to alternation in space: *деревья высаживают через 1.5 м в ряду* 'the trees are planted every 1.5 m in a row', *по всей длине не реже чем через два метра* 'along the entire length no sparser than every two meters'.

The **mediative** as a semantic rubric has a narrow and a wide interpretation. Generally, it is regarded as a particular semantic role in the predicate structure of a verb. In the narrow sense the mediative is understood as a means, that is, a substance or an object used during the performance of an action or a process. In a broader sense the mediative is a tool (the instrumentive meaning) and includes its material and abstract implementations [7]. In the Russian language both the mediative and the instrumentative are regularly expressed by the instrumental case form (*красить стены валиком* 'to paint walls with a paint roller', *рисовать картину красками* 'to paint a picture with paints'), however, we can observe more complex syncretic instances in the form of prepositional constructions.

Other syntaxemes of this preposition are rarely observed.

3 Methodology of Corpus Statistical Analysis

We have developed a procedure for describing the continuum of prepositional meanings basing on the corpus data starting from the bottom – that is, textual analysis of sense distribution in random context samples from different corpora. The stages of our procedure are as follows:

- Acquisition of sets of prepositional constructions from corpora of different types and different functional styles;
- Acquisition of a number of statistical characteristics for each preposition from corpora of different types and functional styles, namely:
  - ipm in a corpus (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.

The semantic-grammatical analysis of relations between lexical items certainly cannot be performed entirely automatically and requires participation of linguists. To estimate the percentage of each meaning of the preposition *через* we have resorted to expert evaluation of the selected constructions (contexts). Those were annotated according to the meaning realised in the given context, for example:

- *он наблюдал бы прохождение Юпитера* **через** *диск Солнца* → transitive;
- **через** *минуту дверь открылась* → temporative;
- **через** *богослужение мы действительно достигаем истины* → mediative.

The Russian National Corpus (RNC) was used as the base material source.

We found that out of the 5 meanings ascribed to the preposition *через* by G.A. Zolotova [10] only 3 appear to be present in real texts (Fig. 1).

All of the observations suggest that the bottom-up corpus-based approach is imperative in the task of studying preposition semantics. However, the context window method used originally in the study was discovered to be insufficiently effective as the actual governor and governee, which are crucial in prepositional phrase identification, are not always captured by the window. The quality of automatically extracted prepositional constructions could be improved through the use of full syntax parsing. Furthermore, it is impossible to process and annotate large text arrays without using reliable automatic analysis tools.

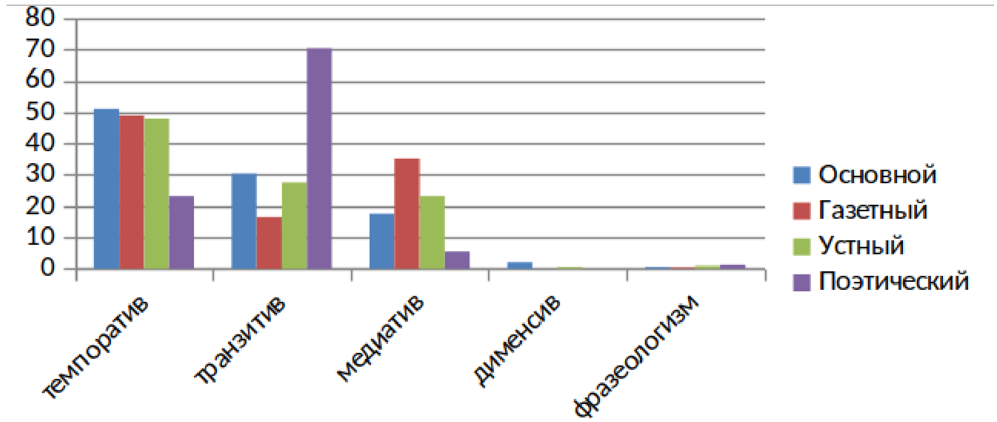

Fig. 1. Ratios of *через* meanings in RNC subcorpora

4 Automatic Parsing of Prepositional Phrases with *Через*

Unfortunately, the choice of openly available tools capable of performing a detailed semantic analysis is very limited for Russian. Identification of prepositional phrase semantics is one of the most difficult tasks for a parser due to some specific features of prepositions. Firstly, prepositions are not declinable, and many consist of just one or two letters. That makes it impossible to rely on the morpheme (root or word ending) during the analysis. Secondly, a high degree of governee homonymy does not allow for unambiguous semantics identification through the analysis of just the prepositional phrase and often demands a full sentence analysis.

In order to study semantic meanings of prepositions we used the SemSin parser [3], which builds a dependency tree for each sentence and detects types of relations between its nodes. The parser relies on a semantic-syntactic dictionary and a classifier, both of which are extensions of the semantic dictionary by V.A. Tuzov [9]. The dictionary currently contains about 200 000 lexemes belonging to 1700 classes. An important element of the system dictionary is a table of prepositions containing over 2200 combinations of semantic categories of nouns with which prepositions can interact as well as the names of relations between governors and prepositional phrases.

Each sentence undergoes morphological analysis involving tokenization, after which the lexical analyser transforms the linear sequence of tokens into a dependency tree with the help of a system of production rules [2].

In order to test the accuracy of the automatic semantics identification we have used the parser to analyse corpora of around 100 000 tokens each of texts representing different functional styles: newspaper and magazine articles, fiction, scientific papers and texts, legal documents and oral speech transcripts. All the sentences containing the preposition *через* (50 to 100 depending on the corpus) were then selected for further inspection. The correctness of prepositional phrase governor and governee detection was checked by experts. In some complex cases parsing results were checked against those made by the ETAP-4 parser [6].

The following sentence is an example from the Russian National Corpus: *Не хочется, чтобы через определённый промежуток времени у нашей молодёжи настолько поменялись приоритеты* 'It would be undesirable if the priorities of our youth changed that much in a certain span of time'. The word *промежуток* 'span' has two semantic meanings: temporal and locative, which is reflected in the RNC annotation. In the analysis of this sentence with ETAP-4 the prepositional phrase is linked to the governor, the verb *поменялись* 'changed', through the "adverbial adjunct" relation, which does not help to resolve the semantic homonymy. However, the *Когда-relation* ('When') of the tree built by SemSin (Fig. 2) unambiguously detects the meaning of the preposition as temporative and allows for the temporal interpretation of the word *промежуток* 'span' only.

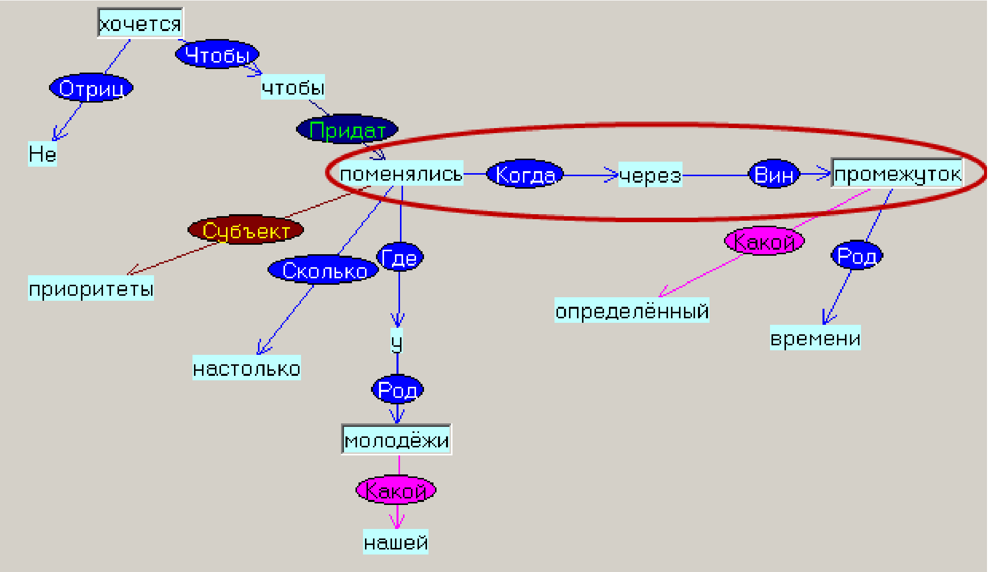

Fig. 2. The preposition *через* in the temopative meaning

Cases in which the preposition is found at a significant distance from its governor or governee present a considerable challenge for automatic parsing. This is especially characteristic of legal texts, such as laws, statutes, etc. For instance, in the following sentence fragment (full sentence length: 71 words) *при наличии... необходимого оборудования* **передавать** *в соответствии со стандартными процедурами Всемирной метеорологической организации в основные международные синоптические сроки* **через** *береговой* **радиоцентр**... *оперативные данные...* 'in the presence of... necessary equipment **transmit**... live data in accordance with the standard procedures of the World Meteorological Organization in the main international synoptic hours **through** the coastal **radio centre**' the preposition *через* is located 14 words away from its governor *передавать* 'transmit'. ETAP-4 wrongly links the prepositional group to the word *сроки* 'hours', while SemSin correctly detects the true governor (Fig. 3). The fact that the governee *радиоцентр* 'radio centre' belongs to the semantic class of establishments allows to infer that the preposition has the mediative meaning in this context.

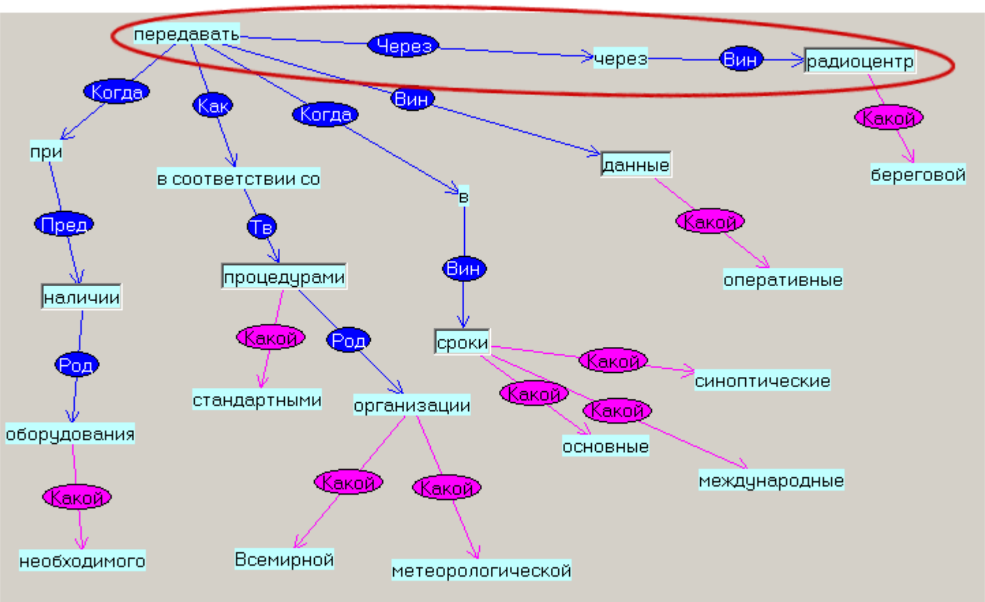

Fig. 3. The preposition *через* in the mediative meaning

Although the maximum distance from a preposition to its governee is much shorter than to its governor, issues do occur in some cases, especially when there are punctuation marks between the preposition and its governee, like in the case of parenthetical phrases marked off by commas, e.g. *Дина понуро шла через широкий, как площадь, двор гаража* 'Dina was walking gloomily through the wide as a square court'. ETAP-4 considers the governee to be the word *широкий* 'wide'. SemSin locates the governee *двор* 'court' correctly while also resolving the semantic homonymy of the word *двор* (plot of land vs. social category). Therefore, the semantics of the preposition is identified as transitive.

The results of the accuracy evaluation of the automatic formation of prepositional phrases with the preposition *через* and the detection of the relation type, e.g. the semantics, by the parser are provided in Table 1.

Thus, we can conclude that the SemSin parser provides prepositional phrase semantics detection of sufficient quality. That being said, the accuracy could be improved by means of improving parsing rules and detalization of interactions between prepositions and nouns of various semantic classes.

One of the objects of our current research is the preposition *с* ('with', 'from'). The phase requiring the most detailed examination is the extraction of syntaxemes, which are listed below (Table 2).

Table 1. Accuracy of formation of prepositional phrases with preposition *через* by the parser

Text type	Governor/governee accuracy	Relation type accuracy
Oral transcripts	87 %	87 %
Newspaper articles	87 %	85 %
Scientific texts	92 %	73 %
Fiction	92 %	87 %
Legislative texts	65 %	78 %

Table 2. Syntaxemes of the preposition *с* ('with', 'from')

Syntaxeme	Examples
Directive	*съёмка со спутников* 'images from satellites'
Instrumentive	*кормить с ложки* 'to feed from a spoon'
Source	*перевод с китайского* 'translation from Chinese'
Object	*помочь с задачей* 'to help with a problem'
Comitative	*пирог с начинкой* 'pie with a filling'
Cause	*закричал с радости* 'cried for joy'
Comparison	*длиной с полметра* 'half a meter long'

The least frequently occurring syntaxemes of the preposition *с* are "cause" and "comparison", which, in combination with their syntactic complicacy, makes them the most difficult for parser identification. Still, with some of the grammatical particularities determined, it is highly possible to bring the level of automatic identification above the current threshold.

5 Conclusion and Future Work

All of the observations presented in the paper suggest that our approach is suitable for use in the task of studying preposition semantics. Further stages of our research include expansion of the application of the methodology presented in this paper to other prepositions.

Additionally, research on the prepositional use in fixed phrases and idioms has been started.

To improve the quality of extracted constructions and to reduce human participation and labor costs a syntactic parser operating on a base of semantic categories is to be used in further studies. We strive for automatic identification of preposition meanings as demonstrated in the current paper on the preposition *через*.

The conducted research shows that the SemSin parser successfully finds prepositional groups with the preposition *через* as well as others and determines the type of semantic connection with a high degree of accuracy. However, to automatically determine the semantics of prepositions that have a greater variety of semantic meanings, additional research is needed on the compatibility of prepositions with nouns.

Acknowledgements. This work has been supported by the Russian Foundation for Basic Research, project No. 17-29-09159.

We express our heartfelt gratitude to Irina Azarova, Andrei Masevich and Anna Moskvina for their help in processing corpus materials.

References

1. Azarova, I., Zakharov, V., Khokhlova, M., Petkevič, V.: Ontological description of Russian prepositions. In: CEUR Workshop Proceedings, 2552, pp. 245-257 (2020).
2. Boyarsky, K., Kanevsky, Ye. A system of production rules for the building of a sentence syntax tree [Sistema produkcionnykh pravil dlya postroyeniya cintaksicheskogo dereva predlozheniya]. In: Applied linguistics and linguistic technology [Prikladna lingviskika ta lingvistichni tekhnologii], MegaLing-2011. Kyiv, Dovira, pp. 73-80 (2012).
3. Boyarsky, K., Kanevsky, Ye. Semantic-syntactic parser SemSin [Semantiko-sintaksicheskiy parser SemSin]. In: Scientific and technical bulletin of information technologies, mechanics and optics [Nauchno-tekhnicheskiy vestnik informacionnykh tekhnologiy, mekhaniki i optiki]. Vol.15. No. 5, pp. 869-876 (2015).
4. Dictionary of the Russian language: in 4 volumes / Russian Academy of Sciences, Institute of Linguistic Research. 4th edition, stereotyped. Moscow: Russian language: Polygraph resources (1999)
5. Fillmore, Ch. J. The Case for case. In: Universals in Linguistic Theory. Bach and Harms (Ed.). New York: Holt, Rinehart, and Winston, pp. 1-88 (1968).
6. Language processor ETAP-4 [Lingvisticheskiy protsessor ETAP-4]. URL: `http://proling.iitp.ru/ru/etap4` (last access: 28.10.2020).
7. Mustajoki, A. Theory of functional syntax [Teorija funtsionalnogo sintaksisa]. Moscow (2006).
8. Plungyan, V.A., Rakhilina, E.V. Function word polysemy: the prepositions *через* and *сквозь* [Polisemiya sluzhebnykh slov: predlogi cherez i skvoz]. In: Russian studies today [Rusistika segodnya]. No. 3, pp. 1–17 (1996).
9. Tuzov, V.A. Computational semantics of the Russian language. Saint Petersburg: SpbU publishing (2004).
10. Zolotova, G.A.: Syntactic Dictionary: a set of elementary units of the Russian syntax [Sintaksicheskiy slovar': repertuar elementarnykh edinits russkogo sintaksisa]. 4th edition. Moscow (2011).