Russian Secondary Prepositions: Methodology of Analysis

Victor Zakharov[0000-0003-0522-7469], Anastasia Golovina[0000-0002-9239-2050], Elena Alexeeva[0000-0002-3841-3199], Vadim Gudkov[0000-0003-2152-9598]

St. Petersburg University, Saint Petersburg 199034, Russia
v.zakharov@spbu.ru

Abstract. The present study proposes a methodology of a corpus-based analysis of Russian secondary prepositions, primarily focusing on multiwords. Secondary prepositions are units motivated by content words (nouns, adverbs, verbs), which may be combined with primary prepositions to form multiword prepositions (MWPs). 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. A strict division between secondary multiword prepositions and equivalent free word combinations is not specified. This presents an issue in the task of building a language model as compound prepositional units are commonly mislabeled as free combinations or are labelled inconsistently, thus leading to parsing errors with far-reaching consequences. Our larger study aims at solving this problem by identifying, describing and eventually formalizing the full inventory of Russian MWPs, which demands a special corpus-based research. This paper is devoted to statistical analysis of the use of secondary multiword prepositions in corpora using prepositions expressing causal relations as the base material. 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.

1 Introduction

This study is part of a large project with the goal of creating the first corpus-driven semantic-grammatical description of Russian prepositional constructions. A number of tasks are planned to achieve this goal, with the following at the base level:

- development of a high-precision language model for extracting prepositional constructions;
- development of a high-precision language model for morphological analysis of structural elements.

However, the creation of such models, or even use of the existing ones, is hampered by one circumstance. While primary prepositions in Russian are well-studied and described (most importantly, there exist exhaustive lists of these entities), the same cannot be said about secondary prepositions. In fact, even the volume of this subclass is unknown. In language models available for Russian, secondary prepositions are handled inconsistently; more on this in Section 2.2. At the same time, it is obvious that accurate and consistent annotation is crucial for building a language model that represents real language use (which is, after all, its main purpose). We believe that secondary prepositions, and MWPs in particular, should be given more attention in language model development. We are addressing this by building our own models for the designated purposes with special focus on secondary prepositions as part of our project. However, to begin with, we must first find out what these units are characterized by as a subclass and what entities it includes as their detailed formal description is not yet available in Russian linguistics.

Generally speaking, the preposition is a common part of speech found in many languages. Its frequency naturally varies but tends to be quite high. In Russian, prepositions have been found to constitute on average 10% of all tokens [1]. That makes the preposition a regular constituent of the language system. Consequently, automatic recognition and analysis of prepositions is crucial in numerous NLP tasks, such as prepositional phrase attachment [2], syntactic role acquisition, word sense disambiguation for discriminating between senses of polysemous word [3], information retrieval [4] and automatic ontology extraction [5].

Russian linguistic tradition implies subdivision of the class of prepositions into primary and secondary ones by origin as well as simple (one word) and complex (multiword) units by structure. While the primary preposition subclass is relatively well-studied and sufficiently documented, secondary prepositions, especially multiword ones, have not enjoyed equal attention in linguistic literature despite making up a large part of prepositions as a class. The reason for the ambiguous status of most of these units lies largely in the issue of their identification and the overall lack of agreement over the base features of a preposition among linguists.

The vague and complex prepositional semantics lies at the centre of many disputes on the nature of prepositions. Primary prepositions are highly polysemous. For instance, the Russian preposition в ‘in’ has 23 meanings in the Dictionary of the Russian Language [6]. The majority of them are quite rare, in some cases the preposition is a part of an idiom. The meanings of prepositions in explanatory dictionaries are usually expressed descriptively or by other synonyms, forming a “vicious circle”. Prepositional ambiguity is manifested in the complex nature of the prepositional meaning and in selective preferences of certain prepositions, depending on context. That alone makes the systematization of the prepositional class a very complicated and tedious undertaking.

The existing schemes of lexical and syntactic structuring of the Russian prepositional system found in [7–9] have led us to the conception of the prepositional ontology. The main problem of such an ontology is its inherent inconsistency stemming from the nature of the base material since the ontological structure presupposes logical analysis of concepts, while prepositions are usually interpreted as elements that have no lexical meaning. Therefore, a prepositional ontology has a significant difference from a classic

one. It is an ontology of lexico-grammatical relations implied in prepositional constructions. Thus, in agreement with [9] 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.

We regard this notion as a semi-grammatical language component linking fuzzy lexico-semantic word classes by the hierarchical set of grammatical relations. These relations are established by the combination of a particular preposition, the semantic type of the lexeme attaching the prepositional construction, and the semantic class and grammatical form of the governee (dependent). An additional factor in the proposed view on the prepositional meaning is the case of the governee. We believe that the preposition should be studied in conjunction with the associated case as the case is often the key factor in identifying the meaning of the preposition in identical contexts (compare e.g.: *маршировали в зале* (Loc) vs. *маршировали в залу* (Acc) "marched in the hall" vs. "marched into the hall").

We believe that such an ontology cannot be built from top to bottom. We advocate a data-based, bottom-up corpus approach and focus on usage models. A similar approach was adopted in building the dictionary of English preposition templates (PDEP) [10]. The connections and relationships between the objects of our ontology (syntaxemes), in turn, can also be identified by means of the corpus approach. Such relations are usually calculated using the vector space model [11]. Our approach is closer to [12], where machine learning is used. We rely on corpus statistics. The corpus-based semantic and grammatical description of Russian prepositional constructions uses empirical data from various corpora of the modern Russian language to identify and then formalize the main ontological semantic patterns of "prepositional grammar".

In [13] we suggest that the ontology of prepositions has a hierarchical structure. The most abstract concepts are semantic rubrics that are implemented in the form of syntaxemes on the second level. This term was proposed by G.A. Zolotova [9] as a designation of minimal syntactic-morphological prepositional constructions that have certain meanings. Syntaxemes can be divided into subtypes (subsyntaxemes) that convey lexical and grammatical meanings and can be expressed by primary or secondary prepositions in various text forms. Concepts from all ontological levels have a primarily grammatical nature, which requires a special quantitative and grammatical approach for further structuring.

A prepositional syntaxeme is characterized by a morphological arrangement (preposition plus noun case form) that has a unity of form and meaning that functions as a constructive and meaningful component of a phrase or a sentence. Syntaxemes in Zolotova's original description resemble semantic roles or specification of arguments: locative, temporal, directive, destinative, correlation, quantitative, mediative, qualitative etc. A typical syntaxeme is expressed in several prepositional templates.

An important step in the building of the proposed ontology is the identification of the entities comprising it. As has already been mentioned, Russian prepositions are a fuzzy class, with secondary prepositions being its most problematic subset, which is why further specification and analysis of this subclass is crucial to our understating of how the prepositional system could be organized.

2 Russian Secondary Prepositions

2.1 Related Work

While much research has been dedicated to primary prepositions, the same cannot be said about secondary prepositions. In fact, as of now we have yet to obtain an exhaustive list of the elements of this set. This is, however, not to say that secondary prepositions have not been examined and catalogued altogether.

Perhaps the most well-structured inventory of Russian secondary prepositions can be found in the Russian Grammar [14]. A sizeable list of secondary prepositions along with their case government is presented after the description of each subtype (according to the part of speech they derive from and their structural type) in the source mentioned. However, no summary list of secondary prepositions is provided. Even more importantly, it is noted by the author 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 [14: §1661].

The Explanatory Dictionary of Functional Parts of Speech of the Russian Language [15], another work touching on the subject, contains less than 300 secondary prepositions. Much fewer – just 157 – are found in the Explanatory Dictionary of Combinations Equivalent to a Word [16].

Overall, secondary prepositions tend to be overlooked in favour of primary ones in most of the relevant sources.

2.2 Secondary Prepositions in UD Models

Prepositions are naturally included in most language models, but their handling, as has been mentioned previously, is quite inconsistent.

For the purposes delineated in the introduction, we have studied the available language models for Russian and developed our own prepositional phrase extraction tool [17] based on the Universal Dependencies (UD) models available in the CONLL-U format [18]. However, while prepositional constructions with primary prepositions can generally be extracted with little trouble, those with secondary prepositions present a serious problem in the task of identifying a prepositional phrase. This is best demonstrated on the following example. The UD_Russian-SynTagRus model, based on the syntactically annotated part of the Russian National Corpus, SynTagRus, contains 73 simple prepositions, primary and secondary, as identified by the *ADP* tag, and 26 MWPs, identified as a sequence of tokens connected by the *fixed* relation with at least one *ADP* token among them. The UD_Russian-Taiga model, based on the Taiga corpus, contains 81 simple prepositions and 25 MWPs. The numbers indicate an obvious discrepancy in the lists of entities recognized as prepositions in the models: while the prepositional inventories intersect, they are not identical. Additionally, the numbers of prepositions, MWPs in particular, annotated as such appear to be alarmingly low in both cases. Thus, the extraction of prepositional phrases is only available for those entities which are recognized as simple or multiword prepositions in a given model.

The annotation of prepositions in UD models is debatable in general. It appears that secondary prepositions, especially multiword ones, tend to be neglected when developing an annotation scheme. [19] name the flat internal structure, lack of common POS tag, discrepancies between lists of such units among the main issues of multiword entities in the current UD annotation standard. In [Kahane and Gerdes, 2016] the authors point out that the preference for relations between notion words as per the UD standard leads to inconsistencies in the case of one-word secondary prepositions which retain the features of notion words as well as incorrect annotation of MWPs, which tend to be encoded compositionally instead of as a semantic unit.

Poor definition of the secondary preposition subclass and especially the subpar handling of prepositional multiword entities make the use of the existing language models in the task of prepositional phrase extraction and analysis a risky endeavor. As our own study aims to describe the Russian prepositional system as a whole, we cannot rely on the available resources blindly while being aware of the issues mentioned. This prompted us to formulate a more in-depth base description of the secondary preposition subclass as well as form our own list of these units.

2.3 Secondary Prepositions: An Overview

Secondary prepositions are words and word combinations that have taken on the function of a preposition at some point of language development. Structurally these units can be subdivided into simple and complex (multiword) ones.

Simple secondary prepositions are usually fully homonymous with some word form of their motivating content word or a different part of speech sharing the same root. The same words and word units may perform as prepositions as well as other parts of speech (e.g.: *силами* 'by force of' – noun, *снаружи* 'outside' – adverb, *исключая* 'excluding' – verb (participle).

Multiword prepositions (MWPs) make up a large part of secondary prepositions. Structurally speaking, a multiword preposition is a combination of a content word and one or two simple adpositions. MWPs can be divided into nominal, adverbial or verbal units based on the part of speech of the motivating content word. Most MWPs contain only one adposition preceding or following the content word (e.g.: *рядом с* 'close to', *в результате* 'as a result'), but some include two adpositions enclosing the content element (e.g.: *в соответствии с* 'in accordance with', *по направлению к* 'toward, in the direction of'). The most commonly observed structural patterns of MWPs are Prep+N, Prep+N+Prep and Adv+Prep, where Prep stands for preposition, N for noun, Adv for adverb. Much like simple secondary prepositions, multiword prepositional units perform as prepositions in some contexts and as free word combination in others (e.g.: *в форме* 'in the form of': preposition + noun, *что до* 'as for': conjunction + preposition; *начиная с* 'starting with': verb + preposition).

As a rule, the distinction between these ambiguous entities is outlined neither in grammar books nor in dictionaries. The homonymy presents an additional issue for a corpus-driven study, which is why we have decided to organize the investigation of secondary prepositions by their structural type. Our current paper is devoted mainly to secondary multiword prepositions.

Overall, the MWP subclass is quite diverse, which is a direct consequence of its size. Our research has uncovered a great number of multiword prepositions, with some of them having up to four morphonological (e.g.: *в сравнении/сравненье с/со* 'compared to') and spelling (e.g.: *в счет/счёт* 'on account of') variations. The great variety of MWPs on all language levels implies the necessity of an in-depth analysis of their common features. In other words, we need to understand, firstly, what unites such diverse entities in order to be able to discern free combinations from MWPs.

2.4 Characteristic Features of Multiword Prepositions

As has already been stated, 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 limits, we have formulated the following preliminary list of 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.

The presented list was initially meant to serve as a guideline. However, we soon realized the importance of relying on more clearly and precisely defined and formalized features. Our current study is intended to be a step towards clarifying some of them and defining others more narrowly through studying how these features manifest (or do not manifest) themselves in the potential MWPs in real texts. In order to demonstrate that our proposals are based on real language use we mostly focus on the features that lend themselves to statistical description and analysis on corpus material in this paper.

3 Materials

Our main research is dedicated to the entire class of prepositions. In order to obtain the fullest inventory of the units in question we have compiled a table of Russian prepositions totalling 740 entries (including variations) based on a number of linguistic sources (dictionaries, grammar guides, corpora, syntax parsers), including those mentioned in Section 2.1. Naturally, the degree of "prepositionality" of these entities varies. The contents of this table were used as the base material of our study.

As the current study has multiword prepositions as its main focus, the pool of relevant prepositions has been narrowed down to 445 multiword units. A general statistical overview of the whole set has been provided by us in [20]. However, as our goal was to study the main features relevant to all MWPs, it has been concluded that a smaller selection would be sufficient for the stated purpose. Therefore, we have settled on a subset of the original list that only contained prepositions expressing causal relations. While approaches to semantic classification of prepositions naturally vary, it is generally agreed upon that prepositions can be used to express cause and effect. Our selection consists of 13 multiword preposition candidates that have been observed to express causal or causal-adjacent relations:

- *в зависимости от* 'depending on'
- *в ответ на* 'in response to'
- *в преддверии* 'on the eve of, at the forefront of'
- *в результате* 'as a result of'
- *в свете* 'in light of'
- *в связи с* 'due to'
- *в силу* 'by force of'
- *за счёт* 'on account of'
- *исходя из* 'drawing from'
- *на основании* 'on the basis of'
- *на основе* 'based on'
- *на почве* 'on the ground of'
- *по причине* 'because of, for the reason of'

The results of the statistical analysis presented in this article have been acquired mainly on the Araneum Russicum III Maius corpus (1.25 billion tokens) created by Vladimír Benko (Comenius University in Bratislava, Ľ. Štúr Institute of Linguistics, Slovakia) (www.unesco.uniba.sk). Those features that are more lexically or semantically inclined are subject to future qualitative and quantitative studies.

The Russian National Corpus (www.ruscorpora.ru), the joint project of the Russian Academy of Sciences and multiple research institutions, was also used for more detailed research on some of the features. The main corpus (over 320 million tokens) was chosen due to its considerable size as well as the inclusion of the subcorpus with manually resolved morphological homonymy, which is relevant to the task at hand.

4 Results

4.1 Structural Features of Causative MWPs

The first set of features to observe is the structure of the MWPs in question. 10 out of 13 units are bigrams of a content word and a simple adposition, which appears to be the most typical MWP structure. 9 of the bigram units follow the structural pattern of Prep+Noun, one has the less common pattern of Verb+Prep. The remaining 3 out of 13 units are trigrams consisting of a noun between two adpositions.

As has been noted by us in [11], the three simple adpositions most commonly used as elements of multiword prepositions are *в, на, по*. Out of the 13 items under current study, 7 contain the preposition *в*, 4 contain *на*, 1 contains *по*. Also present are the less frequently found in MWPs but nonetheless typical prepositions *от, с, за, из*.

Most of the content words in the causative MWPs refer to the two nodes of causal relations: the reason (*основание* 'foundation', *основа* 'base', *почва* 'ground', *причина* 'reason') and the effect (*ответ* 'response', *результат* 'result'), as well as the relation itself (*зависимость* 'dependency', *связь* 'connection', *сила* 'force'). The semantics of the motivating content words correspond with the observed tendency of MWP component nouns to lean towards abstraction.

Another point of interest is the use of the content words as MWP components in comparison to their general corpus frequency. The table below demonstrates the relative frequencies (in ipm) of the content words in question as well as the frequencies of the MWPs themselves.

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

Base word	MWP	Base word, ipm	As MWP component, ipm	% of MWP use
Преддверие	*В преддве-рии/ рьи/рие/рье*	11.30	10.79	95
Зависимость	*В зависимости от(о)*	171.10	111.62	65
Исходить	*Исходя из(о)*	77.10	44.70	58
Счёт	*За счё/ет*	303.00	137.05	45
Связь	*В связи с(о)*	346.90	119.00	34
Основа	*На основе*	314.60	105.38	33
Основание	*На основании*	174.90	57.32	33
Результат	*В результате*	604.70	173.78	29
Сила	*В силу*	452.70	52.70	12
Причина	*По причине*	341.90	19.70	6
Ответ	*В ответ на*	239.80	11.35	5
Почва	*На почве*	60.90	2.84	5
Свет	*В свете*	196.10	7.68	4

As demonstrated by the table, most of the content words retain their relative independence as they are not bound to the other parts of the MWPs. Only one of them, *в преддверии*, appears to be a set phrase in which the motivating word has fallen out of use.

4.2 Statistical Analysis of Causative MWPs

In order to gain a clearer understanding of how prepositional units perform as MWPs and free combinations as well as how frequently either of the states occurs we have studied the use of the causative prepositional units in the corpus. In order to do that we have obtained the top 50 highest frequency ngrams for each of the prepositional units in question and their immediate context, which in most cases consisted of the nearest left or right neighbour token(s). The resulting construction lists have been studied and tagged by hand according to the function the prepositional unit performs in the given context: "MWP" or "free combination". The frequencies of these two states have been then translated into percentages of prepositional use in the studied selection. The results are presented in Table 2.

Table 2. Percentage of prepositional use of MWP candidates among top 50 frequency ngrams in Araneum Russicum Maius and 50 random contexts in the RNC

MWP	**% of prepositional use, AR**	**% of prepositional use, RNC**
В преддверии/-рьи/-рие/-рье	100	98
В зависимости от(о)	100	98
Исходя из(о)	100	100
За счё/ет	88	100
В связи с(о)	100	100
На основе	82	82
На основании	100	100
В результате	100	54
По причине	86	96
В ответ на	100	98
На почве	100	96
В свете	84	78
В силу	90	80

The results show that the distribution of prepositional use of the word combinations in question is quite similar in the two corpora despite the difference in the sample and corpus parameters. In addition to the stable occurrence of prepositional uses in different textual contexts, the results demonstrate that these word combinations are mostly used as prepositions and not free word combinations.

Among the non-prepositional uses discovered in the samples most were cases of free combinations of a simple preposition and the content word used in one of its primary meanings. For example, the MWP candidate *на основе* 'on the basis of' was found to be a free word combination in contexts referring to the physical basis of an entity, e.g. [X] *"на основе гиалуроновой кислоты"* ('hyaluronic acid-based [X]'); similarly, word combination *в свете* 'in light of' was used literally in contexts where the governee belonged to the semantic class of objects capable of emanating light, e.g. *"в свете заходящего солнца"* ('in the light of the setting sun'). One MWP candidate, *в результате* 'as a result of', is homonymous with the adverbial modifier *в результате* 'as a result', which led to the inclusion of the adverbial modifier in the context sample as the case of the governee was not predefined in the corpus search. The context window approach used in the work on the Araneum Russicum Maius corpus was therefore unable to yield useful data on the distribution of prepositional uses of the word combination in question. The MWP candidate *в силу* 'by force of, due to' was found to be used occasionally as a free combination (*[верить] в силу* '[believe] in the force') and as part of an adverbial idiom (*[вступить] в силу* 'come into power'), which led to the relatively lower observed percentage of its prepositional use as well.

4.3 General Observed Tendencies

Some general observations were made in the course of the ngram frequency analysis described in the previous section. While the data provided in Table 2 is primarily based on the analysis of the left and right context windows, some additional procedures were used in order to study the variability of the selected prepositional units. Thus, in addition to obtaining data on the immediate context neighbours of the MWP candidates we have also studied whether the prepositional unit allows for modifier insertion, such as an adjective or an adverb before or after the content word, as well as whether restraining the query by adding case markers to the potential prepositional phrase governee token makes a meaningful difference to the proportion of prepositional uses of the MWP candidate in the resulting concordance. Therefore, the procedures taken for most of the units under study were the following:

- Frequency analysis of the immediate neighbour tokens of the prepositional unit with no case restraint
- Frequency analysis of the immediate neighbour tokens of the prepositional unit with a case restraint
- Frequency analysis of the node components in a modifier-enabled query for the prepositional unit with no case restraint
- Frequency analysis of the node components in a modifier-enabled query for the prepositional unit with a case restraint

In some cases the POS tag distribution within a query of any kind was also taken into consideration.

The following tendencies of some MWPs were uncovered as a result.

Firstly, a high number of the causative prepositional units were found to take the initial position in a sentence or a clause as evidenced by the inclusion of punctuation

marks and conjunctions in the top frequency lists of their immediate left neighbour tokens. Periods, commas and the conjunction и 'and' were found among the top 5 left neighbours of the prepositional units *в результате, в свете, исходя из, за счёт, в силу*. The tendency for the initial sentence/clause position can be explained by the semantics of the relation expressed by means of the MWPs in question. Causative prepositions serve to connect events rather than objects and their features, which is why more complex syntax structures, such as clause or sentence sequences, are needed to express cause-and-effect relations.

Secondly, the identification of prepositional vs. free uses of some MWP candidates was found to be a difficult task in the absence of either the governor or the governee of the prepositional phrase. For instance, resolution of the contextual homonymy of the prepositions *в зависимости от, в результате, в силу* and their free equivalents appears to rely mainly on their governors as their semantic classes seem to be more restricted than those of the governees. The most frequent governors in prepositional usage cases belong to the group of verbs and verbal nouns expressing change of state, e.g. *получить* 'receive', *возникать* 'appear', *образоваться* 'form' for *в результате* 'as a result of', or expressing difference, e.g. *меняться* 'change', *варьироваться* 'vary', *отличаться* 'differ' for *в зависимости от* 'depending on'. For *в силу* 'due to, by force of' the governors were useful in identifying free usage cases, e.g. *вступление* 'entry', *вступить* 'come', *верить* 'believe' [in(to) (the) power]. Inversely, the homonymy resolution of the MWP candidates *в свете, в преддверии* was more successful in the presence of their governees. As such, contexts with the governees *фары* 'headlights', *фонари* 'street lamps', *луна* 'the moon' for *в свете* 'in light of' and *рот* 'mouth', *влагалище* 'vagina' for *в преддверии* 'at the forefront of' were found to be free word combinations.

A few of the MWP candidates were discovered to frequently serve as components of conjunctional phrases with the demonstrative pronoun *то* 'that'. *В зависимости от* (ipm 111.6) and *в связи с* (ipm 114.6) are particularly remarkable examples boasting ipm frequencies of 8.1 for the conjunctional phrase *в зависимости от того, …* 'depending on…', and 8.8 for *в связи с тем, …* 'due to…'. The fact that these prepositional units have become part of a more complex structure implies two things. First of all, it is indirect evidence of the prepositionality of these units as they have apparently created a very stable prepositional bond with the pronoun. Secondly, it is necessary for us to decide whether such structures should be regarded as an inseparable whole and therefore excluded from the statistical analysis of the MWP components in question.

Finally, two curious usage cases were discovered while examining the context samples of MWP components *исходя из* 'drawing from' and *по причине* 'because of, for the reason of'. *Исходя из* was found to be used prepositionally in the syntax structure *'исходя из:*

– [point 1],
– [point 2], ...'.

The structure itself is not uncommon in real written texts but is atypical from the point of view of traditional prepositional syntax, which presupposes the positioning of the governee(s) without any punctuation marks after the preposition. Primary prepositions

(*в, из-за, к, о, с* and others) have been observed to occur in this structure as well, which supports the prepositional status of the word combination *исходя из*. Another interesting case is that of the MWP candidate *по причине*, which has been found to bind commonly enough with direct quotes, e.g. *по причине "не повышают зарплату"* 'for the reason "[they are] not raising the pay"'. As the quoted verbal structure is not declinable (has no case marker) and serves as an attribute for the content word *причина* 'reason', the prepositional unit is most likely a special case of a free word combination rather than a preposition in such contexts.

4.4 Separability of Causative MWPs

A special point of interest is the separability of MWPs, that is, the allowance for modifier insertion into the MWP structure. In order to study this phenomenon, we have examined context samples of the causative MWP candidates with and without content word modifiers.

Overall, our presupposition that insertion is atypical for MWPs has been proven true. Since most of the content words in our MWP candidate selection are nouns, it was primarily adjectival modifiers that were found splitting the original prepositional unit structure. As per the list of characteristic features of MWPs, the motivating nominal component loses its lexical meaning to a degree when the unit is used as a preposition to allow it to perform the basic prepositional function of conveying a relation between the governor and the governee of the prepositional phrase. However, when modifier insertion takes place, the semantic weight of the whole construction figuratively shifts back to the modified noun, which retains its original lexical meaning. Therefore, the resulting structure can no longer perform the prepositional function and can only be regarded as a free word combination. The table below demonstrates this phenomenon observed in the 10 most frequent modified structures for the MWP candidate в свете.

Table 3. 10 most frequent prepositional structures with an adjectival modifier for *в свете* 'in light of' (ipm 7.68, Araneum Russicum Maius)

Modified structure	**Translation**	**Frequency in corpus, ipm**
В новом свете	In a new light	0.33
В этом свете	In this light	0.28
В лучшем свете	In the best light	0.24
В выгодном свете	In a favorable light	0.22
В лунном свете	In the moonlight	0.22
В ином свете	In another light	0.12
В другом свете	In a different light	0.11
В таком свете	In such light	0.09
В солнечном свете	In the sunlight	0.09
В негативном свете	In a negative light	0.09

The examples make it apparent that the modified construction can no longer perform as a causative preposition.

However, two types of modifier insertion were found to not disturb the prepositional function of the studied MWP candidates.

Firstly, some nominal causative MWPs do not seem to lose their function in the case of anaphoric use of personal pronouns, such as *на его почве* 'on his [its] ground', *в её преддверии* 'on her [its] eve'. Out of a sample of 34 contexts of the phrase *в его результате* 'as its result' 26 uses were found to be prepositional, 8 were not. The MWP retained its function when the inserted pronoun referred to a previously named entity. The structure is typically used to avoid repetition, e.g. *беспрецедентное журналистское расследование и всплывшие в его результате ужасающие факты* "an unprecedented journalistic investigation and the horrifying facts that have emerged as its result".

Secondly, исходя из, a verbal MWP, retained its function when the modifying part of speech was a particle, with the most common ones being *только* 'only', *лишь* 'merely', *не* 'not', *именно* 'precisely', *уже* 'already', *исключительно* 'exclusively'. As Russian particles are function words with no independent lexical meaning and serve to express modality or impart shades of meaning to the modified notion word, their function does not appear to interfere with the prepositional function of the modified phrase.

Therefore, we can conclude that causative MWPs generally do not allow for insertion (modification of the content component) except when the modifier is a personal pronoun modifying the nominal component or a particle modifying the verbal component. Whether this rule applies to the entirety of the MWP class is subject to further investigation.

5 Conclusion and Future Work

The current paper deals with the complex aspects of Russian multiword prepositions (MWPs). These units may perform as prepositional entities with particular grammatical semantics or manifest themselves as free combinations in which each word has its own meaning and syntactic function. While MWPs are a large and diverse subclass, they are nonetheless characterised by a number of common features and, therefore, lend themselves to description, definition and measurement.

The aim of the study presented in the paper was to test out the proposed methodology of determining the prepositional status of multiword prepositional units using a set of units expressing causal relations. 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.

All of the observations presented in the paper suggest that the bottom-up corpus-based approach is indispensable in the task of studying multiword units of ambiguous status owing to the direct focus on real patterns of usage. However, the context window method used in the study appears 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 specialized corpus tools, such as the Word Sketches tool of the Sketch Engine system. An even more effective approach would involve full syntax parsing, even though it also does not guarantee errorless extraction. An alternative approach is the use of treebanks, although their limitations in volume and annotation present a challenge of its own.

Further stages of our research include expansion of the application of the methodology presented in this paper to the entirety of the MWP subclass. The separability of the components of a multiword preposition is to be examined on a wider variety of MWPs.

Additionally, research on the prepositional use in fixed phrases and idioms as well as clusters of conditional synonymy has been started, which will hopefully help in defining the status of MWP candidates more precisely.

Automatic recognition and analysis of MWPs plays an important role in a number of key NLP tasks: prepositional phrase attachment, syntactic role acquisition, corpus annotation, multiword unit recognition, word sense disambiguation, etc. The results obtained promise to help solve these tasks with greater accuracy given the volume of the MWP subclass in the already sizeable class of prepositions. Being the first of its kind for Russian secondary multiword prepositions, our study provides an insight into the ambivalent nature of these entities and will hopefully contribute both to the theoretical description of the Russian prepositional system and to the solution of the practical problems of computational linguistics.

6 Acknowledgements

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

Authors wish to express their sincere gratitude to the 2nd year students of the SPbU Mathematical Linguistics Department for their valuable help in annotating the data.

References

1. Lyashevskaya, O.N., Sharov, S.A.: Frequency Dictionary of the Modern Russian Language (on the Materials of the National Corpus of the Russian Language). Azbukovnik, Moscow (2009).
2. Delecraz, S., Nasr, A., Béchet, F., Favre, B.: Correcting prepositional phrase attachments using multimodal corpora. In: Proceedings of the 15th International Conference on Parsing Technologies, 72–77. Association for Computational Linguistics, Pisa, Italy (2017).
3. Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: 33rd annual meeting of the association for computational linguistics, 189–196. Association for Computational Linguistics, Cambridge, MA (1995).
4. Ballestros, L., Croft, W.W.: Dictionary-based methods for cross-lingual information retrieval. In: Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, 791–801. Springer, New York, NY (1996).
5. Jensen, P.A., Nilsson, J.F.: Ontology-based semantics for prepositions. In: Syntax and semantics of prepositions, 229–244. Springer, Dordrecht (2006).

6. Dictionary of the Russian Language, 4 volumes. 3rd edn. Russkie Yaziki [Russian Languages], Moscow (1988).
7. Solonitskiy, A.V.: Problems of semantics of Russian primitive prepositions [in Rus]. Far Eastern Federal University Publishing, Vladivostok (2003).
8. Filipenko, M.V.: Problems of the description of prepositions in modern linguistic theories [in Rus]. In: Research on the semantics of prepositions, 12–54. Russkie Slovari [Russian Dictionaries], Moscow (2000).
9. Zolotova, G. A.: Syntactical Dictionary: a set of elementary units of the Russian syntax. 4th edn. Moscow (2011).
10. Litkowski, K.: Notes on grilled opakapaka: Ontology in preposition patterns. Technical Report 15-01. CL Research, Damascus, MD (2015).
11. Zwarts, J., Winter, Y.S.: A semantic characterization of locative PPs. In: A. Lawson (ed.), Proceedings of Semantics and Linguistic Theory, 294–311. CLC Publications, Ithaca, NY (1998).
12. Lassen, T.: An Ontology-Based View on Prepositional Senses. In: Proceedings of the Third ACL-SIGSEM Workshop on Prepositions, 45–50. Association for Computational Linguistics, Trento, Italy (2006).
13. Zakharov, V., Azarova, I.: Semantic structure of Russian prepositional constructions. In: K. Ekstein, V. Matousek (eds.). Lecture Notes in Computer Science, 11697 (Text, Speech, and Dialogue – 22th International Conference, TSD 2019 Proceedings), 224–235. Springer International Publishing AG (2019).
14. Shvedova, N.Ju.: Russian Grammar. Vol. 1: Phonetics. Phonology. Word Stress. Intonation. Word Formation. Morphology [in Russian]. Nauka [Science], Moscow (1980).
15. Efremova, T.F.: Explanatory dictionary of functional parts of speech of the Russian language. AST, Moscow (2004).
16. Rogozhnikova, R.P.: Explanatory dictionary of combinations equivalent to a word: Approx. 1500 fixed phrases of the Russian language. AST, Moscow (2003).
17. Gudkov, V., Golovina, A., Mitrofanova, O., Zakharov, V.: Russian prepositional phrase semantic labelling with word embedding-based classifier. CEUR Workshop Proceedings, 2552, 272–284. RWTH Aahen University (2020).
18. Zeman, D., Nivre, J., Abrams, M. et al.: Universal Dependencies 2.5, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3105, last accessed 2020/11/11.
19. Kahane, S., Courtin, M., Gerdes, K.: Multi-word annotation in syntactic treebanks: Propositions for Universal Dependencies. In: Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16), 181–189, Prague, Czech Republic (2018).
20. Zakharov, V., Golovina, A., Azarova, I.: Statistical analysis of Russian multiword prepositions. In: NordSci International Conference Proceedings, 1, 191–200. Saima Consult LTD, Sofia, Bulgaria (2020).