sciendo
DOI 10.2478/jazcas-2023-0035

TOWARDS A CORPUS-BASED DICTIONARY OF VERBAL GOVERNMENT FOR THE RUSSIAN LANGUAGE

EDUARD KLYSHINSKY[1] – ANNA BOGDANOVA[1]
– MIKHAIL KOPOTEV[2,3]
[1] Independent researchers
[2] Department of Languages, University of Helsinki, Helsinki, Finland
[3] Department of Slavic and Baltic Studies, Stockholm University, Stockholm, Sweden

KLYSHINSKY, Eduard – BOGDANOVA, Anna – KOPOTEV, Mikhail: Towards a Corpus-based Dictionary of Verbal Government for the Russian Language. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 173 – 181.

Abstract: This paper introduces a technique for automatic verbal governance extraction in the Russian language, which encapsulates information on the grammatical features of verb-noun co-occurrences, encompassing both prepositional and non-prepositional dependencies. The construction of the dictionary, a corpus of approximately 3.5 billion words was used. The proposed method involves syntactic parsing of the texts, filtering of resultant outputs, and creating a dictionary of prepositional government. After error filtering, the dictionary contains ca. 18,000 verbs along with NP/PPs governed by these verbs.
Keywords: verbal government, automatic extraction, Russian language

1 INTRODUCTION

Collocational dictionaries hold a significant position in Russian lexicography. However, their utility is often constrained by their relatively modest size and sub-optimal accuracy. Prokopovich et al. (1981) offer combinations for 1219 Russian words, alongside theoretical discourse on nominal and verbal governance. Denisov et al. (1983) furnish a comprehensive description of the 2506 most frequent Russian words, including 727 verbs. Meľčuk and Zholkovsky (1984/2016), among others, supply exhaustive, standardized details concerning word government structure. Regrettably, the dictionary encompasses only 283 entries.

Presently, extensive efforts are being directed towards the development of the Russian Active Dictionary (Apresjan et al. 2014–2017, which encapsulates not only word meanings but also data pertinent to speech production, such as combinatorial characteristics and pragmatic conditions of word usage. Among a variety of features, the dictionary incorporates information on verbal governance, including the most commonly used cases and prepositions marking prepositional and non-prepositional

dependencies. This data is invaluable for researchers and learners of Russian but it may be insufficient for the creation of Natural Language Processing (NLP) systems, which aim to present a comprehensive and complete list. Consequently, existing dictionaries, while they contain a limited number of entries, are infrequently presented in a machine-readable format.

From a Computational Linguistics perspective, the compilation of robust electronic lexicographic resources for the Russian language is a labor-intensive process, and it has seen the development of numerous automated techniques. Most of these methods leverage data from the Russian National Corpus (RNC), a compendium of Russian texts with sophisticated linguistic annotation. An example of these is represented in Biryuk et al. (2008), encompassing 10,015 verb combinations with abstract nouns filling the patterns 'Noun + Verb', 'Verb + Noun', and 'Verb + Adjective + Noun'. This dictionary signifies an attempt to establish a digital resource, grounded in Meľčuk's Meaning-Text theory, yet employing large corpus data. All combinations within the dictionary are classified according to lexical functions—semantic labels corresponding to case roles, as proposed by Ch. Fillmore. Another online resource for word co-occurrences is the database of Russian lexical constructions known as FrameBank (Lyashevskaya et al. 2011). It comprises a corpus of the 2,500 most frequent Russian verbs and verbal constructions, accompanied by a description of their governance patterns (syntax annotation and semantic role labeling).

Two projects were initiated by the authors of this article. The project 'CoCoCo: Collocations, Colligations, and Constructions' (Kopotev et al. 2015; Kormacheva et al. 2014), utilizes three corpora, RNC, I-RU, and Taiga, to showcase not only collocations but—as indicated by the title—also colligations (grammatical patterns) and constructions (grammatical patterns supplemented by lexical variables). Another resource, called CoSyCo, is introduced in Klyshinsky et al. (2018). It comprises syntactic patterns extracted from a corpus of approximately 17 billion tokens. The resource capitalized on the grammatical and semantic relations between tokens and is primarily devised for NLP and CL tasks.

Consequently, among the current resources targeted at word combinations, those specifically designed for verbal patterns are underrepresented. Both print and online dictionaries are limited in size, online resources are somewhat unspecific in this aspect since they focus on any word collocations for all parts of speech (POS), and they do not incorporate information specific to verbal governance.

With this in mind, the verbal collocation dictionary would serve many users. Firstly, it could be utilized by foreigners learning Russian, as it provides information not only about combinations but also the frequency along with a comprehensive list of examples. Secondly, it may captivate researchers in the fields of theoretical and descriptive linguistics, as it provides a distribution of verb combinations among different genres and semantic classes. Finally, the outcomes of this work may be harnessed by researchers in the field of Natural Language Processing (NLP) as

a benchmark for the evaluation of automated systems. The primary objective of this paper is to describe a corpus, which fulfils specific requirements. Firstly, the corpus must encompass as many diverse verbs as possible to be applicable to NLP tasks; secondly, the corpus must contain less than 5% of incorrect combinations of verbal patterns.

The rest of the paper is structured as follows: Section 2 elucidates the proposed method for the automatic composition of the dictionary of verbal governance. Section 3 furnishes details about the experimental data and results used. The remainder of the paper encompasses a brief discussion of the experimental results and the limitations of the method.

2 PROPOSED METHOD

To fulfil our research objectives, we constructed a syntactically annotated corpus of Russian texts. This enabled the computation of frequencies for syntactic patterns featuring verbal heads and nominal/prepositional dependencies. Nevertheless, the annotated corpus manifested a fairly high degree of errors, necessitating painstaking preprocessing. Consider, for instance, the following sentence:

(1) *Раз за разом в библиотеки приходили новости, которые давали всё меньше подробностей.*
'There came news in the libraries, again and again, which provided fewer and fewer details.'

Theoretically, parsing this sentence could engender multiple errors, thereby compromising statistical data. Initially, the expression *раз за разом* (lit. 'time-NOM for time.INS'; 'again and again' is linked to the verb *приходили* 'come-PAST.PL', which does not govern the nominative case in this position; thus, it would be erroneous if linked to the verb. Secondly, the fixed expression *всё меньше* 'less and less' necessitates the genitive case for the noun *подробностей* 'detail-GEN.PL'; however, the parser could incorrectly associate the noun with the verb *давали* 'give-PAST.PL'. Lastly, the phrase *в библиотеки* 'in the library' can be parsed as в 'to/in' + library-ACC.PL (the most probable analysis) or as 'to/in' library-GEN.SG (incorrect, but still available in a parser). Consequently, the raw frequencies failed to provide relevant data for dictionary construction. This realization necessitated the development of a novel method for text preprocessing.

In the first phase, the lexicon of prepositional governance was established by assembling a comprehensive list of all Russian prepositions and the cases governed by them. The statistics were calculated over the syntactically annotated corpus and normalized solely for prepositions. All instances that exhibited a relative frequency of

less than 1% for a given preposition were considered marginal and, thus, were omitted. These data were inspected and corrected by an expert. The result is a list of prepositions and the cases they govern, henceforth referred to as the Lexicon of Prepositional Governance (hereafter LPG). The list incorporates 132 prepositions, including some compounds such as *за счёт* 'by means of' or *в течение* 'during'. These prepositions may govern, in different combinations, five cases: genitive, dative, accusative, instrumental, and locative. The two marginal cases in Russian, the second genitive (partitive) and second locative, are considered the genitive and locative cases, respectively.

In the second phase, we compiled statistics on the co-occurrence of all prepositional phrases for all verbs in the corpus. Any instances of the cases, which are not attested in the LPG, were eliminated. It is important to note that filtering is not the best decision because many low-frequency co-occurrences are known to be rare, mainly idiomatic, word combinations, necessitating further investigation. Conversely, the data frequently contain a substantial number of errors induced during both text production and parsing. For our project, a classic trade-off between precision and recall is that we elected to decline instances according to the frequency-based filtering.

In the third phase, we examined the non-prepositional case government, which presents challenges due to the specific features of Russian syntax. For instance, a direct object is marked, albeit non-automatically, with the genitive case if a verb is used under negation. This variation isn't exclusive to a specific verb but rather applies to any negated verb. After a thorough preliminary investigation, we opted to apply the following filters to avoid standard variations in Russian syntax:

- All verbs including auxiliary ones should not be negated: *предвещать беду* 'to portend trouble-ACC' becomes *не предвещать беды* 'not to portend trouble-GEN';
- A noun should not form part of a numeral group and, therefore, should not be governed by or agreed with numerals or quantitative adverbs: *трое грустных мужчин* 'three sad men-GEN' VS *три счастливые женщины* 'three happy women- NOM'.

Returning to example (1) above, the genitive case for *библиотеки* 'library-PL. GEN' can be filtered out because it is below the threshold for the preposition в 'to/ in'. The phrase *давали всё меньше подробностей* 'gave less and less detail-PL. GEN' is also out of the scope here, since it contains a quantitative adverb.

Despite these filters proving effective in reducing noise, they have not entirely eradicated it. Consequently, we decided to filter out all combinations with a relative frequency below a certain threshold: 5% for genitive, 0.2% for dative, and 1% for accusative and instrumental cases. The locative case has no threshold as it cannot be used without a preposition. Similarly, the nominative case was completely excluded due to its standard syntactic linkage with virtually any verb, save a few exceptions. To encapsulate, our approach comprises several stages:

- assembling the syntactically tagged corpus;
- computing statistics of co-occurrence for verbs and dependent NP/PPs;
- formulating a lexicon of prepositional governance using calculated statistics and implementing expert filtering;
- creating a list of dependent prepositional phrases for all verbs and filtering it following the lexicon of prepositional governance;
- and constructing and filtering a list of nominal dependencies for all verbs.

3 EXPERIMENTAL SETUP AND THE RESULTS OF EXPERIMENTS

3.1 Used texts and tools

The textual corpora employed in this study are detailed in Tab. 1. These collections, sourced from the Internet, were tagged using the DeepPavlov parser (Burtsev et al. 2018) due to its superior accuracy. For our computations, we omitted non-dictionary words, given that DeepPavlov utilizes a neural network for lemmatization, which invariably generates a significant volume of unknown 'lemmas'. For the filtration process, we employed the OpenCorpora (Bocharov et al. 2011) lexicon, implemented in the Pymorphy2 library.

Collection	Number of sentences	Number of words	% of sentences	% of words
Ph.D. and doctoral thesis	26,764,667	560,907,459	14.45%	16.06%
General news	42,572,120	819,591,304	22.98%	23.47%
Thematic news	20,679,206	413,693,321	11.16%	11.85%
Fiction texts	57,230,401	799,596,093	30.89%	22.90%
Wikipedia	25,932,220	544,161,541	14.00%	15.58%
Official texts	12,082,916	354,259,132	6.52%	10.14%
Total	185,261,530	3,492,208,850	100.00%	100.00%

Tab. 1. Size of the text collections

3.2 Results of experiments

Two distinct representations of the dictionary are proposed, each contingent on its specific application. The first is geared toward Russian language acquisition. For this purpose, we have selected the 80%-quantile of the most frequent combinations of verbs and nouns in the nominative case. These combinations encapsulate the most practical elements of verbal governance for a student to master. The remaining combinations, though less common, comprise up to 4% of usage for the verbs under consideration.

A portion of the dictionary is exhibited in Tab. 2, where "-" and prepositions followed by grammatical cases represent a syntactic connection between a noun phrase (NP) and a prepositional phrase (PP), respectively, and a verb. The figures

represent the proportion of such syntactic patterns for a specific verb. The nominative case is not represented in Tab. 2; consequently, the total for a single row may not equal 80%.

verb	NP/PP	%	NP/PP	%	NP/PP	%
организовать 'to organize'	-Acc	37.9	*в* 'in' -Loc	10.4	-Ins	9.6
существовать 'to exist'	*в* 'in' Loc	15.2	*на* 'at' -Loc	4.7		
выполнять 'to execute'	-Acc	69.8				
разработать 'to develop'	-Acc	32.9	-Ins	10.5	*в* 'in' -Loc	8.3
рассмотреть 'to consider/to view'	-Acc	53.0	*в* 'in' -Loc	8.5		
обладать 'to possess'	-Ins	72.4				
поднять 'to lift'	-Acc	58.5	*на* 'at' -Loc	6.0	*в* 'in' -Loc	3.3
использоваться 'to be used'	*для* 'for' -Gen	20.9	*в* 'in' -Loc	16.5	-Ins	4.7
закрыть 'to close'	-Acc	44.7	-Ins	6.9	*в* 'in' -Loc	4.4
пользоваться 'to use'	-Ins	61.2				

Tab. 2. Examples of verbal government in the human-readable format

Our proposed methodology facilitates the compiling of a dictionary comprising 17,367 verbs. Furthermore, it includes 1,510 verbs wherein only the nominative case is attested, i.e., instances where the nominative case is utilized in over 80% of all co-occurrences. A few examples of such verbs are *засориться* 'to clog', *обесточиваться* 'to de-energize', *настать* 'to come' (e.g. the time has come), *расцениваться* 'to apprise', and *прозвенеть* 'to ring out'.

It is important to note that we have only utilized combinations that are represented in our corpus at least twice; there are examples with a frequency of less than 1 instance per million, which means they are extremely rare in the data. They necessitate a more flexible threshold and/or the need for more comprehensive data for their effective handling. Two further filtering criteria that we applied include the requirement that a specific verb-noun pattern must have a minimum frequency of 4–5% among all patterns for that verb and that the number of verbal arguments is limited to four or five.

3.3 Evaluation of results

To evaluate recall, we compared our results with three hundred of the most frequently used verbs in the Active Dictionary of Russian (AD; Apresjan et al. 2014–2017), which consists of 1,073 verbs, while our dictionary provides 9,045 verbs. Depending on frequency, 78–90% of combinations from the AD are found in our dictionary (the lower the frequency, the higher the recall), and only 55% from our dictionary are found in the AD. The AD provides an average of 2.28 combinations per verb compared to 19.32 in ours, but the former includes semantic relations such as purpose and direction introduced with dependent clauses, which are omitted in our dictionary. The recall can be seen as rather low, however, however our methodology identifies statistically significant discrepancies in the usage of aspectual pairs that are consistent with the findings of previous research (Janda et al. 2013). Adopting a more lenient threshold for noun frequencies could potentially facilitate the ex- traction of these patterns as well, thus the recall will be higher.

To assess the precision, we manually examined the 300 most commonly used verbs and discovered that our method commits less than 5% errors, with one stipulation. We considered noun phrases indicating time, frequency, logical inference, and similar concepts as appropriate usage. This can be observed in the phrase *прийти вечером* 'to arrive in the evening.' The word *вечером* can be categorized as a Noun.Ins or an Adverb, contingent upon the adopted theoretical framework. Such instances straddle the boundary between a proper noun phrases and adverbial ones.

After deeper analysis, we discovered that many errors in combinations pertained to the dative case in noun phrases – 17 out of top 100 verbs were errorneously associated (4 times) or disassociated (13 times) with the dative. The second most common errors were in the instrumental case (6 errors among top 100 verbs). Some of these errors can be ascribed to the parser's preferences in creation of syntactic dependencies in case of unconventional word order. For instance, a word in the instrumental case might be erroneously linked to a regular verb at a close distance, rather than to the auxiliary verb it should actually connect with. This could be attributed with the lower threshold value; however, we are faced with the classic precision VS recall conundrum, which should be addressed in regards to the purposes, which we discuss in the conclusion.

4 CONCLUSION

In this study we introduced a method for the automated construction of a dictionary of verbal governance. This dictionary includes a compilation of verbs accompanied by details about governed prepositional and non-prepositional phrases. A corpus of 3.5 billion tokens was employed to accumulate representative data for the construction of this dictionary. The efficacy of the method hinges on the quality of the syntactic parser utilized.

The approach generates approximately 5% non-attested errors for non-prepositional and roughly 3% for prepositional syntactic groups. It's worth noting that many contemporary NLP tools demonstrate a comparable level of accuracy and are still successfully employed in practical applications.

However, our project proves particularly beneficial for learners of the Russian language, as verbal governance is a common classroom topic. In this respect, the 5% error rate could be considered unsuitable for learners of the Russian language, as the resource we offer could potentially lead them astray. To circumvent this limitation, we have devised a condensed version of the dictionary that only includes the most prevalent combinations. This streamlined version boasts fewer errors and is more user-friendly. The application of additional filters in the future could serve to further refine these results.

Furthermore, verbal governance is intrinsically tied to verbal semantics, which we did not delve into in this study. We believe that our work lays the groundwork for future investigations by providing relatively unambiguous data for further exploration. For this project, we collected all noun phrases connected to the verb directly or through prepositions, along with their grammatical information. Clustering nouns into semantic groups is a matter for our future research. Another issue we in- tend to address in our work is the distinction between verbal arguments, which are semantically linked to a specific verb (e.g., to read a book), and adjuncts, which apply to an entire class of verbs (e.g., *to read at the table*). The resulting dictionary will undergo proofreading and will be made accessible in the Git repository of the project: https://github.com/klyshinsky/Slovko-2023.

ACKNOWLEDGEMENTS

We extend our sincere gratitude to Prof. Olga Lyashevskaya for her insightful discussions on Russian syntax, as well as to ChatGPT for proofreading this paper.

References

Apresyan, Y. D. et al. (2014–2017). Active Dictionary of Russian [Activnyj slovar' russkogo yazyka]. Moscow: Yazyki slavjanskoj kultury, vol. 1–3.

Biryuk, O. L., Gusev, V. Y., and Kalinina, E. Y. (2008). Dictionary of Russian Abstract Nouns' Verbal Collocability [Slovar' glagol′noj sochetaemosti nepredmetnyh imen russkogo yazyka] Accessible at: http://dict.ruslang.ru/abstr_noun.php?.

Bocharov, V. V., and Granoskiy, D. V. (2011). Software for Collaborative Work on Morphological Tagging of a Corpus [Programmnoe obespechenie dlya kollektivnoj raboty nad morfologicheskoj razmetkoj korpusa]. Proc. of "Corpus Linguistics – 2011" (27-29 June 2011, Saint-Petersburg), p. 348.

Burtsev, M., Seliverstov, A., Airapetyan, R. et al. (2018). DeepPavlov: Open-Source Library for Dialogue Systems. Proc. of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pages 1–6.

Denisov, P. N., and Morkovkin, V. V. (1983). Combinatorial Dictionary of the Russian Words [Slovar' sochetaemosti slov russkogo yazyka]. Moscow: Russkij yazyk, 2nd ed., 688 p.

Klyshinsky. E. S., Lukashevich, N. Y., and Kobozeva, I. M. (2018). Creating a corpus of syntactic co-occurrences for Russian. Computational Linguistics and Intellectual Technologies. Proc. of "Dialogue 2018" (30 May–2 June 2018, Moscow), pages 305–316.

Kopotev, M., Escoter, L., Kormacheva, D., Pierce, M., Pivovarova, L., and Yangarber, R. (2015). CoCoCo: Online Extraction of Russian Multiword Expressions. The 5th Workshop on Balto-Slavic Natural Language Processing (10–11 September 2015, Hissar, Bulgaria), pages 43–45.

Kormacheva, D., Pivovarova, L., and Kopotev, M. (2014). Automatic Collocations Extraction and Classification of Automatically Obtained Bigrams. Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations, Tübingen, 2014, pages 27–33.

Lyashevskaya, O., and Kashkin, E. (2015). FrameBank: a database of Russian lexical constructions. Proc. of Analysis of Images, Social Networks and Texts. 4th Int. Conf., AIST 2015, pages 337–348.

Meľčuk, I. A., and Zholkovsky, A. K. (2016). Explanatory Combinatorial Dictionary of Contemporary Russian [Tolkovo-kombinatornyj slovar' sovremennogo russkogo yazyka]. Moscow: LRC Publishing House, 2nd ed.

Prokopovich, N. N., Deribas, L. A., and Prokopovich E. N. (1981). Nominal and verb government in modern Russian language [Imennoe i glagol′noe upravlenie s ovremennom russkom yāzyke]. Moscow: Russkij yazyk, 2nd ed.

Richardson, K. R. (2007). Case and Aspect in Slavic. Oxford, OUP, 288 p.