Verb-Object Collocations in the Russian Collocations Database: Linguistic and Statistical Properties

Maria Khokhlova

St Petersburg State University, Universitetskaya emb. 7-9-11,
199034 St Petersburg, Russia
m.khokhlova@spbu.ru

Abstract. Russian Collocations Database comprises collocations extracted from nine dictionaries. The examples provide additional statistical information based on text corpora. The paper deals with those new characteristics that have been added to the database, and how the verb-object collocations that are represented in it intersect with corpus data. The database offers two kinds of interfaces that imply a simple or an advanced search. The former is aimed at language users while the latter can be used by linguists and show a wide range of quantitative characteristics. The paper also presents results of correlation analysis made between collocation lists extracted from dictionaries and corpora. Verb-object collocations from the top of the list of any association measure used in the database proved to be described in several dictionaries compared to the bottom of the list. Verbs tend to be more productive than nouns and produce more examples.

Keywords: Collocation, database, Russian language, dictionaries, corpora, gold standard

1 Introduction

The Russian Collocations Database is a collection of collocations extracted from Russian explanatory and specialized dictionaries, supplemented with statistical information based on text corpora [1]. Since creating a resource is always a process of trial and error, this paper will focus on what features have been added to the database since the launch of the project and how it has been changed and improved. As an example, we will consider verb-object collocations registered in the database, and they will also be compared with corpus data. Collocations were extracted from a number of acknowledged dictionaries. However, the question arises: how do these dictionary collocations correlate with corpora (first of all, with large web ones).

The paper has the following structure. The Introduction explains the motivation of the paper. Section 2 gives an overview of the enhanced database and its interfaces. The next two sections discuss statistical properties of collocations and analyze their structure, paying attention to quantitative properties. The last section concludes the paper and proposes plans for future work.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2023, pp. 121–129, 2023. © Tribun EU 2023

2 Database

2.1 Overview

Since the Russian Collocations database was launched, it has been enriched with further examples from other Russian dictionaries. The initial volume was equal to 20,000 units that were described in five dictionaries [2]. At the moment, the database has doubled its size and has about 40,000 collocations, which were extracted from nine lexicographic resources.

Below we will discuss the example of verb-object collocations. The database comprises 20,145 entries of such a type that were obtained from the following six dictionaries (Table 1) [3,4,5,6,7,8].

Table 1: The number of the extracted data per dictionaries.

Borisova, 1995	Mel'čuk et al., 1984	MAS	Reginina et al., 1980	Biriuk et al, 2008	Deribas, 1983
3,908	1,797	3,308	1,832	5,951	8,607

It can be seen that the dictionary of verb-noun collocations by Deribas [5] is the most numerous source in its examples. The maximum number of verb-object collocations is 5 (that is, no collocation occurs in 6 dictionaries), while the maximum value in the case of adj-noun collocations is 6 [9]. The introduced dictionary index indicates the number of dictionaries in which collocations are presented (Table 2). Thus, in case of verb-object collocations, the index ranges from 1 up to 5.

In all 5 dictionaries, we find the following 8 examples: *oderzhat' pobedu* 'to win', *pol'zovat'sya doveriyem* 'to enjoy confidence', *prinyat' mery* 'take measures', *vesti bor'bu* 'to struggle', *ispytyvat' chuvstvo* 'to feel', *nesti otvetstvennost'* 'to be responsible', *pol'zovat'sya uvazheniyem* 'to be held in respect' and *brat' primer* 'to follow the example'. 181 collocations have the dictionary index equal to 4. Almost all of them are given in the dictionaries of collocations by Borisova [4] and of Russian verbal collocability compiled by Biriuk et al [3]. Three dictionaries present 759 common examples, while two resources produce 3,165 phrases. 80% of the total number of verb-object collocations (16,032) is described only in one dictionary.

The items are also represented by longer collocations with objects represented by prepositional or noun phrases. For example, *ispol'zovat' administrativnyy resurs* 'to use administrative resource', *nakopit' opredelennyy opyt* 'to gain certain experience', *podvergnut'sya radiatsionnomu vozdeystviyu* 'to be exposed to radiation', *poluchit' finansovyyu podderzhku* 'to receive financial support', *prinyat' okonchatel'noye resheniye* 'to make a final decision', *slat' serdechnyy privet* 'to send warmest regards'. It is peculiar that all these examples of distance collocations are given in [3].

Table 2: Examples of collocations with different dictionary indices.

Dictionary Collocation	Borisova, 1995	Mel'čuk et al., 1984	MAS	Reginina et al., 1980	Biriuk et al, 2008	Deribas, 1983	dictionary index
oderzhat' pobedu 'to win'	1[1]	1	1	0	1	1	5
nesti otvetstvennost' 'to be responsible'	1	1	0	1	1	1	5
brat' primer 'to follow the example'	1	0	1	1	1	1	5
stavit' zadachu 'to put the problem'	1	0	0	1	1	1	4
proizvesti vpechatleniye 'to make an impression'	1	1	0	0	1	1	4
otdavat' dan' 'to pay tribute'	0	0	1	1	1	1	4
propvat' blokadu 'to run the blockade'	0	1	0	0	1	1	3
vyderzhat' ekzamen 'to pass the exam'	1	0	0	1	1	0	3
otvodit' vzglyad 'to look away'	1	0	0	0	1	1	3
dostignut' tseli 'to succeed'	1	1	0	0	0	0	2
zavyazat' besedu 'to make a talk'	1	0	0	0	0	1	2
sgladit' ugly 'to smooth things over'	0	0	1	0	0	1	2
zaklyuchit' dogovor 'to enter into a contract'	0	0	0	0	1	0	1
nosit' otpechatok 'to imprint'	0	0	0	0	0	1	1
razgonyat' tosku 'to dispel gloom'	1	0	0	0	0	0	1

Pairwise comparison between the dictionaries has shown that the following resources have the largest intersection (Table 3): 1) dictionaries [5] and [4]; 2) dictionaries [5] and [3].

Table 3: Pairwise comparison between the dictionaries.

	Mel'čuk et al., 1984	MAS	Reginina et al., 1980	Biriuk et al, 2008	Deribas, 1983
Borisova, 1995	124	65	359	313	502
Mel'čuk et al., 1984		11	14	88	222
MAS			9	72	345
Reginina et al., 1980				74	261
Biriuk et al, 2008					706

2.2 Interfaces

Since the database can be in demand by different groups of users, there are two kinds of interfaces, namely, a linguistic search and a statistical one. The first type of interface makes it possible to view the collocations for either a node or collocate. The results contain a list of collocations, in which the following linguistic information is presented:

- definition of lemmata from the Wiktionary;
- type of syntactic structure (i.e., adj-noun, verb-noun, etc.);
- a link to an example of usage in the Russian National Corpus [10];
- presence/absence of a collocation in the SynTagRus [11] and Taiga [12] corpora;
- intersection with other collocations.

The results involve a dictionary index as well. The larger it is, the greater the probability of using a collocation is. We introduced a graphical interpretation of dictionary indices to indicate that a collocation is typical. Figure 1 shows a bar plot for the results for the verb *obratit'* 'to turn'. One can note that the most common examples shown in the dictionaries are *obratit' vnimaniye* 'to draw attention, to give attention' (in three dictionaries), *obratit' vzor* 'to look' and *obratit' v begstvo* 'to put to flight' (both collocations are described in two dictionaries).

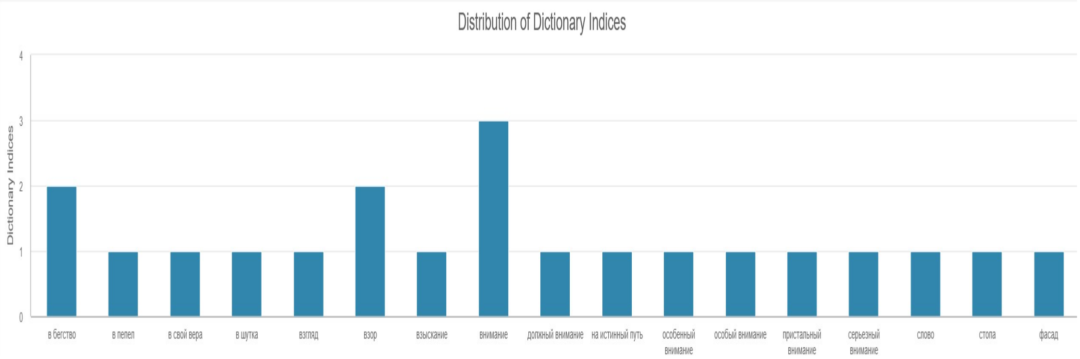


Fig. 1: Distribution of dictionary indices for the verb obratit' 'to turn'

Figure 2 shows visualisation used in the database to present the node and its collocates. The examples found in several dictionaries are marked with dark arrows between the verb and its collocates.

A statistical search offers a more specialized way to present results aimed at advanced users. Each entry is supplied with the following statistical information:

- presence of a particular collocation into dictionaries (9 dictionaries in total);
- a dictionary index;

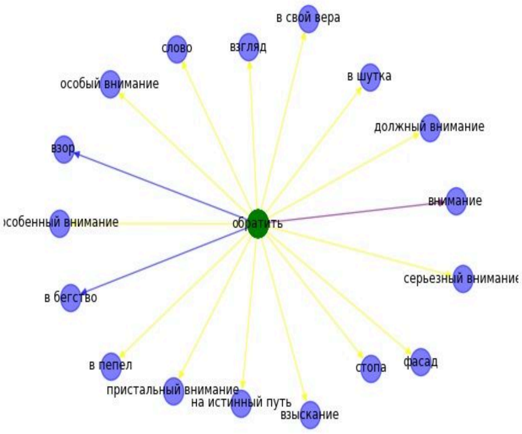
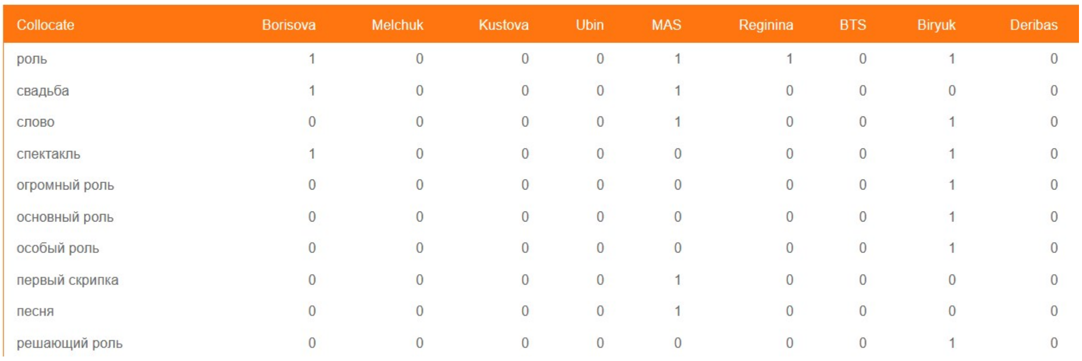


Fig. 2: Distribution of dictionary indices for the verb *obratit′* ‘to turn’

- relative frequency (ipm) based on Russian National corpus and Araneum Russicum Maximum corpus [13];
- values of association measures based on Araneum Russicum Maximum (MI, MI3, log-likelihood, logDice, t-score).

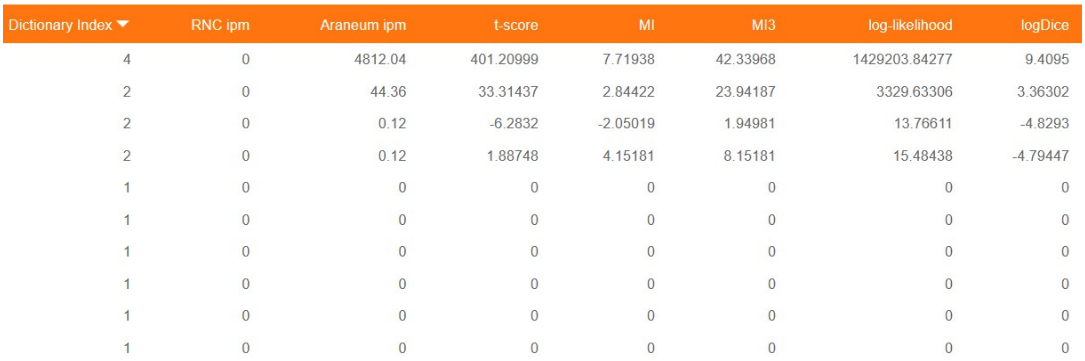
Figures 3 and 4 show examples of the collocations for the verb *igrat′* ‘to play’. Here we find the following examples: *igrat′ rol′* ‘to play a role’ (in four dictionaries), *igrat′ slovami* ‘to play on words’ (in two dictionaries), *igrat′ spektakl′* ‘to play a play’ (in two dictionaries), *igrat′ svad′bu* ‘to celebrate a wedding’ (in two dictionaries).



Collocate	Borisova	Melchuk	Kustova	Ubin	MAS	Reginina	BTS	Biryuk	Deribas
роль	1	0	0	0	1	1	0	1	0
свадьба	1	0	0	0	1	0	0	0	0
слово	0	0	0	0	1	0	0	1	0
спектакль	1	0	0	0	0	0	0	1	0
огромный роль	0	0	0	0	0	0	0	1	0
основный роль	0	0	0	0	0	0	0	1	0
особый роль	0	0	0	0	0	0	0	1	0
первый скрипка	0	0	0	0	1	0	0	0	0
песня	0	0	0	0	1	0	0	0	0
решающий роль	0	0	0	0	0	0	0	1	0

Fig. 3: The first part of the output of the statistical search for the verb *igrat′* ‘to play’.

0 and 1 indicate if the collocation is present or absent in the dictionary. The table is the same for all types of collocations and hence shows many zeros if the



Dictionary Index ▼	RNC ipm	Araneum ipm	t-score	MI	MI3	log-likelihood	logDice
4	0	4812.04	401.20999	7.71938	42.33968	1429203.84277	9.4095
2	0	44.36	33.31437	2.84422	23.94187	3329.63306	3.36302
2	0	0.12	-6.2832	-2.05019	1.94981	13.76611	-4.8293
2	0	0.12	1.88748	4.15181	8.15181	15.48438	-4.79447
1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0

Fig. 4: The second part of the output of the statistical search for the verb *igrat'* 'to play'.

example is not given in a dictionary. Six collocations given in Figure 2 were not found in the following dictionaries [5,7,14] and [15].

The second part of the statistical interface shows quantitative data for collocations (see Figure 4 for possible collocates with the verb *igrat'* 'to play').

3 Statistical Properties and Representation

Statistical validation of the gold standard is an essential step in a database design, as the dictionaries are the product of introspection. Association measures and dictionary indices were used to determine the typical character of word combinations. As statistical indicators, we used highly widespread association measures based on the Araneum Russicum Maximum corpus, namely, MI, MI3, log-likelihood, logDice and t-score. These measures belong to different classes and therefore can produce different results. By interpreting quantitative data, an advanced user can get more thoughtful data. Below we will show which verb-noun collocations are the most frequent when using statistics.

T-score and MI tend to show opposite results [16], and our study confirms this statement. T-score ranges the following collocations as the most typical ones: *igrat' rol'* 'to play a role', *pol'zovat'sya populyarnost'yu* 'to be popular, in favour', *udelyat' vnimaniye* 'to give attention', *prinyat' resheniye* 'to make a decision', *oderzhat' pobedu* 'to win'. Top-50 includes collocations with the nouns *vnimaniye* 'attention' (6 [2]), *populyarnost'* 'popularity' (2) and *rol'* 'role' (3). The most frequent collocations, selected according to the values of this measure, are recorded on average in two dictionaries. One can suggest a correlation between dictionary and corpus data. In other words, t-score can be used to select data for compiling a dictionary and will show the most frequently occurring examples. MI made it possible to find collocations that occur on average in one dictionary. They are represented by the following examples : *tochit' lyasy* 'to chat', *zamorit' chervyachka* 'to have a snack', *zamolvit' slovechko* 'to put in a word (for)', *porot'*

[2] Henceforth, the number of collocations is shown in parentheses.

goryachku ‘to be in a hurry’, *smorozit′ glupost′* ‘to say stupid things’. Both a node and a collocate in each example have a low frequency in the corpus and hence collocations are closer, rather, to idioms or phraseological units due to their non-compositionality. As for the three remaining measures (MI3, log-likelihood, logDice), they produce similar results. On average, collocations occur in one or two dictionaries. The most frequent ones are as follows: *oblizyvat′ pal′chiki* ‘about smth delicious’, *igrat′ rol′* ‘to play a role’, *pol′zovat′sya populyarnost′yu* ‘to be popular, to be in favour’, *privlech vnimaniye* ‘to attract attention’, *vyzvat′ interes* ‘to provoke interest’.

One can notice the following trend: top results for the measure are phrases recorded in different dictionaries. These are more frequent collocations both in terms of statistics and in terms of their reproducibility in speech.

For a visual evaluation of collocations, we used bar plots. Figure 5 shows the values of statistical measures (logDice, MI, MI3, t-score) obtained by collocations with the noun *glupost′* ‘stupidity, stupid things’. The highest value for MI3 (the second bar in each group) is equal to 21.94 and corresponds to the collocation *nadelat′ glupost′* ‘to have stupidity’.

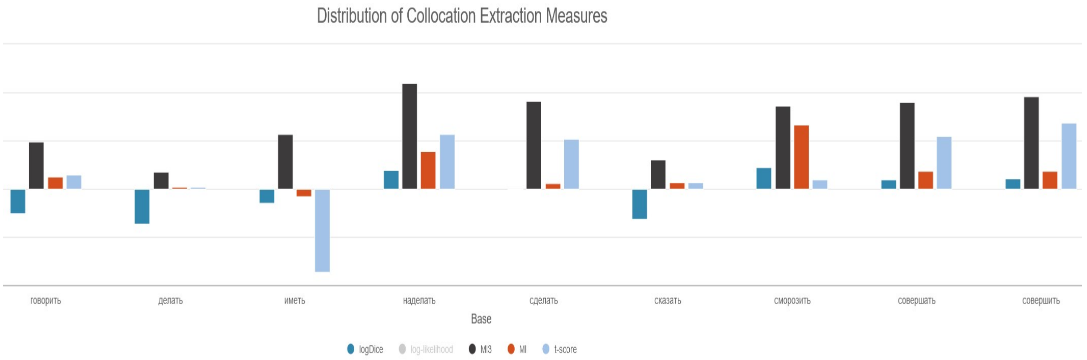


Fig. 5: Distribution of association measures for the noun *glupost′* ‘stupidity, stupid things’

Based on the bar height of the corresponding measure, one can judge how typical a collocation is. For example, *nadelat′ gluposti* ‘to do stupid things’ or *sovershat′/sovershit′ gluposti* ‘to do stupid things’ are frequent, while *smorozit′ glupost′* ‘to say stupid things’ (with the highest value for the MI measure) is almost a phraseological unit. In contrast, negative values for bar plots indicate that collocation is not as common in corpora, despite being registered in dictionaries. Here we can also name *imet′ glupost′* ‘to have stupidity’ (t-score is equal to -17.16) or *delat′ glupost′* ‘to do stupid things’ (logDice is equal to -7.21).

4 Analysis of Dictionary Collocations

Verb-object collocations involves 2,722 verbs, 1,028 (about 38%) among them produce only one collocation. Opposed to adj-noun collocations, verbs are highly productive. Five verbs form more than 200 collocations : *byt'* 'to be' (223), *davat'* 'to give' (216), *dat'* 'to give' (270), *poluchit'* 'to receive' (271), *sdelat'* 'to do, to make' (223). Other productive verbs can be exemplified by the following ones: *imet'* 'to have' (195), *poluchat'* 'to receive' (192), *delat'* 'to do, to make' (187), *prinyat'* 'to receive' (181), *brat'* 'to take' (177), *vzyat'* 'to take' (137), *provesti* 'to conduct, to lead' (134), *provodit'* 'to conduct, to lead' (132), *vesti* 'to conduct, to lead' (131), *vyzvat'* 'to cause' (124), *vyzyvat'* 'to cause' (123), *videt'* 'to see' (121), *proyavlyat'* 'to display, to show' (120), *prinimat'* 'to receive' (119) and *proyavit'* 'to display, to show' (117).

The list of collocates includes 5,665 nouns in total, of which 1,030 (i.e., about 18%) are unique and form only one collocation. The rest of nouns suggest various collocations, exceeding several dozens. The most productive lexemes are, for example, *zhizn'* 'life' (115), *sila* 'force, power' (103), *delo* 'case, matter' (103), *slovo* 'word' (97), *rabota* 'job, work' (85), *vremya* 'time' (83), *vzglyad* 'glance, opinion' (83), *vopros* 'question' (75), *vozmozhnost'* 'opportunity, possibility' (71), *pravo* 'right' (68), *vnimaniye* 'attention' (64), *interes* 'interest' (63), *polozheniye* 'position' (62), *otnosheniye* 'attitude, relation' (58), *chuvstvo* 'feeling' (56), *nadezhda* 'hope' (56), *mysl'* 'idea, thought' (55), *glaz* (53), *initsiativa* 'initiative' (52) and *vlast'* 'authority' (51).

It should be noted that, unlike nouns, verbs show more significant variability in producing collocations. On average, there are 7.4 collocations per verb, while there are 3.6 collocations per noun.

5 Conclusion and Future Work

In the paper, we discussed the features of the Russian collocations database and analyzed the examples of verb-object collocations. We traced the possible correlation between dictionaries and statistical coefficients. It can be noted that collocations from the top of the list of any measure are more stable and are described in several dictionaries. In verb-object collocations, verbs tend to be more productive than nouns and produce more examples.

We have shown some technical details concerning the database. Visualisation helps users to understand the usage of collocations in speech: how frequent and typical they are. However, it is necessary to enhance the results. Output collocations are shown for lemmas, while it is better to display them as token collocations. There are many zeros in the tables indicating that many phrases are recorded only in one dictionary. Such a table view can be questioned if it is appropriate and user-friendly and might be changed in future.

Acknowledgements. The presented research was supported by the Russian Science Foundation, project No. 22-18-00189 "Structure and functionality of stable multiword units in Russian everyday speech".

References

1. Russian Collocations Database, `http://collocations.spbu.ru`. Last accessed 5 Nov 2023
2. Khokhlova, M.: Collocations in Russian Lexicography and Russian Collocations Database. In: Proceedings of The 12th Language Resources and Evaluation Conference. Marseille, France, pp. 3198–-3206. European Language Resources Association (2020).
3. Biriuk, O.L., Gusev, V.Yu., Kalinina, E.Yu.: Dictionary of Russian Abstract Nouns' Verbal Collocability. A Dictionary based on the Russian National Corpus [Slovar' Glagol'noj Sochetaemosti Nepredmetnykh Imen Russkogo Yazyka. Slovar' na osnove Natsional'nogo Korpusa Russkogo Yazyka] (2008)
4. Borisova, E.G.: A Word in a Text. A Dictionary of Russian Collocations with English-Russian Dictionary of Keywords [Slovo v tekste. Slovar' kollokatsiy (ustoychivykh sochetaniy) russkogo yazyka s anglo-russkim slovarem klyuchevykh slov]. Filologiya, Moscow (1995)
5. Deribas, V.M.: Verb-Noun Collocations in Russian. [Ustojchivyje glagol'no-imennyje slovosochetanija russkogo jazyka]. Russkij jazyk, Moscow (1983)
6. Dictionary of the Russian Language in 4 volumes [Slovar' russkogo yazyka v 4 tomakh] (MAS) (1999), Yevgen'yeva A. P. (ed.-in-chief). Vol. 1–4, 4th edition, revised and supplemented. Russkiy yazyk, Moscow.
7. Mel'čuk, I., Zholkovsky, A.: Explanatory Combinatorial Dictionary of Modern Russian [Tolkovo-kombinatornyy slovar russkogo yazyka]. Vienna (1984)
8. Reginina, K.V., Tjurina, G.P., Shirokova, L.I.: Set Expressions of the Russian Language. A Reference Book for Foreign Students [Ustoychivye slovosochetaniya russkogo yazyka: Uchebnoye posobiye dlya studentov-inostrantsev], Shirokova, L.I. (ed.). Moscow (1980)
9. Khokhlova, M.: Attributive collocations in the Russian gold standard and their representation in dictionaries and corpora [Attributivnyje kollokatsii v zolotom standarte sochetajemosti russkogo jazyka i ih predstavlenije v slovarjah i korpusah tekstov]. Questions of Lexicography (21), 33–68 (2021)
10. Russian National Corpus, `http://ruscorpora.ru`. Last accessed 5 Nov 2023
11. SynTagRus, `https://ruscorpora.ru/new/search-syntax.html`. Last accessed 5 Nov 2023
12. Taiga, `https://tatianashavrina.github.io/taiga_site`. Last accessed 5 Nov 2023
13. Benko, V.: Aranea Yet Another Family of (Comparable) Web Corpora. In: Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS, vol. 8655, pp. 257–264. Springer, Heidelberg (2014)
14. Kustova, G.I.: Dictionary of Russian Idiomatic Expressions [Slovar' russkoyj idiomatiki. Sochetaniya slov so znacheniyem vysokoy stepeni] (2008).
15. Big Russian explanatory dictionary [Bol'shoy tolkovyy slovar' russkogo yazyka] (BTS), Kuznetsov, S.A. (ed.). Norint, St. Petersburg (1998)
16. Gablasova, D., Brezina, V., McEnery, T.: Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. Language Learning, 67(S1), 155–179 (2017)