Baltic J. Modern Computing, Vol. 10 (2022), No. 3, pp. 481–489
https://doi.org/10.22364/bjmc.2022.10.3.20

Approaches and Resources for Lithuanian Collocation Research

Jolanta KOVALEVSKAITĖ, Loïc BOIZOU, Ieva BUMBULIENĖ, Erika RIMKUTĖ, Jurgita VAIČENONIENĖ

Vytautas Magnus University, K. Donelaičio str. 58, 44248, Kaunas, Lithuania

jolanta.kovalevskaite@vdu.lt, loic.boizou@gmail.com, ieva.bumb@gmail.com, erika.rimkute@vdu.lt, jurgita.vaicenoniene@vdu.lt

Abstract. This paper gives an overview of the conducted research on Lithuanian multi-word expressions, particularly, collocations, and presents the developed resources. Beginning from the identification, analysis, and lexicographic description of MWEs in general, the focus of the research was narrowed down to collocations and their certain features, such as arbitrariness, during the later stages. Arbitrary collocations were seen as having lexically motivated relations and a certain degree of restricted collocability of constituents. Within the framework of the two projects, PASTOVU (2016-2018) and ARKA (2020-2022), corpus-driven approaches were applied to extract and document Lithuanian collocations and develop a number of lexical resources to be discussed in this paper.

Keywords: Lithuanian multi-word expressions, Lithuanian collocations, arbitrary collocations, Gravity Counts, hybrid methods for collocation identification, lexical resources.

1 Introduction

Prior to the corpus-driven research of multi-word expressions (MWEs), the related Lithuanian lexical resources were mostly devoted to the description of figurative expressions as idioms or sayings. Such MWEs would typically be marked in general dictionaries or published in specialized dictionaries of phraseology (*Frazeologijos žodynas*)[1], collections of sayings, and other publications (for example, (Lipskienė, 2008); (Kašėtienė and Kudirkienė, 2016)). Research on collocations, which form a significant part of the formulaic language, began with the automatic identification of MWEs in corpora, specifically, the *Corpus of Contemporary Lithuanian Language* and the DELFI.lt corpus. In addition to collocation research, lexical bundles defined by Biber (2009) have been studied in Lithuanian texts (for example, in academic discourse) on the basis of

[1] https://ekalba.lt/frazeologijos-zodynas/

which special databases representing the academic formulaic language were developed (e.g., the *Compendium of Academic Phrases*)[2].

The aim of this article is to present the developments of research on Lithuanian collocations and related corpus-driven resources. In Section 2, research on the Lithuanian MWE identification in the general corpus and the retrieved data on formulaic language are described. We narrow the focus down to collocation research in Section 3, which presents the outcomes of the PASTOVU and ARKA projects. During the PASTOVU project, hybrid methods were applied for the collocation extraction in non-general corpus; as a result, lexical database PASTOVU was compiled. This database was used to extract arbitrary collocations during the ARKA project. A demonstration of the PASTOVU database which integrates the results of the two projects is presented in Section 4.

2 Extraction of Multi-Word Expressions from the *Corpus of Contemporary Lithuanian Language*

First attempts to extract Lithuanian MWEs automatically were made using the general *Corpus of Contemporary Lithuanian Language*[3] (Rimkutė et al., 2010), particularly, its unannotated version consisting of 100 million running words. As expected, a rich diversity of MWEs was detected because of the corpus structure and extraction methods described further on.

2.1 MWE Identification Using Gravity Counts

The method of Gravity Counts (GC) (Daudaravičius and Marcinkevičienė, 2004) was used for the extraction of collocational chains. The method measured the combinability and collocability for each pair of words in the corpus within the moving span of three words to detect frequent, recurrent, and uninterrupted strings of word-forms. GC was based on a full text approach, where the corpus is seen as a changing curve of lexical combinability (see Daudaravičius and Marcinkevičienė, 2004, Fig. 5): "A collocational chain is [. . .] a segment of text where the combinability of constituent adjacent word pairs is above the arbitrary chosen point of collocability. The lower combinability of word pairs preceding and following the segment marks the boundaries of a chain" (see Daudaravičius and Marcinkevičienė, 2004, p. 334). GC was based on a statistical approach with no machine learning involved.

2.2 *Lexical Database of the Dictionary of Lithuanian Phrases*

The statistical collocational chains obtained from the corpus were processed manually. Certain chains were autonomous and grammatically well-formed phrases, whereas other chains were deficient. Thus, during the linguistic evaluation, some chains were deleted, shortened or augmented to select only grammatical and meaningful phrases.

[2] `http://www.frazynas.flf.vu.lt`

[3] `http://corpus.vdu.lt/lt/`

Phrases with at least one noun were included in the database. The collected database contains phrases of different lengths, from 2 words to 46 words, in total 69,000 phrases. 2-word, 3-word, and 4-word phrases make up 84 percent of all the data (for more details on the lexical and grammatical structure of the phrases (Marcinkevičienė and Grigonytė, 2004); (Rimkutė et al., 2012, p. 19-23). 2-4-word phrases revealed many collocations; other types of MWEs (idioms, sayings, formulas, etc.) were also represented. Accordingly, in the *Database of the Dictionary of Lithuanian Phrases*[4] (2012), the user could see the whole spectrum of the Lithuanian formulaic language and to search the data by lexical and grammatical criteria for the first time. This database became a basis for further Lithuanian MWE research (see (Marcinkevičienė and Grigonytė, 2004); (Boizou et al., 2015)).

3 Extraction of Multi-Word Expressions from the DELFI.lt Corpus

In 2016-2018, the project "Automatic Identification of Lithuanian Multi-word Expressions (PASTOVU)"[5] became a second step in deeper investigations of the Lithuanian collocations. The project aimed to create a methodology for the MWE (collocations and idioms) analysis in contemporary written Lithuanian, to investigate Lithuanian collocations, to create or adapt tools necessary for the Lithuanian collocation research, and to compile the *Database of Lithuanian MWEs* and the first *Dictionary of Lithuanian Collocations*.

For the collocation extraction task, the DELFI.lt corpus[6] was compiled to test statistical, machine learning, and hybrid methods. This corpus consists of 70 million running words of articles published in 2014-2016 in the DELFI.lt news portal. Texts were collected from the 12 topical categories (science, sport, people, automobiles, and others). The DELFI.lt corpus was automatically morphologically annotated using the morphological analyser of the webservice Semantika.lt[7].

3.1 Extraction of Multi-Word Expressions Using Hybrid Methods

The data provided in the *Database of Lithuanian MWEs* (further on, lexical database PASTOVU) is a small part of a large set of MWEs that were extracted from the DELFI.lt corpus by applying machine learning (ML) approaches. Training was carried out using a small subset of the corpus manually annotated by a group of linguists. The validity of the identified MWEs was checked by cross-validation. The characteristics used to build the ML model included MWE forms, as well as morphological information and a set of 17 statistical measures: mutual information (MI), pointwise mutual information, MI2, MI3, log likelihood ratio, Dice coefficient, logarithmic Dice coefficient, t-test, Pearson's chi-square test, phi coefficient, odds ratio, z-score, Poisson-Stirling measure, geometric

[4] `https://klc.vdu.lt/fraziu-zodynas/`

[5] `https://pastovu.vdu.lt/en/`

[6] `https://klc.vdu.lt/pastovuSearch.html`

[7] `http://semantika.lt/`

mean score, relative risk, Liddel coefficient, and minimum sensitivity (MS). In order to make the values comparable, all measures that are not originally expressed from 0 to 1 were re-scaled accordingly.

Using these features, two types of ML methods were used to extract potential MWEs from the whole corpus:

- n-gram classifier: Naive Bayes;
- sequence classifiers: conditional random fields (CRF) and long short-term memory (LSTM).

The Naive Bayes method was used to identify bigrams, while trigrams were recognized by either sequence classifier. The use of ML approach (with statistical measures as features) significantly improved the F1 score (0.6 for bigrams) in comparison to the usual statistical approaches (0.31 for Dice coefficient, other measures or combination of measures giving even lower F1 scores) (for more details, see (Bielinskienė, Boizou, Bumbulienė, Kovalevskaitė, Krilavičius, Mandravickaitė, Rimkutė and Vilkaitė-Lodzienė, 2019)).

This approach was also used in the development of the *Colloc* tool[8], which is designed to identify MWEs in a given text. This tool is freely available, thus users can extract collocations in their own texts by uploading the file to the tool which returns the annotated text. The experimental prototype includes all the steps of linguistic analysis, namely: text pre-processing; PoS tagging; n-gram generation and calculation of their statistical properties; calculation of Lexical Association Measures (LAMs); word embedding generation; MWE identification using filtering (gazetteers, dictionaries), the application of LAMs, the application of machine learning, and the hybrid methods. The tool has been statistically trained on the GloVe word vectors and uses neural networks.

3.2 Data Selection for the PASTOVU Database

The procedure described in Section 3.1 allowed us to obtain a list of co-occurrences, i.e., potential MWEs, that amounted to hundreds of thousands. Since the purpose of the PASTOVU database was to provide rich and linguistically sound information, the list had to be heavily reduced to be manually reviewed by the linguists.

In order to have a synthetic view of some lexical variations and to avoid considering only the highly frequent combinations, a decision was made to select the items containing the most frequent 97 nouns (three nouns of the original list of the first 100 nouns were discarded because they belonged to several parts of speech) in the automatically retrieved potential MWEs. In addition, only MWEs with a frequency of 10 or more occurrences in the corpus were selected to ensure a minimal diversity among concordances. This new list (about 30,000 items) was manually filtered out according to the selected grammar patterns (noun + noun, adjective + noun, verb + noun), thus leaving a final list of about 12,000 MWEs. As MWEs were only 2- and 3-word units, most of them were collocations, with a rather low percentage of idioms.

Additional automatic steps were performed to enrich information for the selected MWEs using information from the DELFI.lt corpus. For each MWE, grammatical forms, their frequency, and a set of concordances were retrieved (see Section 3.4 for

[8] `https://resursai.pastovu.vdu.lt/atpazintuvas`

more information about the PASTOVU database). The concordances allowed to select usage examples.

3.3 Identification of Arbitrary Collocations

In the lexical database PASTOVU, the largest part of the data consists of collocations, understood here as usage-based grammatically well-formed word combinations. Nevertheless, a significant part of the data form trivial collocations – semantically motivated and largely predictable word combinations depicting the nonverbal reality (e.g., *saulėta diena 'a sunny day'*, *atidaryti langą 'to open the window'*). Accordingly, only a part of collocations can be seen as arbitrary (further on ACs), characterized by lexically motivated relations between the constituents and a certain degree of usage restrictedness: e.g., although there may be several close synonyms, a particular one is preferred in a certain word combination (e.g., *broad/wide outlook* vs.**big outlook*). ACs are particularly important in foreign language learning, teaching, translating, or text editing.

Within the framework of the project "Arbitrary Collocations of Lithuanian: Identification, Description and Usage (ARKA)"[9], a methodology for the recognition of Lithuanian ACs was developed. The source for AC extraction was the PASTOVU lexical database that encompasses more than 12,000 collocations from the DELFI.lt corpus. The data included adjectival (e.g., *greitas sprendimas 'a quick decision'*), verbal (e.g., *sekti įvykius 'to follow events'*), and nominal collocations (e.g., *žemės sklypas 'a land plot'*). The AC identification workflow consisted of manual and semi-automatic data analysis.

Manually, collocations were marked as arbitrary if they met (1) lexical restrictedness (Nesselhauf, 2003) and (or) (2) meaning transfer criteria (Marcinkevičienė, 2010). Lexical restrictedness was evaluated by applying a synonym substitution of pre-modifier and (or) semantic field comparison of the head noun test:

1. Synonym substitution: a collocation was seen as arbitrary if the pre-modifying component of the collocation could not be replaced by a close synonym without a change in the meaning (e.g., *palanki aplinka 'favourable environment'* vs. **naudinga aplinka 'useful environment'*);
2. Semantic field comparison: the potential collocability of collocation constituents with a wider or narrower range of nouns was discussed (e.g., *parkuoti automobilį 'to park a car'* vs *prišvartuoti laivą 'to moor a ship'*);
3. Meaning transfer: collocations with restricted collocability words, abstract nouns, and metaphorization potential were seen as arbitrary (e.g., *skaudus klausimas 'a painful question'*).

The semiautomatic approach consisted of three stages: (1) automatic generation of vector strings with potential synonyms (Pennington et al., 2014)[10]; (2) manual vector string editing to reduce the noise, as the same vector string could include both close and distant synonyms (sometimes several pairs) or antonyms; (3) semiautomatic process during which the collocations were compared to particular synonym pairs in vector

[9] Project conducted in 2020-2022, `https://arka.pastovu.vdu.lt/en/`.

[10] To minimize noise in data extraction, the similarity index was modified to no less than 0,5 (the maximum similarity index is 1).

strings to be approved or not as arbitrary. An example of AC recognition workflow is as follows:

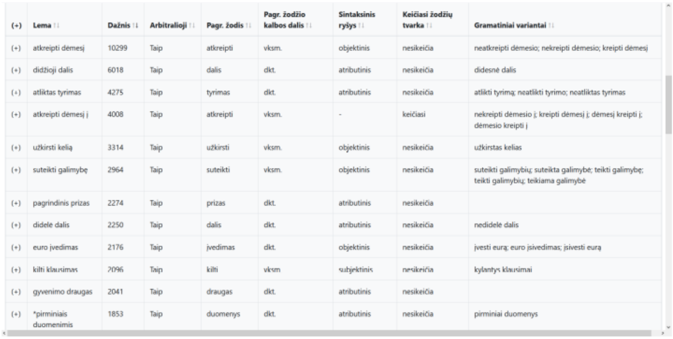
1. synonym pairs in the adjective vector string: *svarbus 'important'*; *reikšmingas 'significant'*;
2. available collocations with the noun *asmuo 'person'*: *svarbus asmuo 'important person'*;
3. decision: *svarbus asmuo 'important person'* is an AC.

In the first stage, out of 12,000 collocations, 2,000 were identified as arbitrary in the PASTOVU database. Approximately half of ACs were detected manually, one-third semiautomatically, and one fifth using a combination of both methods (see (Kovalevskaitė et al., 2021b); (Kovalevskaitė et al., 2021a)). In the next stage, to reduce manual work, new ACs were extracted semi-automatically from the dataset of 2-grams (freely available at https://clarin.vdu.lt/xmlui/handle/20.500.11821/25) generated on the basis of the DELFI.lt corpus. As a result, 7000 new ACs were added to the PASTOVU database.

3.4 *The Database of Lithuanian MWEs*

At present, the PASTOVU database contains 19,000 collocations: 10,000 trivial collocations or idioms consisting of 2- or 3-words (see section 2.2) and, approximately, 9,000 ACs (see Section 2.3.). Although some MWEs are idioms, their number is low.

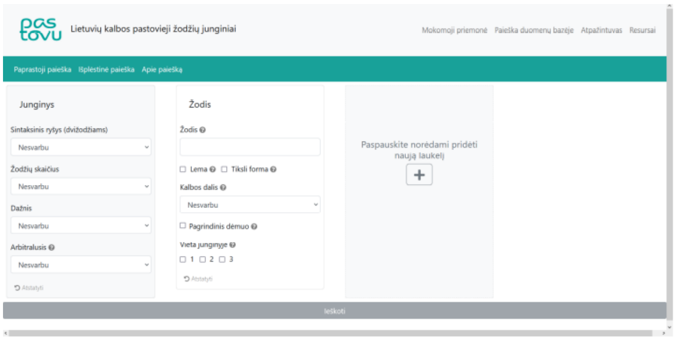
The database provides diverse information about collocations: lemmas, word forms and their frequency, morphological information, and syntactic relations of 2-word collocations (attributive, subject, object, or adverbial). The collocation entry also includes usage-related information: grammatical variants (Bielinskienė, Kovalevskaitė, Rimkutė and Vilkaitė, 2019), examples of concordances, information on possible insertions of words between the MWEs parts, and changes in word order:



(+)	Lema	Dažnis	Arbitralioji	Pagr. žodis	Pagr. žodžio kalbos dalis	Sintaksinis ryšys	Keičiasi žodžių tvarka	Gramatiniai variantai
(+)	atkreipti dėmesį	10299	Taip	atkreipti	vksm.	objektinis	nesikeičia	neatkreipti dėmesio; nekreipti dėmesio; kreipti dėmesį
(+)	didžioji dalis	6018	Taip	dalis	dkt.	atributinis	nesikeičia	didesnė dalis
(+)	atliktas tyrimas	4275	Taip	tyrimas	dkt.	atributinis	nesikeičia	atlikti tyrimą; neatlikti tyrimo; neatliktas tyrimas
(+)	atkreipti dėmesį į	4008	Taip	atkreipti	vksm.	-	keičiasi	nekreipti dėmesio į; kreipti dėmesį į; dėmesį kreipti į; dėmesio kreipti į
(+)	užkirsti kelią	3314	Taip	užkirsti	vksm.	objektinis	nesikeičia	užkirstas kelias
(+)	suteikti galimybę	2964	Taip	suteikti	vksm.	objektinis	nesikeičia	suteikti galimybių; suteikta galimybė; teikti galimybę; teikti galimybių; teikiama galimybė
(+)	pagrindinis prizas	2274	Taip	prizas	dkt.	atributinis	nesikeičia	
(+)	didelė dalis	2250	Taip	dalis	dkt.	atributinis	nesikeičia	nedidelė dalis
(+)	euro įvedimas	2176	Taip	įvedimas	dkt.	objektinis	nesikeičia	įvesti eurą; euro įsivedimas; įsivesti eurą
(+)	kilti klausimas	2096	Taip	kilti	vksm	subjektinis	nesikeičia	kylantys klausimai
(+)	gyvenimo draugas	2041	Taip	draugas	dkt.	atributinis	nesikeičia	
(+)	*pirminiais duomenimis	1853	Taip	duomenys	dkt.	atributinis	nesikeičia	pirminiai duomenys

Fig. 1. Example of search results in the database

The database can be queried through simple or advanced search (by selecting such parameters as syntactic relations between collocation constituents, the frequency of use, arbitrariness, etc.) (see Figure 2).



Junginys

Sintaksinis ryšys (dvižodžiams): Nesvarbu

Žodžių skaičius: Nesvarbu

Dažnis: Nesvarbu

Arbitralusis: Nesvarbu

Žodis

Žodis:

□ Lema □ Tiksli forma

Kalbos dalis: Nesvarbu

□ Pagrindinis dėmuo

Vieta junginyje: □ 1 □ 2 □ 3

Paspauskite norėdami pridėti naują laukelį +

Fig. 2. An example of advanced search in the PASTOVU database

The database was used as a resource for the compilation of the *Lithuanian Collocation Dictionary*((Bielinskienė, Boizou, Bumbulienė, Kovalevskaitė, Krilavičius, Mandravickaitė, Rimkutė and Vilkaitė-Lodzienė, 2019)) which includes approximately 12,000 collocations with the 97 most frequently used nouns in the word list (see Section 2.2). The dictionary presents the collocates of these nouns, classified according to the part of speech and syntactic relations. All inflectional forms of each MWE are provided as a dictionary appendix. As one the results of the ARKA project, a *Dictionary of Lithuanian Arbitrary Collocations* (Boizou et al., 2022) was compiled. In the main part of the dictionary, ACs are arranged according to the head noun and accompanying adjectives, nouns, and verbs (syntactic relations are also reflected for the latter category). The first appendix lists all ACs included in the dictionary in alphabetical order. The second appendix lists all AC constituents (2318 different words in total) in alphabetical order. Both dictionaries are freely available for users via the project websites.

4 Conclusions and Further Research

The aim of the article was to briefly overview the most recent Lithuanian collocation research and developed resources. At present, two Lithuanian MWE databases are freely available online: the *Database of the Dictionary of Lithuanian Phrases* and the *Database of Lithuanian Multiword Expressions*. The PASTOVU database, which was described in more detail, makes a distinction between the trivial and arbitrary collocations. For example, of the 7300 adjectival collocations, 3500 are marked as arbitrary. It would be relevant to continue research by investigating the collocational strength differences of the constituents of these collocations. It might be argued that ACs would be more likely to fall within the medium collocability range combinations of words (in comparison to trivial collocations or free combinations of words). To test these predictions, relevant collocation strength measures should be selected as, for example, directionality ((Brezina et al., 2015); (Gries, 2013); (Handl, 2008), cited in (Gablasova et al., 2017, p. 160)). An important question is whether collocational strength can contribute to automatic AC recognition and, accordingly, minimize manual work in the AC identification task. In sum, the PASTOVU database includes a large number of collocations described in terms of composition and usage and could be further developed by including the data (ACs) not only from journalistic, but also form other types of discourse (both written and spoken).

Acknowledgements

This research was funded by a grant (No. S-LIP-20-18) from the Research Council of Lithuania. The project "Automatic Identification of Lithuanian Multi-word Expressions (PASTOVU)" was funded by a grant (No. LIP-027/2016) from the Research Council of Lithuania.

References

Biber, D. (2009). A corpus-driven approach to formulaic language in English. Multi-word patterns in speech and writing, *International Journal of Corpus Linguistics* **14**(3), 275–311.

Bielinskienė, A., Boizou, L., Bumbulienė, I., Kovalevskaitė, J., Krilavičius, T., Mandravickaitė, J., Rimkutė, E., Vilkaitė-Lodzienė, L. (2019). *Lietuvių kalbos kolokacijų žodynas*, Vytauto Didžiojo universitetas, Kaunas.
`https://doi.org/10.7220/9786094673733`

Bielinskienė, A., Kovalevskaitė, J., Rimkutė, E., Vilkaitė, L. (2019). Lietuvių kalbos pastoviųjų junginių gramatinis variantiškumas, *Kalbų studijos* **34**, 91–110.

Boizou, L., Bumbulienė, I., Kovalevskaitė, J., Rimkutė, E., Vaičenonienė, J. (2022). *Lietuvių kalbos arbitraliųjų kolokacijų žodynas*, Vytauto Didžiojo universitetas, Kaunas.
`https://doi.org/10.7220/9786094675232.`

Boizou, L., Kovalevskaitė, J., Rimkutė, E. (2015). Automatic lemmatisation of Lithuanian MWEs, *NODALIDA 2015: proceedings of the 20th Nordic conference of computational linguistics*, pp. 41–49.

Daudaravičius, V., Marcinkevičienė, R. (2004). Gravity Counts for the boundaries of collocations, *International Journal of Corpus Linguistics* **9**(2), 321–348.

Gablasova, D., Brezina, V., McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence, *Language Learning* **67**(1), 155–179.

Kašėtienė, R., Kudirkienė, L. (2016). *Vilką minim, vilkas čia. Lietuvių situaciniai posakiai*, Lietuvių literatūros ir tautosakos institutas, Vilnius.

Kovalevskaitė, J., Rimkutė, E., Vaičenonienė, J. (2021a). Arbitraliųjų lietuvių kalbos kolokacijų nustatymas, *Bendrinė kalba* **94**, 1–37.

Kovalevskaitė, J., Rimkutė, E., Vaičenonienė, J. (2021b). Automatizuotas arbitraliųjų kolokacijų atpažinimas: būdvardžių ir daiktavardžių kolokacijos, *Studies About Languages* **39**, 71–84.

Lipskienė, J. (2008). *Vaizdingieji lietuvių kalbos posakiai*, Lietuvių kalbos institutas, Vilnius.

Marcinkevičienė, R., Grigonytė, G. (2004). The dictionary of Lithuanian phrases, *Proceedings of the 2nd International Conference on BALTIC HUMAN LANGUAGE TECHNOLOGIES*, pp. 299–304.

Marcinkevičienė, R. (2010). *Lietuvių kalbos kolokacijos*, Vytauto Didžiojo universitetas, Kaunas.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching, *Applied Linguistics* **24**(2), 223–242.

Pennington, J., Socher, R., Manning, C. D. (2014). GloVe: Global vectors for word representation, *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Rimkutė, E., Bielinskienė, A., Kovalevskaitė, J. (2012). *Lietuvių kalbos daiktavardinių frazių žodynas*, Vytauto Didžiojo universitetas, Kaunas. http://donelaitis.vdu.lt/lkk/pdf/daikt_fr.pdf

Rimkutė, E., Kovalevskaitė, J., Melninkaitė, V., Utka, A., Vitkutė-Adžgauskienė, D. (2010). Corpus of contemporary Lithuanian language - the standardised way, *The Fourth International Conference on HUMAN LANGUAGE TECHNOLOGIES — THE BALTIC PERSPECTIVE*, pp. 154–160.

Received August 18, 2022 , accepted August 22, 2022