From Lexical Entries to Corpus Queries: Retrieving Multiword Expressions in Serbian

Cvetana Krstev
Association for Language Resources and Technologies
Belgrade, Serbia
cvetana@jerteh.rs

Ranka Stanković
University of Belgrade
F. of Mining and Geology
Belgrade, Serbia
ranka@rgf.rs

Aleksandra Marković
Institute for the Serbian Language SASA
Belgrade, Serbia
malexa39@gmail.com

Relevant UniDive working groups: WG2

1 Introduction

Several linguistic phenomena complicate the retrieval of MWEs in corpora. This research aims to propose a methodology for transforming lexicographic encoding of multiword expressions (MWEs) into effective corpus queries. The work is motivated by the development of the Dictionary of the Modern Serbian Language (RSSJ) (Stanković et al., 2025), although the issues discussed are also relevant to other lexical resources. Particular attention is devoted to the representation of MWEs and to the need for their precise encoding, including complements, modifiers, optional elements, morphosyntactic and word-order variation, while preserving established lexicographic traditions to some extent. These issues and their consequences for NLP are discussed in more detail in (Krstev and Vitas, 2017). The automatic construction of a lemma is based on a set of manually crafted rules, designed following expert analysis of available MWU lemmas (Krstev et al., 2013). Based on such encoding, the study explores how to derive corpus queries that are both linguistically precise and sufficiently comprehensive. This includes determining appropriate part-of-speech constraints and designing queries that capture relevant morphosyntactic and other variation in corpus data (Barbu Mititelu et al., 2025). The MWE-Finder is a tool for searching flexible MWEs in corpora that automatically generates XPath-based query patterns from canonical forms, enabling retrieval and analysis of MWEs while leveraging syntactic structure in treebank data (Odijk et al., 2024).

The traditional practice for encoding MWEs in Serbian descriptive dictionaries is not appropriate for their encoding in digital lexicographic resources linked with corpora. The usual approach is to use only one type of brackets: "()" in expressions, denoting several different things (alteration, less frequent variants, omissible or facultative components, etc.). For example: *baciti / bacati (podmetnuti / podmetati, staviti / stavljati, gurnuti / gurati) [nekome] klip\|klipove pod noge (pod točkove, u točkove)* lit. throw (frame, put, push) [someone] chock\|chocks under someone's legs (wheels) 'to put a spoke in someone's wheel'. Some dictionaries use "()", "/" and "[]" (Fig 1).

Figure 1: An illustration of meta elements in MWE

The main research questions addressed in this work are therefore:

RQ1 How can MWEs encoded in lexical resources be retrieved from large corpora?

RQ2 How can corpus queries account for morphosyntactic variation and tagging inconsistencies?

RQ3 How can MWE entries, often containing optional elements, variants, or obligatory complements, be transformed into effective corpus queries?

Addressing these questions improves the integration of lexicographic resources with corpus-based evidence and supports more systematic phraseological analysis.

2 Methodology

The proposed methodology begins by specifying modifications to current Serbian lexicographic

practice in encoding MWEs, followed by the formulation of a set of rules that systematically transform such encodings into corpus queries. The development and application of these rules rely on the analysis of MWEs in the lexical database Leximirka (Stanković et al., 2018; Lazić and Škorić, 2020), where nominal multiword units are represented through their canonical forms and linked to their constituent lexical elements. This structure enables the automatic generation of corpus queries by decomposing expressions into their components and retrieving possible lemma variants. The resulting CQL queries capture morphosyntactic variation and potential tagging inconsistencies, prioritizing recall to ensure that valid corpus attestations of MWEs are not missed. However, several linguistic phenomena complicate the retrieval of MWEs in corpora.

One major issue is free word order, which allows the components of a phraseological unit to appear in different configurations. E.g., the expression *kvadratni metar* ('square meter') may also appear as *metar kvadratni* (lit. meter square) A. To capture both variants, the query must allow alternative component orders (Fig. 2).

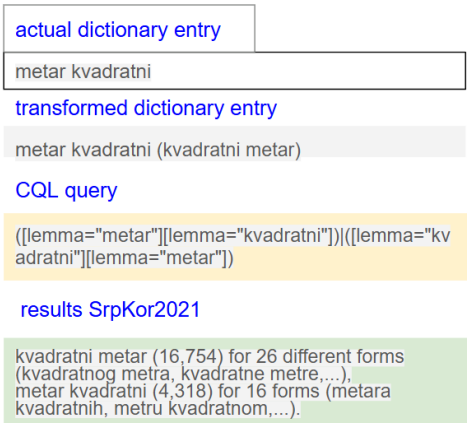

Figure 2: An illustration of capturing alternative WO

Another challenge arises from discontinuous expressions, where the components of an idiom may be separated by intervening words. For example, the idiom *ispod brka | pod brkom (na)rugati se, (na)smejati se* (lit. to smile under the mustache, 'to be secretly amused') can appear in several discontinuous forms, one of them with the noun *mustache* modified: *smešeći se ispod podšišanih brkova* (lit. to smile under the trimmed mustache). Such variability requires corpus queries that allow flexible patterns rather than strictly adjacent tokens.

Metalanguage in dictionary descriptions may cause further complications. Bracket usage may refer to prefixes, alternative prepositions, or optional components (Fig. 3).

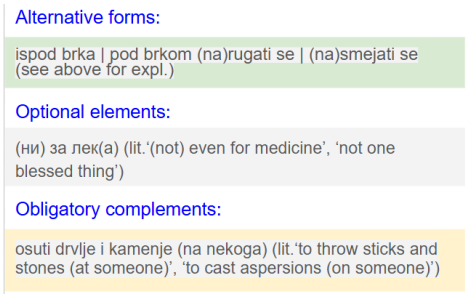

Figure 3: Brackets for different purposes

It is well known that describing colligational preferences makes a lexicographic description more reliable, and this holds especially for MWEs. There are many phraseological MWEs whose use is restricted to one or just a few possible morphological forms, not in the whole paradigm. For example, the Serbian idiom *ne podnosim ga/je* 'I can't stand him/her' is restricted to negative form, while some verbs are used predominantly or only in imperative, past participle, etc (Marković, 2018). This information is not systematically presented in Serbian dictionaries, thereby hindering the retrieval of MWEs from corpora. Since dictionaries use different metalanguages and we still lack a unified encoding of MWEs, transforming lexicographic MWE encodings into corpus queries is not straightforward.

3 Data

Leximirka currently contains more than 250,000 lexical entries, of which approximately 23,000 are MWEs (NMWEs). These include idioms, collocations, terminological expressions, and named entities. Each lexical entry is related to its possible inflected forms. For example, the idiom *ljubavna priča* ('love story') appears in forms such as *ljubavne priče*, *ljubavnih priča*, *ljubavnim pričama*, *ljubavnoj priči*, *ljubavnom pričom*, and *ljubavnu priču* (according to case and number). These forms are taken into account when generating corpus queries for the Serbian corpora SrpKor (Vitas et al., 2025). An additional resource presents adjective and verbal similes in tabular form, along with their possible variations. It now includes more than 900 such descriptions, which can be transformed into queries in the form of finite automata or, alternatively, CQP (Krstev et al.,

2023). One such example is *pištati kao (ljuta) zmija|guja (u procepu)* 'to hiss like an (angry) snake|viper (in a crevice)'.

In addition, the study examines phraseological data from three Serbian dictionaries: a Serbian–English idioms dictionary, a phraseological dictionary, and a descriptive dictionary. These resources provide additional MWEs, including nominal, verbal, adverbial, and functional, as well as additional examples of phraseological descriptions using metalanguage conventions that influence the transformation of lexical entries into corpus queries.

4 Conclusion

This work contributes to phraseology and lexicography by proposing a systematic approach to linking lexical databases and corpus data, and to decomposing and transforming MWEs into flexible corpus queries that account for morphological variation and tagging inconsistencies. The analysis highlights the importance of balancing recall and precision when designing corpus queries for phraseological research. By prioritizing recall, the approach ensures that valid occurrences are not missed, while subsequent filtering can remove irrelevant examples.

Acknowledgements

This research was supported by the Science Fund of the Republic of Serbia, #GRANT 7276, Text Embeddings - Serbian Language Applications – TESLA, the Ministry of Science of the Republic of Serbia (451-03-33/2026-03/200174), and COST ACTION CA21167 - Universality, Diversity, and Idiosyncrasy in Language Technology (UniDive).

References

Verginica Barbu Mititelu, Voula Giouli, Gražina Korvel, Chaya Liebeskind, Irina Lobzhanidze, Rusudan Makhachashvili, Stella Markantonatou, Alexandra Markovic, and Ivelina Stoyanova. 2025. The Challenges of Syntactic Descriptions of Multiword Expressions in Electronic Lexicography. In *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 Conference*, pages 253–273.

Cvetana Krstev, Ivan Obradović, Ranka Stanković, and Duško Vitas. 2013. An approach to efficient processing of multi-word units. In *Computational Linguistics: Applications*, pages 109–129. Springer.

Cvetana Krstev, Ranka Stanković, and Aleksandra Marković. 2023. Multiword expressions – comparative analysis based on aligned corpora. In *Book of Abstracts of the UniDive 1st general meeting, 16-17 March 2023, Paris-Saclay University, France*.

Cvetana Krstev and Duško Vitas. 2017. Multi-Word Expressions in Serbian – Properties, Typology and Classification for Natural Language Processing. In *The International Jubilee Conference of the Institute for Bulgarian Language "Prof. Lyubomir Andreychin"*, pages 298–310, Sofia, Bulgaria.

Biljana Lazić and Mihailo Škorić. 2020. From DELA based dictionary to Leximirka lexical database. *Infotheca - Journal for Digital Humanities*, 19(2):81–98.

Aleksandra M. Marković. 2018. Značaj nekih koligacionih sklonosti glagolskih subleksema za semantiku i sintaksu (na građi iz opisnih rečnika srpskog jezika) = The Semantic and Syntactic Importance of Some Colligational Preferences of Verbal Sublexemes (Verbal Lexical Units). *Naučni sastanak slavista u Vukove dane*, 47(1):117–125.

Jan Odijk, Martin Kroon, Tijmen Baarda, Ben Bonfil, and Sheean Spoel. 2024. MWE-Finder: a Demonstration. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12027–12031.

Agata Savary. 2009. Multiflex: A multilingual finite-state tool for multi-word units. In *International Conference on Implementation and Application of Automata*, pages 237–240. Springer.

Ranka Stanković, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. Electronic dictionaries–from file system to lemon based lexical database. In *6th Workshop on Linked Data in Linguistic (LDL-2018), Towards Linguistic Data Science*.

Ranka Stanković, Rada Stijović, Mihailo Škorić, and Cvetana Krstev. 2025. The Dictionary of Contemporary Serbian Language (RSSJ): Advanced Automation and Other Challenges. In *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 Conference*, pages 59–75, Bled.

Duško Vitas, Ranka Stanković, and Cvetana Krstev. 2025. O familiji korpusa savremenog srpskog jezika SrpKor. In *Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig: Thematic Collection of Papers*.

A Appendix: example explanation

The lemma in Leximirka for *kvadratni metar* is represented as *kvadratni(kvadratni.A2:adms1g) metar(metar.N3:ms1q)*, while the compound inflectional class is *NC_AXNr*.

This class automatically generates 28 distinct inflected forms, covering 104 combinations of grammatical categories (e.g., case, number, gender, and definiteness), thereby enabling systematic modeling of morphosyntactic variation in corpus queries. For example:
metru kvadratnom:ms3q:ms7q
metrom kvadratnim:ms6q
metrima kvadratnima:mp3q:mp6q:mp7q
...
kvadratnom metru:ms3q:ms7q
kvadratnoga metra:ms2q
kvadratnog metra:ms2q

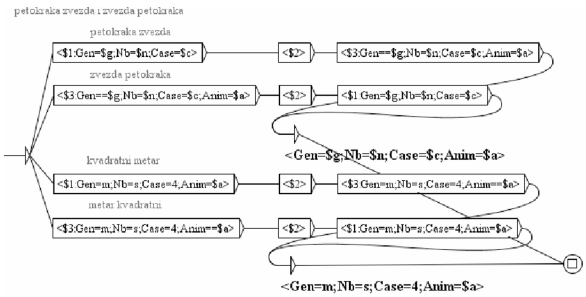

Figure 4: A simplified transducer for compounds of the type *metar kvadratni*

This example highlights the complexity of encoding all relevant information about a single MWE within its lemma representation. The most challenging aspect lies in consistently specifying agreement constraints. These issues are largely addressed by a specialized type of inflectional transducer proposed by Savary (2009) and implemented in the Multiflex system. The inflectional graph shown in Figure 4 demonstrates that the MWE used as a lemma is first tokenized. Each token is assigned to a variable: for example, in *petokraka zvezda*, the variables are 1 = *petokraka*, 2 = <*space*>, and 3 = *zvezda*, while in *kvadratni metar*, they correspond to 1 = *kvadratni*, 2 = <*space*>, and 3 = *metar*. When a simple pattern of the form <*i*> appears in the inflectional graph, it indicates that the corresponding token is copied unchanged across all inflected forms of the MWE; in these examples, the space (token 2) is consistently preserved in every inflectional variant.