Ivana Filipović Petrović and Kristina Kocijan

CREATING THE DATASET OF CROATIAN VERBAL IDIOMS

Automatic Identification in a Corpus and Lexicographic Implementation

Abstract This research proposes a step forward in the automatic identification and analysis of verbal idioms in Croatian. The use of the NooJ automated text processing tool, along with the MaCoCu corpus and the *Online Dictionary of Croatian Idioms* (ODCI), provides a robust framework for recognizing and categorizing these multi-word expressions (MWEs). The research comprises two parts: (a) creation of a dataset by utilizing the ODCI that allowed for a set of 898 verbal idioms to be compiled and annotated with linguistic features, including structure, morphological features, and variation patterns; (b) analysis of extracted data that provides insights into the lexicographical and linguistic significance of the idioms, such as variability, modification, and frequency of use. The study highlights the challenges posed by idiomatic variations and the verb's role as the most variable component in idioms. For instance, the idiom *soliti pamet komu* ('to give unsolicited advice') is often modified for expressiveness, such as in the phrase "having a big saltshaker to salt everyone's mind." The dataset aims for lexicographic integration into ODCI and supports the creation of electronic language resources. It also contributes to theoretical and cross-lingual research, with the CLARIN repository expected to enhance data reusability in NLP. The study's findings offer a deeper understanding of verbal idioms' dynamics and their computational processing.

Keywords verbal idioms; automatic identification of multi-word expressions; linguistic features of idioms; NooJ; lexicographic analysis

1. Introduction

The computational processing of figurative language, particularly multi-word expressions (MWEs), is a significant area of interest in nowdays computational linguistics and phraseology. The goal is to create language resources and effectively use linguistic data in various applications such as natural language processing (NLP), machine translation, and large-scale language models (LLMs).

This has brought the following methodological issues to the fore: how to accurately describe the syntactic and semantic features of MWEs, given their complex structure, variability, and figurative meanings, the set of features well known as a pain in the neck of NLP (Sag et al., 2002). In other words, how to formulate correct search criteria, computationally process the data, and finally, use the results for practical applications like lexicography.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

In our research, we focus on addressing these issues by introducing a method for the automatic identification of sentences containing verbal idioms in Croatian, with the aim of creating a lexicon of Croatian verbal idioms.[1]

Like many other low to middle-resourced languages, Croatian has lacked digital phraseological resources until the recent publication of the *Online Dictionary of the Croatian Idioms*[2] (ODCI) version 1.0 freely accessible since 2022. This dictionary is corpus-driven and has been processed by linguists and lexicographers, making it a reliable source for the initial description of verbal idioms and their usage.

Previous research in this field (Pasquer et al., 2020a; 2020b) has shown that verbal idioms pose a challenge as they occur in different variants, such as word order or lexico-syntactic forms. Corpus studies on Croatian verbal idioms (Filipović Petrović & Parizoska, 2017; Parizoska, 2022) have indicated that the verb is the most varied component, while the other components are relatively fixed, e.g., *gurati/pomesti/staviti pod tepih* ‘to hide a problem’ (lit. push/sweep/put sth under the carpet=N Acc sg). Some idioms also have open slots that can be filled with various adjectives, adding to ther complexity, e.g., *poprimiti* ADJ *razmjere* which translates to ‘reach ADJ proportions’ (epidemic/dramatic/unprecedented).

To tackle these challenges, we have utilized the NooJ tool (Silberztein, 2016) and automatic identification (Ramisch et al., 2020; Gantar & Krek, 2022) for the second methodological issue. NooJ is a rule-based automated text processing tool that supports the detection of MWEs in running texts. Previous studies (Machonis, 2012; Kocijan & Librenjak, 2016; Najar et al., 2017; Rhazi & Boulaalam, 2018) have validated the NooJ tool’s efficacy in identifying multi-word expressions, which is a testament to its utility in linguistic research. It combines a dictionary and grammar to identify elements that co-occur in specific positions. This approach allows for the identification of both known verbal idioms and potential new lexical and syntactic realizations thus facilitating us in extraction of MWEs from the large corpora for the purpose of lexicon building. For this specific task, either a list of MWEs from existing lexicons or manually annotated corpora are used as the basis, and the result is a set of corpus sentences containing MWEs in all typical syntactic and semantic realizations.[3]

Based on the described features that were implemented in NooJ, we automatically identified the sentences that include verbal idioms in Croatian, resulting in over 132,000 corresponding examples from the corpus. The final step involved linguistic evaluation of the results and performance of the automatic procedure.

[1] In this paper, we use the term *idiom*, which is equivalent to the term *phraseological unit* in the terminologies of some languages, and it refers to MWE in the narrowest sense. We rely on the division of MWEs from Kosem, Krek & Gantar (2020), according to which phraseological units are multiword lexical units with their own metaphoric meaning that speakers most often use when they want to express something in a noticeable and expressive way, differently from the neutral, for example *stealing days from God* instead of *wasting time.*

[2] https://lexonomy.elex.is/#/frazeoloskirjecnikhr

[3] The other method of automatic MWEs detection refers to the identification of MWEs in corpora regardless of existing MWEs. Both methods have already been researched, for example, for Slovenian (Škvorc, Gantar & Šikonja, 2022; Gantar & Krek, 2022).

In addition to describing the method that can be used for research in the field of computational phraseology, the intention of this study was also to produce a dataset of automatically identified Croatian verbal idioms, which, as a freely available linguistic resource, can be used for various purposes in the field of NLP, theoretical phraseology and cross-lingual research. For this paper, we utilized the obtained dataset for lexicographical purposes, specifically to supplement the existing ODCI with data from a more contemporary corpus.

The paper is structured as follows. The introductory section is followed by Section 2 which outlines the methodology for identifying verbal idioms within the corpus, including the initial list creation, formal description of linguistic features, design of the dictionary and grammar in NooJ, details of the corpus, and the extraction method employed. Section 3 discusses the evaluation of the obtained data. In Section 4, we delve into the lexicographic analysis and implementation of the dataset, particularly its use in enhancing existing phraseological dictionaries. Finally, in Section 5, we conclude and discuss implications for future research.

2. Methodology of Verbal Idioms Identification in the Corpus

The method used for the automatic identification of verbal idioms in the corpus comprises four steps. First, an initial list of verbal idioms is compiled from freely available phraseological resource for the Croatian language (Filipović Petrović & Parizoska, 2022) which serves as a starting point for the description and classification of morpho-syntactic models of verbal idioms. This step includes the preparation of the data for the second step – the creation of the dictionary and grammar in the NooJ tool. The default settings in NooJ are then used to search the corpus and identify phraseological expressions. The final step involves manually evaluating the results to create a dataset of verbal idioms.

The chosen method has several advantages. Firstly, the ODCI is freely available which is an important prerequisite for using the data. Also, it is regularly updated, so the results of this research can be practically applied to this dictionary. Additionally, previous studies have demonstrated the effectiveness of the NooJ tool for identifying multi-word expressions in Croatian text by combining information provided at the level of a dictionary with algorithms developed over syntactic grammar (Kocijan & Librenjak, 2016; 2018). Moreover, a new resource, the CLASSLA web corpus has become available in 2023, providing a large and diverse collection of contemporary Croatian texts suitable for phraseological research.

2.1 The Initial List

To create the initial list of Croatian verbal idioms, we used data from ODCI. It is accessible through the Lexonomy Elexis platform and based on hrWaC, the most extensive Croatian corpus at the time of its initial release (Ljubešić & Klubička, 2016). This dictionary was created using a corpus-driven method, where lexicographers collected idioms by extracting the thousand most frequent verbs in the hrWaC corpus using the Sketch Engine tool. Then, using collocation, filter, and word sketch functions, they extracted conventionalized multi-word expressions with figurative meanings (Filipović Petrović & Parizoska, 2022). In the extraction of verbal idioms, we adhered to the morpho-syntactic and semantic criteria (Fink-Arsovski, 2002). Namely, if the meaning of the idiom is expressed by a verb or a group of verbs, it is a verbal idiom (e.g., 'to manage, control, operate' as in *vući konce* 'pull the strings'). Also, when incorporated into discourse, verbal idioms are in the function of a predicate. In this way, we utilized a set of 898 verbal idioms for the initial list.

2.2 Formal Description of Linguistic Features

To create a NooJ dictionary and grammar, it was necessary to formulate the linguistic features of idioms. The list of 898 idioms was analyzed and relevant linguistic characteristics were attached to each idiom. Based on these characteristics, verbal idioms were categorized according to their grammatical structure into several basic models and submodels, as exemplified in Table 1[4].

Table 1: Excerpt of the classification of verbal idioms by grammatical structure into models and submodels

Model	**Extended submodel**	**Idiom**
V + N		*raditi frku (*'make a fuss about sth') *puniti kesu (*'line your pockets')
	+ PREP with N	*nositi gaće na* štapu ('on the breadline') *tjerati mak na konac* ('pigheaded')
V + PREP with N		*ispasti iz kolotečine* ('to get off track') *kipjeti od bijesa* ('to boil with rage')
	+ REC	*opaliti po džepu* koga ('hit smb's pocketbook') *nasmijati se u brk* komu ('laugh in smb's face')
V + *KAO* ('AS') + N		*spavati kao klada* ('sleep like a top') *piti kao smuk* ('drink like a fish')
	+ PREP with N	*nicati kao gljive poslije kiše* ('to mushroom') *snalaziti se kao riba u vodi* ('take to sth like a duck to water')

Each component of the idiom is assigned additional grammatical description, as shown in Table 2.[5]

Table 2: Example of models with additional grammatical description

Model	**Idiom**
TRANS V, all tense, pers. + N Acc pl	*vući konce* 'pull the strings'
NEG V form, all tense, pers. + N Acc sg + PREP with N Instr sg	*ne vidjeti prst pred nosom* 'one can't see a thing'

[4] In Table 1, the following symbols are used: V – verb; N – noun; PREP – preposition; REC – rection (a grammatical relationship where a verb or preposition determines the case of a noun or noun phrase).

[5] In Table 2, the following symbols are used: TRANS V – transitive verb; NEG V form – negative verb form; all tense, pers. – all tenses and persons; N – noun; Acc – accusative case; Instr – instrumental case; pl – plural; sg – singular; PREP – preposition.

Some idioms are completely fixed, with the verb appearing in a single person and tense form, as in *nije o glavu* (‘it’s not a hanging matter’) and *nigdje ne gori* (‘what’s the rush?’). In such cases, the components of idioms are entered into the NooJ dictionary as uninflected words. The NooJ dictionary also contains all elements of an idiom, while the grammar defines the order of idiom components. This is because the verb does not necessarily need to be the first component in the expression. For example, in the idiom *od govna napraviti pitu* (‘make a silk purse out of a sow’s ear’), the verb is positioned in the middle.

2.3 Dictionary and Grammar Design in the NooJ

Two main engines in NooJ are the dictionary and grammar. The dictionary determines the elements that occur together, and grammar defines the position of each element. The grammar complements the dictionary and the two work in unison.

A special feature, denoted as +FXC (Frozen Expression Components), is employed to aid in distinguishing between individual word occurrences and their idiomatic usages within multiword expressions (Kocijan & Librenjak, 2016). For instance, the verb *objesiti* (meaning ‘*to hang*’) is a transitive verb and when used as in sentence (1), we aim to treat it as a cohesive unit. To accomplish this, we will introduce an entry (2) in the NooJ dictionary. However, in sentence (3), the focus is not on analyzing *objesiti* as a standalone verb but rather on its idiomatic usage. Consequently, we will supplement the dictionary with an additional entry for this verb, augmented with an explanation of its idiomatic structure (4).

1) *Ana je objesila rublje.* ‘Ana hung the laundry outside to dry.’

2) objesiti,V+FLX=UGASITI

3) *Ana je objesila nos.* ‘Ana became dejected’ (lit. hung her nose)

4) objesiti,V+FXC+FLX=UGASITI+Model=V_N+N=<nos,N+Case=Acc>

When the noun *nos* (meaning ‘nose’) in the accusative case and the verb *objesiti* (‘to hang’) appear together in a sentence, NooJ prioritizes the annotation from the second dictionary entry (4).

The rule variation is specified through the **+Model** attribute assigned to each dictionary entry. This attribute clarifies potential ambiguity when the same verb is used in various idioms, such as *pobrati batine* (‘to receive a beating’) compared to *pobrati vrhnje* (‘reap the benefits’). In the same vein, the verb *ići* (‘to go’) forms part of idiomatic expressions such as *ići* čijim *stopama* (‘to follow in someone’s footsteps’) shown in example (5), and *ići preko leševa* (literally, ‘to walk over corpses’) which translates to ‘stop at nothing’, illustrated in example (6).

5) ići,V+FXC+FLX=IĆI+**Model=V_XN**+N=<stopa,N+I+p>

6) ići,V+FXC+FLX=IĆI+**Model=V_PP**+S=preko+NP=<leš,N+G+p>

In the dictionary, we specify which elements can coexist, and in grammar design, we determine the placement of each element and their broader context when necessary. For example, the algorithm depicted in Figure 1 is designed for two patterns:

1. Verb (**V**) + Noun (**N**) [upper branch] and
2. Verb (**V**) + intermediary element (**X**) + Noun (**N**) [lower branch].

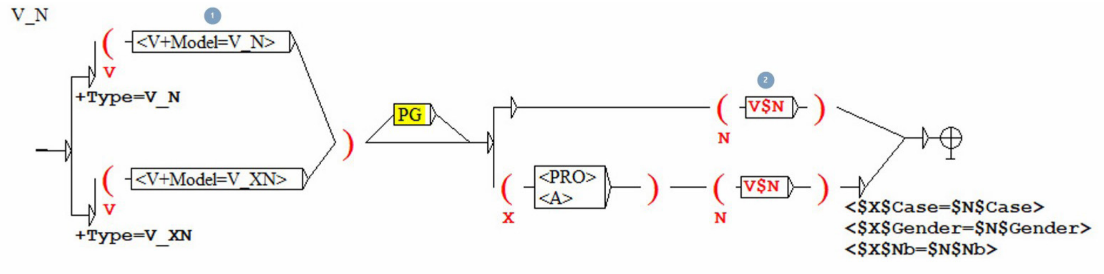


Fig. 1: Algorithm that recognizes models V_N and V_XN

The first pattern aligns with example (3) and connects to dictionary entry (4). The algorithm moves from the leftmost side to the rightmost node. It starts by pinpointing the verb defined in the dictionary with the attribute Model=V_N (Figure 1, marker 1). This verb is then stored as variable $V. The next node (Figure 1, marker 2) uses this variable to locate the noun that pairs with that specific verb, distinguishing it from other verbs in the dictionary. Thus, if the verb *objesiti* ('to hang') is followed by the noun *nos* ('nose'), the algorithm recognizes both words as an idiomatic pair, rather than as separate entities.

Furthermore, Croatian grammar's nuances allow for an auxiliary verb between the main verb and noun in idiomatic expressions, especially when forming complex verb tenses (as in example 3). An optional node [PG] is included to account for this.

The same algorithm applies to the pattern Verb (V) + intermediary element (X) + Noun (N), as seen in *ići* čijim *stopama* ('follow in someone's footsteps') and *poprimiti epidemijske razmjere* ('reach epidemic proportions'). The grammar's lower branch (Figure 1) describes idiomatic structures with the attribute Model=V_XN, permitting a pronoun or adjective before the noun, stored as variable $X. In examples like *poprimiti epidemijske razmjere* ('reach epidemic proportions'), the noun must match the preceding variable $X in case, gender, and number as defined in the last node of the lower branch.

The ODCI corpus data underscored the prevalence of inversion in Croatian idioms, necessitating its inclusion in our description. For instance, *od govna napraviti pitu* ('turn sth worthless into sth of value') commonly appears as PP V N in the hrWaC corpus, with variations like *od govna pitu napraviti* (PP N V), *napraviti od govna pitu* (V PP N), and *napraviti pitu od govna* (V N PP). Despite multiple possible versions, the dictionary contains only one entry for all the listed variations of the idiom (7).

7) napraviti,V+FXC+FLX=CIJEPITI+Model=V_NP_
PP+N1=<pita,N+Acc>+S=od+NP=govna.

The entry (7) consists of the verb *napraviti* ('to make'), the accusative noun *pitu* ('a pie'), and the prepositional phrase *od govna* ('from shit'), which can appear in positions 1, 2, and 3 (Figure 2). Although all components are listed in the dictionary, their possible arrangements are defined through the grammar (Figure 2).

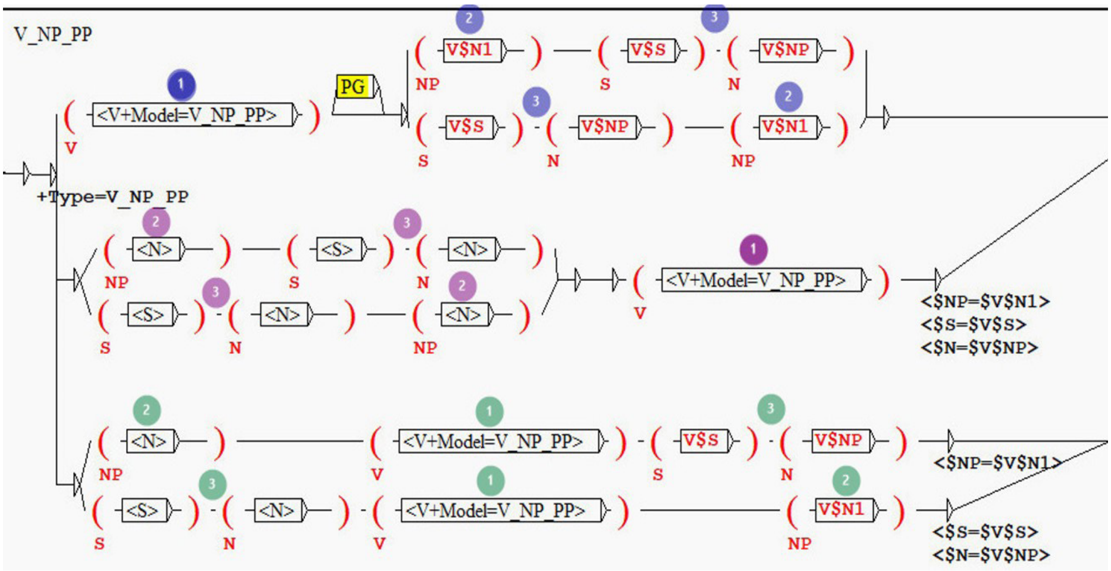


Fig. 2: Excerpt of an Augmented algorithm for V + N + Prepositional Phrase idiomatic structure and its inversions

2.4 Corpus

The CLASSLA web corpus for Croatian stands as one of the seven newly established corpora for South Slavic languages at the CLASSLA Knowledge Centre of the Slovenian CLARIN consortium. Originating from the MaCoCu project (Bañón et al., 2022), web crawling techniques were employed to amass monolingual and parallel datasets, as described in Ljubešić & Kuzman (2024). The corpus's resency, volume, and the text variety render it exceptionally apt for phraseological studies. The final CLASSLA-web corpus for Croatian is hosted on the CLARIN.SI repository, labelled as the Croatian CLASSLA-web.hr corpus (Ljubešić, Rupnik & Kuzman, 2024), encompassing 2,575 million tokens across 5,422 thousand documents.

2.5 Extraction Method

Based on our input described in sections 2.2. and 2.3., NooJ's data processing identified over 132,000 examples that correspond to corpus sentences. The unpredictability of the file sizes necessitated sophisticated technical support. Specifically, the data extraction, which took more than 49,000 working hours across four servers, occasionally exceeded the resources available to us.

Furthermore, the data obtained indicated the need for additional manual cleaning and removal. This was due to the method of corpus collection and the nature of the source data, which were web pages. Therefore, the results included various types of data. These included machine-generated text such as postdates and coded repetitions, as well as constructions like "learn more", "read more", "sent from my…". The results also contained sentences truncated into multiple lines and significant text repetition characteristic for forums, especially when authors referred to previous writings. Considering all these factors, along with the time constraints of our conference paper submission deadline, we decided to present the results for the hundred most frequent verbs, using approximately 80% of the corpus data. We will continue with data extraction and analysis until we create a final dataset to be a stand-alone resource in open access.

3. Data Evaluation

When 70% of the files were processed, we compiled a list of the 100 most frequently occurring verbs in the results. This resulted in a list of 50,987 lines that we manually evaluated linguistically. Figure 3 shows the top 10 most frequent models, with the model „V S N" at the top, occurring 12,920 times.

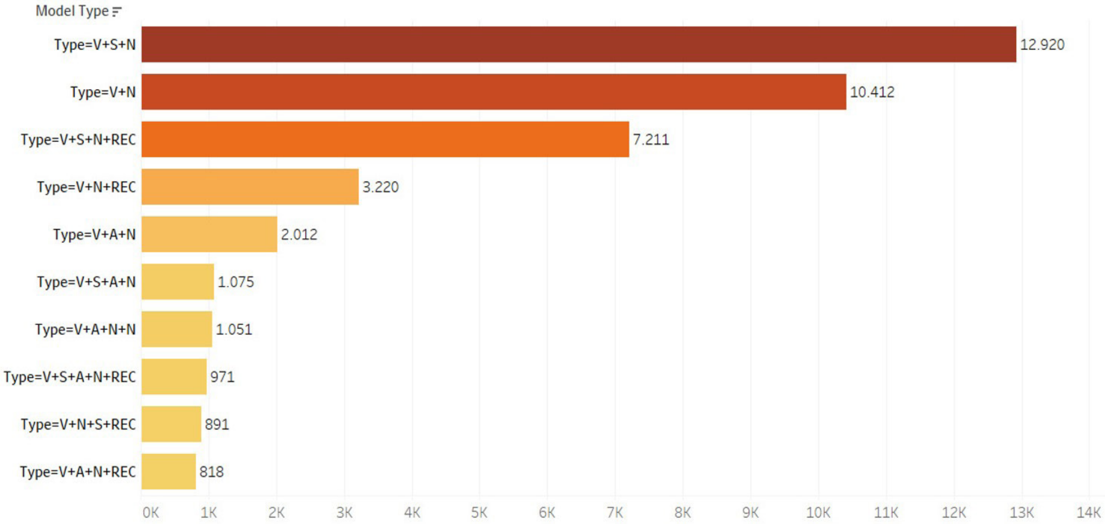


Fig. 3: Distribution of idioms within the top 10 model types

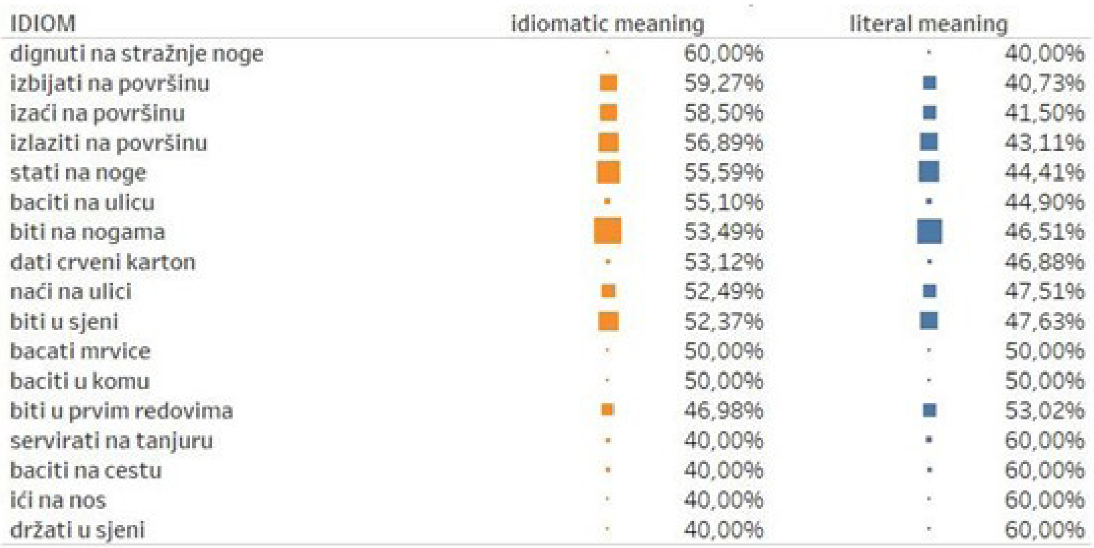
Among these verbs, *biti* ('to be'), *imati* ('to have'), *poći* (in this specific context, meaning 'to get'), *postaviti* ('to set'), and *ići* ('to go') yielded the most idiomatic examples overall. Verb *biti* with a total of 38 idioms, has the greatest variety of idiomatic usage, ranging from only one occurrence (*biti kao bubregu u loju* lit. to be like a kidney in fat 'to be in clover') up to 655 occurrences (*biti na izmaku* lit. to be nearing the end 'to be running out of steam'). On the other hand, *poći* is detected only in one idiom, *poći za rukom* ('to manage to do something'), but it occurs 2,466 times, making it the most frequent idiom in use within our corpus so far. It is followed by 1,816 occurences of *postaviti temelje* ('to lay foundations'), 1,222 occurrences of *pasti u zaborav* ('to be forgotten'), 1,051 occurrences of *imati pune ruke posla* ('to have hands full of work' meaning 'to be very busy'), and 1,040 occurrences of *uzeti maha* ('to gain ground'). The majority of the manual evaluation involved removing instances that contained the desired components but formed examples of literal usage in the language. For example, we obtained 1073 instances of the idiom *isplivati na površinu* ('to come to light, to surface'). Out of these, in 703 instances, it appeared as an idiom, meaning 'to appear, suddenly become visible, noticeable' as in example (8), while in 370 instances, it had the literal meaning 'to emerge from a liquid' (as in example 9).

8) *Nije trebalo dugo da* ***na površinu isplivaju*** *dobre i loše strane takvog načina rada.* 'It didn't take long for the good and bad sides of such a working method to come to light.'

9) *Njoke pažljivo stavite u posudu i kuhajte par minuta dok ne* ***isplivaju na površinu***. 'Carefully place the gnocchi in a pot and cook for a few minutes until they float to the surface.'

On the other hand, the construction *postaviti temelje* ('to lay foundations') appeared in 8 literal contexts out of a total of 1824 appearances. Table 3 shows which idioms are used both with their idiomatic and literal meaning ranging from 40% up to 60% of times. The size of the square before the percentage corresponds to the number of occurrences in the corpus. For example, *biti na nogama* (lit. to be on one's feet) 'to be up and about' and *dati crveni karton* (lit. to give a red card) 'to eject' are both used in 53% of times as an idiom. However, the total number of occurrences of *biti na nogama* in the corpus is 873 which is much greater compared to the number of occurences of *dati crveni karton* which appears only 32 times.

Table 3: Excerpt of multiword expressions that appear in their idiomatic meaning between 40% and 60%



IDIOM	idiomatic meaning		literal meaning	
dignuti na stražnje noge	·	60,00%	·	40,00%
izbijati na površinu	■	59,27%	■	40,73%
izaći na površinu	■	58,50%	■	41,50%
izlaziti na površinu	■	56,89%	■	43,11%
stati na noge	■	55,59%	■	44,41%
baciti na ulicu	▪	55,10%	▪	44,90%
biti na nogama	■	53,49%	■	46,51%
dati crveni karton	·	53,12%	·	46,88%
naći na ulici	▪	52,49%	▪	47,51%
biti u sjeni	■	52,37%	■	47,63%
bacati mrvice	·	50,00%	·	50,00%
baciti u komu	·	50,00%	·	50,00%
biti u prvim redovima	▪	46,98%	▪	53,02%
servirati na tanjuru	·	40,00%	·	60,00%
baciti na cestu	·	40,00%	·	60,00%
ići na nos	·	40,00%	·	60,00%
držati u sjeni	·	40,00%	·	60,00%

XXI EURALEX

4. Lexicographic Analysis and Implementation

In this section, we analyze extracted data from a lexicographic perspective. To describe a language in a dictionary, a lexicographer needs to gather linguistic data, specifically data on how the language is used. Idioms, which are multi-word expressions with figurative meanings, pose a challenge not only for NLP but also for lexicographic processing. Our research has shown that automated procedures for identifying idioms in a corpus can help lexicographers by providing comprehensive lists of sentences with idiomatic usage and statistical data that can be extracted from them. The data revealed insights about verbal idioms in terms of linguistically significant aspects such as variability, modification, and frequency of use. These insights have a direct impact on lexicographically substantial categories such as macrostructure, entry structure, usage notes and examples of use.

4.1 Macrostructure and Entry Structure

As presented in Section 3, we gathered data on the frequency of verbal idioms in Croatian, identifying the most common ones and the most productive ones. Three verbs seem to dominate in the productivity domain, namely verbs *biti* ('to be') with 38 idioms, *imati* ('to have') with 23 idioms and *ići* ('to go') with 19 idioms (Figure 4).

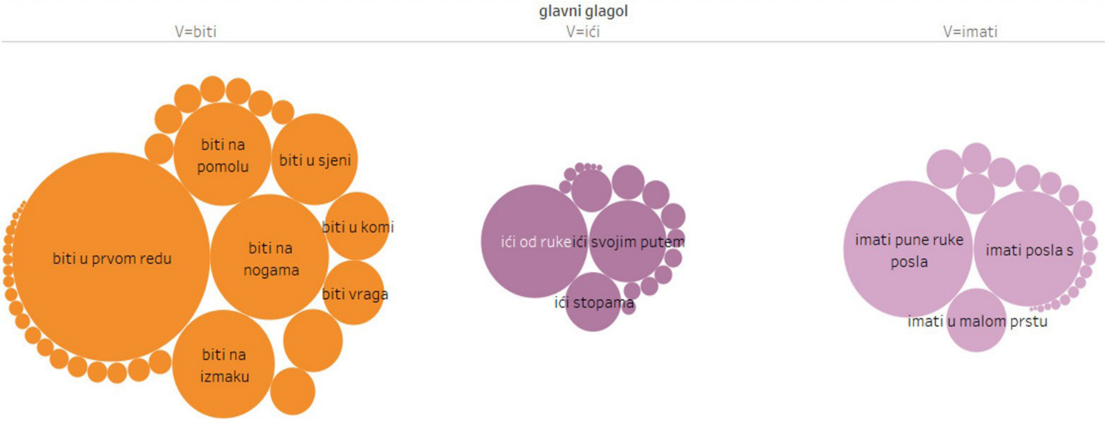


Fig. 4: Distribution of the top 3 most productive verbs: *biti*, *imati*, *ići*

Further analysis revealed that for idioms with a variable verbal component, the frequency of occurrence can serve as a guide for selecting the canonical form and determining the lexical variants to be included in the entry. For instance, construction *na svojoj koži* ('on one's skin'), mostly detected with the idiom *osjetiti na svojoj koži* ('to feel on one's skin') appeared with the verb *osjetiti* ('to feel') in 472 occurrences in the corpus, and with the verb *iskusiti* ('to experience') in 34 occurrences. Component that occurs with the largest variety of verbs is *na površinu* (literally meaning 'on the surface') which in most instances appears with the verb *izaći* ('to come out') (703 occurrences) and in somewhat smaller number of occurrences with verbs *izbijati, izbiti, izlaziti, izvući* and *izbaciti* as shown in Figure 5.

Based on the frequency criterion, which is used for selecting the canonical form of the idiom in ODCI, the form *izaći na površinu* should occupy the position of the headword. Depending on the lexicographic decision, two other variants that are most common after *izaći* can be included as variants in the entry.

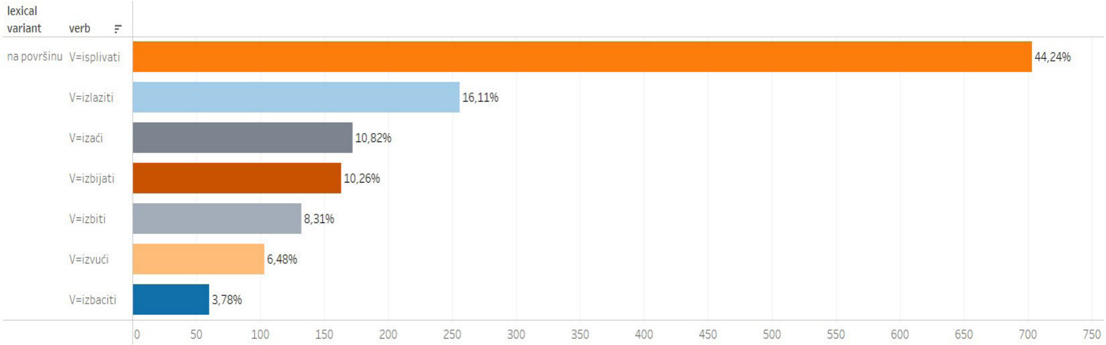


Fig. 5: Diversity of verbs occurring with component ‘na površinu’

4.2 Usage Notes

4.2.1 Negation

The data also showed a pattern of frequent use of certain verbal idioms in specific contexts. For example, the idiom *imati i ovce i novce* ‘to have it all’ (lit. ‘to have both sheep and money’) often appears with a negation: *you cannot have both sheep and money.* This type of data, known as usage notes, has undergone significant development in lexicography thanks to corpus data upon which dictionaries are based, as large corpora provide evidence about contexts in which idioms are typically used. Usage notes particularly come into focus in phraseological dictionaries, taking into account the importance of context for understanding the meaning of idioms. Additionally, they are highly valuable for language learners. A significant number of verbal idioms appear with negation in corpus, some of which are more commonly used with negation than in the affirmative form, such as *poći za rukom* ‘to manage to do something’, with only 3,75% appearances in the affirmative form. Usage notes may be grammatical, such as the data on negation, and sociolinguistic (see Rundell, 2023), such as data on intentional *ad hoc* modifications of idioms in the speech act, which we discuss in the following paragraph.

4.2.2 Idiom Modifications

A comprehensive list of sentences has also revealed certain patterns in the use of variants and modifications of idioms. This is yet another phenomenon of language in use that can only be documented in a corpus-based dictionary (cf. Langlotz, 2006), and this research has shown that automatic identification is a method that significantly enhances the detection of intentional alterations of idioms in language. During the linguistic evaluation of the results, we observed several modified uses of the idiom *soliti pamet* ‘to advise someone in a condescending way’ (lit. to salt someone’s mind) in sentences from the corpus. Speakers deliberately change idiom for expressiveness by playing with the literal meaning of the idiom components, as in the example (10), or expand it with new components semantically close to the literal use of the idiom's components, as in the example (11).

10) *Gradonačelnik je bio u pravu kad je kazao kako su to* ***ljudi koji imaju jednu veliku soljenku i vole svima soliti pamet.*** 'The mayor was right when he said that these are people who have a big saltshaker and love to salt everyone's mind.'

11) *Gdje bi nam bio kraj da nemamo vas da nam* ***solite pamet i biberite rane****?* 'Where would we be without you to salt our mind and pepper our wounds?'

5. Conclusion

In this study, we tackled the complex challenges related to the computational processing of figurative language, particularly MWEs, by focusing on Croatian verbal idioms. Our primary goal was to develop a reliable method for the automatic identification of these idioms, thereby creating a valuable linguistic resource for various applications in NLP, machine translation, and large language models (LLMs). By doing so, we hope to significantly enhance the sophistication and contextual awareness of language technologies and enable more accurate translations, nuanced sentiment analysis, and more natural interactions in chatbots and virtual assistants, among other applications.

Utilizing the NooJ tool, we successfully identified over 132,000 sentences containing Croatian verbal idioms from a large corpus. This significant achievement was made possible through a combination of dictionary-based and grammar-based approaches. These methods enabled us to capture both confirmed and potentially novel lexical and syntactic realizations of idioms.

The resulting dataset can enhance the existing Online Dictionary of Croatian Idioms (ODCI) by providing detailed information on the canonical forms of idioms and their most frequent variant forms. Additionally, it offers insights into specific usage patterns that are relevant to dictionary users and highlights intentional modifications made by speakers during communication. Finally, the rules and methodologies established for Croatian verbal idioms in this study have the potential to be adapted for use in other Slavic languages, thereby broadening the scope and applicability of our findings.

References

Bañón, M., Esplà-Gomis, M., Forcada, M. L., García-Romero, C., Kuzman, T., Ljubešić, N., van Noord, R., Pla Sempere, L., Ramírez-Sánchez, G., Rupnik, P., Suchomel, V., Toral, A., van der Werff, T., & Zaragoza, J. (2022). *Croatian web corpus MaCoCu-hr 1.0*. Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1516

Filipović Petrović, I., & Parizoska, J. (2017). Leksikografska obrada frazema s promjenjivom glagolskom sastavnicom u hrvatskome. *Jezikoslovlje*, *18*(2), 245–278.

Filipović Petrović, I., & Parizoska, J. (2022). *Frazeološki rječnik hrvatskoga jezika.* Hrvatska akademija znanosti i umjetnosti. https://lexonomy.elex.is/#/frazeoloskirjecnikhr

Fink-Arsovski, Ž. (2002). *Poredbena frazeologija: pogled izvana i iznutra.* FF press.

Gantar, P., & Krek, S. (2022). Creating the lexicon of multi-word expressions for Slovene. Methodology and structure. In A. Klosa-Kückelhaus, S. Engelberg, Ch. Möhrs, & P. Storjohann (Eds.), *Dictionaries and Society. Proceedings of the XX EURALEX International Congress* (pp. 549–562). IDS-Verlag.

Kocijan, K., & Librenjak, S. (2016). Recognizing verb-based Croatian idiomatic MWUs. In T. Okrut, Y. Hetsevich, M. Silberztein, & H. Stanislavenkapp (Eds.), *Automatic Processing of Natural-Language Electronic Texts with NooJ* (pp. 96–106). Springer.

Kocijan, K., & Librenjak, S. (2018). The quest for Croatian idioms as multiword units. In R. Mitkov, J. Monti, G. Corpas Pastor, & V. Seretan (Eds.), *Multiword units in machine translation and translation technology* (pp. 202–221). John Benjamins.

Kosem, I., Krek, S., & Gantar, P. (2020). Defining collocation for Slovenian lexical resources. In I. Kosem, & P. Gantar (Eds.), *Kolokacije v leksikografiji: Collocations in lexicography: existing solutions and future challenges,* Letn. 8, št. 2 (pp. 1–27). Znanstvena založba Filozofske fakultete, Slovenščina 2.0. https://revije.ff.uni-lj.si/slovenscina2/article/view/9338/9069, DOI: 10.4312/slo2.0. 2020.2.1-27

Langlotz, A. (2006). *Idiomatic creativity: a cognitive-linguistic model of idiom-representation and idiom-variation in English.* John Benjamins.

Ljubešić, N., & Klubička, F. (2016). *Croatian web corpus hrWaC 2.1.* Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. http://hdl.handle.net/11356/1064

Ljubešić, N., & Kuzman, T. (2024). *CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation.* https://arxiv.org/abs/2403.12721

Ljubešić, N., Rupnik, P., & Kuzman, T. (2024). *Croatian web corpus CLASSLA-web.hr 1.0.* Slovenian language resource repository CLARIN.SI.

Machonis, P. (2012). Sorting NooJ out to take Multiword Expressions into account. In K. Vučković, B. Bekavac, & M. Silberztein (Eds.), *Formalising Natural Languages with NooJ: Selected Papers from the NooJ 2011 International Conference* (pp. 152–165). Cambridge Scholars Publishing.

Najar, D., Mesfar, S., & Ben Ghezela, H. (2017). Inflectional and Morphological Variation of Arabic Multi-Word Expressions. In T. Okrut, Y. Hetsevich, M. Silberztein, & H. Stanislavenkapp (Eds.), *Automatic Processing of Natural Language Electronic Texts with NooJ: Selected Papers from the International Conference NooJ 2016* (pp. 37–47). Springer CCIS Series #667.

Parizoska, J. (2022). *Frazeologija i kognitivna lingvistika.* Srednja Europa.

Pasquer, C., Savary, A., Antoine, J.-Y., Ramisch, C., Labroche, N., & Giacometti, A. (2020a). *To Be or Not To Be a Verbal Multiword Expression: A Quest for Discriminating Features.* Retrieved May 10, 2024, from arXiv:2007.11381

Pasquer, C., Savary, A., Ramisch, C., & Antoine, J.-Y. (2020b). Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 3333–3345). International Committee on Computational Linguistics.

Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., Güngör, T., Jiang, M., Lichte, T., Liebeskind, Ch., Monti, J., Ramisch, R., Stymne, S., Walsh, A., & Xu, H. (2020). Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In S. Markantonatou et al. (Eds.), *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons* (pp. 107–118). Association for Computational Linguistics.

Rhazi, A., & Boulaalam, A. (2018). Corpus-based Extraction and Translation of Arabic Multi-Words Expressions (MWEs). In S. Mbarki, M. Mourchid, & M. Silberztein (Eds.), *Formalising Natural Languages with NooJ 2017 and its Natural Language Processing Applications* (pp. 143–155). Springer CCIS Series #811.

Rundell, M. (2023). Automating the creation of dictionaries: are we nearly there? In *Lexicography, Artificial Intelligence, And Dictionary Users. Proceedings of the 16th International Conference of the Asian Association for Lexicography: Lexicography* (Asialex 2023 Proceedings) (pp. 1–9). Yonsei University.

Sag, I., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword Expressions: a Pain in the Neck for NLP. In A. Gelbukh (Ed.), *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics* (CICLing-2002) (pp. 1–15). Springer-Verlag.

Silberztein, M. (2016). *Formalizing Natural Languages: The NooJ Approach.* Wiley-ISTE.

Škvorc, T. Polona, G., & Robnik-Šikonja, M. (2022). MICE: mining idioms with contextual embeddings. *Knowledge-based systems*, *235*, 1–11. https://doi.org/10.1016/j.knosys.2021.107606

Contact information

Ivana Filipović Petrović
Croatian Academy of Sciences
Linguistic Research Institute
ifilipovic@hazu.hr

Kristina Kocijan
University of Zagreb
Faculty of Humanities and Social Sciences
Department of Information and Communication Sciences
krkocijan@ffzg.unizg.hr