Croatian Idioms Integration: Enhancing the LIdioms Multilingual Linked Idioms Dataset

Ivana Filipović Petrović, Miguel López Otal, Slobodan Beliga
Croatian Academy of Sciences and Arts, Zagreb, Croatia
University of Zaragoza, Aragon Institute for Engineering Research, Zaragoza, Spain
University of Rijeka, Faculty of Informatics and Digital Technologies, Rijeka, Croatia
ifilipovic@hazu.hr, mlopezotal@unizar.es, sbeliga@inf.uniri.hr

Abstract

Idioms, also referred to as phraseological units in some language terminologies, are a subset within the broader category of multi-word expressions. However, there is a lack of representation of idioms in Croatian, a low-resourced language, in the Linguistic Linked Open Data cloud (LLOD). To address this gap, we propose an extension of an existing RDF-based multilingual representation of idioms, referred to as the LIdioms dataset, which currently includes idioms from English, German, Italian, Portuguese, and Russian. This paper expands the existing resource by incorporating 1,042 Croatian idioms in an Ontolex Lemon format. In addition, to foster translation initiatives and facilitate intercultural exchange, these added Croatian idioms have also been linked to other idioms of the LIdioms dataset, with which they share similar meanings despite their differences in the expression aspect. This addition enriches the knowledge base of the LLOD community with a new language resource that includes Croatian idioms.

Keywords: Croatian, idioms, multilingual, LLOD, Ontolex Lemon

1. Introduction

The growing availability of linguistic datasets and resources has increased the potential for natural language processing (NLP) development. However, this has also increased the diversity of data formats and metadata. To address this, the Linguistic Linked Open Data (LLOD) movement made an effort to achieve interoperability between datasets and promote data reuse (Cimiano et al., 2020). In this sense, there is great potential for the development of knowledge bases containing data on two rather challenging linguistic phenomena: idioms and multilingualism.

Idioms, also referred to as phraseological units in some language terminologies, are a subset of a larger community of multi-word expressions (MWEs). They have a figurative meaning that can only sometimes be inferred from the meaning of their components. Sometimes, idioms are culturally specific and have meanings based on everyday cultural references. For example, idioms like 'when pigs fly' in English, 'nem que a vaca tussa' in Portuguese, 'quando gli asini voleranno' in Italian, and 'kad na vrbi rodi grožđe' in Croatian all appear as different language realizations at the level of expressions. However, they all convey the same meaning, indicating a situation in which the speaker wants to tell someone that there is no chance that something will happen. In other words, the chances of it happening are as likely as pigs flying (EN), grapes growing on willow trees (HR), cows coughing (PT), or donkeys flying (IT).

This is the reason why one of the primary concerns in phraseology is understanding the meaning of idioms for non-native speakers. In addition, the correct usage indicates language proficiency for non-native speakers (Miller, 2018). In this sense, the possibilities of natural language processing can significantly enhance conditions for linguistic research and language learning, but it is necessary to ensure appropriate resources.

Therefore, this paper aims to propose an extension of the multilingual linked idioms dataset, integrating Croatian idioms into an existing dataset. The contributions of this work include (1) the creation of a dataset of Croatian idioms in LLOD format, (2) linking with idioms from an existing multilingual dataset that includes idioms from five languages, and (3) the creation of bilingual definitions for Croatian idioms in both Croatian and English.

In the following section of the paper, we will discuss the related work and the motivation for researching Croatian idioms in LLOD. Section 3 will explain the methodology of a new language resource development. In the last part, we will conclude the paper with final remarks and provide guidelines for future work.

2. Background and Motivation

Croatian idioms have not been extensively covered in terms of linguistic resources, especially in the LLOD format. Škvorc et al. (2022) presents an approach called MICE that uses contextual embeddings to detect idioms. In addition, they created a new dataset of multi-word expressions with literal

LREC-COLING 2024, pages 4106–4112

and idiomatic meanings and trained a classifier based on two state-of-the-art contextual word embeddings: ELMo and BERT. They confirmed that deep neural networks using both embeddings are suitable for detecting idiomatic expressions. In the conducted experiments, they used the PARSEME Croatian dataset, which is part of the shared task on automatic identification of verbal multiword expressions described in Ramisch et al. (2018), and contains sentences with MWEs annotations. The Croatian part of PARSEME dataset contains only 131 Croatian idioms. However, PARSEME later underwent modifications and extensions (Savary et al., 2023).

Orešković et al. (2018) has created a network lexicon for the Croatian language based on a syntactic and semantic computational framework. The proposed framework facilitates the retrieval and filtering of information related to the searched word in a paradigmatic and syntagmatic sense. The created lexicon is stored in the Croatian Linguistic Linked Open Data (CroLLOD) cloud, and incorporates three main types of lexicons. These are the lexicon of subatomic lexical entries (morphs, syllables, or syllable-morphs), the lexicon of words (central lexicon), and the lexicon of MWEs. The latter comprises about 120,000 MWEs, distributed among several meaningful groups, such as collocations, phrases, and synonyms. Idioms, however, were not of primary interest in this work.

To the best of our knowledge, we have not been able to find any comprehensive and compact research primarily investigating Croatian idioms in LLOD. However, a multilingual linked idioms dataset called LIdioms (Moussallem et al., 2018) was developed in 2018 to support natural language processing applications by linking idioms from various languages. At the time, LIdioms included idioms from English, German, Italian, Portuguese, and Russian, intending to show the possibility of correct translations among idioms independent of their language, family, syntax, or culture.

To fill the gap of the lack of Croatian idioms in LLOD, we introduce an extension of the LIdioms knowledge base with a Croatian dataset. Croatian belongs to the South Slavic branch of the Slavic language family and is generally considered a low- to middle-resource language (Beliga and Martinčić-Ipšić, 2017), especially lacking representation in linguistic datasets as Linked Data (Orešković et al., 2018). The motivation for this work stemmed from the availability of Croatian idioms in open access as an XML (Extensible Markup Language) file and the existence of a data resource of idioms in other languages, represented in the linguistic linked data standard Ontolex Lemon (McCrae et al., 2017). Our goal is to adhere to the existing format proposed for LIdioms in Moussallem et al. (2018) and adapt it to the Croatian dataset. The implementation will use the properties for representing multilingual linked data and relationships between translations proposed by Gracia et al. (2014). The aim is to create an enriched resource that can enhance its reusability in language learning and translation and contribute to a broader understanding of universalities and differences between languages and cultures.

3. Methods and Results

3.1. The Initial Dataset

This section provides a more detailed description of the linguistic resources used for our research. Namely, Croatian idioms have recently become more accessible through the Online Dictionary of Croatian Idioms[1] (Filipović Petrović and Parizoska, 2022). This is a corpus-driven dictionary, version 1.0 of which was released in October 2022. It is published on Elexis Lexonomy, a platform for creating and publishing dictionaries, under the Creative Commons Attribution 4.0 International license. This machine-readable dictionary currently contains 513 headwords, comprising a total of 1,042 idioms and their variants. In the Croatian language, there are older phraseological dictionaries, but they can only be found in printed form. There is a related online resource, the Online Database of Croatian Idioms[2], but it is unsuitable for our purposes due to its lack of definitions.

The Online Dictionary of Croatian Idioms is a highly reliable source of lexical data, which has been meticulously curated by linguists and lexicographers based on real language usage. It is worth noting that our use of this resource allowed us to avoid a common challenge in NLP research, which is the categorization of different types of multiword expressions and their differentiation from non-figurative language (Moussallem et al., 2018; Pasquer et al., 2020; Gantar and Krek, 2022). This dictionary provides us with pre-evaluated idioms and carefully designed definitions, which have been processed by linguists and lexicographers at previous stages, making them a ready-made resource for this study.

However, similar to many other lexicographical sources, it presents its contents linearly, limiting its accessibility to human readers, primarily native Croatian speakers. Furthermore, while the data is open-source, it lacks the structuring necessary for benefiting NLP applications.

[1] https://lexonomy.elex.is/frazeoloskirjecnikhr
[2] http://frazemi.ihjj.hr/

3.2. Semantic Representation Model and RDF Generation

In the previous section, we discussed the linguistic source for this dataset, which was made available to us in the form of an exported XML document. Therefore, our initial step in further research was to convert this information into an Ontolex Lemon (McCrae et al., 2017) format. This conversion was essential to ensure its accessibility through Semantic Web services and to pave the way for linking this dataset with others containing idioms in different languages. To accomplish this task, as a first step, we created a monolingual RDF-based (Resource Description Framework) dataset for Croatian idioms, using information from the previously mentioned Lexonomy platform. We chose to model our data closely following the specifications outlined by Moussallem et al. (2018) for their LIdioms dataset. We made that choice not only because this work provided an already well-developed model that is compatible with the Ontolex specifications, and tailor-made for idiom representation, but also because by closely following this format, we can link Croatian data with confidence to those of the LIdioms dataset in a further step.

However, the Croatian data in the exported XML file, despite its machine-readable format, presented a challenge due to inconsistent tags and structuring. This made it difficult, initially, to devise a strategy for its automatic conversion to RDF. One major issue in this sense was with the amount and type of encoded linguistic information in our source file. On the one hand, the XML file consisted of a series of entries (marked with an `<entry>` tag) that corresponded to actual individual entries in a traditional dictionary. For instance, our XML-based dictionary contains an entry for the idiom 'naći se u škripcu', whose meaning is that of 'to be in a difficult position, an awkward situation'. An entry like this could in theory be easily mapped to a single Ontolex `LexicalEntry` instance, following the scheme proposed by Moussallem et al. (2018). This would make the conversion process straightforward.

This process, however, is hindered by the fact that each of the individual entries in the source XML file can also optionally list other idiom forms that have the same meaning but exhibit minor differences. In the case of the idiom mentioned above, this specific example presents the following alternative forms:

- 'biti u škripcu' – 'be in a corner'
- 'naći se u škripcu' – 'find yourself in a corner'
- 'izvući se iz škripca' – 'get out of a corner'

These forms, while different, are all variations of the same base idiom. The differences in this

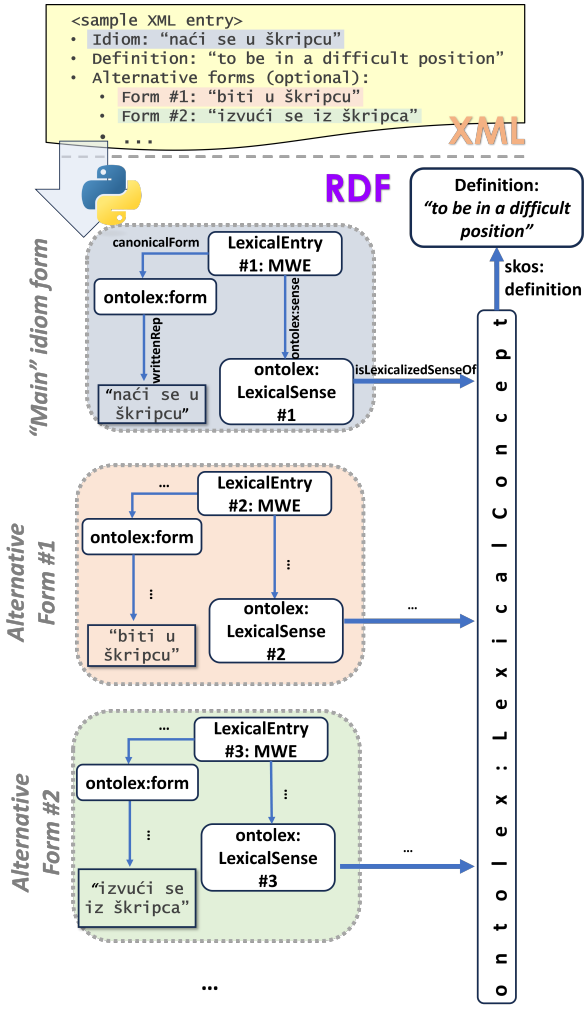

Figure 1: An example of the mapping of an idiom – with multiple surface forms – from our source XML file to a set of Ontolex Lemon format entries compliant with the standard devised for LIdioms.

specific case have their origin in a different conceptualization of the same event,[3] (Langlotz, 2006; Parizoska and Omazic, 2020) which is why they all share the same meaning.

Representing these alternative forms in an Ontolex format, while still complying with the template proposed by Moussallem et al. (2018), became a challenge, as the latter work did not foresee the inclusion of this type of variation in idiom forms. As interoperability with the LIdioms dataset was highly desired, however – as we will see in more detail in section 3.3 – a compromise had to be made where these related forms would need to be modeled as distinct units, thus being assigned their

[3]The differences in the forms of an idiom can be due to other factors, like, for instance, the expression aspect, such as the use of perfective and imperfective verbs (for example, 'doliti / dolijevati ulje na vatru' - 'making a bad situation even worse')

own standalone Ontolex `LexicalEntry` entities. This marked a difference with our source dictionary, which had been built from a lexicographical standpoint, where these forms were considered lexico-grammatical variants of the same idiom – and not different entries. To comply, on the one hand, with the schema presented by Moussallem et al. (2018), we decided to represent each form as a different instance of an Ontolex `LexicalEntry`, with their own associated `ontolex:LexicalSense` entry, the same format as used in LIdioms. On the other hand, however, in order to keep the information that these forms are related to each other – as realizations of the same base idiom – in the final dataset, we also linked each one of those `ontolex:LexicalSense` entries to a single, underlying `ontolex:LexicalConcept` entry. This `LexicalConcept` would include the definition – via a `skos:definition` label – shared by all the related idiom forms. This format closely matches the standard presented by LIdioms, hence ensuring interoperability between the two resources. In Figure 1 we can see an example of this adopted Ontolex representation, as suited for our purposes, for the idiom 'naći se u škripcu' and its different forms. This figure also shows how the initial XML representation of this idiom influenced its final Ontolex Lemon presentation.

Another related issue we had to solve in this regard was the fact that, other than additional forms, some XML entries could also contain alternative definitions listed under some of those forms. These definitions were given alongside the main definition for the idiom, and they were not meant to replace the main one, but rather add a specific explanation point about the form in question. An example of this would be the idiom 'doći pameti,' whose main meaning is 'to reason, to realize what is important', which presents an alternative form, 'dozvati pameti koga'. This form conveys the same meaning as the base idiom form, but it adds the specific remark of not only reasoning, but 'reasoning **with someone**'. In order to represent this additional information in a suitable format, we decided to add a link, originating from the affected form's `LexicalSense` entry, to its attached 'secondary' definition, connected via an `ontolex:usage` label. The aforementioned `LexicalSense` entry, at the same time, would still be connected to an `ontolex:LexicalConcept` entry containing the main definition which, as explained above, is a requirement to ensure interoperability with LIdioms.

The conversion from XML to RDF was done via a series of Python scripts, which contained a series of hand-crafted rules that were the result of close examining the source XML file. Our final, monolingual Croatians dataset consists on an RDF Turtle file comprising more than 17,000 triples.

3.3. Linking Idioms

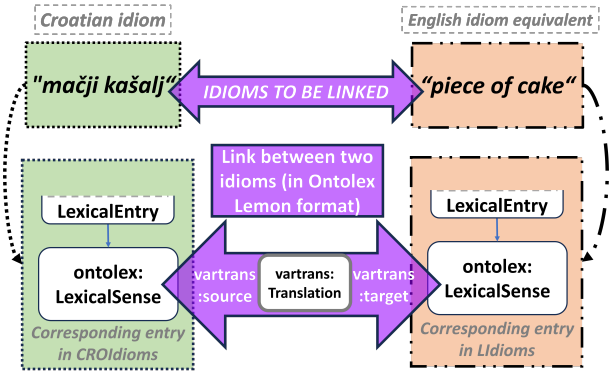

Figure 2: Linking example of an idiom from the Croatian idioms dataset to an equivalent idiom in English found in LIdioms, making use of the same linking strategy as the latter for pairing idioms – by relying on the use of the vartrans:Translation module.

To establish a connection between the Croatian dataset and LIdioms, which contains various expressions across multiple languages, we manually looked for Croatian idioms that closely resemble those in other languages. Having in mind that idioms rarely match on the level of expression, but relatively often on the level of meaning, we focused on identifying idioms in other languages whose meanings align closely with those in Croatian. An example of this would be the Croatian idiom 'mačji kašalj', which carries the same meaning as the English expression 'piece of cake' i.e., 'something that is considered easy to carry out'.

LIdioms achieves the pairing of related idioms from different languages by internally linking multilingual idioms based on a shared semantic concept, specifically using the Ontolex category `trcat:culturalEquivalent` from the `vartrans:translation` module. The link between two related idioms is done by leveraging the `ontolex:LexicalSense` entries of each of the two idioms, and pairing them together with two labels: `vartrans:source` and `vartrans:target`. Given that the Croatian dataset was modeled based on the approach presented by LIdioms, we chose to model the linkage between these datasets as an extension of LIdioms.

Initially, the linguist on the team carried out a manual search task to identify Croatian idioms with equivalents in different languages. To facilitate this process, we translated the existing idiom definitions – all written in Croatian – into English. The set of definitions was machine-translated on the first iteration and then each of them was manually reviewed and linguistically edited.

The idea of translating all the idioms' definitions

	EN	PT	IT	DE	RU	HR
Idioms	291	114	175	130	105	**1,042**
Transl.	192	79	73	60	82	**49**

Table 1: Number of idioms and pairs of idioms translations in the combined LIdioms and Croatian idioms datasets.

into English was also borrowed from Moussallem et al. (2018), which used this language as a common ground to help with the task of linking idioms with similar meanings that originated from different languages. In this way, in our work a manual search for the meanings of idioms was enabled to identify equivalent meanings in other languages. The search was conducted using keywords representing fundamental life concepts, such as love, fear, difficulty, disease, etc. Finally, numerous idioms were identified in the searched languages as well as in Croatian, varying in expression but aligning in meaning. More specifically, 49 Croatian idioms were located, comprising 28 English-Croatian pairs, 13 Italian-Croatian pairs, and 8 Portuguese-Croatian pairs. These translation pairs serve as an addition to the preexisting set of idioms translations already present in the LIdioms dataset, whose numbers (alongside that of our addition) are reflected in Table 1. The last column, in bold letters, illustrates Croatian idioms extension to the initial LIdioms dataset (added 1,042 Croatian idioms, and 49 direct links). Other columns contain data about the rest of the languages, where the number of links represents the total number of direct and indirect[4] links.

After collecting these multilingual idioms, we employed SPARQL queries to automatically retrieve `ontolex:LexicalSense` entries from LIdioms that could be linked to the `LexicalSense` entries in our dataset. This aligns with the multilingual idiom linking strategy outlined by Moussallem et al. (2018). Using Python scripts, we established connections between these entries, tagging them with `vartrans:translation`, an example of which can be seen in Figure 2 for the idiom 'mačji kašalj' and its English equivalent 'piece of cake'. This resulted in another RDF Turtle file with a total of more than 180 triples.

4. Discussion and Future Work

In conclusion, this study presents an extension to the multilingual linked idiom dataset by incorporating Croatian idioms into the existing LIdioms dataset. The process involved creating a dataset of Croatian idioms in LLOD format (CroIdioms), providing bilingual definitions in both Croatian and English, and linking them to idioms from a LIdioms five-language dataset, resulting in a larger multilingual six-language linked idioms dataset (LIdioms2). The outcome of this effort is publicly available in the CroIdioms repository[5].

A potential avenue for future research involves expanding the number of Croatian idioms linked to those of other languages. The limitations of this research include, on the one hand, the inability to incorporate German and Russian idioms at this moment due to the absence of proficient linguist experts in these languages. On the other hand, our initial dataset is a dictionary that is continuously being supplemented with new data manually, representing a work in progress where the number of available idioms is still growing.

Future enhancements to the CroIdioms dataset can occur in two main directions. Firstly, contextual embeddings like MICE, as proposed in Škvorc et al. (2022), could be utilized for mining procedures to identify idioms outside our current dataset. Secondly, leveraging NLP techniques, especially large language models, can aid in calculating semantic similarities between definitions of Croatian idioms and idioms from other languages. This approach can uncover new links between idioms that might not have been recognized in our study, assisting language experts in finding intricate links among a multitude of idioms across different languages. Finally, in the future, we plan to evaluate the quality of the expanded CroIdioms dataset to ensure its reliability in linguistic research.

5. Acknowledgements

This paper is based on work from the COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology). Additionally, it has been partially supported by the Spanish project PID2020-113903RBI00 (AEI/FEDER, UE) and by DGA/FEDER, as well as by the project uniri-mladidrustv-23-33 of the University of Rijeka and the project Online Dictionary of Croatian Idioms of the Croatian Academy of Sciences and Arts.

We are particularly grateful to Professor Jorge Gracia from the University of Zaragoza, Spain, as well as Manuel Fiorelli, Armando Stellato, Diego Moussallem, and Thierry Declerck, for their valuable advice and guidance during the 5th Summer Datathon on Linguistic Linked Open Data, organized by COST Action Nexus Linguarum.

[4] Indirect links are those that arise due to the property of transitive relations between already directly linked idioms.

[5] Available at: https://github.com/CroIdioms/CroIdioms

6. Bibliographical References

Slobodan Beliga and Sanda Martinčić-Ipšić. 2017. Network-enabled keyword extraction for under-resourced languages. In *Semantic Keyword-Based Search on Structured Data Sources*, pages 124–135, Cham. Springer International Publishing.

Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data: Representation, Generation and Applications*, 1st edition. Springer Publishing Company, Incorporated.

Ivana Filipović Petrović and Jelena Parizoska. 2022. Frazeološki rječnik hrvatskoga jezika. Hrvatska akademija znanosti i umjetnosti.

Polona Gantar and Simon Krek. 2022. Creating the Lexicon of Multi-Word Expressions for Slovene. Methodology and Structure. In *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*, pages 549–562, Mannheim. IDS-Verlag.

Jorge Gracia, Elena Montiel-Ponsoda, Daniel Vila-Suero, and Guadalupe Aguado-de Cea. 2014. Enabling language resources to expose translations as linked data on the web. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 409–413, Reykjavik, Iceland. European Language Resources Association (ELRA).

A. Langlotz. 2006. *Idiomatic Creativity: A cognitive-linguistic model of idiom-representation and idiom-variation in English*. Human Cognitive Processing. John Benjamins Publishing Company.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: Development and applications. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference, in Leiden, Netherlands*, pages 587–597. Lexical Computing CZ s.r.o.

Julia Miller. 2018. Research in the pipeline: Where lexicography and phraseology meet. *Lexicography*, 5:23–33.

Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. 2018. LIdioms: A multilingual linked idioms data set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Marko Orešković, Sandra Lovrenčić, and Mario Essert. 2018. Croatian Network Lexicon within the Syntactic and Semantic Framework and LLOD Cloud. *International Journal of Lexicography*, 32(2):207–227.

Jelena Parizoska and Marija Omazic. 2020. Sheme dinamike sile i promjenjivost glagolskih frazema [force-dynamic schemas and variability of verbal idioms]. *Jezikoslovlje*, 21:179–205.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020. Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–3345, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze, and Abigail Walsh. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop*

on Multiword Expressions (MWE 2017), pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Tadej Škvorc, Polona Gantar, and Marko Robnik-Šikonja. 2022. Mice: Mining idioms with contextual embeddings. *Knowledge-Based Systems*, 235:107606.