Multiword Expressions between the Corpus and the Lexicon: Universality, Idiosyncrasy and the Lexicon-Corpus Interface

Verginica Barbu Mititelu, Voula Giouli, Kilian Evang, Daniel Zeman, Petya Osenova, Carole Tiberius, Simon Krek, Stella Markantonatou, Ivelina Stoyanova, Ranka Stankovic, Christian Chiarcos

Multiword Expressions between the Corpus and the Lexicon: Universality, Idiosyncrasy and the Lexicon-Corpus Interface

Verginica Barbu Mititelu[1] , **Voula Giouli**[2], **Kilian Evang**[3], **Daniel Zeman**[4], **Petya Osenova**[5], **Carole Tiberius**[6], **Simon Krek**[7], **Stella Markantonatou**[8], **Ivelina Stoyanova**[9], **Ranka Stankovic**[10], **Christian Chiarcos**[11]
[1]Romanian Academy Research Institute for Artificial Intelligence, vergi@racai.ro
[2]Insitute for Language and Speech Processing, Athena Research Centre, voula@athenarc.gr
[3] Heinrich Heine University Düsseldorf, evang@hhu.de
[4] ÚFAL MFF, Charles University, zeman@ufal.mff.cuni.cz
[5]Institute of Information and Communication Technologies, BAS, petya@bultreebank.org
[6]Leiden University, c.p.a.tiberius@hum.leidenuniv.nl
[7]Jožef Stefan Institute, simon.krek@ijs.si
[8]Insitute for Language and Speech Processing, Athena Research Centre, marks@athenarc.gr
[9]Institute for Bulgarian Language, BAS, iva@dcl.bas.bg
[10]University of Belgrade, ranka@rgf.rs
[11]University of Augsburg, christian.chiarcos@uni-a.de

Abstract

We present ongoing work towards defining a lexicon-corpus interface to serve as a benchmark in the representation of multiword expressions (of various types – nominal, verbal, etc.) in dedicated lexica and the linking of these entries to their corpus occurrences. The final aim is the harnessing of such resources for the automatic identification of multiword expressions in a text. The involvement of several natural languages aims at the universality of a solution not centered on a particular language, and also accommodating idiosyncrasies. Challenges in the lexicographic description of multiword expressions are discussed, the current status of lexica dedicated to this linguistic phenomenon is outlined, as well as the solution we envisage for creating an ecosystem of interlinked lexica and corpora containing and, respectively, annotated with multiword expressions.

Keywords: multiword expression lexicon, corpus, proof-of-concept lexicon encoding

1. Introduction

In the last decade, the PARSEME COST Action (Savary et al., 2015) created the prerequisites for annotating corpora with multiword expressions (MWEs), mainly verbal ones. Consistent guidelines[1] and an infrastructure for ensuring annotation consistency were developed, while the interaction among the members of the community was made possible by the COST Action and extended even beyond its duration. A corpus was created for 26 languages (Savary et al., 2023), in which verbal MWEs (VMWEs) were annotated according to the established guidelines. Meanwhile, a new COST Action, UniDive[2], is gathering the community again, simultaneously increasing in size and allowing for the development of guidelines for annotating MWEs of other parts of speech, and eventually for further annotation of corpora with the new MWE types, as well as for increasing the number of languages represented in the corpus so far. At the same time, UniDive builds on Universal Dependencies (UD) (de Marneffe et al., 2021), which posits standardized guidelines for tokenization, lemmatization and morphosyntactic annotation in treebanks of languages.

Despite the abundance of large bodies of annotated corpora and large language models, systems still fail to adequately identify MWEs and thus the need for lexica that are specifically designed to handle MWEs within the context of Natural Language Processing (NLP) (Savary et al., 2019b). Within UniDive, Working Group 2[3] seeks to take this further and to schematize the steps needed towards creating an ecosystem in which annotated corpora and MWE lexica are linked together, intra- and interlingually and are used to facilitate MWE identification in a way that universality and idiosyncrasy are taken into account.

In this paper, we report on original (ongoing) work towards designing this lexicon-corpus interface. The paper is structured as follows: we first outline our goals and the challenges we need to face (Section 2); then, an overview of the current

[1] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=030_Categories_of_VMWEs
[2] https://unidive.lisn.upsaclay.fr/
[3] https://unidive.lisn.upsaclay.fr/doku.php?id=wg2:wg2

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@*LREC-COLING-2024*, pages 147–153. May 25, 2024.

MWE dedicated lexica and the results of a survey aimed at better accounting for universality are presented (Section 3 and Section 4 respectively). The initial steps towards designing the lexicon-corpus interface, in a standardized manner with all its advantages are presented in Section 5. We outline the minimal requirements for encoding MWEs in computational lexica, with an eye to their interlinking with annotated corpora, in Section 6. Our conclusion is presented in Section 7.

2. Towards a lexicon-corpus interface: goals and challenges

For many decades, MWE-aware lexica have contributed a much larger set of MWEs than (annotated) corpora can do, as MWEs are rather rare in texts (Savary et al., 2019a), and to model their linguistic properties, namely, non-compositionality, lexical fixedness, discontinuity, potential modifiers of components, word order variation, etc. However, the representation of MWEs in hand-crafted lexica is far from homogeneous and even incomplete. At the same time, annotated corpora have been used as major operational tools for language modelling and the backbones of data-driven NLP methods. Yet, they seem inadequate when unseen MWEs are at stake, as these unseen ones may well be characterised by lexical combinations or syntactic structures that did not occur in annotated corpora and are thus hard to be identified automatically. Therefore, linking corpora and lexica would be beneficial for the robust MWE identification (Savary et al., 2019b). As of now, corpora and lexica remain to a great extent disconnected, with a few exceptions (Odijk, 2013; Markantonatou et al., 2019; Autelli, 2020) in which examples are extracted from corpora and added to the lexicon to illustrate the use of the MWEs.

Our goal is to design a lexicon-corpus interface that leverages MWE identification cross-linguistically. Three are the major challenges: (a) the harmonisation of corpora and lexica by also accounting for universality and diversity, (b) the efficient encoding of MWEs of all grammatical categories cross-linguistically, and (c) the adoption of the appropriate mechanisms and tools for linking lexica and corpora. Our work has been organised along three axes:

i capturing universality via cross-language unification of lexical features,

ii designing a lexicon-corpus interface usable for several languages, and

iii proof-of-concept encoding of MWEs based on the outcomes of (i) and (ii).

3. MWEs in computational lexica: state-of-the-art

In order to overview the state-of-the-art in the development of computational lexica of MWEs, we collected information about resources in a structured and systematic way, focusing on those published since 2016, as those published before this year were included in the survey performed within the COST Action PARSEME (Losnegaard et al., 2016). We have retrieved information for 75 resources from the following sources: European Language Grid repository, using the keyword “expressions” in the category Lexical/Conceptual resources; the ACL Anthology, in which we also used a keyword search (*multiword, idiom, phraseology*, etc.); the Phraseology and Multiword Expressions book series published by Language Science Press, and Europhras conferences, which were manually examined.

The data was harmonised aiming at a uniform and comprehensive description of the identified resources. It was organized in the following sections: General information (general or dedicated lexicon, mono- or multi-lingual), Corpus (in cases where the resource is related to a corpus), Resource (size, owner, licensing), Lemma & Representations (whether the resource provides information about the “lemma” of the MWE and its morphosyntactic properties), Syntax (details about syntactic information about the MWEs), Semantics (whether the resource provides semantic information about the MWEs and of what type) and References (major publication(s) about the identified resource).

The general picture obtained so far shows that:

- 72% of the resources are aimed for NLP use.
- More than 40 languages and dialects are represented, mostly Indoeuropean ones.
- 70.7% of the resources are monolingual, 18.7% bilingual and 10.6% multilingual.
- Most datasets were acquired manually or semi-automatically (automatically collected and manually verified).
- Only 24% of the resources are linked to a corpus and 12% are linked to other resources. The resources are usually linked to small purpose-built corpora. Usage examples are sometimes collected from a large representative corpus (without linking to the corpus).
- With regards to the encoded information, 45% of the resources provide comprehensive description of MWEs (including morphological, syntactic and semantic information). Semantic information, in particular, is extremely diverse.

The survey on MWE lexica raises several significant questions related to handling universality and diversity. First, most resources assume that a MWE entry is the coupling of a "lemma" form with a meaning. The definition of the "lemma" form is an open issue (see also section 4). In addition, often MWEs have "lemma" variants due not to grammatical phenomena but, for instance, to mutually exclusive choices of functional words or to the optionality of articles, and still, all these forms correspond to one meaning. It has been up to each resource's authors to decide which of these forms represents the MWE as its "lemma form" and how all these forms are related among them. As a result, different resources encode essentially the same MWE under different entries, as shown in Ex. 1 for Greek. Guidelines are needed even at this elementary level.

(1) [el]

vazo	(ti)	thilia	sto	lemo	kapiou	
put	(the)	noose	to.ADP.the	neck	someone.GEN	
vazo	(ti)	thilia	giro	apo	to	lemo
put	(the)	noose	around.ADV	from.ADP	the	neck
kapiou						
someone.GEN						

'to force someone to be involved in an unpleasant situation'

Second, various resources encode a different set of morphosyntactic and semantic features, in some cases with different degree of granularity, which poses a problem for their combined use and mutual enrichment. Guidelines handling the diversity among languages, in terms of morphological and syntactic properties of MWEs would facilitate their uniform representation and boost their NLP applications.

4. Universality: on cracking hard nuts

The notion of "word" is central to UD, but it is hard to define it in the context of the various typologically diverse languages. Thus, as a starting point of comparison, the strategy proposed by Haspelmath (2023) is followed. According to Haspelmath, 'A word is (i) a free morph, or (ii) a clitic, or (iii) a root or a compound possibly augmented by nonrequired affixes and augmented by required affixes if there are any.'. He also defines all the terms that constitute this definition: a free morph, a clitic, roots of various kinds, a compound, required/nonrequired affixes. Even with this typologically friendly approach, there exist a number challenges in a cross-lingual context. The main ones are: demarcation of clitics (words) vs. affixes (non-words), analysis of the compounds, marking the places of contraction splits.

For better modeling of data on the word level, a survey was conducted with Haspelmath's criteria. Responses for 43 languages were received. Based on that, a second version of the survey is being prepared that will allow for better comparison among language-specific properties. This new survey will target UD and non-UD languages and ask for examples of all of Haspelmath's word types that occur in the language. For UD languages, it will also ask for divergences between Haspelmath words and treebank words.

Although lemmatization may seem a very straightforward process and a solved task, this is quite misleading, because there exists a number of problems both in the lemmatization of words and in that of MWEs. The guidelines from UD and PARSEME say relatively little about lemmatization from a linguistic point of view. The focus there has been predominantly on tokenization and morphosyntactic analysis before the application of various linguistic tests and proposed classifications. For example, the relation between a token and a word is discussed in Savary et al. (2018): a token coincides with a word, several tokens constitute a multiword and one multiword contains several tokens. In UD the following is said: "The LEMMA field should contain the canonical or base form of the word, which is the form typically found in dictionaries. If a language is agglutinative, this is typically the form with no inflectional affixes; in fusional languages, the lemma is usually the result of a language-particular convention. If the lemma is not available, an underscore ("_") can be used to indicate its absence.". It means that the majority of decisions are left within the hands of treebank providers. Also, the guidelines say that "Except perhaps in rare cases of suppletion, one form should be chosen as the lemma of a verb, noun, determiner, or pronoun paradigm".

Various frameworks and annotation schemes apply different strategies to lemmatization and identify various issues. For example, Mambrini and Passarotti (2019) point to the following challenges in relation to Latin: the graphical representation, the spelling, the word ending, the representative paradigmatic slot, the homographic lemmas, the ambiguity in choosing the lemma, for example for participles that are hybrid forms and can be viewed either as verb forms by origin or as adjectives in some of their usages. The same holds for the deadjectival adverbs that can be viewed as part of the adjective paradigm or have their own lemmas. In (Mubarak, 2018) it is shown that the lemmatization task is quite complex for Arabic. The main linguistic problem is the mismatch between a word with a diacritic and its context (e.g. nouns and adjectives).

We outline only some of the challenges here. They refer to the issues of selecting the right form as a lemma, the existence of two options, the graphic representation varieties, the spelling specifics, the relation between inflection and derivation, the relation between orthographic words, their meaning

and their spelling. The presented examples below feature some frequent lemma assigning problems across annotation schemes – within a single language and among languages. The list is not exhaustive, but it reflects the situation in many other languages and frameworks. Since this task is work in progress, the plan is to study the lemmatization decisions in the various UD treebanks and in PARSEME corpora as being already very multilingual and as sources of integration of these two frameworks and data, and also beyond them – through investigating papers on different language families, as well as through questionnaires.

Lemmatization challenges of some words and tokens

- *Pronouns*. In some languages (like Bulgarian, Czech, Maltese) there are short and long forms of some pronouns (e.g. personal), or strong and weak ones (like in Greek and Italian). Thus, the following possibilities for lemmatization exist for the short 3rd person pronouns in Czech, for example: a) the lemma equals the wordform itself ([cs]: *ho*-3P.MASC.SG.ACC.SHORT 'him'), b) the lemma goes to the long 3rd person form ([cs]: něho-3P.MASC.SG.ACC.SHORT 'him'), c) the lemma goes to the nominative, masculine, 3rd person form ([cs]: *on*-3P.MASC.SG.NOM 'he'), while in d) the lemma is the pronoun in 1st person, singular, nominative as the less marked form ([cs]: *já*-1P.SG.NOM 'I'). Thus, different strategies can be applied with varying depth until reaching the lemma.
- *Doublets*. There are doublet verbs that share the same paradigm. For example, the same lemma verb with two different endings ([bg]: *zna-m* and *zna-ya* (lit. *know-I*) 'to know'); or the same lemma adjective with two different variants ([bg]: *sasht* and *sashti* 'same-M.SG'). Thus, one of the doublets might be selected as representative, but it is sometimes hard to make such a selection.
- *Numbers*. In text data, numbers can occur as words or as digits. Should both representations of the same number have the same lemma? And if so, then which one?
- *Negated words*. This problem relates also to graphic conventions. In some languages, the negation of a word is written together, for example – as a prefix. In Bulgarian, this holds for the nominals, in Czech this holds also for verbs, while in Romanian it holds for some nominals and for three out of the four non-finite forms of a verb (only for participle, supine and gerund, but not for infinitive). Should the lemma of the negated word be its positive counterpart (meaning that negation is treated rather like inflection than derivation)?
- *Diminutives*. Although the process of making diminutives is derivational, it is still not clear whether the lemma of the word should be the diminutive or the original word. According to the current UD guidelines, the lemma does not remove derivational morphology. If such a strategy is followed, the lemma should be the diminutive. However, if most of the diminutives are not part of the dictionary, then there might be problems during the next NLP processing tasks.

Lemmatization challenges of some MWEs

- *Compounding*. In many languages, a compound (traditionally a word with (at least) two roots) can be written differently: as two words, as one word or with a hyphen. Compare in Bulgarian the double spelling: *biznes plan* (two words) and *biznesplan* (one word), in English *business plan* (two words) and in German *Businessplan* (one word). A problem arises when trying to offer a uniform analysis of these compounds within a language and across languages.
- *(Quasi)reflexive verbs*. Even within one language family like the Slavic languages, the quasi-reflexive particle can be either a separate word ([bg]: *smeya se*, [cs]: *smát se* 'to laugh') or part of the word ([uk]: *smijatysja* 'to laugh'). The reflexive pronouns are part of the word also in some Romance languages ([es]: *lavarse* 'to wash oneself') and not in others ([ro]: *se spăla* 'to wash oneself'), but in the non-reflexive meaning they lose this clitic (*lavar* 'to wash something/someone'). The question is whether the lemma is defined within each language/language family on formal criteria, or there might be possibilities to create some cross-linguistic strategies.

5. Linking MWE lexicon entries with their occurrences in corpora

Publishing language resources as Linked Data enhances accessibility, interoperability, semantic enrichment, community collaboration, and the promotion of open science. These contribute to the advancement of linguistic research, language technology, and cross-disciplinary insights.

Analyzing unique language patterns across different languages can benefit from sharing aligned and annotated corpus data in a format that complies with community standards like the NLP Interchange

Format (NIF) (Hellmann et al., 2012, 2013) and CoNLL-RDF (Chiarcos and Fäth, 2017; Chiarcos and Glaser, 2020). CoNLL-RDF is a simplified version of NIF that aligns with tab-separated formats, such as CoNLL, CoNLL-U for Universal Dependencies, and Parseme-TSV for PARSEME.

Working towards the objective of designing a lexicon-corpus interface and prove its functionality, we will expand the existing ELEXIS-WSD Parallel Sense-Annotated Corpus (Martelli et al., 2023). Currently at version 1.1, it can be accessed from the CLARIN.SI repository[4] and contains 2,024 sentences across 10 languages, along with a sense repository for each language. The expansion of the corpus will involve adding new languages (Krstev et al., 2024) and upgrading the annotation to enable linking MWE lexicon entries with their occurrences in the corpora.

Moreover, these resources should also be published as Linked Data (using NIF) to facilitate linking with the sense repository of the corpus. For the ELEXIS dictionary data, the OntoLex vocabulary[5], a widely used community standard for machine-readable lexical resources in the context of RDF, Linked Data, and Semantic Web technologies (McCrae et al., 2017), will be considered, as it is currently the foundation for the majority of lexical data available on the web of data.

Apart from the core module **Lemon** with general data structures, OntoLex modules relevant to MWEs include the module for the internal structure and combinatory semantics of MWEs **Decomp**, and MWE morphology **Morph** module. The new module for Frequency, Attestations, and Corpus-based Information (**FrAC**)[6] (Chiarcos et al., 2022a,b) supports linking lexica with corpora in many aspects of information relevant to the joint work with corpora and dictionaries. **Lexicog** (Bosque-Gil et al., 2019) is a module for lexicography that addresses structures and annotations commonly found in lexicography. It is designed to operate in combination with OntoLex for the representation of dictionaries and any other linguistic resource containing lexicographic data.

An attempt at leveraging Linked Data, NIF, and CoNLL-U for Enhanced Annotation in Sentence Aligned Parallel Corpora is reported in the literature and could be followed (Stanković et al., 2024).

[4] https://www.clarin.si/repository/xmlui/handle/11356/1842

[5] https://www.w3.org/2016/05/ontolex

[6] The current draft version of the FrAC specification is found under https://github.com/ontolex/frequency-attestation-corpus-information/

6. Proof-of-concept lexical encoding of MWEs

Taking the above into consideration, a proof-of-concept lexical encoding of MWEs in NLP lexica, that also maintains the lexicon-corpus interface, should minimally abide by the following requirements:

- a definition of the notion of "word" that is as universal as possible,
- a shared understanding of MWEs that can be annotated in corpora and then linked with lexicon entries (both the MWE as a whole and its components), including all types of MWEs (not only nominal and verbal),
- centralised guidelines for lexicon encoding regarding, i.e., the notions of lemma, canonical form, lexical features, etc.,
- a uniform representation of the syntactic properties of MWEs, and
- tools and mechanisms for linking MWE entries with their occurrences in corpora.

7. Conclusion

In an effort to create an ecosystem of interlinked MWE-dedicated lexica and annotated corpora, with an eye to universality and accommodating the languages specificities, we have already painted the current landscape of this field and are striving to find solutions for cracking the hard nuts (syntactic word definition, word and MWE lemmatization, lexical features, etc.) and to create guidelines for MWE lexicographic description. Development of linguistic resources for various languages in a harmonized way and their interlinking using standardization methods can only lead to the progress of language technology, as well as serve as a model for low-resourced languages in their endeavour to catch up with domain's evolution, speeding this process due to the benefits that Linked Data can offer (Bosque-Gil et al., 2022).

8. Acknowledgments

This paper is funded by the CA21167 COST Action UniDive, supported by COST (European Cooperation in Science and Technology).

9. Bibliographical References

Erica Autelli. 2020. *Phrasemes in Genoese and Genoese-Italian lexicography*, page 101–127.

University of Białystok Publishing House., Białystok.

Julia Bosque-Gil, Dorielle Lonke, I Kernerman, and J Gracia. 2019. Validating the ontolex-lemon lexicography module with k dictionaries"multilingual data. In *Electron. lexicogr. 21st cent., Proc. eLex conf.*, ART-2019-123124.

Julia Bosque-Gil, Verginica Barbu Mititelu, Hugo Gonçalo Oliveira, Maxim Ionov, Jorge Gracia, Liudmila Rychkova, Giedre Valunaite Oleskeviciene, Christian Chiarcos, Thierry Declerck, and Milan Dojchinovski. 2022. Balancing the digital presence of languages in and for technological development, a policy brief on the inclusion of data of under-resourced languages into the linked data cloud.

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022a. Modelling Frequency, Attestation, and Corpus-Based Information with OntoLex-FrAC. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4018–4027.

Christian Chiarcos and Christian Fäth. 2017. CoNLL-RDF: Linked corpora done in an NLP-friendly way. In *Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings 1*, pages 74–88. Springer.

Christian Chiarcos, Katerina Gkirtzou, Maxim Ionov, Besim Kabashi, Fahad Khan, and Ciprian-Octavian Truică. 2022b. Modelling Collocations in OntoLex-FrAC. In *Proceedings of Globalex Workshop on Linked Lexicography within the 13th Language Resources and Evaluation Conference*, pages 10–18.

Christian Chiarcos and Luis Glaser. 2020. A tree extension for CoNLL-RDF. In *Proceedings of the 12th LREC*, pages 7161–7169.

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Martelli Federico, Navigli Roberto, Krek Simon, Kallas Jelena, Gantar Polona, Veronika Lipp, Tamás Váradi, András Győrffy, and László Simon. 2021. Designing the elexis parallel sense-annotated dataset in 10 european languages. In *Proceedings of the eLex 2021 conference*, pages 377–395. Lexical Computing.

Martin Haspelmath. 2023. Defining the word. *WORD*, 69(3):283–297.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating nlp using linked data. In *The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II 12*, pages 98–113. Springer.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Marcus Nitzschke. 2012. NIF Combinator: Combining NLP Tool Output. In *Knowledge Engineering and Knowledge Management: 18th International Conference, EKAW 2012, Galway City, Ireland, 2012. Proceedings 18*, pages 446–449. Springer.

Cvetana Krstev, Ranka Stanković, and Aleksandra Marković. 2024. Towards the semantic annotation of sr-elexis corpus: Insights into multiword expressions and named entities. In *Proc. of Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)*.

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. PARSEME survey on MWE resources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2299–2306, Portorož, Slovenia. European Language Resources Association (ELRA).

Francesco Mambrini and Marco Passarotti. 2019. Harmonizing different lemmatization strategies for building a knowledge base of linguistic resources for Latin. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 71–80, Florence, Italy. Association for Computational Linguistics.

Stella Markantonatou, Panagiotis Minos, George Zakis, Vassiliki Moutzouri, and Maria Chantou. 2019. IDION: A database for Modern Greek multiword expressions. In *Proceedings of Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), Workshop at ACL 2019*, pages 130–134, Florence, Italy. Association for Computational Linguistics (ACL).

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: Development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

Hamdy Mubarak. 2018. Build fast and accurate lemmatization for Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jan Odijk. 2013. *Identification and lexical representation of multiword expressions*, pages 201–217. Springer Berlin Heidelberg, Berlin, Heidelberg.

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, Marie Candito, Verginica Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten Van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke Der, Behrang Qasemi Zadeh, Carlos Ramisch, and Veronika Vincze. 2018. *PARSEME multilingual corpus of verbal multiword expressions*.

Agata Savary, Silvio Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta, and Voula Giouli. 2019a. Literal occurrences of multiword expressions: Rare birds that cause a stir. 112:5–54.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019b. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91, Florence, Italy. Association for Computational Linguistics.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Mathieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland.

Ranka Stanković, Christian Chiarcos, and Milica Ikonić Nešić. 2024. Leveraging Linked Data, NIF, and CONLL-U for Enhanced Annotation in Sentence Aligned Parallel Corpora. In *Book of Abstracts of the UniDive 2nd general meeting, 8-10 February 2024, Naples*.

10. Language Resource References

Martelli, Federico and Navigli, Roberto and Krek, Simon and Kallas, Jelena and Gantar, Polona and Koeva, Svetla and Nimb, Sanni and Sandford Pedersen, Bolette and Olsen, Sussi and Langemets, Margit and Koppel, Kristina and Üksik, Tiiu and Dobrovoljc, Kaja and Ureña-Ruiz, Rafael and Sancho-Sánchez, José-Luis and Lipp, Veronika and Váradi, Tamás and Győrffy, András and Simon, László and Quochi, Valeria and Monachini, Monica and Frontini, Francesca and Tiberius, Carole and Tempelaars, Rob and Costa, Rute and Salgado, Ana and Čibej, Jaka and Munda, Tina and Kosem, Iztok and Roblek, Rebeka and Kamenšek, Urška and Zaranšek, Petra and Zgaga, Karolina and Ponikvar, Primož and Terčon, Luka and Jensen, Jonas and Flörke, Ida and Lorentzen, Henrik and Troelsgård, Thomas and Blagoeva, Diana and Hristov, Dimitar and Kolkovska, Sia. 2023. *Parallel sense-annotated corpus ELEXIS-WSD 1.1*. Jožef Stefan Institute. Slovenian language resource repository CLARIN.SI.