Linguistic review of the PARSEME model of multiword expressions

Anna Latusek[1], **Alina Wróblewska**[1], **and Agata Savary**[2]

[1]Institute of Computer Science PAS, Warsaw, Poland
[2]Université Paris-Saclay, CNRS, LISN, France
a.latusek@ipipan.waw.pl, alina@ipipan.waw.pl, agata.savary@lisn.fr

Relevant UniDive working groups: WG1

1 Introduction

PARSEME is an initiative bringing together researchers of multiword expressions (MWEs) – a community that develops unified guidelines enabling consistent annotation of text corpora across many languages. One of the key features of the PARSEME model is tractability, which notably means that the corpus must be easily usable in natural language processing. Our goal here, however, is to move beyond the NLP perspective and try to put ourselves in the shoes of a linguist who does not necessarily create computational resources. Key questions are: what linguistic foundations PARSEME is built on, and whether MWEs can be successfully identified with the help of a unified set of tests, taking into consideration the diversity and complexity of grammatical categories? What kinds of linguistic research PARSEME can be used for?

2 About PARSEME

PARSEME is both an annotation framework – defining a consistent procedure (Savary et al., 2017) for identifying MWEs across 30 typologically diverse languages – and a multilingual corpus implementing this approach. The corpus has been developed during the PARSEME COST Action (2012–2017, Savary et al., 2015) and continues to evolve within the UniDive COST Action (2022–2026, Savary et al., 2024). Based on the PARSEME corpus, four shared tasks have been organized. The first three (Savary et al., 2017; Ramisch et al., 2018, 2020) focused on the identification of verbal MWEs, while the last one (Scholivet et al., 2026) addresses MWEs across all syntactic categories and includes a paraphrasing component.

MWEs are defined as word combinations that exhibit lexical, syntactic, semantic and/or pragmatic idiosyncrasies (Baldwin and Kim, 2010). The PARSEME framework prioritizes consistency in MWE annotation across languages. This may limit the level of granularity typically achieved for individual languages, but it enables the framework to be applied to typologically diverse systems and ensures cross-linguistic comparability.

3 Between linguistics and computational linguistics

Establishing a common ground In PARSEME, linguists, computational linguists and computer scientists cooperate at the level of both modeling and implementation. Importantly, the PARSEME definition of a MWE builds upon the linguistic theories developed on the French and American ground (Gross, 1986, 1988; Nunberg et al., 1994). The major hypothesis here is that semantic non-compositionality is hard to test directly but correlates with lexical and morpho-syntactic inflexibility. This translates in the PARSEME guidelines into flexibility tests. For instance, in (pl) *Havel* ***czuł się źle*** (lit. 'Havel felt himself badly') 'Havel felt unwell', the expression in bold does not allow *się* 'himself' to be replaced by a non-reflexive direct object (**Havel czuł go źle* 'Havel felt him unwell'), which is considered an evidence of its semantic non-compositionality.

Identifying MWEs – principles of linguistic tests PARSEME pushes the operationalization of the previously mentioned theories towards strong reproducibility, required in computational linguistics. Namely, the linguistic tests are organized in a decision diagram with a single entry point. In this way, if two annotators give the same answers to atomic tests, their final decision is necessarily the same[1].

The tradeoff of this approach is the binary understanding of the MWE-hood. Even if linguistic work shows that compositionality is a matter of

[1]All the tests can be found here: https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0.

scale (Gross, 1986, 1988), the PARSEME corpus draws a line between MWEs on one side and regular expressions on the other. This follows 3 major principles. For nominal, adjectival, adverbial and functional MWEs, as well as for verbal idioms, one passed inflexibility test is enough to give the candidate the status of a MWE. For light verb constructions and inherently reflexive verbs, conversely, flexibility (rather than inflexibility) tests are used and roughly all of them have to be passed for a candidate to be annotated as an MWE.

One important perspective for PARSEME is to relate it more strongly to the theory of metaphor and other rhetorical figures. Many MWEs are lexicalized metaphors, but those two categories overlap only partially.

Do the tests work for all types of MWEs? Because PARSEME annotates all syntactic types of MWEs, this raises the question of whether a similar set of tests can be applied to both content and functional words. They differ from each other both grammatically and semantically, which makes the distinction between them one of the most prominent in linguistics. Content words can occur independently as clauses or clause-like units, carrying lexical meaning and referring to extralinguistic reality, while functional words lack referential meaning and serve intra-textual functions in discourse.

To check how well the tests capture this diversity, we selected Polish PARSEME corpus units from each MWE category that were not clearly idiomatic and required careful verification. We also added units identified as MWEs by only some annotators. Then, we went through all tests, noting where and why *yes/no* decisions were most difficult or which tests worked most effectively for each MWE type. The conclusions are as follows:

- The tests, due to their multidimensionality, work well for both content and functional words;
- Each type of MWE has its own specific properties, and in consequence, annotation difficulties appear in different places/tests. For instance, for content words, such as nominal MWEs, the main challenge is to distinguish collocations from MWEs – collocations in PARSEME are not counted as MWEs (***niezawisły sąd*** 'an independent court' is a collocation, while ***sąd ostateczny*** 'the last judgment' is a MWE). For functional words, irregular syntactic structure tests are difficult to apply due to the nature of these expressions, as in ***jako że*** (lit. 'as that') 'since'. For prepositional expressions, usually annotated as AdpID or AdvID, tests such as prohibited modification or lexical inflexibility are not always straightforward to apply, as in ***z pewnością*** (lit. 'with certainty') 'certainly', or ***z powołania*** (lit. 'from vocation') 'by vocation' (see more examples in the Appendix 1);
- Answering some of the tests requires additional verification (in dictionaries, corpora, and even in inflectional paradigms for morphological tests), or modifications involving operations at a high level of linguistic proficiency, which sometimes makes the tests highly specialized;
- Some of the test questions introduce an element of subjectivity (e.g. whether replacement by "words taken from a relatively large semantic class" is possible). Despite difficulties, which are rather case- or class-specific, the tests prove effective.

The second issue, which can lead to misunderstanding, is that linguists understand the *lexicalization* term differently from NLP researchers. For linguists, it is primarily the result of a diachronic process in which the meaning of a word shifts and becomes detached from its morphology. It is an "adoption into the lexicon" (Brinton and Traugott, 2005, p. 20), and its results may include, among others, MWEs. NLP experts in PARSEME (Savary et al., 2018, p. 94) adopt the linguistic definition of lexicalization but also extend it so that it applies not only to a MWE as a whole but also to its components. This is because, annotation of MWEs in text requires "specifying the precise span of a MWE, i.e. pointing at those words which are considered its inherent, lexically fixed components. Precisely these components are referred to as lexicalized within the given MWE."[2]

There is also a process of the opposite kind, grammaticalization, which means "the increase of the range of a morpheme advancing from a lexical to a grammatical or from a less grammatical to an inflectional one" (Kuryłowicz, 1975, p. 52). Here, again, grammaticalization may, but does not have to, lead to MWEs. In linguistics, this dichotomy between content and function words may be expected

[2]The precise wording in (Savary et al., 2018, p. 94) refers to verbal MWEs (VMWEs), but it can be considered applicable to MWEs of all syntactic types.

to align with the distinction between content and functional (multi-)words. But this principle does not underline the PARSEME guidelines.

PARSEME in practice The PARSEME corpus has been used in several linguistic studies dedicated to: (i) statistical cross-language commonalities and particularities in MWEs (Savary et al., 2018), (ii) rarity of literal reading of MWEs, despite their semantic plausibility (Savary et al., 2019), (iii) morpho-syntactic flexibility of French MWEs (Pasquer, 2017), (iv) connectivity patterns in Maltese and Lithuanian light-verb constructions (Savary et al., 2018), (v) correlations between semantic non-compositionality and co-reference in French (Savary et al., 2023), (vi) MWE identification for foreign language learning in French (Überrück-Fries et al., 2024).

These studies, however, are mostly internal to the PARSEME project. It would be interesting to understand how useful the PARSEME corpus and infrastructure would prove to newcomers who wish to use them for linguistic studies. The Grew-Match corpus browser (Guillaume, 2023), with its multi-layer query language, could automate many corpus investigations.

It would be particularly interesting to further exploit the cross-linguistic potential of the corpus, including studies of "statistical universals", as sketched by Savary (2023).

4 Conclusions

Even though linguistic research (for example in phraseology, comparative phraseology, or glotto-didactics) does not constitute the core application domain of PARSEME, the project's expanding resources and strong theoretical grounding within linguistics make it potentially useful for researchers working in these areas.

Acknowledgments

This research was funded by the CA21167 COST Action UniDive (Universality, diversity and idiosyncrasy in language technology).

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2 edition, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA.

Laurel J. Brinton and Elizabeth Closs Traugott. 2005. *Lexicalization and Language Change*. Research Surveys in Linguistics. Cambridge University Press.

Gaston Gross. 1988. Degré de figement des noms composés. *Langages*, 90:57–71. Paris: Larousse.

Maurice Gross. 1986. Lexicon-grammar: The Representation of Compound Words. In *Proceedings of the 11th Conference on Computational Linguistics*, COLING '86, pages 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bruno Guillaume. 2023. Graph-based multi-layer querying in Parseme Corpora. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 58–64, Dubrovnik, Croatia. Association for Computational Linguistics.

Jerzy Kuryłowicz. 1975. The evolution of grammatical categories. In *Esquisses linguistiques*, pages 38–54. Originally published in "Diogenes" 1965: 55–71.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70:491–538.

Caroline Pasquer. 2017. Expressions polylexicales verbales : étude de la variabilité en corpus (verbal MWEs : a corpus-based study of variability). In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. 19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)*, pages 161–174, Orléans, France. ATALA.

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.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Agata Savary. 2023. NLP-based Study of Universals of Linguistic Idiosyncrasy. volume 13, pages 100–102.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, and Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 87–147. Language Science Press., Berlin.

Agata Savary, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa I nurrieta, and Voula Giouli. 2019. Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir. *The Prague Bulletin of Mathematical Linguistics*, 112:5–54.

Agata Savary, Jianying Liu, Anaëlle Pierredon, Jean-Yves Antoine, and Loïc Grobol. 2023. We thought the eyes of coreference were shut to multiword expressions and they mostly are. *Journal of Language Modelling*, 11(1):147–187.

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.

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, Matthieu 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.

Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz, and Alina Wróblewska. 2024. UniDive: A COST action on universality, diversity and idiosyncrasy in language technology. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 372–382, Torino, Italia. ELRA and ICCL.

Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Mitrofan, and Vasile Pais. 2026. Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, pages 254–275, Rabat, Marocco. Association for Computational Linguistics.

Till Überrück-Fries, Agata Savary, and Agnieszka Dryjańska. 2024. Sailing through multiword expression identification with Wiktionary and Linguse: A case study of language learning. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 248–262, Rennes, France. LiU Electronic Press.

A Appendix 1

Category	Polish / English	MWE?	Test	Problem
Nominal	***woda sodowa*** *carbonated water*	yes	Applied syntactic inflexibility test – We cannot say **woda z sodem*.	A subtle semantic boundary between the fixed construction and a collocation (*woda sodowa* is sometimes mixed with baking soda in addition to CO_2, whereas woda gazowana takes its name from the gas itself).
Nominal	***woda gazowana*** *sparkling water*	no	None of the tests applies – this is a collocation.	
Adverbial	*chodzić* ***po ciemku*** *walking in the dark*	yes	Applied cranberry word test – there is no *ciemek, ciemko* in contemporary Polish.	An identical syntactic structure and similar meaning of the expressions, yet each of these cases is subject to different tests. The preposition *po* has different temporal interpretations: in *po zmierzchu* it denotes a subsequent time ('after dusk has fallen'), whereas in *po nocy* it denotes a time span within the night.
Adverbial	*chodzić* ***po nocy*** *be out at night*	yes	Applied prohibited modification test – We cannot say **chodzić po długiej/ciemnej nocy*.	
Adverbial	*chodzić* ***po zmroku*** *walking at dusk*	no	None of the tests applies.	
Adverbial	*Należę do stowarzyszenia* ***z wyboru***. *I belong to an association by choice.*	yes	Applied morphological inflexibility test – We cannot use the plural form here. This is an AdvID MWE.	A similar syntactic structure and a causative character of the expressions. Difficulty in testing prepositions in prepositional phrases due to the preposition's (here: *z/ze*) invariability and its high degree of combinability.
Functional	*Jestem burmistrzem* ***z wyboru*** *mieszkańców.* *I am the mayor by the choice of the residents.*	yes	Applied morphological inflexibility test – We cannot use the plural form here. This is an AdpID MWE.	
Functional	*Nie możesz tam wejść* ***ze względów*** *bezpieczeństwa.* *You cannot go in there for safety reasons.*	no	None of the tests applies.	

Table 1: Examples of MWEs and not-MWEs that caused difficulties during the tests.