A taxonomy proposal for multiword expressions

Carlos Ramisch
Aix Marseille Univ, CNRS, LIS, Marseille, France
first.last@lis-lab.fr

Relevant UniDive working groups: WG1

Multiword expressions (MWEs) defy attempts to categorise them, but it may be useful to group similar expressions into *categories*, both for lexicon creation and corpus annotation.[1] Several categorisation proposals exist, e.g. in construction grammar (Fillmore et al., 1988), meaning-text theory (Mel'čuk and Polguère, 1987; Mel'čuk, 2023), and computational approaches (Smadja, 1993; Sag et al., 2002). Ramisch (2015) proposed a typology based on two orthogonal axes: morphosyntax and "difficulty". Escartín et al. (2018) proposed a categorisation for Spanish, including an interesting comparative overview of the proposals mentioned above. Categorisations can also be elicited from annotation guidelines, e.g. Candito et al. (2021).

A unique *cross-lingually valid and operational MWE categorisation*, covering all phenomena that match the PARSEME MWE definition (Savary et al., 2017), remains an open question. An important milestone is the PARSEME guide for *verbal MWEs*, covering many languages (Savary et al., 2018). Its generalisation to other categories is part of future work in Unidive. To make progress on this front, we submit for discussion a typology proposal first presented in Ramisch (2023). We believe that it could inspire the new PARSEME guidelines for all-category MWEs. Our proposed taxonomy is based on external syntactic distribution (i.e. role/function in the sentence). We choose to ignore the internal syntactic structure, motivated by the existence of syntactically irregular (or exocentric) MWEs, e.g. a verb phrase acting as a pronoun fr ***n'importe quoi*** (lit. 'no matter what') 'anything' (Kahane et al., 2017). It would be tricky to model such MWEs in a taxonomy based on word POS and within-MWE dependencies.[2]

MWEs share properties with both single words and phrases. Thus, their (morpho-)syntactic characterisation asks the question: should coarse MWE categories rely on single-word POS tags (NOUN, VERB, etc.), or rather on phrasal structure (NP, VP, PP, etc.)? Adopting single-word POS is tempting, as there would be no need to create new tagsets. It would also correspond to the intuition that (some) MWEs are "words with spaces" (Sag et al., 2002). However, many categories would contain MWEs whose component POS tags are different from the whole POS (e.g. *at stake* is a ADP+NOUN, but acts as ADJ). Moreover, some MWEs are complex, e.g. pt ***quem vê cara não vê coração*** (lit. 'who sees face doesn't see heart') 'one can lie/omit their true feelings'. It might sound artificial to call these "verbs" instead of VPs. Finally, the criteria to distinguish some categories might not be clear cut (e.g. ADV vs. ADJ for some PP expressions). This would require either having multiple POS for the same MWE, or arbitrary categorisation.

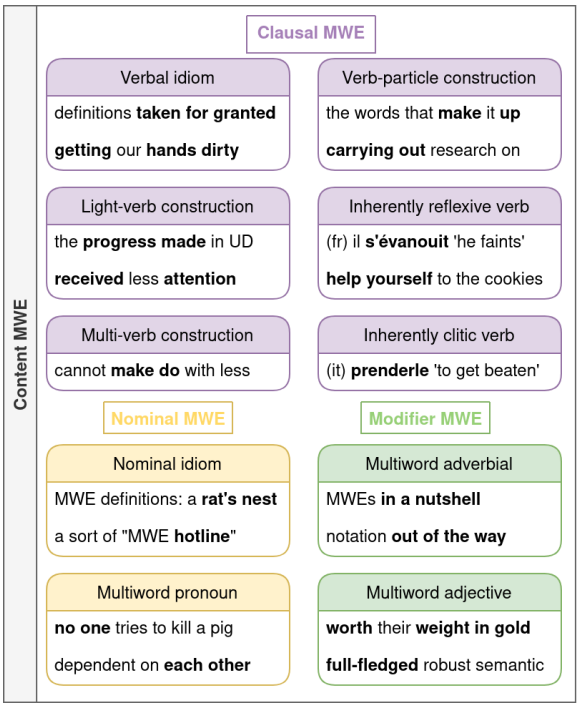

Figure 1: Taxonomy proposed for content MWEs.

Our proposal relies on phrasal structure rather than POS tags, taking advantage of the significant progress made in UD (de Marneffe et al., 2021). One advantage of adopting UD's view is that is has been put to a test for treebanking in many languages, favouring cross-lingual plausability. In

[1] We prefer "*category*" over of the ambiguous term "*type*".
[2] We assume, though, that MWEs form connected dependency subtrees, as implied by the PARSEME MWE definition.

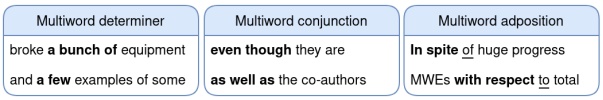

Figure 2: Taxonomy proposed for functional MWEs.

UD, linguistic units are classified as *nominals* referring to entities (usually nouns), *clauses* referring to events or states (usually verbs), and *modifiers* used to specify the attributes of nominals, clauses or other modifiers. In addition, a set of *functional* items such as determiners and adpositions are not independent, but act as specifiers of the 3 main categories. Our typology, presented in Figures 1 and 2, extends these four notions to MWEs.[3]

Clausal MWEs This broad category is roughly equivalent to PARSEME's verbal MWEs, but:

- We exclude the category *inherently adpositional verb* (e.g. *rely on*). Adopting UD, it becomes hard to annotate selected prepositions governing non-lexicalised complements: this would not form a connected subtree.
- We assume that the language-specific category *inherently clitic verb* can be generalised.

Nominal MWEs *Nominal idioms* correspond to combinations functioning as nominals in the sense of UD. In line with Cordeiro et al. (2019), we propose not distinguishing nominal idioms by the type of modifier, including bare nouns (e.g. ***science fiction***, ***dataset***), genitive nouns (e.g. ***rat's nest***), PPs (e.g. ***bed of roses***, ***pain in the neck***), etc. Compounding is a word formation process orthogonal to MWEs, so nominal idioms include both closed (***chatbot***), and open MWEs (***science fiction***). Nominal MWEs can also be nominal pro-forms, i.e. *multiword pronouns*. Since most of the time multiword pronouns contain no content word, idiomaticity tests are hard to define, so they are probably better defined as closed lists. Nominal MWEs can be exogenous, i.e. their syntactic head does not need to be a noun (e.g. ***merry-go-round***). We propose that MWEs functioning as nominals, but derived from clausal MWEs, are annotated as clausal (e.g. *the* ***progress made***), in line with PARSEME, and differently from UD guidelines, which would categorise this combination as a "nominal". This also applies to nominals acting as modifiers, importantly covering prepositional phrases such as ***from time to time*** and ***by the way***. In UD, prepositional phrases are considered nominals, even when they act as modifiers. It seems more convenient for MWEs to take a more semantic-oriented position and assume that the tests characterising modifier MWEs are more appropriate for nominals behaving so. Finally, we exclude multiword terms and named entities for the sake of simplicity.

[3]Notice that collocations are considered out of scope.

Modifier MWEs This includes *adjectives* (modify nominals) and *adverbials*, (modify clauses and other modifiers). Like for words, the distinction between multiword adjectives and adverbials is tricky. Beyond obvious adjectives (e.g. ***old school***), most MWEs here can modify both nominals and clauses (e.g. *get it* ***out of the way*** vs. *with this* ***out of the way***). We can classify as adjectives only MWEs that cannot modify anything other than nominals, and the others as adverbials. However, modifier MWEs also stand somehow in between content and functional MWEs. Thus, many PPs (e.g. ***in addition***) can take complements (e.g. ***in addition*** *to*). We propose treating all these as adverbials, and categorise as adpositions only those MWEs that cannot occur without complements (e.g. ***with respect*** *to* but not **with respect*). Beyond PPs, adverbial MWEs can be nominals (e.g. ***day after day***), coordinated adjectives (e.g. ***safe and sound***), adverbs, etc.

Functional MWEs cover *multiword adpositions*, *determiners* and *conjunctions*. Although apparently simple, this category has its share of challenges, e.g. syntactic irregularity alone is not sufficient to categorise them (Kahane et al., 2017; Savary et al., 2023). Functional MWEs are sometimes considered as completely frozen or `flat`, but non-functional MWEs may also exhibit these properties. We propose using syntactic distribution to classify these MWEs. Multiword adpositions are usually PPs that cannot occur without a complement. Determiners include idiosyncratic quantifiers (e.g. ***a few*** *examples*), but they can also be ambiguous with multiword adverbials, as in ***a lot*** *of examples* vs. *we eat* ***a lot***. In this case, we propose that the latter should be preferred, that is, considering both as adverbials, one of them taking a prepositional complement. To date, it is unclear whether numerals should be seen as multiword determiners, as tests hardly apply. Multiword conjunctions are an exception to the connected subtree rule, since they usually contain no content word. The trick here is to make them into a connected tree using UD's `flat`. Also, some conjunctions may introduce their complements using prepositions (***as***

well as) or complementisers (***now that***), which are exceptionally considered as MWE parts. Complex conjunctions are hard to delimit because traditional MWE tests are usually designed for MWEs containing at least one content word. Multiword interjections may be necessary to annotate speech, but are omitted for now. Complex tests for functional MWE categories may sound odd, as a simpler solution would consist in making closed lists. However, the criteria to include a construction in the list still needs to rely on some reasonable justification to maximise cross-lingual compatibility. Moreover, closed lists are not sufficient for ambiguous cases.

From taxonomy to guidelines If the categorisation proposed above is used as a backbone for future annotaiton guidelines, we will need to specify how to apply the criteria. For instance, the examples in the text above are mostly prototypical (or "MWE lemmas"), but defining such normalized forms can be tricky in running text (Schmitt and Constant, 2019). A possible solution would be to define *canonical forms*, as formalized by Savary et al. (2019). Working at the level of canonical forms also atenuates the complexity syntactic alternations, e.g. we hypothesise that canonical MWEs always form connected dependency UD subtrees (catena). However, the feasibility of defining canonical forms still needs to be tested empirically for non-verbal MWEs.

Acknowledgements

This work has been funded by the French Agence Nationale pour la Recherche, through the SELEXINI project (ANR-21-E23-0033-01), and benefitted from discussions with the participants of the UNLID Dagstuhl seminar (Baldwin et al., 2023).

References

Timothy Baldwin, William Croft, Joakim Nivre, Agata Savary, Sara Stymne, and Ekaterina Vylomova. 2023. Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics (Dagstuhl Seminar 23191). *Dagstuhl Reports*, 13(5):22–70.

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier, and Silvio Cordeiro. 2021. A french corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*, 8(2):415–479.

Silvio Ricardo Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

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

Carla Parra Escartín, Almudena Nevado Llopis, and Sánchez Martínez. 2018. Spanish multiword expressions: Looking for a taxonomy. In *Multiword expressions: Insights from a multi-lingual perspective*, pages 271–323. Language Science Press.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64:501–538.

Sylvain Kahane, Marine Courtin, and Kim Gerdes. 2017. Multi-word annotation in syntactic treebanks - propositions for Universal Dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 181–189, Prague, Czech Republic.

Igor Mel'čuk and Alain Polguère. 1987. A formal lexicon in the meaning-text theory or (how to do lexica with words). *Computational Linguistics*, 13(3-4):261–275.

Igor Mel'čuk. 2023. *General Phraseology: Theory and Practice*, volume 36 of *Lingvisticae Investigationes Suplementa*. John Benjamins, Amsterdam/Philadelphia.

Carlos Ramisch. 2015. *Multiword Expressions Acquisition: A Generic and Open Framework*, volume XIV of *Theory and Applications of Natural Language Processing*. Springer.

Carlos Ramisch. 2023. *Multiword expressions in computational linguistics: down the rabbit hole and through the looking glass*. Ph.D. thesis, Aix Marseille University, Marseille, France.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, volume 2276/2010 of *Lecture Notes in Computer Science*, pages 1–15, Mexico City, Mexico. Springer.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, van Gompel Maarten, 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*, volume 2 of *Phraseology and Multiword Expressions*. Language Science Press, Berlin, Germany. http://langsci-press.org/catalog/view/204/1344/1319-1.

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, 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 MWEs*, pages 31–47, Valencia, Spain. ACL. http://aclweb.org/anthology/W17-1704.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023. PARSEME meets universal dependencies: Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*, 9:14. https://nejlt.ep.liu.se/article/view/4453.

Marine Schmitt and Mathieu Constant. 2019. Neural lemmatization of multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 142–148, Florence, Italy. ACL.

Frank A. Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.