Multiword Expressions Under the Microscope

Aline Villavicencio[1]

[1] Department of Computer Science, University of Sheffield (UK)
[2] Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
a.villavicencio@sheffield.ac.uk

Almost 2 decades have passed since the publication of the paper *Multiword Expressions: a pain in the neck for NLP*[15]. In this time there have been considerable advances in representing Multiword Expressions (MWEs) in various languages, resulting from several projects, events and initiatives devoted to them[13] [12][11][16]. Ranging from idioms (*make ends meet*), light verb constructions (*take a shower*) and verb particle constructions (*shake up*) to noun compounds (*loan shark*), MWEs have provided new challenges and opportunities for language processing[5]. Their integration in tasks and applications like parsing[9][6], information retrieval[1], machine translation[4] has brought improvements for language technology, providing a degree of precision, naturalness and fluency. Any amount of interest they have attracted is justified as they account for an important part of human languages with estimates that they appear in the mental lexicon of native speakers with the same order of magnitude as single words[10], and with about four MWEs being produced per minute of discourse[8], in all languages and domains from informal to technical contexts[3]. After all this time, should they still be considered as an *open problem*[17] and *hard going*[14], or is it all *plain sailing*[14]?

In this talk I present an overview of advances in the identification of multiword expressions, that often capitalize on the various degrees of idiosyncrasy they display, including lexical, syntactic, semantic and statistical[2][18]. I will concentrate on techniques for identifying their degree of idiomaticity and approximating their meaning, as their interpretation often needs more knowledge than can be gathered from their individual components and their combinations[7] (Fillmore, 1979) to differentiate combinations whose meaning can be (partly) inferred from their parts (as *apple juice*: *juice made of apples*) from those that cannot (as *dark horse: an unknown candidate who unexpectedly succeeds*).

Acknowledgements

This talk includes joint work with Carlos Ramisch, Marco Idiart, Silvio Cordeiro, Rodrigo Wilkens, Felipe Paula and Leonardo Zilio.

References

1. Costa Acosta, O., Villavicencio A., Pereira Moreira, V.: Identification and treatment of multiword expressions applied to information retrieval. In: Kordoni, V., Ramisch,

EUROPHRAS 2019, pages 181–183,
Malaga, Spain, September 25-27, 2019.
https://doi.org/10.26615/978-2-9701095-6-3_023

C., Villavicencio, V. (eds.) Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, MWE@ACL 2011, pp. 101-109. Association for Computational Linguistics, Portland, Oregon, USA (2011).
2. Baldwin, T., Nam Kim, S.: Multiword expressions. In: Indurkhya, N., Damerau, F. J. (eds), Handbook of Natural Language Processing, 2nd edn., pp. 267-292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA (2010).
3. Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E.: Longman Grammar of Spoken and Written English. 1st edn. Pearson Education Ltd, Harlow, Essex (1999).
4. Carpuat, M., Diab, M. T.: Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, pp. 242-245. The Association for Computational Linguistics, Los Angeles, California, USA (2010).
5. Constant, M., Eryğit, G., Monti, J., Plas, L., Ramisch, C., Rosner, M., Todirascu, A: Multiword expression processing: A survey. Computational Linguistics **43**(4), 837–892 (2017).
6. Constant, M., Nivre, J.: A transition-based system for joint lexical and syntactic analysis. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016. The Association for Computer Linguistics, Berlin, Germany (2016).
7. Fillmore, C. J.: Innocence: A second idealization for linguistics. In: Annual Meeting of the Berkeley Linguistics Society (1979).
8. Glucksberg, S.: Metaphors in conversation: How are they understood? Why are they used?. Metaphor and Symbolic Activity **4**(3), 125-143 (1989).
9. Green, S., de Marneffe, M-C., Manning, C. D.: Parsing models for identifying multiword expressions. Computational Linguistics **39**(1), 195–227 (2013).
10. Jackendoff, R.: Twistin' the night away. Language **73**, 534–559 (1997).
11. Corpas Pastor, G., Colson, J-P.: Computational and Corpus-based Phraseology. John Benjamins, (2019).
12. Ramisch, C., Cordeiro, S.R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljn, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaité, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Escartín, C. P., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A., Welsh, A.: 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), pp. 222–240. Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018).
13. Ramisch, C., Villavicencio, A.: Computational treatment of multiword expressions. In: Mitkov, R. (Ed.) The Oxford Handbook of Computational Linguistics. 2nd edn. Oxford University Press (2018).
14. Rayson, P., Piato, S., Sharoff, S., Evert, S., Villada Moirón, B.: Multiword expressions: hard going or plain sailing?. Language Resources and Evaluation **44**(1-2), 1-5 (2010).
15. Sag, I., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword expressions: A pain in the neck for NLP. In: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, pp. 1-15. Springer-Verlag, Berlin, Heidelberg (2002).
16. Savary, A., Parra Escartín, C., Bond, F., Mitrovic, J., Barbu Mititelu, V.: Proceedings of the Joint Workshop on Multiword Expressions and WordNet, MWE-

WN@ACL 2019. Association for Computational Linguistics, Florence, Italy (2019), https://www.aclweb.org/anthology/volumes/W19- 51/
17. Schone, P., Jurafsky, D.: Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2001. ACL, Pittsburgh, PA USA (2001).
18. Villavicencio, A., Idiart, M.: Discovering multiword expressions. Natural Language Engineering, 1-19 (2019). https://doi.org/10.1017/S1351324919000494.