When the whole is greater than the sum of its parts: Multiword expressions and idiomaticity

Aline Villavicencio
Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
and
Computer Science and Electronic Engineering, University of Essex (UK)
avillavicencio@inf.ufrgs.br

Abstract

Multiword expressions (MWEs) feature prominently in the mental lexicon of native speakers (Jackendoff, 1997) in all languages and domains, from informal to technical contexts (Biber et al., 1999) with about four MWEs being produced per minute of discourse (Glucksberg, 1989). MWEs come in all shapes and forms, including idioms like *rock the boat* (as *cause problems or disturb a situation*) and compound nouns like *monkey business* (as *dishonest behaviour*). Their accurate detection and understanding may often require more than knowledge about individual words and how they can be combined (Fillmore, 1979), as they may display various degrees of idiosyncrasy, including lexical, syntactic, semantic and statistical (Sag et al., 2002; Baldwin and Kim, 2010), which provide new challenges and opportunities for language processing (Constant et al., 2017). For instance, while for some combinations the meaning can be inferred from their parts like *olive oil* (*oil made of olives*) this is not always the case, as in *dark horse* (meaning *an unknown candidate who unexpectedly succeeds*), and when processing a sentence some of the challenges are to identify which words form an expression (Ramisch, 2015), and whether the expression is idiomatic (Cordeiro et al., 2019). In this talk I will give an overview of advances on the identification and treatment of multiword expressions, in particular concentrating on techniques for identifying their degree of idiomaticity.

Acknowledgments

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

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.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*, 1st edition. Pearson Education Ltd, Harlow, Essex. 1204 p.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

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

Charles J. Fillmore. 1979. Innocence: A second idealization for linguistics. *Annual Meeting of the Berkeley Linguistics Society*, 5.

Sam Glucksberg. 1989. Metaphors in conversation: How are they understood? why are they used? *Metaphor and Symbolic Activity*, 4(3):125–143.

Ray Jackendoff. 1997. Twistin' the night away. *Language*, 73:534–559.

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

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 Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 1–15, Berlin, Heidelberg. Springer-Verlag.

Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), page 1
Florence, Italy, August 2, 2019.