Quantifying Changes in English Noun Compound Productivity and Meaning

Maximilian Maurer & Chris Jenkins & Filip Miletic & Sabine Schulte im Walde

January 10, 2023

Combinations of words are considered to be multi-word expressions (MWEs) if they are semantically idiosyncratic to some degree, i.e., the meaning of the combination is not entirely (or even not at all) predictable from the meanings of the constituents [Sag et al., 2002, Baldwin and Kim, 2010]. MWEs subsume multiple morpho-syntactic types, including noun compounds such as *flea market*, which have been explored extensively and across research disciplines from synchronic perspectives [Reddy et al., 2011, Bell and Schäfer, 2013, Schulte im Walde et al., 2013, Salehi et al., 2014, 2015, Schulte im Walde et al., 2016, Cordeiro et al., 2019, Alipoor and Schulte im Walde, 2020, i.a.], but state-of-the-art studies are lacking large-scale distributional approaches towards diachronic models of noun compound meaning.

The current study goes beyond the restricted synchronic concept of compound semantics and provides a novel diachronic perspective on meaning changes and compositionality (i.e., meaning transparency) of English noun compounds. We specifically investigate the diachronic evolution of the productivity of compound constituents relative to their degree of compositionality, relying on an established gold standard dataset with human compositionality ratings by Reddy et al. [2011] and a cleaned version of the English diachronic corpus *CCOHA* [Alatrash et al., 2020]. Given that type and token frequencies and probabilities, type-token ratios, entropy, etc. represent key concepts in determining quantitative properties of corpora as well as regarding individual word types and co-occurrences, we compute a range of statistical measures to quantify changes in productivity. These include Baayen's *Large Number of Rare Events (LNRE)* measures [Baayen, 2001], which have become a standard in statistical estimation of productivity, as well as measures that represent textual constants and therefore smooth the effect of different text lengths. For example, Tweedie and Baayen [1998] showed that with the exception of two measures, K suggested by Yule [1944] and Z suggested by Orlov [1983], all constants systematically change as a function of the text length.

In terms of empirical findings, we hypothesise that the current-language degree of compositionality differs for compounds with high- vs. low-productive constituents [Jurafsky et al., 2001, Hilpert, 2015, i.a.]. That is, we expect to find distinct analogical temporal development patterns for compositional compounds (such as *maple tree*, *prison guard*, *climate change*) in comparison to more idiosyncratic compounds (such as *flea market*, *night owl*, *melting pot*), with regard to modifier as well as head productivity. Our results constitute an important step towards a better understanding of compound semantics over time, as well as a reference point for future work deploying other modeling approaches on the same topic.

References

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. CCOHA: Clean Corpus of Historical American English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France, 2020.

Pegah Alipoor and Sabine Schulte im Walde. Variants of Vector Space Reductions for Predicting the Compositionality of English Noun Compounds. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4379–4387, Marseille, France, 2020.

R. Harald Baayen. *Word Frequency Distributions*. Kluwer Academic Publishers, 2001.

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

Melanie J. Bell and Martin Schäfer. Semantic Transparency: Challenges for Distributional Semantics. In *Proceedings of the IWCS Workshop on Formal Distributional Semantics*, pages 1–10, Potsdam, Germany, 2013.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. Unsupervised Compositionality Prediction of Nominal Compounds. *Computational Linguistics*, 45(1):1–57, 2019.

Martin Hilpert. From *hand-carved* to *computer-based*: Noun-Participle Compounding and the Upward Strengthening Hypothesis. *Cognitive Linguistics*, 26(1):1–36, 2015.

Daniel Jurafsky, Alan Bell, Michelle Gregory, and William D. Raymond. Probabilistic Relations between Words: Evidence from Reduction in Lexical Production. In Joan Bybee and Paul Hopper, editors, *Frequency and the Emergence of Linguistic Structure*, Typological Studies in Language, pages 229–254. John Benjamins, Amsterdam / Philadelphia, 2001.

Y. K. Orlov. Ein Modell der Häufigkeitsstruktur des Vokabulars. In *Studies on Zipf's Law*, pages 154–233. Brockmeyer, Bochum, 1983.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand, 2011.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico, 2002.

Bahar Salehi, Paul Cook, and Timothy Baldwin. Using Distributional Similarity of Multi-way Translations to Predict Multiword Expression Compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 472–481, Gothenburg, Sweden, 2014.

Bahar Salehi, Paul Cook, and Timothy Baldwin. A Word Embedding Approach to Predicting the Compositionality of Multiword Expressions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies*, pages 977–983, Denver, Colorado, USA, 2015.

Sabine Schulte im Walde, Stefan Müller, and Stephen Roller. Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, pages 255–265, Atlanta, GA, USA, 2013.

Sabine Schulte im Walde, Anna Hätty, and Stefan Bott. The Role of Modifier and Head Properties in Predicting the Compositionality of English and German Noun-Noun Compounds: A Vector-Space Perspective. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*, pages 148–158, Berlin, Germany, 2016.

Fiona J. Tweedie and R. Harald Baayen. How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, 32:323–352, 1998.

G. Udny Yule. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, 1944.