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


