Breakaway compounds: diachronic change of noun compounds sharing a head constituent

Chris Jenkins, Filip Miletić, Sabine Schulte im Walde
Universität Stuttgart
{christopher.jenkins, filip.miletic, schulte}@ims.uni-stuttgart.de

Noun compound formation is a productive process in both German and English, facilitating the expression of new concepts by combining existing words in a compositional manner (cf. Schulte im Walde et al., 2016; Cordeiro et al., 2019). However, observing a shift in available senses from e.g. high to low compositionality (testing a hypothesis from Bybee, 2015, that this is a general tendency) poses a challenge for lexical semantic change detection. For example: *brick wall* can be used metaphorically to describe the concept of 'obstacle', as well as to describe a literal wall made of bricks. We propose to observe compound nouns in German and English diachronic corpora (Davies, 2012; BBAW, 2022) before and after discontinuities in frequency, counted by year. Specifically, we use unsupervised methods (Giulianelli et al., 2020; Cassotti et al., 2023) to identify changes in sense distributions in subcorpora before and after these inflection points. We compare representations derived both from a unified space (i.e. a language model exposed to the full extent of the data) and between separate models trained on progressively larger chronological slices of the corpora. We are primarily interested in identifying the emergence of new senses (likely more metaphorical senses) and contracting inventories of senses (i.e. the death of an earlier sense). Uses of each target compound within each sub-corpus are compared against uses of the compound's head constituent (e.g. *wall* for *brick wall*), as well as a set of noun compounds in the same constituent family meeting a minimum frequency threshold within the corpus (e.g. *brick wall, stone wall, city wall, castle wall…*) in order to ground any changes in the target compound against these other terms. In addition to developing a model to *predict* sense expansion / contraction, we identify common features of compounds undergoing changes (such as frequency relative to compounds sharing a constituent). *In-context* compositionality ratings are collected to validate our model predictions.

References: BBAW. (2022). *Deutsches Textarchiv*. • J. Bybee. (2015). *Language Change*. CUP. • P. Cassotti, L. Siciliani, M. DeGemmis, G. Semeraro, P. Basile. (2023). XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proc. ACL*. • S. Cordeiro, A. Villavicencio, M. Idiart, C. Ramisch. (2019). Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1). • M. Davies. (2012). Expanding horizons in historical linguistics with the 400-million-word COHA. *Corpora*, 7(2). • M. Giulianelli, M. Del Tredici, R. Fernández. (2020). Analysing lexical semantic change with contextualised word representations. In *Proc. ACL*. • S. Schulte im Walde, A. Hätty, S. Bott, N. Khvtisavrishvili. (2016). GhoSt-NN: A representative gold standard of German noun-noun compounds. In *Proc. LREC*.