VerbCompoCor: A German Corpus with Compositionality Judgments for Verb-Dependent Pairs

Fabienne Cap[1], Rafael Ehren[2], Maximilian Köper[3], Timm Lichte[2],
Sabine Schulte im Walde[3] and Heike Zinsmeister[4]
[1] *Uppsala Univ.*, [2] *Univ. Düsseldorf*, [3] *Univ. Stuttgart*, [4] *Univ. Hamburg*

Tag
Datum
Zeit
Raum

VerbCompoCor is a German corpus where verb-dependent relations are manually annotated with compositionality judgments. It includes 7,500 sentences that are taken from the German part of the PARSEME shared task corpus (Savary et al. 2017). Prior to annotation, we parsed the corpus using the MATE parser (Bohnet 2010) and then extracted verb-dependent relations following Scheible et al. (2013). For the manual annotation, we use the WebANNO interface (Eckart de Castilho et al. 2016). The automatically extracted verb-dependent relations are highlighted in WebANNO in order to facilitate and speed up the annotation process. Our annotators have to decide whether a relation at hand is valid (i.e. not a preprocessing error) and if so, which compositionality score to assign to the relation. We use a 6-value scale to express compositionality from 0 (completely compositional) to 5 (completely non-compositional) and provide the annotators with annotation guidelines to ensure coherent annotations. The guidelines are to a large extent adapted from PARSEME and formulated as a decision tree with multiple tests in order to determine if the present verb-dependent relation is fully compositional or not. Contrary to the annotation found in the PARSEME shared task corpus, which uses a linguistic taxonomy of multi-word expressions (light-verb constructions, idioms, inherently reflexive verbs and particle verbs), VerbCompoCor follows an empirically more neutral approach. The annotated corpus may be used to develop tools for the prediction of compositionality for verb-dependent relations, which in turn then may improve the performance of larger applications.

References: • Bohnet, B. (2010): Top accuracy and fast dependency parsing is not a contradiction. In: *Proceedings of COLING*. • Eckart de Castilho, R. et al. (2016): A web-based tool for the integrated annotation of semantic and syntactic structures. In: *Proceedings of LT4DH*. • Savary, A. et al. (2017): The PARSEME shared task on automatic identification of verbal multiword expressions. In: *Proceedings of MWE*. • Scheible, S. et al. (2013): A compact but linguistically detailed database for German verb subcategorisation relying on dependency parses from a web corpus: Tool, guidelines and resource. In: *Proceedings of WAC*.