A Proposal for Explicit Word Formation Annotation in Discourse Corpora

Mark-Christoph Müller
Natural Language Processing Group
Heidelberg Institute for Theoretical Studies (HITS) gGmbH
Heidelberg, Germany

Abstract

Meaningful empirical linguistic analysis requires both machine-readable, annotated corpora of sufficient size and computational methods for processing them. For modern annotated corpora, especially those that cover higher-level phenomena like coreference, discourse connectives, rhetorical structure, etc., *stand-off, multilevel annotation* (Gries & Berez, 2017) is the de-facto standard representation format. It supports the coexistence of different annotations on the same textual data, which makes it a prerequisite for the analysis of the interplay of phenomena on different linguistic levels.

We propose to enhance annotated text corpora by adding dedicated WORD FORMATION annotation levels. We argue that this will greatly improve the way that complex words can be analysed in context, including, but not limited to, the role they play in establishing and maintaining discourse structure and coherence.

Many corpora share a limitation, viz. the *lack of recognizing linguistic structure* ***below the level of the orthographic word***, which - at least for German - includes the entire domain of word formation. The reason is that text is usually tokenized, i.e. split into the smallest units that can be annotated, on the basis of whitespace and punctuation. As a result, complex nominals like (synthetic) compounds, derivations, etc. are treated as unanalysed monoliths, just because they happen to be spelled in one word, even though their discourse function could be described much more accurately on the basis of their constituents.

Consider the following two examples, which are the result of a cursory inspection of the Potsdam Commentary Corpus (PCC) 2.2 (Bourgonje & Stede, 2020) (bracketing and emphasis M.-C.M):

> **1.** "Eine der modernsten Produktionslinien ist gestern offiziell gestartet worden, mit deren Hilfe frische **Milch** [...] auf dem Frühstückstisch [...] landet. Zu Recht ist die Chefetage des [**Milch**]giganten zufrieden. [...] Und das macht Mut. Mut, mit dem die [**Milch**]verarbeiter [...] weitere

Investitionen anvisieren." **(MAZ-11544)**

Of the three occurrences of **Milch**, all of which contribute to lexical cohesion by repetition, two occur in *one-word* compounds, which prevents them from being independent tokens. Also, the analysis of the anaphoric expression **Milchverarbeiter** as a synthetic compound is inaccessible.

> **2.** "Diepensee siedelt um. [...] Dessen konnten sich die vor dem möglichen **Ausbau** des Flughafens Schönefeld weichenden Dorfbewohner gewiss sein. Bis eine Presseinformation der [**Ausbau**]-Planer einschlug [...]." **(MAZ-6993)**

Ausbau-Planer is the first mention of a discourse-new referent. While one could argue that **Planer** alone would require a *bridging* interpretation, the synthetic combination with **Ausbau** renders the whole expression unambiguous.

If there is consensus that, *in principle*, analyses like the above are useful, realisation is straightforward. In practical terms, what is required is a flexible multi-level annotation framework like our tool MMAX2 (Müller & Strube, 2006; Müller, 2020), which supports both manual annotation and automated analysis.

Our proposal is linguistically unbiased, or even naive: In essence – at least when implemented in an adequate corpus data representation framework – it may be merely a technicality, but one that we argue is able to support new, informative analyses.

References

Bourgonje, Peter, and Manfred Stede. 2020. The Potsdam Commentary Corpus 2.2: Extending Annotations for Shallow Discourse Parsing. *Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, 1061–1066. European Language Resources Association. Online: https://aclanthology.org/2020.lrec-1.133/.

Gries, Stefan Th., and Andrea L. Berez. 2017. Linguistic annotation in/for corpus linguistics. *Handbook of linguistic annotation*, ed. by Nancy Ide and James Pustejovsky, 379–409. Dordrecht, The Netherlands:: Springer Netherlands.

Müller, Mark-Christoph. 2020. pyMMAX2: Deep access to MMAX2 projects from python. *Proceedings of the 14th linguistic annotation workshop*, 167–173. Barcelona, Spain: Association for Computational Linguistics. Online: https://aclanthology.org/2020.law-1.16.

Müller, Mark-Christoph, and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, ed. by Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, 197–214. Frankfurt a.M., Germany: Peter Lang.