Syntactical Annotation of an Early New High German Corpus: Pipeline of the LangBank-Corpus

Freitag
10.03.2017
12:00 – 12:30
B3 1, 0.12

Zarah Weiß
Universität Tübingen
zweiss@sfs.uni-tuebingen.de

Gohar Schnelle
Humboldt Universität zu Berlin
kodshajg@cms.hu-berlin.de

Within the framework of the LangBank project, we are developing a semi-automatic pipeline for syntactic annotations and complexity analyses of Early New High German (ENHG). Currently, we use diplomatically transcribed ENHG data and their orthographic and morphological normalization, as provided by the RIDGES corpus (Odebrecht et al. submitted). We use the normalized data as a proxy for the diplomatic ENHG to perform Natural Language Processing (NLP) with contemporary German models, thereby circumventing the lack of stable NLP tools for ENHG. This leads to good performance for most NLP tools. However, the lack of definite patterns of sentence delimitation in ENHG prohibits stable automatic sentence segmentation, which is mandatory for most further NLP. Therefore, we manually segment the data into parseable sentential units, which we defined using a non-graphematic, linguistically and pragmatically motivated approach (Weiß & Schnelle forthcoming). It allows us to obtain dependency, constituency, and topological field parses, based on which we calculate over 200 features of linguistic complexity. For this, we use Weiß & Meurers' (submitted) system for measuring complexity of morphological, lexical, clausal, sentential, cohesion, coherence, and deagentivation domains, which we are going to extend by measures of specific ENHG constructions. Complexity features and all parses are used as new annotation layers. We are also developing a method to additionally derive parses of the diplomatic layer from the normalized layer's parses. All annotations are reintegrated to RIDGES and exported to ANNIS via the Pepper tool (Krause & Zeles 2014) for query and visualisation and will be made publicly available.

References: • Krause, T. & A. Zeldes (2014): ANNIS3: A New Architecture for Generic Corpus Query and Visualization. *Digital Scholarship in the Humanities* 33(1), 118–139. • Odebrecht, C., M. Belz, A. Zeldes, A. Lüdeling, T. Krause (Submitted): RIDGES Herbology - Designing a Diachronic Multi-Layer Corpus. • Weiß, Z. & G. Schnelle (To appear): *Sentence Segmentation Guidelines for Early New High German*. • Weiß, Z. & D. Meurers (Submitted): Fine-Grained Linguistic Modeling of Textual Complexity Improves German L1 Grade Level Assessment. *COLING Workshop on "Computational Linguistics for Linguistic Complexity"*.