Donnerstag
09.03.2017
10:30 – 11:15
B4 1, Foyer

A Digital Infrastructure to Study Latin and Historical German

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

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

We present the current state of the LangBank project, which is developing a web-based corpus infrastructure dedicated to support the study of Latin and Early New High German (ENHG). It provides authentic language input enhanced by computational and corpus linguistic methods: We are working on providing in-line translations and adjusting Weiß & Meurers' (submitted) complexity analysis to Latin and ENHG. The resulting multi-layer corpora may be queried for grammatical constructions, and texts may be grouped together based on similar complexity or vocabulary. Currently, we are designing a small, but expandable data basis: For Latin, we augmented the standardized editions of widely taught classical texts from the LatinLit corpus (Almas & Beaulieu 2016). For ENHG, we use diplomatic and normalized texts with highly variable word order, grammar, and spelling from the RIDGES corpus (Odebrecht et al. submitted). We addressed the lack of standardized punctuation in ENHG by introducing our own guidelines for manual, non-graphematic sentence segmentation (Weiß & Schnelle to appear). Also, we investigate the applicability of automatic normalization approaches to augment the data basis. We started to design two interfaces for our resource, which we will make freely accessible online: For complex linguistic queries, we converted all annotations layers to the ANNIS format using Pepper (Krause & Zeldes 2014). For assisted reading, we are working on another interface featuring in-line translations, vocabulary information, and text selection based on text complexity and topic or specific linguistic constructions.

References: • Almas, B. & M.-C. Beaulieu (2016): The Perseids Platform: Scholarship for all! *Digital Classics Outside the Echo-Chamber: Teaching, Knowledge*, 171–186. • 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"*.