The Parsed Corpus of Historical Low German (CHLG): Corpus building and first results

Anne Breitbarth (Ghent University)

Background. Middle Low German (MLG; ca. 1250-1650) syntax is only recently coming into focus of linguistic research, after it had been neglected for decades. One of the reasons for the lack of syntactic research has no doubt been the lack of suitable resources, despite a rich textual attestation due to the success of MLG as international lingua franca in the heyday of the Hanseatic trade. Since 2014, two reference corpus projects have been funded independently of one another, the Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200-1650) (ReN, https://vs1.corpora.uni-hamburg.de/ren/) and the Corpus of Historical Low German (CHLG, http://www.chlg.ac.uk), which besides MLG includes Old Saxon (Walkden 2016). For the MLG part, both projects now collaborate regarding text selection, PoS-tagging, and morphological annotation. The CHLG adds a layer of syntactic annotations, to facilitate syntactic research.

Corpus building. To automate the PoS-tagging, we have developed a PoS-tagger (Koleva et al. 2017), which achieves an in-domain accuracy of up to 87.7% (training and testing on texts from the same city and genre), despite significant spelling variation. While accuracy dropped in cross-city and cross-genre experiments, we still achieved comparatively good results (75.9%) by combining in-city and cross-city training data. We improved tagging accuracy to around 90% by using morphological information treated as separate features (rather than as strings). For the parsing (currently in progress), we use a pipeline of Python scripts (chunker etc.) and CorpusSearch revision queries, followed by manual correction in Annotald (https://github.com/aecay/annotald). The CHLG (like the HeliPaD) is parsed in the Penn-Treebank format, to allow interoperability (sharing queries) with other parsed corpora, such as the PPCHE, the IcePaHC, the Tycho Brahe corpus of Portuguese, or the MCVF corpus of historical French.

First results. The main advantage of a parsed corpora is that they allow one to reliably find even very infrequent syntactic phenomena. Therefore, they can help detect language change in its incipient and final phases, by which one can achieve a more precise dating and analysis of such phenomena. In the present talk, I will demonstrate this using a number of pilot studies carried out by the CHLG team in the last four years, incl. referential null subjects (Farasyn & Breitbarth 2016), agreement mismatches in relative clauses (Farasyn 2017), double agreement (Farasyn in progr.), and verb placement after left-peripheral adverbial clauses (Breitbarth 2017).

References

Breitbarth, A. (2017). V3 and resumptive strategies in Middle Low German conditional constructions. Paper presented at the Workshop on *V3 and resumptive adverbials*, Ghent, October 2016.

Farasyn, M. (2017). Kongruenzmuster in mittelniederdeutschen Relativsätzen: eine Pilotstudie. M. Glawe et a;. (eds.), *Kleine und regionale Sprachen*, 67–90. Münster: Olms.

Farasyn, M. (in progr.). Fitting in or standing out? Variation in agreement phenomena in Middle Low German. Ph.D. thesis, Ghent University.

Farasyn, M. & A. Breitbarth. (2016). Nullsubjekte im Mittelniederdeutschen. *PBB* 138(4): 524–559.

Koleva, M., M. Farasyn, B. Desmet, A. Breitbarth & V. Hoste. (2017). An Automatic Part-of-speech Tagger for Middle Low German. *International Journal of Corpus Linguistics* 22.1: 108–141.

Walkden, G. (2016). The HeliPaD. A parsed corpus of Old Saxon. *International Journal of Corpus Linguistics* 21(4), 559–571.