New methods for Middle Low German dialect research

Melissa Farasyn
Ghent University
Anne Breitbarth
Ghent University

This paper presents the construction of a parsed corpus of Middle Low German (MLG) and a set of automatized annotation methods. The corpus contains located and dated texts from all main dialect areas and facilitates research in the diachronic and diatopic syntax of this under-researched language. MLG is a group of related dialects spoken in northern Germany from about 1150 until 1600. The language was only partly standardized: Regional scribal languages incorporated features of neighboring dialects, but still displayed much variation, both across regional scribal languages and within one scribal language. The main issue in this work is how to deal with this variation.

The problem of datasets with much variation is avoided by choosing a data-driven approach for the Part-of-Speech tagger, using rich linguistic features, robust machine learning algorithms and genetic algorithms to optimize feature selection, independent of spelling normalization. The morphological tagger uses the fine-grained annotation standard HiNTS, specifically developed for MLG, which improves the tagging performance of the POS-tagger and facilitates automation in the parsing stage. The syntactic layer uses the PENN treebank format, chosen over other formats used for German (historical) treebanks such as TüBa (based on topological fields) or TIGER (allowing crossing branches), not only for interoperability with other historical parsed corpora using the PENN system, but also to keep decisions about structures/positions and fine-grained linguistic analysis out of the annotation. We illustrate this with a case study on the varying position of the finite verb in embedded clauses without verb movement to C. Besides the regular unambiguous T-final order a significant number of clauses is ambiguous between a T-medial and a T-final structure; a small number is unambiguously T-medial. The shallow Penn scheme has the clear advantage of avoiding an analytic decision regarding the verb's position.