PARSING THE CORPUS OF HISTORICAL LOW GERMAN (CHLG)  
(HERCULES / FWO-RI AUGE 13/02)

Motivation

• "Die Syntax des Mittelniederdeutschen ist weitgehend unerforscht [...]. Untersuchungen zur md. Syntax sind ein dringender Desiderat." (Peters 1973: 105)

• "Zur md. Syntax liegen bislang nur Einzelbeobachtungen, keine umfassenden Darstellungen, vor." (Diehl 2002: 56)

(For as far as only some individual observations on Middle Low German syntax are available.)

Historical Low German

• Old Saxon (700-1020)

• Middle Low German (1200-1650)

Workflow

Combining the Penn scheme with HINTS

• We use the morphologically fine-grained HiNTS tagset (Barteld et al. 2018), based on HiNTS (Dopper et al. 2015), but adapted specifically for MLG.

Advantages

• Collaboration between Ghent University and the Universities of Cambridge and Konstanz.

• Part-funded through the Hercules Foundation / an FWO Research Infrastructure grant (2014-2020).

Issues

• Results in some redundancy in the encoding of syntactic information (HiNTS already encodes some syntax).

• 5 POS-tags for conjunctions, based on word order:

  • But V2/V-later order does not fully map onto main/sub-clause in MLG, due to word order variation.

  • So annotating a clause as IP-MAT or IP-SUB cannot be done on word order (KON/KOUS) diagnostics alone.

A closer look at the parsing process

• Divided into interleaved phases of human and computer annotation.

• Maximising on the natural strengths/weaknesses of humans versus computers.

Current stage: parsing

A program for manual syntactic annotation in Penn Treebank style; CHLG-specific customisations.

Contact

anne.breitbarth@ugent.be
hannah.booth@ugent.be
melissa.farasyn@ugent.be
www.chlg.ac.uk

Universiteit Gent
@ugent
Ghent University