

DIALING, LT3 & HPIMS

Anne Breitbarth, Melissa Farasyn & Hannah Booth

PARSING THE CORPUS OF HISTORICAL LOW GERMAN (CHLG)

(HERCULES / FWO-RI AUG 13/02)

Motivation

- “Die Syntax des Mittelniederdeutschen ist weitgehend unerforscht. [...] Untersuchungen zur mnd. Syntax sind ein dringendes Desiderat. (Peters 1973: 105)
(The syntax of Middle Low German is largely unresearched. Investigations of Middle Low German syntax are an urgent desideratum.)
- “Zur mnd. Syntax liegen bislang nur Einzelbeobachtungen, keine umfassenden Darstellungen, vor.” (Dietl 2002: 26)
(So far we only have individual observations on Middle Low German syntax, no comprehensive accounts.)

Historical Low German

- Old Saxon (700-1200)
- Middle Low German (1200-1650)



Middle Low German language area

Parsing decisions I: pronominal adverbs

- Pronominal adverbs (e.g. *darumme*) are treated as PPs.
- The head P has a special POS-tag, PAVAP'.
- The D-element has its own POS-tag, PAVKO', and projects an ADVP which is the complement of the P.

```
(IP-MAT (PP (ADVP (PAVKO Dar))
            (PAVAP (vmme)))
  (VVFIN is)
  (NP-SBJ-1 (DPDS dat))
  (ADJP-PRD (ADJD nutte))
  (CP-THT-1 ...))
`therefore it is useful that...' (Engelhus)
```

- Frequently discontinuous in MLG.

```
(IP-MAT (ADVP-1 (PAVKO dar))
  (VAFIN heuit)
  (NP-SBJ (DDARTA de)
          (NA uogit))
  (NP-OB2 (DNEGA nen)
          (NA recht))
  (PP (ADVP *ICH*-1)
      (PAVAP an)))
`the representative has no right to that' (Duderstadt)
```

Parsing decisions II: conjunction

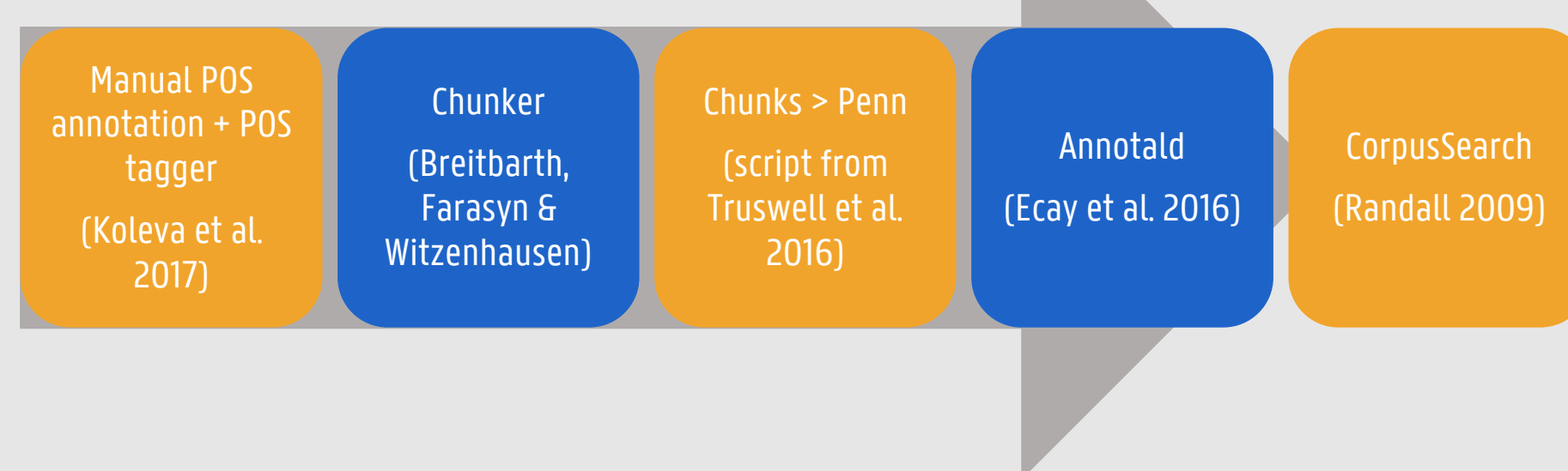
- Conjunction with shared modification.
- Modifier is attached to highest node.
- Treat as phrasal conjunction.

```
(NP-OB1 (DPOSA syn)
  (NP (NA hus)
      (CONJP (KON vnde)
              (NP (ADJA redeste)
                  (NA gud)))))
`his house and (his) available goods' (Schwerin)
```

The CHLG at a glance

- A parsed (= syntactically annotated) corpus of historical Low German in Penn-Treebank style.
- Allows for **efficient, reproducible** searches for a large number of morphosyntactic structures.
- Collaboration between **Ghent University** and the Universities of Cambridge and Konstanz.
- **Part-funded through the Hercules Foundation / an FWO Research Infrastructure grant (2014-2020).**
- Two parts:
 1. **Old Saxon:** HeliPaD (Walkden 2016)
 2. **Middle Low German:** (in progress)

Workflow

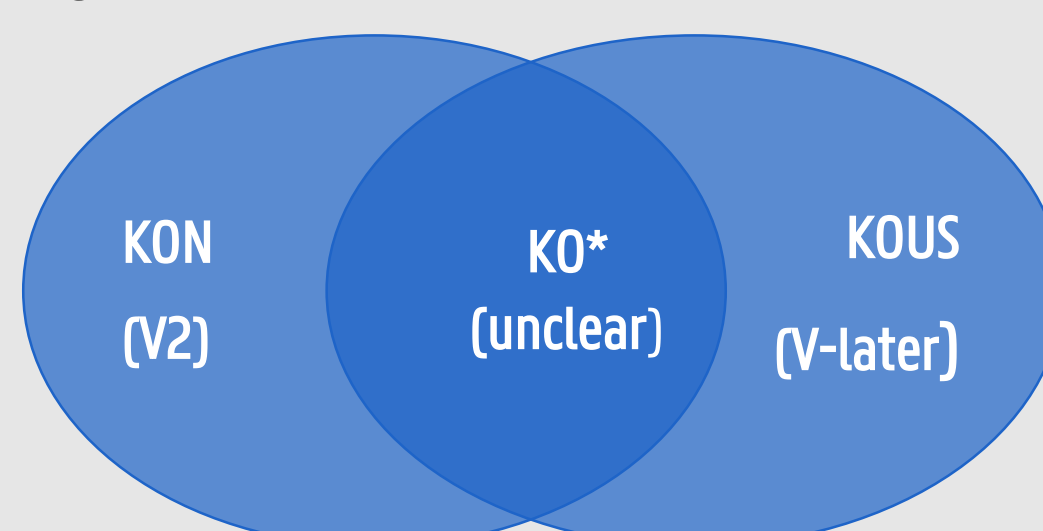


Combining the Penn scheme with HINTS

- We use the morphologically fine-grained HINTS tagset (Barteld et al. 2018), based on HiTS (Dipper et al. 2013), but adapted specifically for MLG.
- Advantages
- Collaboration with the ReN corpus for POS-tagging.
 - In line with other German corpora using Hi(N)Ts, such as the reference corpora for Middle Low German (ReN), Middle High German (ReM) and Old High German (DDD).

Issues

- Results in some redundancy in the encoding of syntactic information (HINTS already encodes some syntax).
- 3 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.

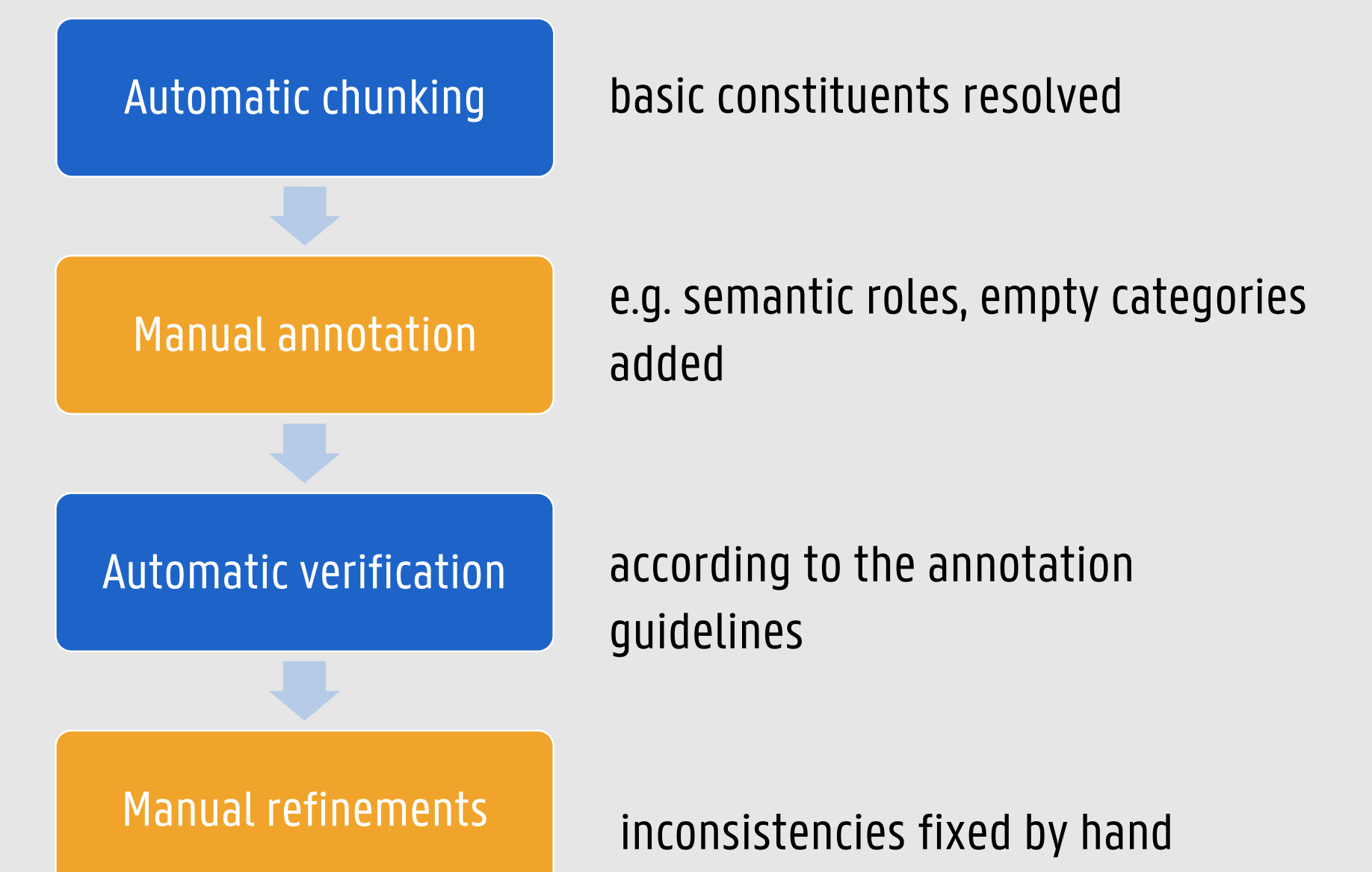
```
(IP-MAT (FM Jtem)
  (NP-SBJ (PPER wy))
  (VVFIN Bekennen)
  (CP-THT (KON dat)
    (IP-SUB (NP-SBJ (NE detmer)
                  (NE Rokeman))
      (VAFIN hefft)
      (VVPP gesettet)
      (NP-OB1 (DPOSA syn)
              (NA hus)
              ...)))
  )
`likewise we recognise that Detmer Rokeman has set his house...' (Schwerin)
```

Current stage: parsing

Scribal language	Text	Genre	No. IP-MATs
Westphalian	ArzneibuchAbdinghof	Science	–
	HerfordRechtsbuch	Law	–
	SoesterSchrae	Law	341
Eastphalian	SpiegelHelen	Religion	1,350
	BraunschweigUrkunden	Law/Charter	170
	Duderstadt	Law	483
	EngelhusChronik	Literature	1,658
North Low German	Zeno	Literature	–
	BremenUrkunden	Law/Charter	58
	BuxtehuderEvangeliar	Religion	2,013
	Griseldis	Literature	586
	OldenburgUrkunden	Law/Charter	309
Eastelbian	WillekenBrief	Private Letter	–
	FiosBlankeFios	Literature	–
	GreifswaldBürgersprache	Law	–
	RostockBürgersprache	Law	51
	SchwerinStadtBuch	Law	304
	StralsundUrkunden	Law/Charter	480
All			7,803

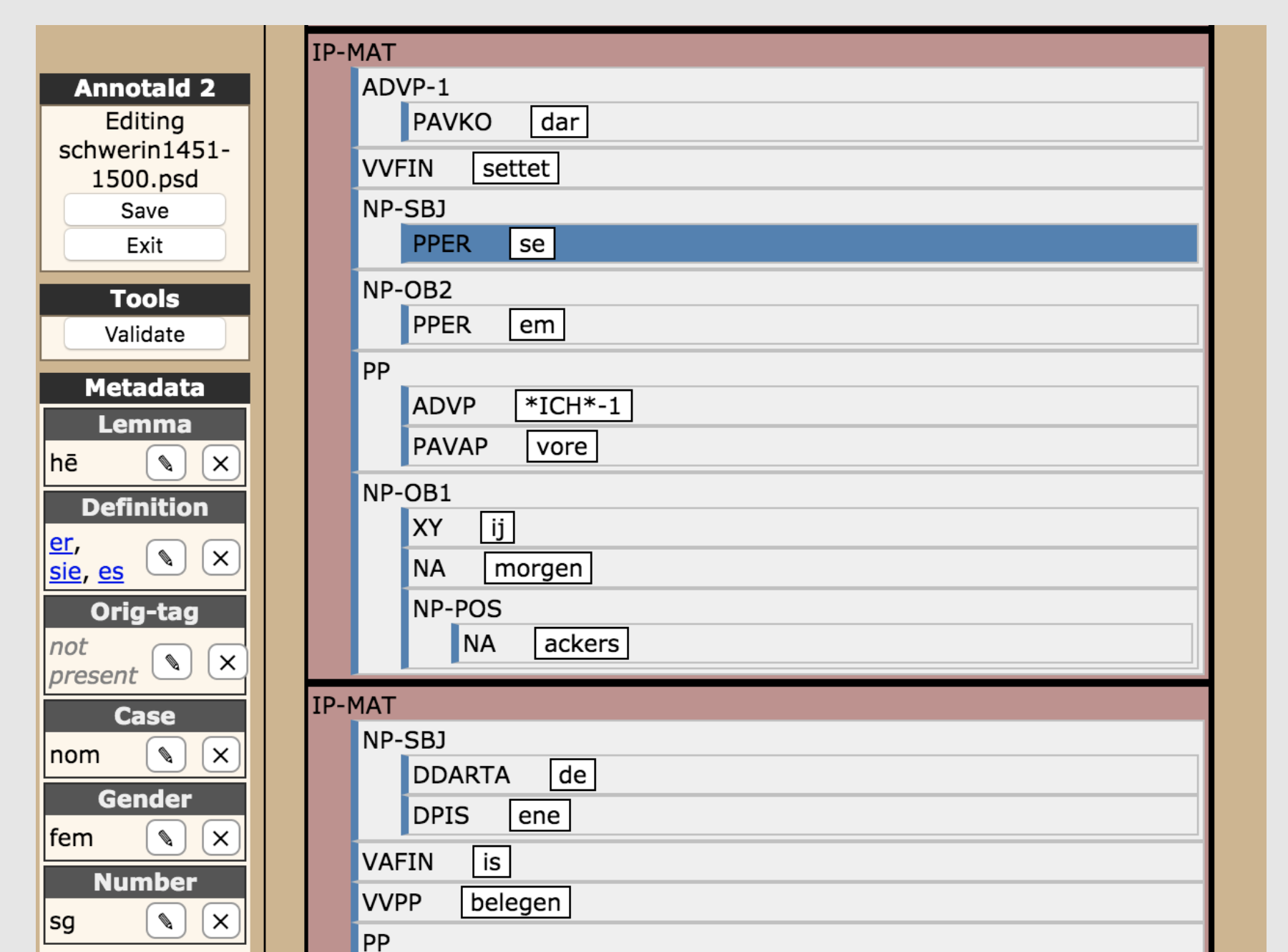
A closer look at the parsing process

- Divided into interleaved phases of human and computer annotation.
- Maximising on the natural strengths/weaknesses of humans versus computers.



Annotald (Ecay et al. 2016)

- 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;
 http://research.flw.ugent.be/en/projects/chlg

- Facebook: Universiteit Gent
- Twitter: @ugent
- LinkedIn: Ghent University

