International Parser and Lemmatizer of Dutch in Retrospect

INPOLDER

Gertjan Postma
Meertens Instituut Amsterdam

gertjan.postma@meertens.knaw.nl
CLARIN 2 - kick off meeting
Utrecht, 9 Febr 2011

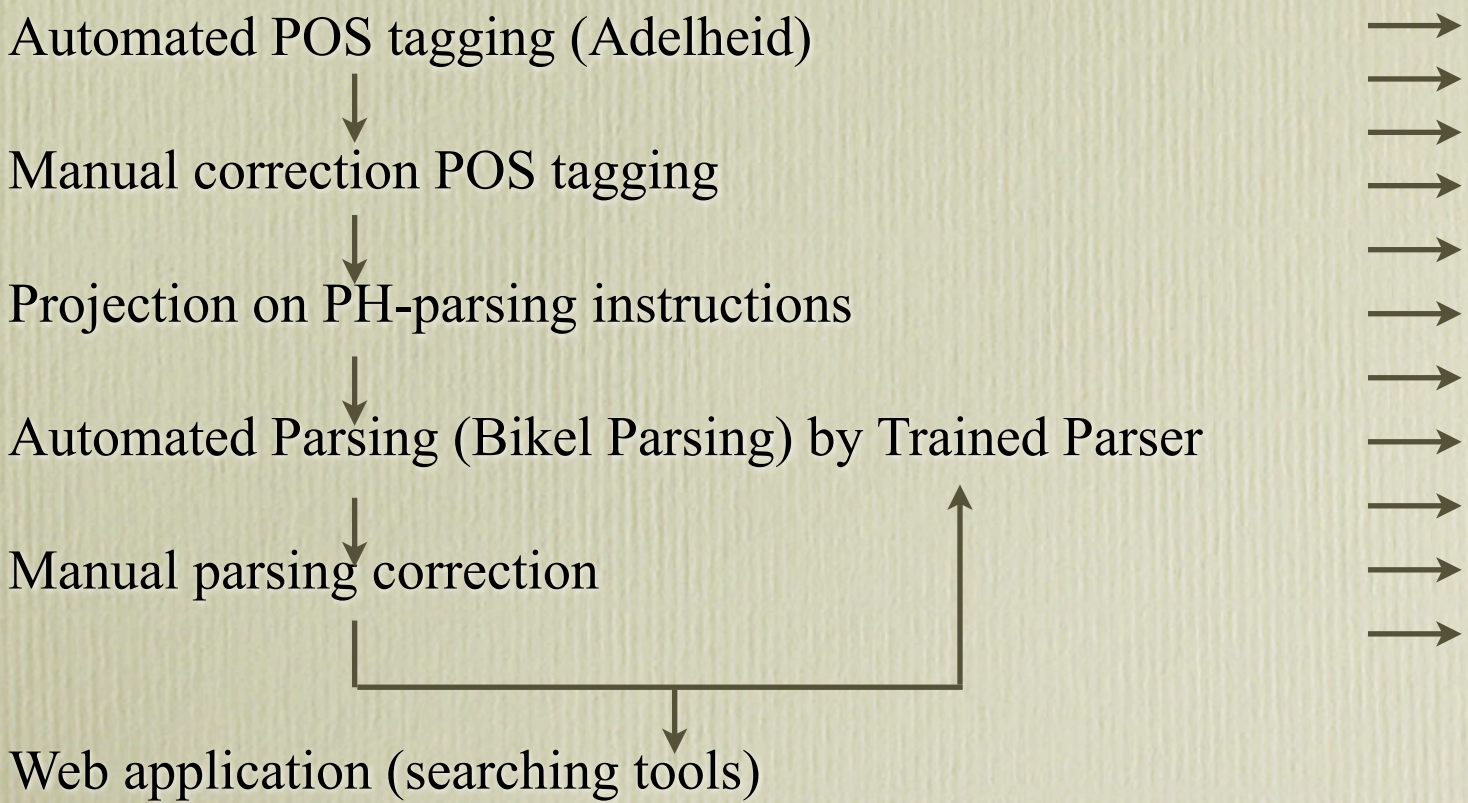
Team

Ans van Kemedade (RU) - hoofdaanvrager
Gertjan Postma (MI)
Ben Hermans (MI)
Hans van Halteren (RU)
Margit Rem (RU)
Marc Kemps-Snijders (MI)

Sister Projects

- York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE)
- Penn-Helsinki Parsed Corpus of Middle English (PPCME2)
- Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)
- Penn Parsed Corpus of Modern British English (PPCMBE)
- Corpus MCVF of French
- Tycho Brahe Parsed Corpus of Portuguese
- Icelandic Parsed Historical Corpus (IcePaHC)
- ....

Flow Diagram



Middle Dutch specimen

DYT SYNT DE ORDELE, DE GEWYST SYN UP DEN
ZWOERENMAENACH IN DEN lAIR VAN Mo CCCo XCIX
BY ZWEDERS TYDEN VAN RECHTEREN

(preambule Vonnissen van de Etstoel van Drenthe, 1399)

POS-tagging & lemmatizing (Adelheid)

<s>
DYT dyt dit 410
SYNT synt zijn 217
DE de de 471
ORDELE ordele oordeel 001
&komma; &komma; &komma; Punc(lk)
DE de die 421
GEWYST gewyst wijzen 273
SYN syn zijn 214
UP up op 700
DEN den de 474
ZWOERENMAENACH zwoerenmaenach ??? 000
IN in in 700
DEN den de 474
IAIR iair jaar 000
VAN van van 700
&Mring; &mring; ??? 020
CC&Cring; cc&cring; ??? Punc(lv)
XCIX xcix xcix 300
BY by bij 700
ZWEDERS zweders sweder 022
TYDEN tyden diede 024
VAN van van 700
RECHTEREN rechteren rechteren 020
<s>

POS-tagging & lemmatizing (corrected)

<s>
DYT dyt dit Pron(dem)
SYNT synt zijn V(fin,pres,aux_cop,formnt)
DE de de Art(def,forme)
ORDELE ordele oordeel N(plur,forme)
&komma; &komma; &komma; Punc(comma)
DE de die Pron(rel,forme)
GEWYST gewyst wijzen V(participle,past,formt)
SYN syn zijn V(fin,pres,aux_cop,formn)
UP up op Adp()
DEN den de Art(def,formn)
ZWOERENMAENACH zwoerenmaendach Zworenmaandag N(sing)
IN in in Adp()
DEN den de Art(def,formn)
lAIR lair letter N(sing)
VAN van van Adp()
Mo_CCCo_XCIX MoCCCoXCIX 1399 Num(card)
BY by bij Adp()
ZWEDERS zweders sweder N(prop,forms)
TYDEN tyden diede N(prop,formn)
VAN van van Adp()
RECHTEREN rechteren rechteren N(prop)

Projection Adelheid to PH-parsing instructions

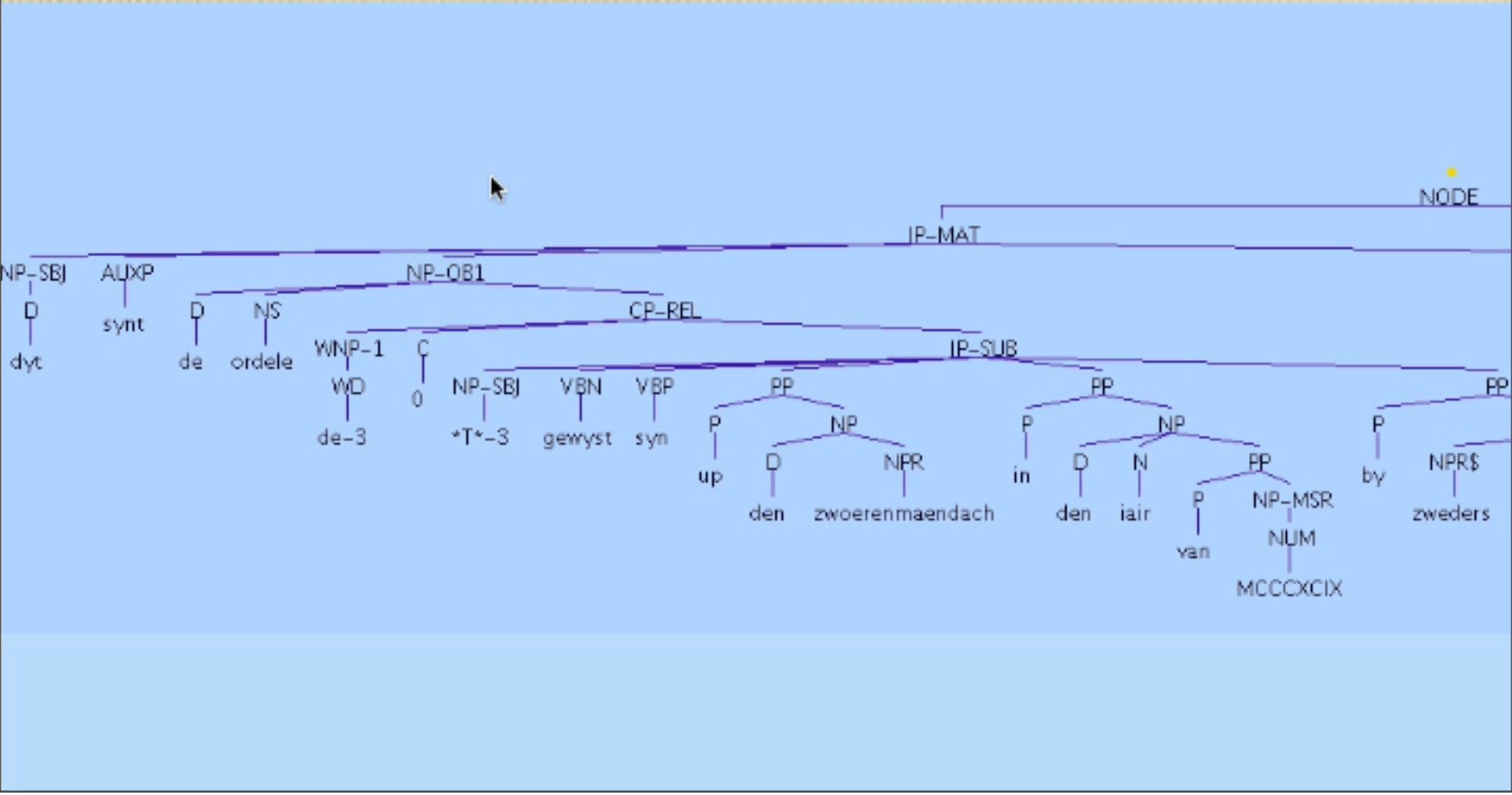
(($word_I$ ($label_I$))…. ($word_n$ ($label_n$)))

((Dyt (D)) (synt (BEP)) (de (D)) (ordele (NS)) (de (WD)) (gewyst (VBN)) (syn (VBP)) (up (P)) (den (D)) (zwoerenmaendach (N)) (in (P)) (den (D)) (iair (N)) (van (P)) (MCCCXCIX (NUM)) (by (P)) (zweders (NPR$)) (tyden (NS)) (van (P)) (rechteren (NPR)) (Etstoel1399 (ID)))

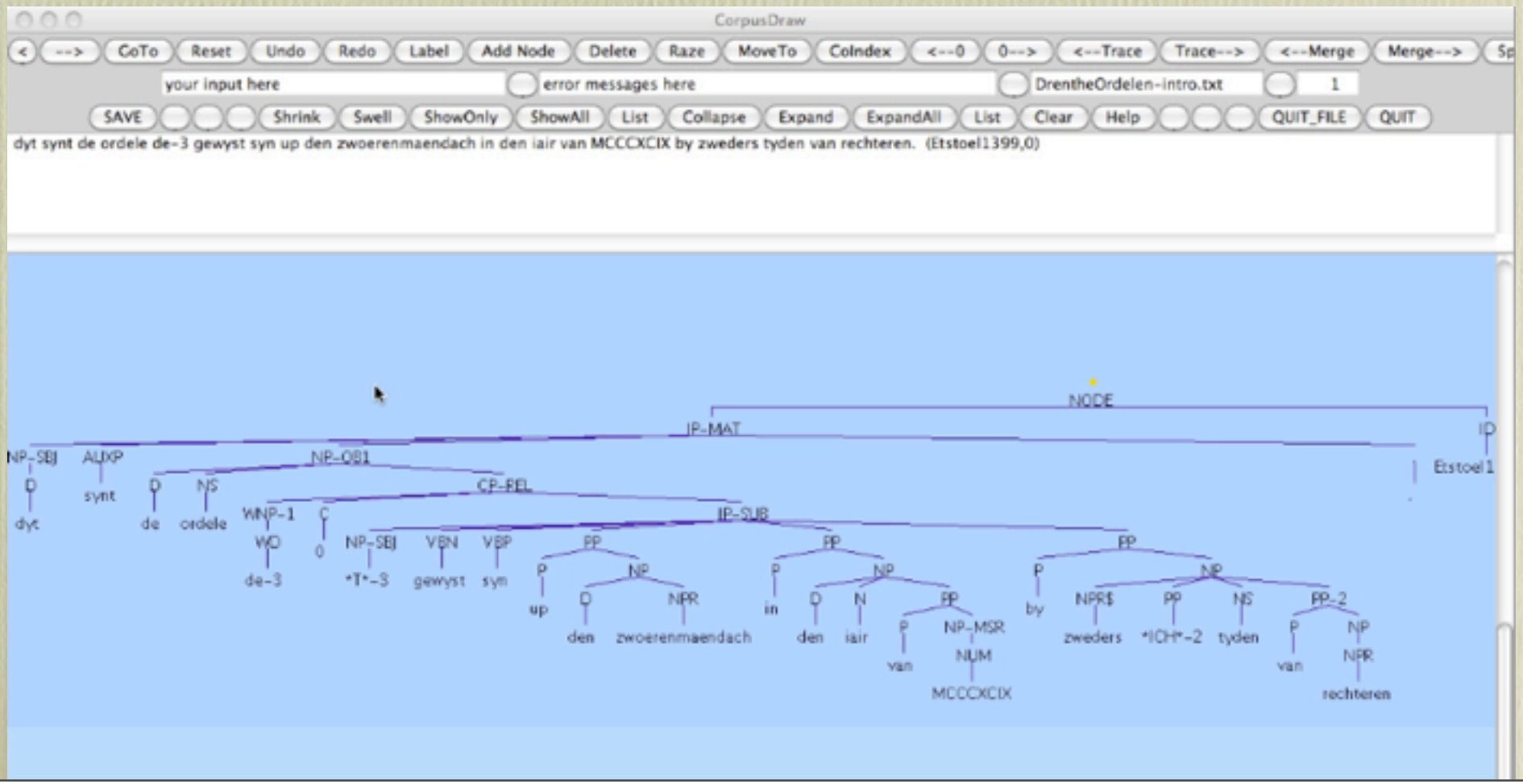
PH-parsing

```
( (NODE (IP-MAT (NP-SBJ (D dyt))
          (BEP synt)
          (NP-OB1 (D de)
               (NS ordele)
               (CP-REL (WNP-1 (WD de-3))
                    (C 0)
                    (IP-SUB (NP-SBJ *T*-3)
                         (VBN gewyst)
                         (VBP syn)
                         (PP (P up)
                            (NP (D den) (NPR zwoerenmaendach)))
                         (PP (P in)
                            (NP (D den)
                              (N iair)
                              (PP (P van)
                                 (NP-MSR (NUM MCCCXCIX)))))
                         (PP (P by)
                            (NP (NPR$ zweders)
                              (PP *ICH*-2)
                              (N tyden)
                              (PP-2 (P van)
                                   (NP (NPR rechteren))))))))
          (ID Etstoel1399,0))
```

Rendering the Tree (CorpusDraw)



CorpusDraw as a correction tool



Webapplicatie (e.g. CorpusSearch)

http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/csquery/csquery.html

CLARIN Infrastructure - metatagging

Go to XML

International Parser and Lemmatizer of Dutch in Retrospect

INPOLDER

Gertjan Postma
Meertens Instituut Amsterdam

gertjan.postma@meertens.knaw.nl
CLARIN 2 - kick off meeting
Utrecht, 9 Febr 2011