Adelheid

A Distributed **L**emmatizer for **H**istorical **D**utch

Hans van Halteren, RU Nijmegen, hvh@let.ru.nl
Margit Rem, RU Nijmegen, M.Rem@let.ru.nl
Daan Broeder, MPI Nijmegen, Daan.Broeder@mpi.nl

Overview

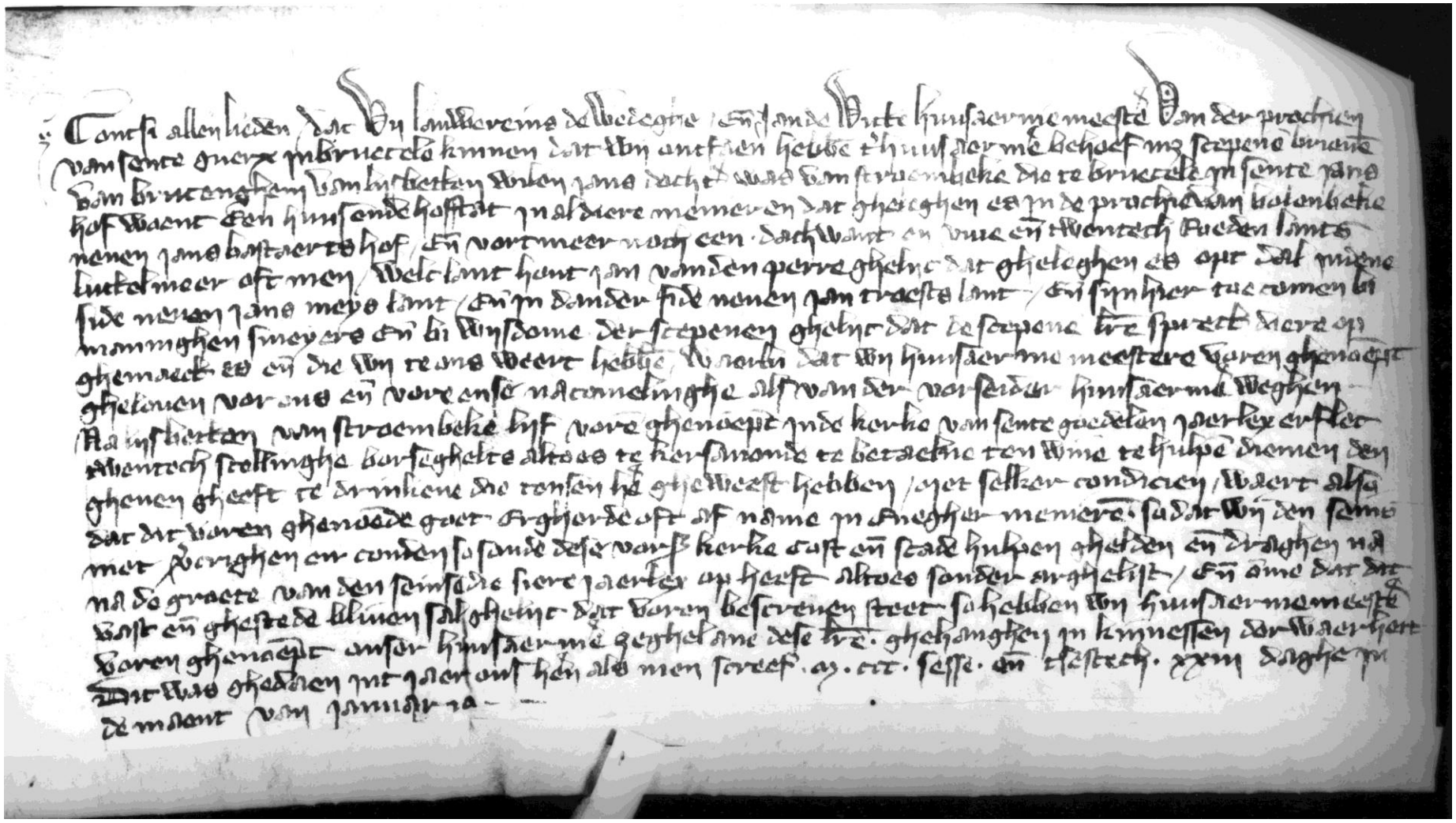
Topic

- Clarin NL demonstrator project
- Tagging and lemmatizing historical text

Structure

- Functionality of the software
- Activities during Adelheid

Software: Task



Software: Task

Input: Transcription

C108p39304 Blok862 gecollationeerd.280394.HD

wy borghermestere ende raet van groninghen bekennen ende betughen met dezen openen breue dat vor ons quam ghelmer storm ende becande dat hie heft vercoft rodetyden vyertyendehalf gras landes met al horen to behoren vor ene summe gheldes de ghelmer vorseit vol ende al betaelt js ende deze vy ertyendehalf gras landes vorseit droech ghelmer vorseit vp rodetyden vorseit ende sinen erfghenamen vrij ende quiit met allen rechte ende eghendome eweliken to bruken ende to besitten dit vorseide land js gheleghen in lywerder wolt vp de noerd zide van den wolt graue daer viif grase landes van gheleghen ziin by rodetyden erue vorseit dat an de oester zide leghet ende viif graze landes daer tette mellens erue by gheleghen js an de oester zide ende vyerdehalf gras landes an de noerd zide van den vorseiden viif grasen daer een sloet en tuschen gaet dat or kunde wy met onser stad seghel . ghegheuen jnt jaer ons heren dusent drehondert dre ende neghentich vp sente nycholaus auond do wicbolt euerdes euerd sickinc johan van den berghe ende jacob schelleghe borghermestere waren onser stad

Software: Task

Output: tags and lemmas

- tags from a reasonably complex tagset
  - based on corpus van Reenen – Mulder
  - 184 **basic** tags, plus **combination** tags for enclitic forms

Token	**Tag**	**Lemma**
och	Conj(coord)	of
en	Adv(neg)	en
betalden	V(fin,past,lex,formn)	betalen
tesen	Adp()+Pron(dem,formn)	te+deze
vorsprokene	Adj(formn)	voorgesproken
tide	N(sing,forme)	tijd
.	Punc(lp)	.

Software: Stages

- Tokenization
  - Reinterpretation of word separation
- Potential tag/lemma assignment
  - Lexicon lookup
    - Known forms
    - Expected variant forms, on basis of known variation
  - Unknown word handling
    - Machine learning
    - Nearest neighbours in expanded lexicon
- Contextual disambiguation

Software: Stages

- Tokenization
- Potential tag/lemma assignment
- Contextual disambiguation: Combination
  - SVMTool: Support Vector Machines
  - TnT: Hidden Markov Models
  - WPDV: Distribution of contextual features
  - All with left-to-right and right-to-left tagging

Software: Quality

10-fold cross-validation on van Reenen-Mulder

Recall for single best tag/lemma

- Token: 99.1%
- Tag: 94.9%
- Lemma: 93.9% or 94.8%
  - Higher percentage with test token adaptation
  - Computationally expensive

Adelheid: Software

Integration into Clarin infrastructure

Aspect	**Currently**	**Adelheid**
User interface	*Linux command line*	Web application
Activating core software	*Linux shell scripts*	Web service(s)
Core sofware	*Linux binaries, Perl scripts*	*Linux binaries, Perl scripts*
Disambiguation data	*Software determined*	*Software determined*
Internal data streams	*Ad hoc*	XML, ISOcat compliant
Input/output formats	*Ad hoc*	XML, ISOcat compliant
Lexicon formats	*Ad hoc*	XML, ISOcat compliant
Lexicons	*(Expanded) charter*	*Charter* + User provided

Adelheid: For the User

Demonstrator scenarios

- Tagging texts with basic system
- Tagging texts with additional own lexicon

Documentation

- Texts
- Tags and lemmas
- Workings and use of the system

Questions?

Later... ?