

Adelheid

A Distributed Lemmatizer for Historical Dutch

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

Consci allen hiden dat Wy landreins delbeleghe en jande Wite hinfierme meeste Van der practien
van sente ouerz inbruerde kinnen dat Wy ontfien hebbe t hinfierme behoefte sepeno lincant
som bruceengham vsmlyt beryn voren jans dacht was som straembekke die te brucele in sente jans
hof waent een hinfierde hofstat in aldiere memere en dat ghelegghen es in de practien van hoienbete
nemen jans bostaez hof En wort meer noch een dach want en vure en vrentech Puden linc
luctelmeer oft men Welc linc hant jans vonden perre ghele dat ghelegghen es opt dat migne
sude nenen jans meye linc En in dander sude nenen jans traeste linc En sijn hier toe comen bi
mannighen simeyere en bi vnsdome der sepenen ghele dat de sepeno tre spuzet diera op
ghemaect es en die Wy te ons weert hebbe voren dat Wy hinfierme meeste varen ghenoept
ghelouen vor ons en vore onse naconclinghe als vanden vorseiden hinfierme wegghen
Na inst beryn van straembekke byf vore ghenoept inde kerke van sente gudeleyn jaerlyc en flec
vrentech scollinghe borsghelke alsoes te kerke vande te betacthe ten Wme te hulpe diemen den
ghenen gheest te drinckene die ronten he gheleest hebben met selker condicyn weert als
dat die voren ghenode guet argherde oft of name in anegher memere. s. dat Wy den seme
met soerghen en cunden s. jande dese vorse kerke dast en scade hulpen ghelden en draeghen na
na de graete van den seme die siere jaerlyc op herte alsoes sander arghelyst. En ome dat die
vorse en gheste de lincen sak ghelinc dat voren bescreuen steet s. hebben Wy hinfierme meeste
voren ghenoept onser hinfierme geytelome dese tre. ghehanghen in kinnestey dor waer hert
die vorse ghedien int jaer ont hen als men screef. ay. cc. sesse. en etestech. xxij daghe in
de maent van jannuar 12

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 vyertyendehalf 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 duserndrehondert dre ende neghentich vp sente nycholaus auond do wicbolt euerdes euerd sickinc johan van den berghe ende jacob schelleghen 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	<i>Linux command line</i>	Web application
Activating core software	<i>Linux shell scripts</i>	Web service(s)
Core software	<i>Linux binaries, Perl scripts</i>	<i>Linux binaries, Perl scripts</i>
Disambiguation data	<i>Software determined</i>	<i>Software determined</i>
Internal data streams	<i>Ad hoc</i>	XML, ISOcat compliant
Input/output formats	<i>Ad hoc</i>	XML, ISOcat compliant
Lexicon formats	<i>Ad hoc</i>	XML, ISOcat compliant
Lexicons	<i>(Expanded) charter</i>	<i>Charter + User provided</i>

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... ?