Processing 14th century Dutch text through Clarin, Step 1: Adelheid

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

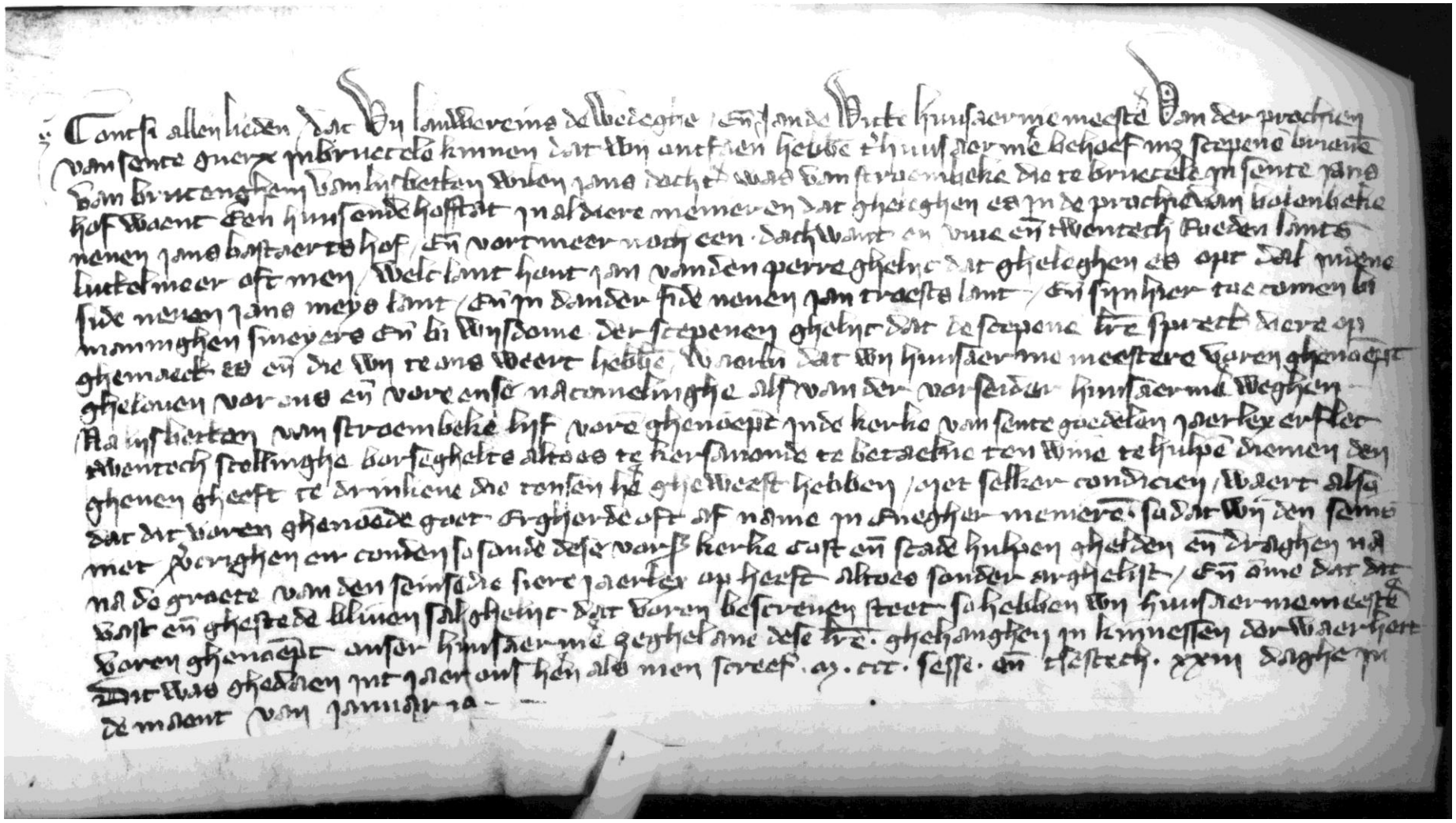
Topic

- Clarin NL demonstrator project
- Tagging and lemmatizing historical text

Structure

- Introduction to task
- Tagging-lemmatization system
- Annotation tool
- Potential uses

Adelheid: Task



Adelheid: 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

Adelheid: Task

Annotation: tags and lemmas

- modern 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)	.

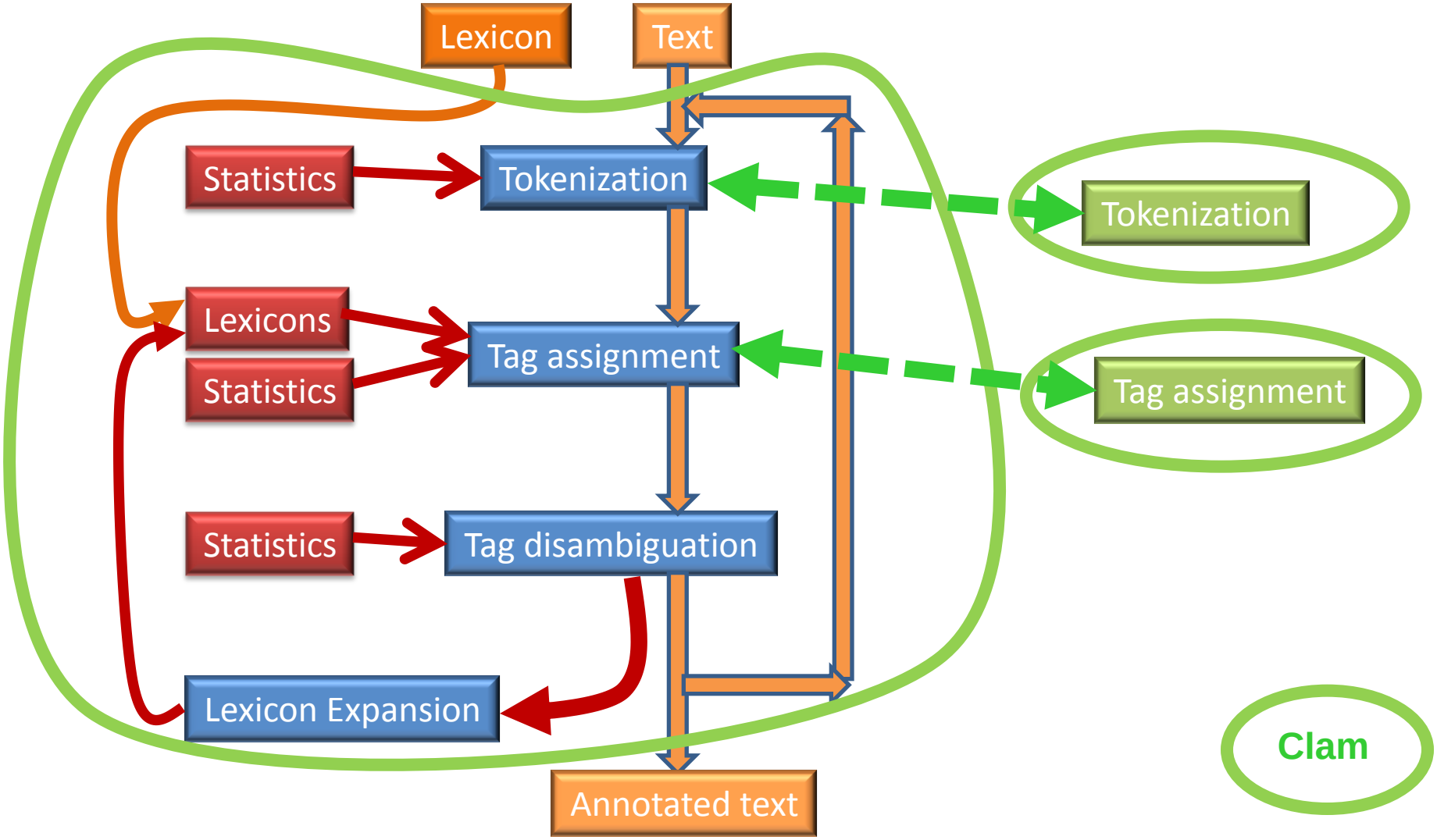
Adelheid: Special Difficulties

- Why not use “normal” existing systems?
  - Not able to properly process older Dutch
  - Assume standardized
    - Spacing
    - Punctuation
    - Spelling
  - None of these are present, thus causing problems
  - Adelheid does provide needed functionality

T-L System: Initial State

- Tagger-Lemmatizer system was available
  - As laboratory experiment
  - On personal laptop
  - Consisting of
    - 5 larger systems
    - 29 smaller Perl, Awk, Sed scripts
    - Controlled by a shell script
  - Using ad hoc data formats
  - Not very documented
- So operational, but not really re-usable

T-L System: New Architecture



T-L System: Final State

- System now available
  - Through Clarin infrastructure
  - More efficient
  - Consisting of
    - One Perl program (±5000 lines)
    - Calling 5 larger external systems
  - Using XML data formats
  - With user manuals, incl. Demonstration scenarios
- Usable by any user without needing us

Annotation tool: Need

- Example of the output

```
<token Tform="dese" Tag="Pron(dem,forme)" Lemma="deze" Tpos="1/25-28" Mform="dese" Aform="dese" Src="sys" Conf="0.7287">
  <tlp ATag="Pron(dem,forme)" ALemma="deze" AProb="0.7287"></tlp>
  <tlp ATag="Art(def,forme)" ALemma="deze" AProb="0.2190"></tlp>
  <tlp ATag="N(prop,forme)" ALemma="dieze" AProb="0.0523"></tlp>
 </token>
 <sep Tpos="1/29" Msep="True" Mform=" " Tsep="True" Asep="True" Src="sys" Conf="0.9992"></sep>
 <token Tform="letteren" Tag="N(plu,formn)" Lemma="letter" Tpos="1/29-36" Mform="lett__en" Aform="letteren" Src="sys" Conf="0.6636">
  <tlp ATag="N(plu,formn)" ALemma="letter" AProb="0.6636"></tlp>
  <tlp ATag="N(sing,formn)" ALemma="letter" AProb="0.3364"></tlp>
 </token>
 <sep Tpos="1/37" Msep="True" Mform=" " Tsep="True" Asep="True" Src="sys" Conf="0.9994"></sep>
 <token Tform="selen" Tag="V(fin,pres,aux_cop,formn)" Lemma="zullen" Tpos="1/37-41" Mform="selen" Aform="selen" Src="sys" Conf="0.6776">
  <tlp ATag="V(fin,pres,aux_cop,formn)" ALemma="zullen" AProb="0.6776"></tlp>
  <tlp ATag="V(infin)" ALemma="zellen" AProb="0.0943"></tlp>
  <tlp ATag="N(prop,forms)" ALemma="seel" AProb="0.0786"></tlp>
  <tlp ATag="V(fin,pres,aux_cop)+Pron(pers,3,sing)" ALemma="zullen+hij" AProb="0.0691"></tlp>
  <tlp ATag="N(plu,formn)" ALemma="ziel" AProb="0.0321"></tlp>
  <tlp ATag="N(prop,formn)" ALemma="seel" AProb="0.0269"></tlp>
  <tlp ATag="N(sing,formn)" ALemma="ziel" AProb="0.0182"></tlp>
  <tlp ATag="N(prop,formn)" ALemma="zelle" AProb="0.0031"></tlp>
 </token>
```

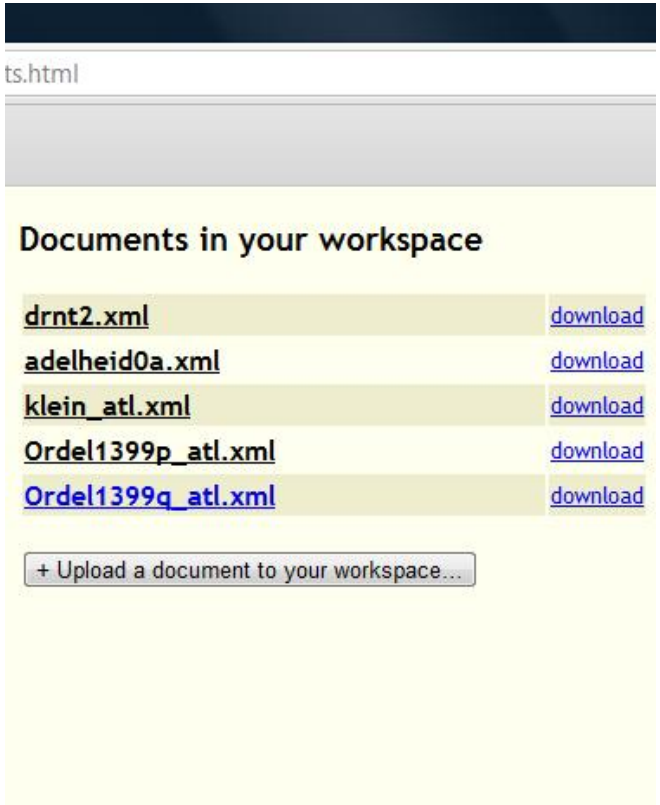
Annotation tool: Provided

- Dedicated tool for
  - Visualization
  - Adjusting annotation
  - Details below
- Also accessible through Clarin infrastructure

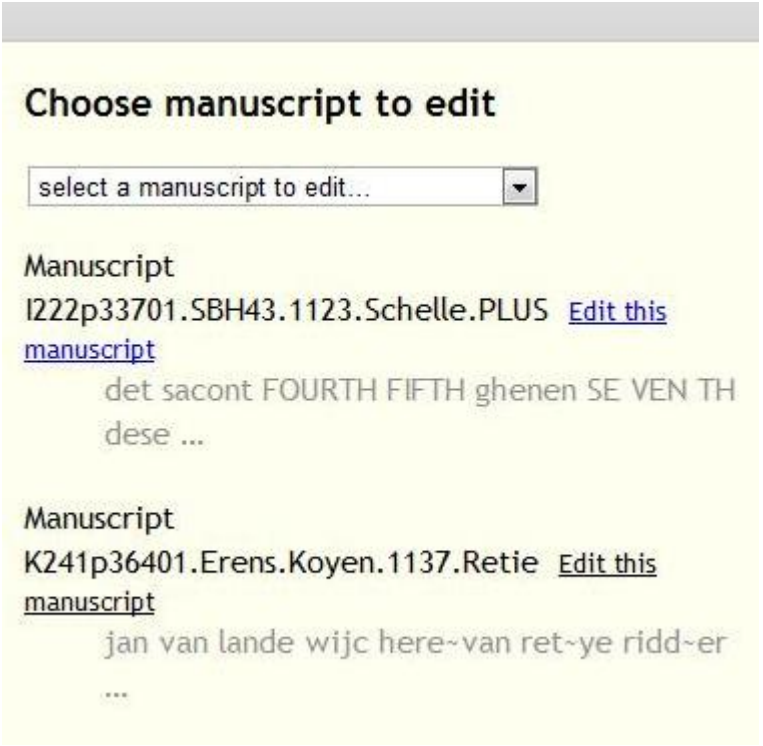
- Tool built by Edia in Amsterdam

Annotation tool: Functionality

- Up- and downloading annotation files

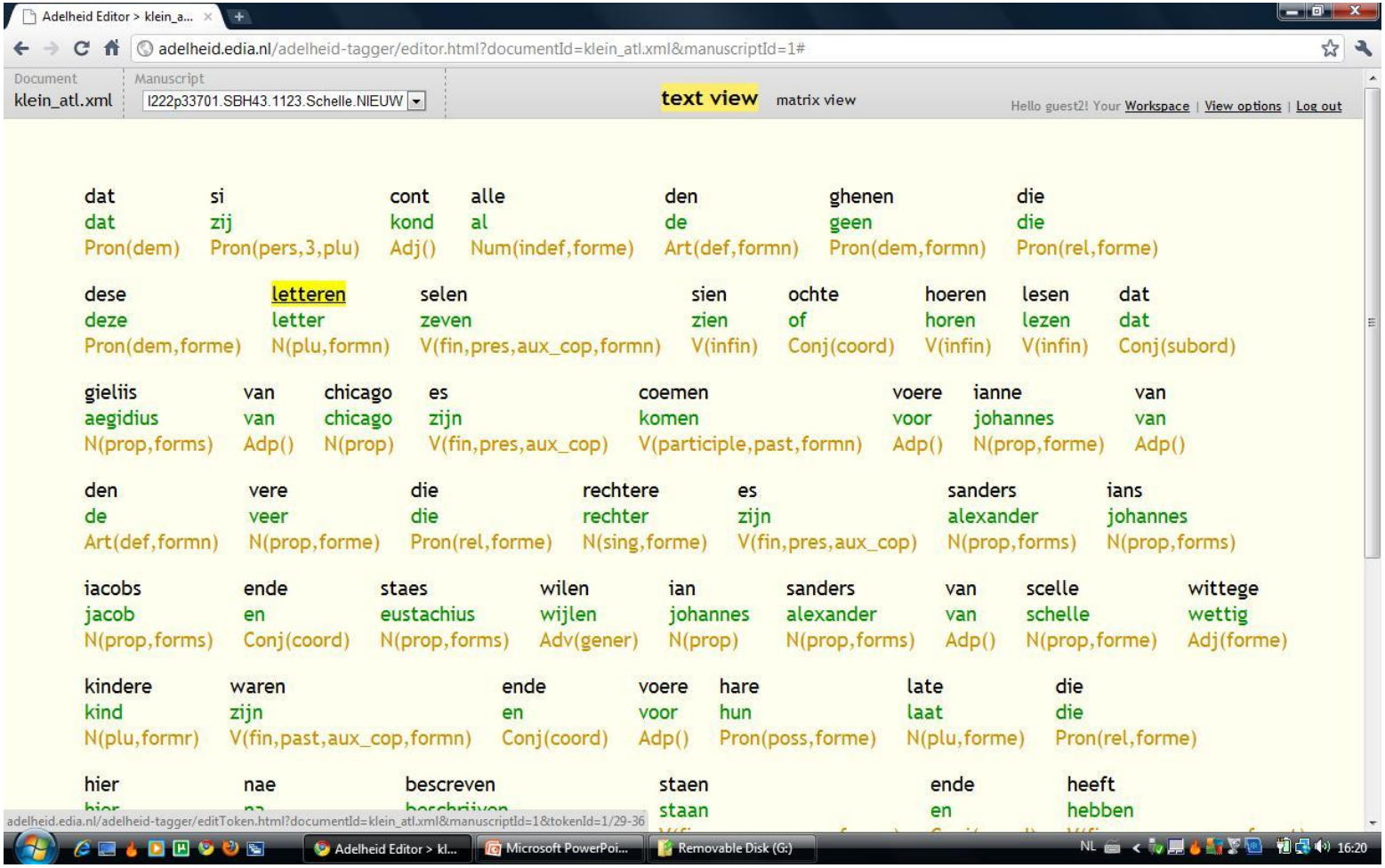


- Selecting manuscripts for processing



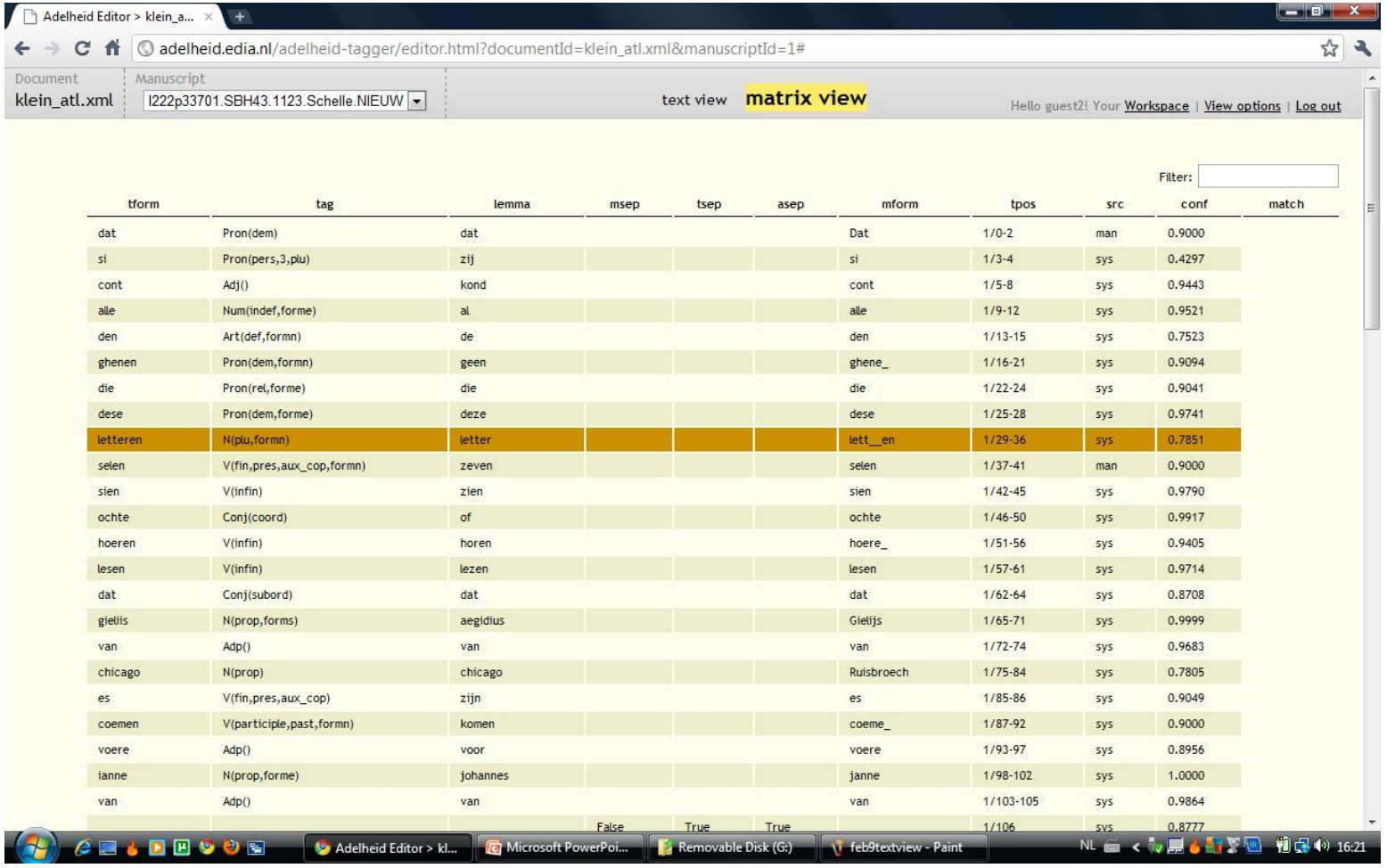
Annotation tool: Functionality

- Seeing tokens, tags and lemmas: Text View



Annotation tool: Functionality

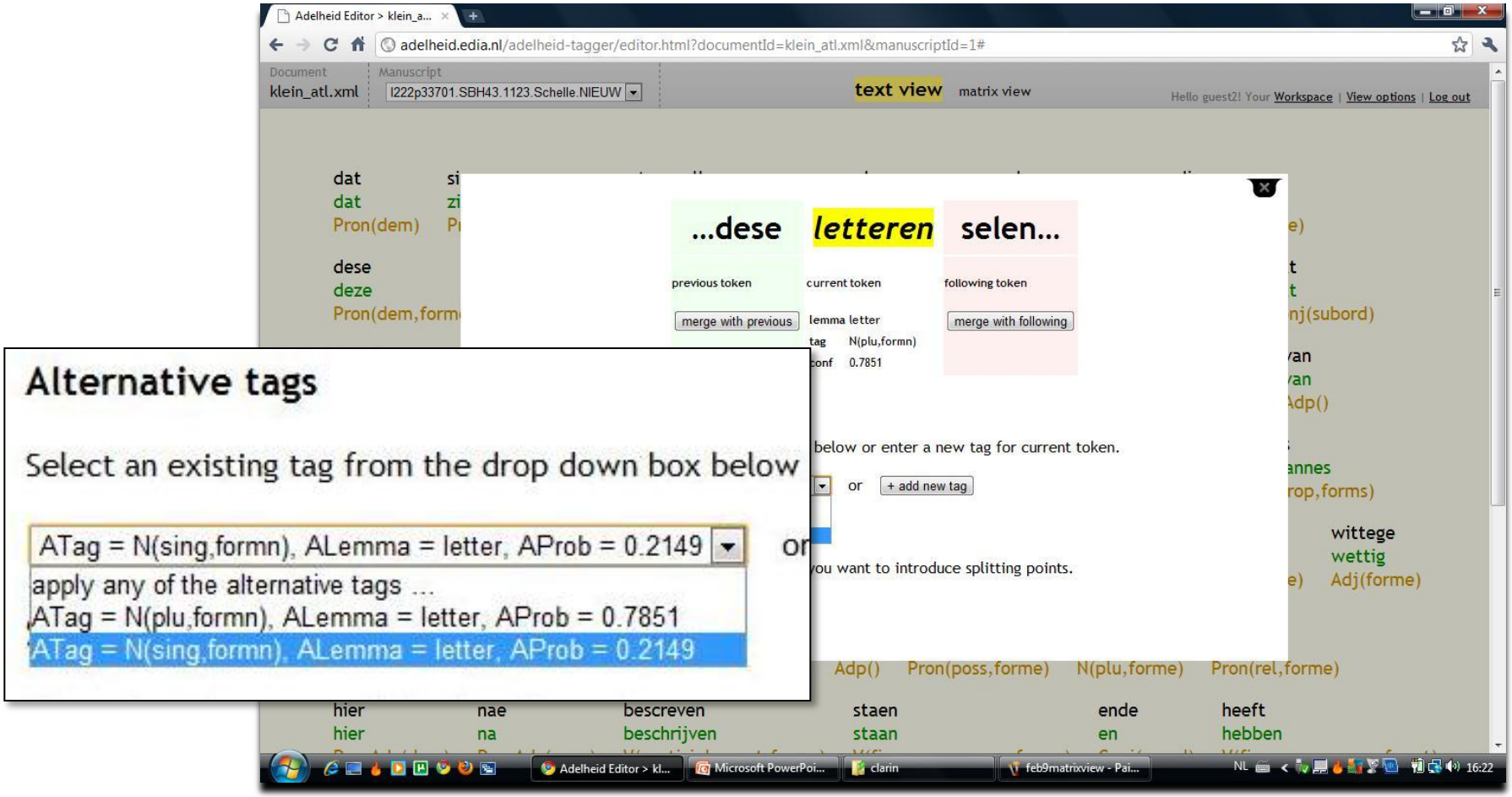
- Seeing tokens, tags and lemmas: Matrix View



tform	tag	lemma	msep	tsep	asep	mform	tpos	src	conf	match
dat	Pron(dem)	dat				Dat	1/0-2	man	0.9000	
si	Pron(pers,3,plu)	zij				si	1/3-4	sys	0.4297	
cont	Adj()	kond				cont	1/5-8	sys	0.9443	
alle	Num(indef,forme)	al				alle	1/9-12	sys	0.9521	
den	Art(def,formn)	de				den	1/13-15	sys	0.7523	
ghenen	Pron(dem,formn)	geen				ghene_	1/16-21	sys	0.9094	
die	Pron(rel,forme)	die				die	1/22-24	sys	0.9041	
dese	Pron(dem,forme)	deze				dese	1/25-28	sys	0.9741	
letteren	N(plu,formn)	letter				lett__en	1/29-36	sys	0.7851	
selen	V(fin,pres,aux_cop,formn)	zeven				selen	1/37-41	man	0.9000	
sien	V(infin)	zien				sien	1/42-45	sys	0.9790	
ochte	Conj(coord)	of				ochte	1/46-50	sys	0.9917	
hoeren	V(infin)	horen				hoere_	1/51-56	sys	0.9405	
lesen	V(infin)	lezen				lesen	1/57-61	sys	0.9714	
dat	Conj(subord)	dat				dat	1/62-64	sys	0.8708	
gielis	N(prop,forms)	aegidius				Gielijs	1/65-71	sys	0.9999	
van	Adp()	van				van	1/72-74	sys	0.9683	
chicago	N(prop)	chicago				Ruisbroech	1/75-84	sys	0.7805	
es	V(fin,pres,aux_cop)	zijn				es	1/85-86	sys	0.9049	
coemen	V(participle,past,formn)	komen				coeme_	1/87-92	sys	0.9000	
voere	Adp()	voor				voere	1/93-97	sys	0.8956	
ianne	N(prop,forme)	johannes				janne	1/98-102	sys	1.0000	
van	Adp()	van				van	1/103-105	sys	0.9864	
			False	True	True		1/106	sys	0.8777	

Annotation tool: Functionality

- Choosing alternative suggested annotation



Annotation tool: Functionality

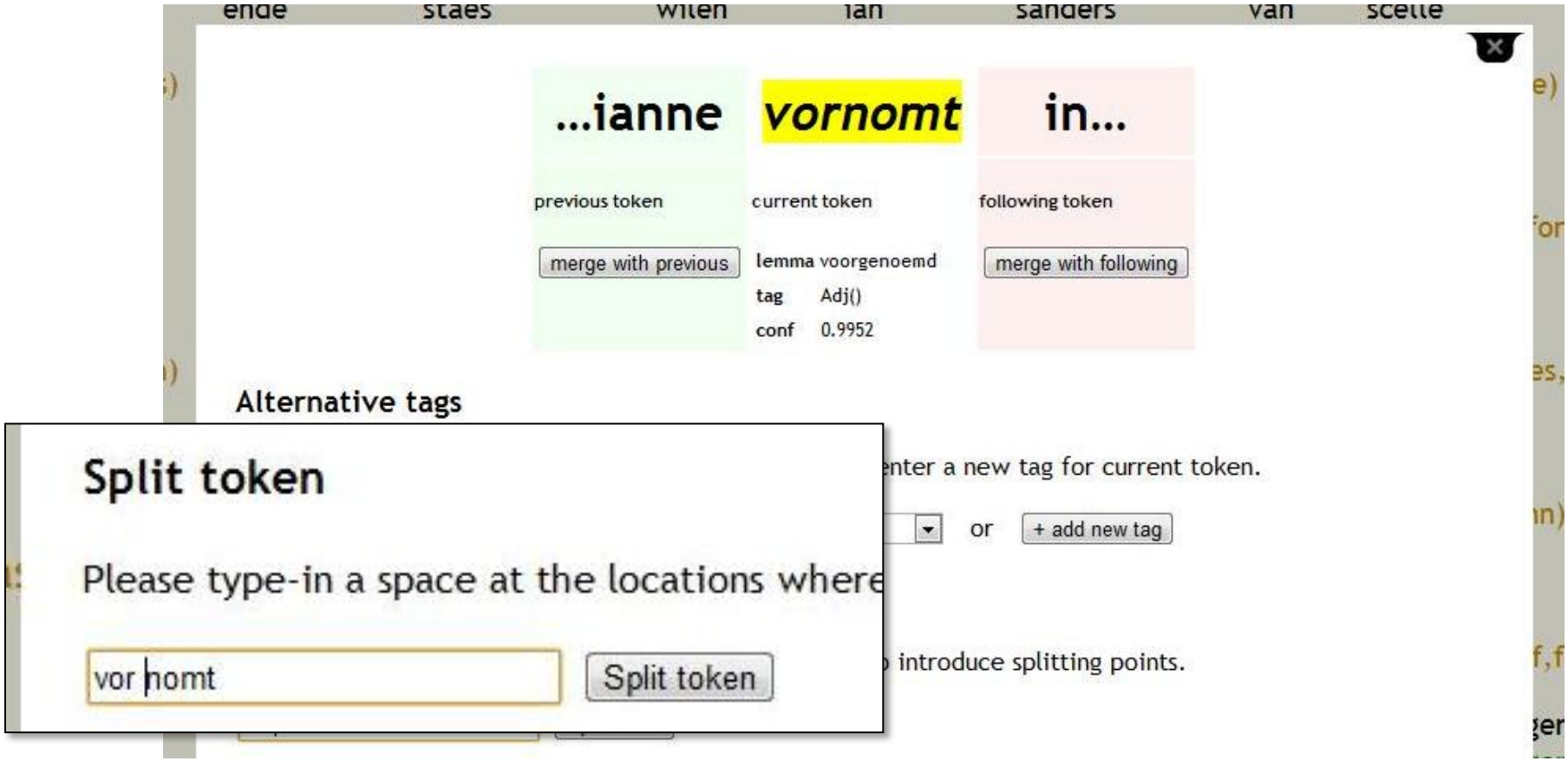
- Entering annotation not suggested by system

Annotation tool: Functionality

- Merging two (or more) tokens

Annotation tool: Functionality

- Splitting tokens into two (or more) parts



Annotation tool: Functionality

- Search for systematic corrections

Potential uses: Linguists

- Lemmas provide handle to access words
  - E.g. to study development of the Dutch vocabulary
  - E.g. to study dialectal variation
- Tags provide handle on syntax
  - E.g. to study verb clusters
  - E.g. to study case system
  - E.g. to study cliticization
- Both allow further linguistic analysis
  - E.g. full syntactic parsing: INPOLDER

Potential uses: Historians, ...

- Lemmas provide handle on the text
  - Especially also where it concerns names
  - Helping to find all relevant passages
  - Helping to read these passages
  - And in the future open these texts up for full NLP/IR processing

Potential uses: The Future
Other periods? Other text types?

- Adelheid also provides an architecture
- Extendable to other than 14th century Dutch

- Customizability in current version
  - By user: Additional lexicons
  - By other developers: Alternative tokenization and/or tag assignment
  - (In cooperation: Alternative tag disambiguation)

Questions?