Adelheid: Tagging and Lemmatizing Historical Dutch Texts through the Clarin Infrastructure

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

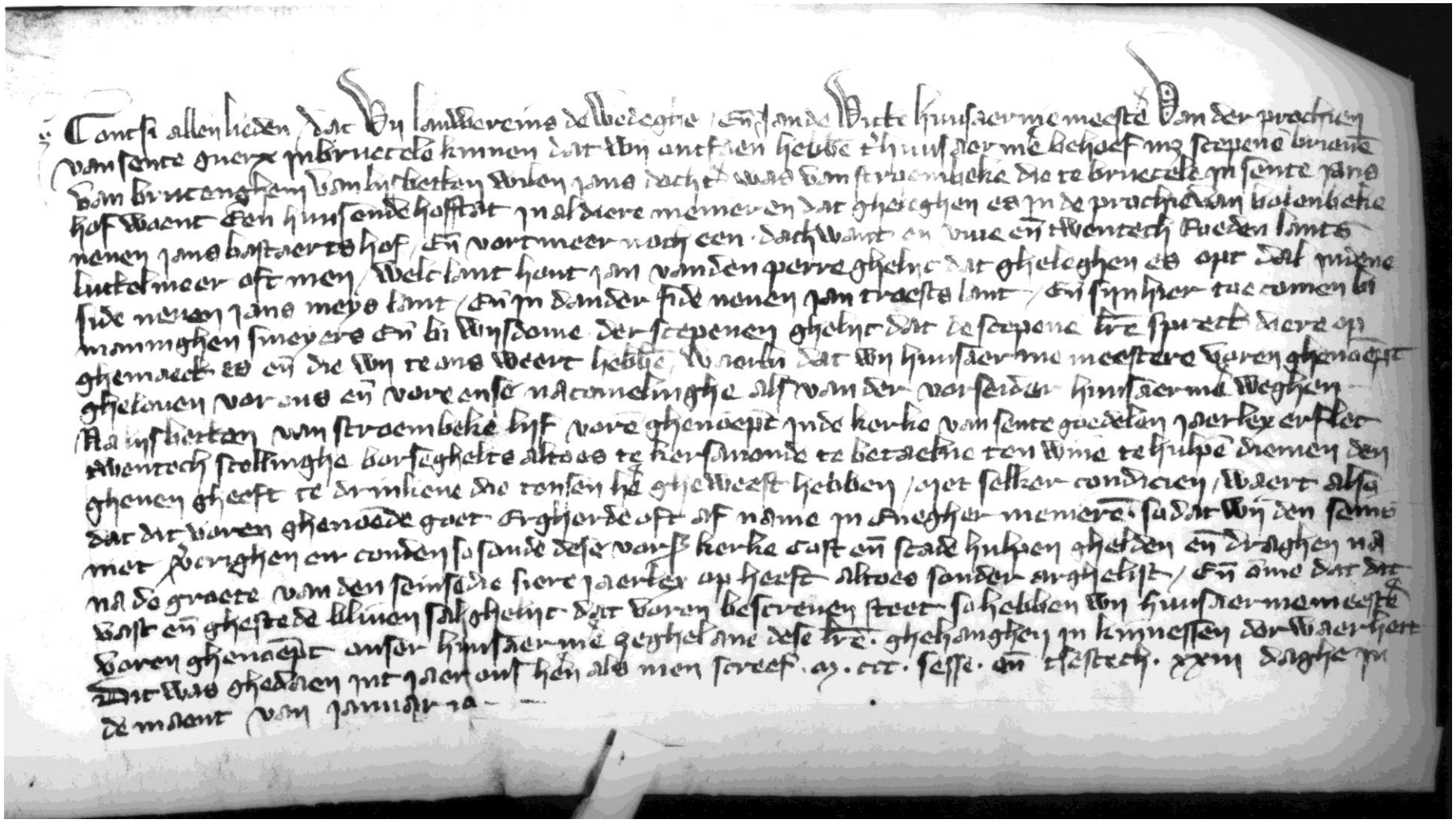
Topic

- Clarin NL demonstrator project: Adelheid
- Tagging and lemmatizing historical text

Structure

- Task
- The basis of the Adelheid system
- Adelheid through Clarin
- Annotation tool

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
- Focus for today: spelling variation

Adelheid before Clarin

Orthographic Variation

Adverb *gelijk*

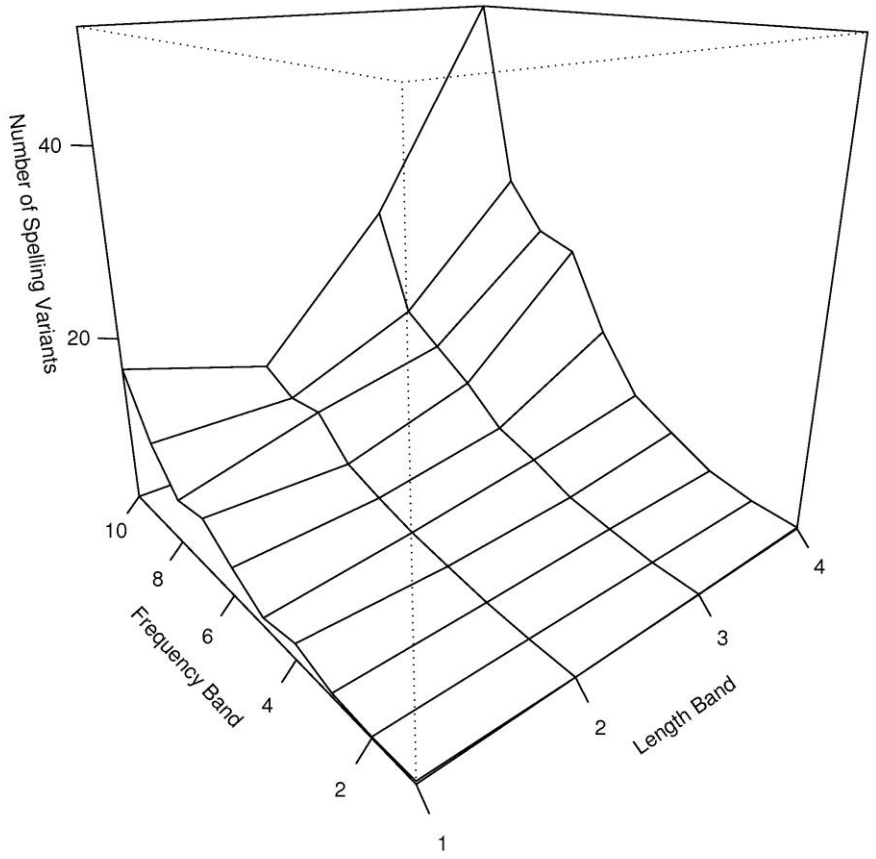
ghelijc (373) gheliic (86) gelijc (64) ghelike (54) ghelijch (33) ghelyc (19) gheliich (10) gelijch (9) gelike (9) gheliken (9) euengheliken (4) gelyck (4) ghelich (4) ghelic (3) gelic (2) geliic (2) ghelijck (2) dinghelike (1) euenghelike (1) evenghelike (1) evenghelyc (1) ghelijcke (1) ghelijct (1) gheljjch (1) ghelljc (1) ghlijc (1) gilycs (1) like (1)

Proper name *Gerard*

gheriit (121) gherijt (111) gherart (84) gheret (70) gherit (70) gherijd (58) gerart (56) gheert (55) gheriid (54) gheraerd (47) gherd (47) ghert (46) ghered (37) gheraet (24) gheeraed (19) gherard (18) gheraert (16) gert (12) gerat (11) gerit (10) gheerd (10) geraert (8) gerd (8) gheryt (8) gerijt (7) gheeraerd (7) gheerard (7) gherret (7) geerd (6) gherid (6) geraet (5) geret (5) gheraed (5) geert (4) gerijd (4) gheeraert (4) gheredt (4) gheryd (4) gherrijd (3) ghierart (3) gered (2) gereet (2) geyrart (2) gheeraerdt (2) gheeraet (2) gheerit (2) gher (2) gherairt (2) gherardt (2) gherat (2) gherrijt (2) gherut (2) garret (1) ger (1) geraed (1) gerairt (1) gerard (1) gerid (1) geriit (1) geryt (1) gheerlec (1) gheraird (1) gherrid (1) gherud (1) gherydijn (1) gierkijn (1)

Orthographic Variation

Number of
Variants
vs
Length/Frequency



Solutions for Orthographic Variation

Phase 1: Determine character-level variation cost

- Based on form pairs with only one difference
- Levenshtein cost reduced every time observed (1→0)

	Substitution
e ↔ i	.050
i ↔ y	.086
d ↔ t	.235
c ↔ k	.598
b ↔ p	.969
b ↔ z	.997

	Insertion	**Doubling**
e	.017	.004
h	.085	.085
n	.459	.339
r	.769	.661
m	.956	.849
b	.979	.949

Solutions for Orthographic Variation

Phase 2: Build token-variation grid

- with alignment software aligning multiple forms

g	h	e	b	o	e	r	t	e	14
g		e	b	v	e	r	d	e	11
			b	o	e	r	t	e	7
g	h	e	b	o		r	t	e	3
g		e	b	o		r	t	e	3
g		e	b	v		r	t	e	2
g	h	e	b	v	e	r	t	e	1
g	h	e	b	v	e	r	d	e	1
g		e	b	o	i	r	d	e	1
			b	o		r	t	e	1

- later on, will generate combination variants

Solutions for Orthographic Variation

Phase 3: Derive rules which appear to be more general

- Character grid positions: focus char + left and right context
- Variant for position character seen for many lemmas

Substitution			**Deletion**			**Insertion**	
#s__<__e → c	.73		eg__h__#	.90		#g__ __el → h	.76
t__z__# → s	.71		f__f__#	.87		#g__ __es → h	.75
an__d__# → t	.70		t__h_en	.85		g__ __el → h	.71
l__d__# → t	.70		en__n__e#	.78		dag__ __e → h	.70
s__<__o → c	.68		ike__n__#	.78		#g__ __e → h	.66

Solutions for Orthographic Variation

Phase 4: Generate variants

- Start from observed variant
- Allow up to √tokenlength changes
- First observed variants for token, then rule-based variants
- Keep change probability over threshold
- Filter out variants with impossible trigrams (and suffix 4-grams)
- Reassign counts, based on C(observed) and P(change)
- Expands number of lexicon tokens from ~50K to ~1.3M

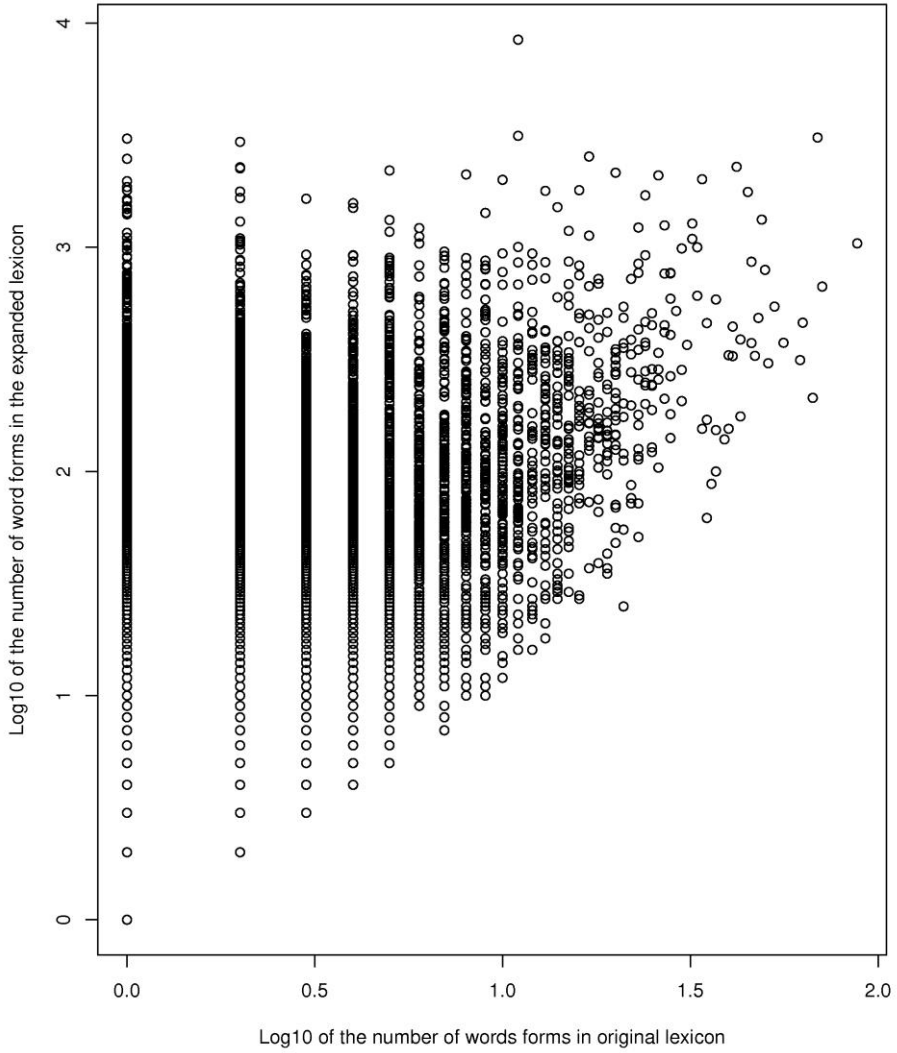
Token	Observed	Generated
gheboerte	14	7.73
gebverte	2	1.94
ghebvrde	0	0.41
boerde	0	0.14
heboeirte	0	0.0005

Solutions for Orthographic Variation

Expansion:

Word forms

per TLP

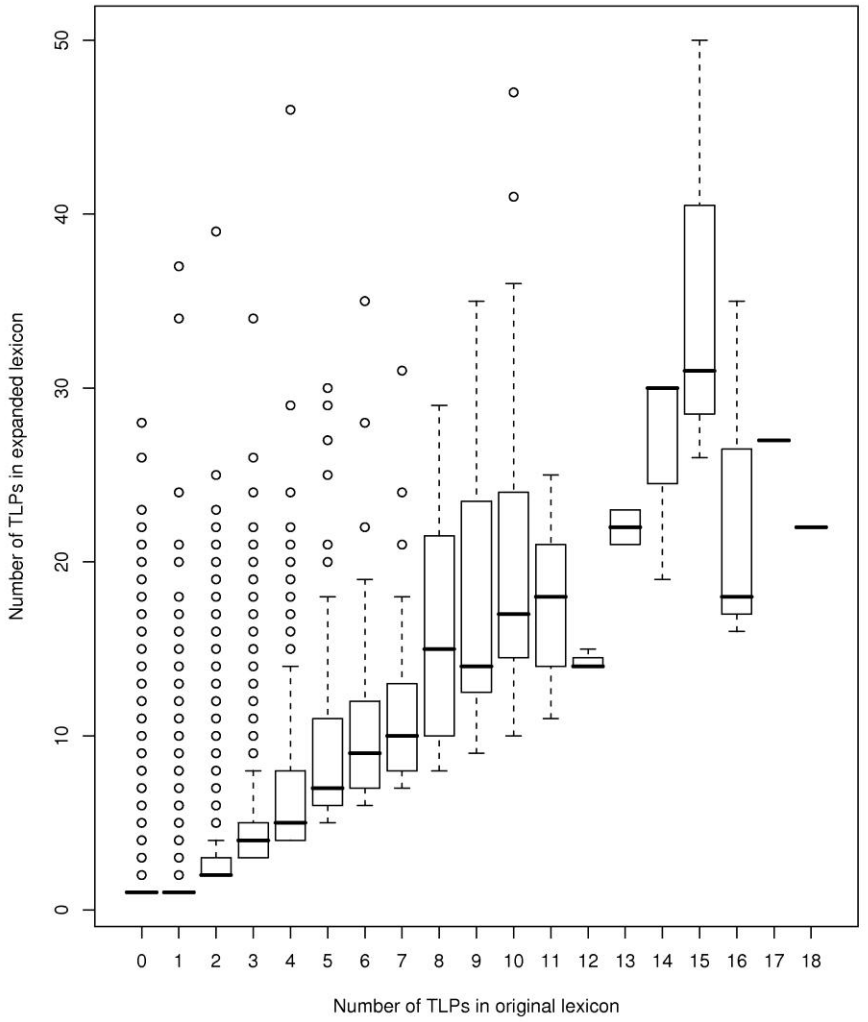


Solutions for Orthographic Variation

Ambiguity:

TLPs

per word form



Solutions for Orthographic Variation

Phase 5: Try to dynamically adapt lexicon

- Token-tag combinations from unknown token module
  - ~ 5K per 80Kw test set
- Find Levenshtein-closest in expanded lexicon
  - With the right tag

E.g. Lemma: *voorgezegd*

- 214 forms observed
- Expanded to 1992 forms
- Identified two further in test: *voerseiit* and *voregeseds*

Evaluation Results

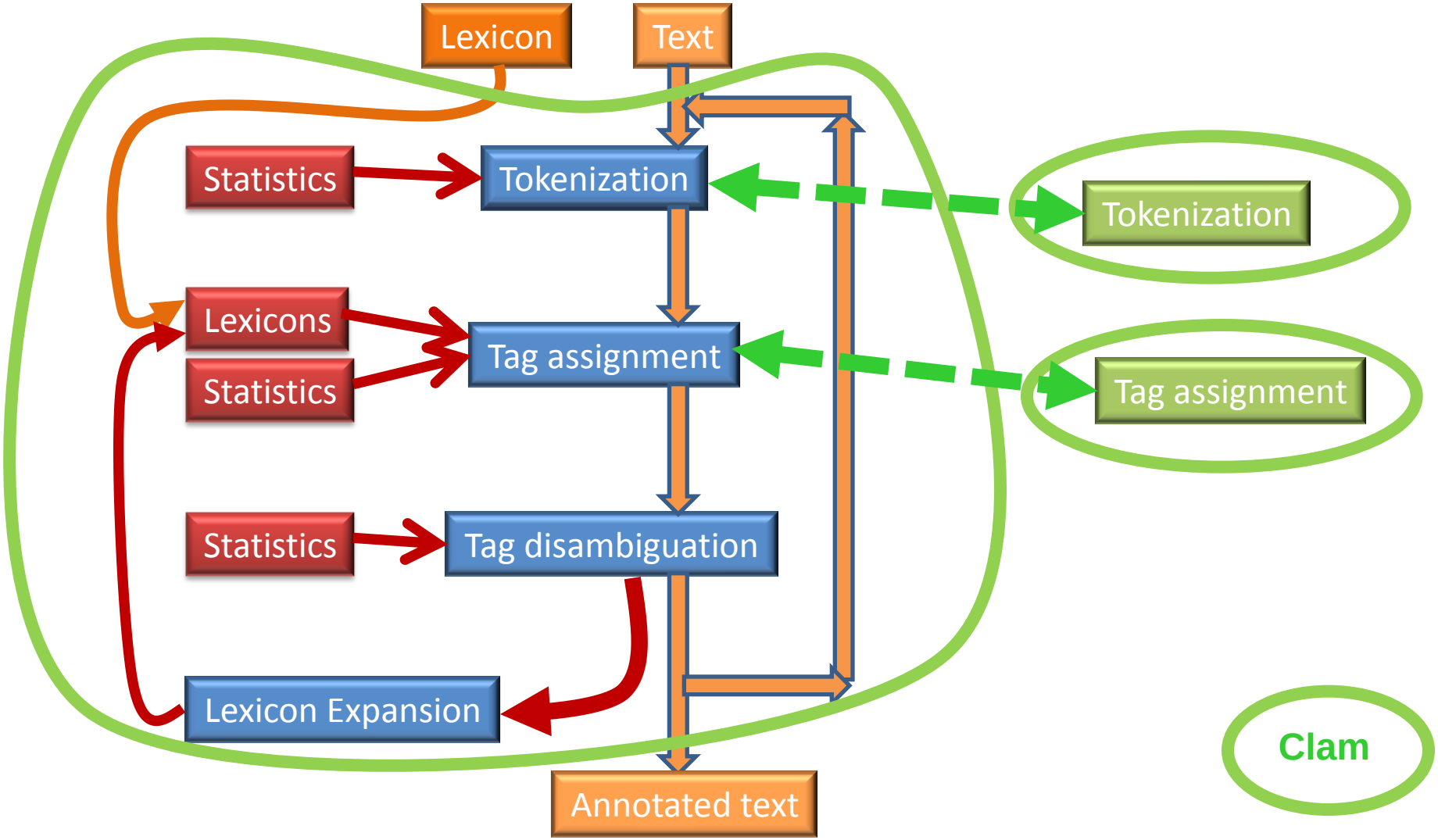
Lexicon improvement

- Recall for re-estimated tokens
  - In 10-fold cross-validation on van Reenen - Mulder

	Tag	Lemma
Known forms only	94.91%	93.11%
Expanded lexicon	94.94%	93.96%
With test token adaptation	94.97%	94.88%

Adelheid in Clarin

Adelheid through Clarin



Adelheid through Clarin

- System now available
  - Through Clarin infrastructure
  - More efficient
  - Using XML data formats
  - With user manuals, incl. Demonstration scenarios
- Interface: Clam
  - http://lux17.mpi.nl/adelheid
  - http://wwwlands2.let.kun.nl/adelheid/
  - Please do not use until release announced

Visualisation and Annotation in Clarin

Annotation tool: Why?

- Example of the (XML) 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>
```

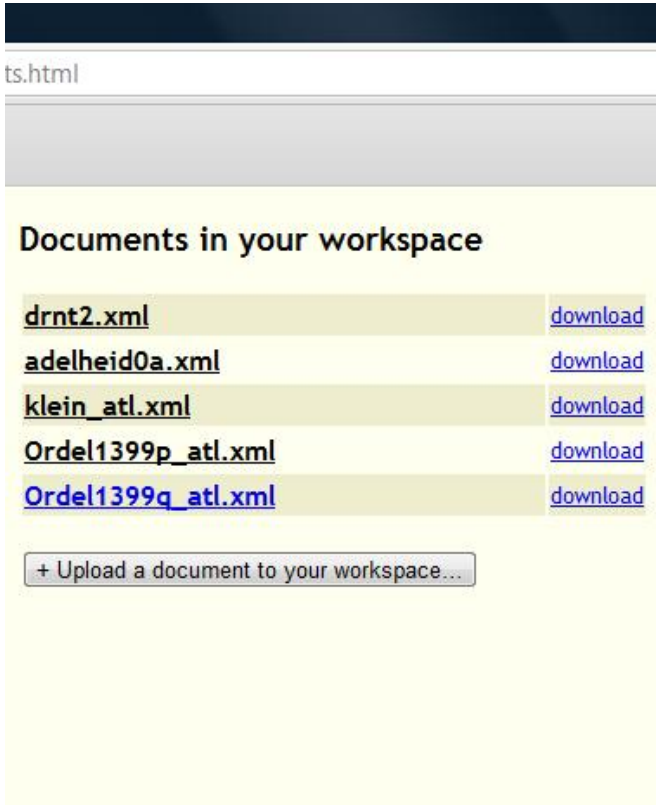
Annotation tool

- Dedicated tool for
  - Visualization
  - Adjusting annotation
  - Details below
- Tool built by Edia in Amsterdam
- Also accessible through Clarin infrastructure
  - http://lux17.let.kun.nl/adelheidanntool
  - http://adelheid.edia.nl/adelheid-tagger
  - Please do not use until release announced

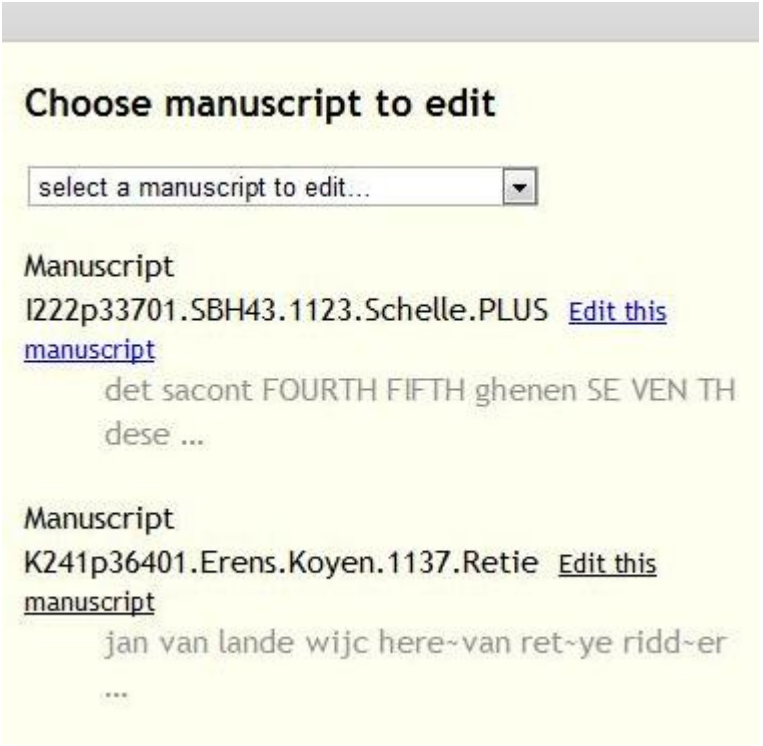
In case life demo not available:
Screenshots
(with some silly debugging artifacts)

Annotation tool: Functionality

- Up- and downloading annotation files

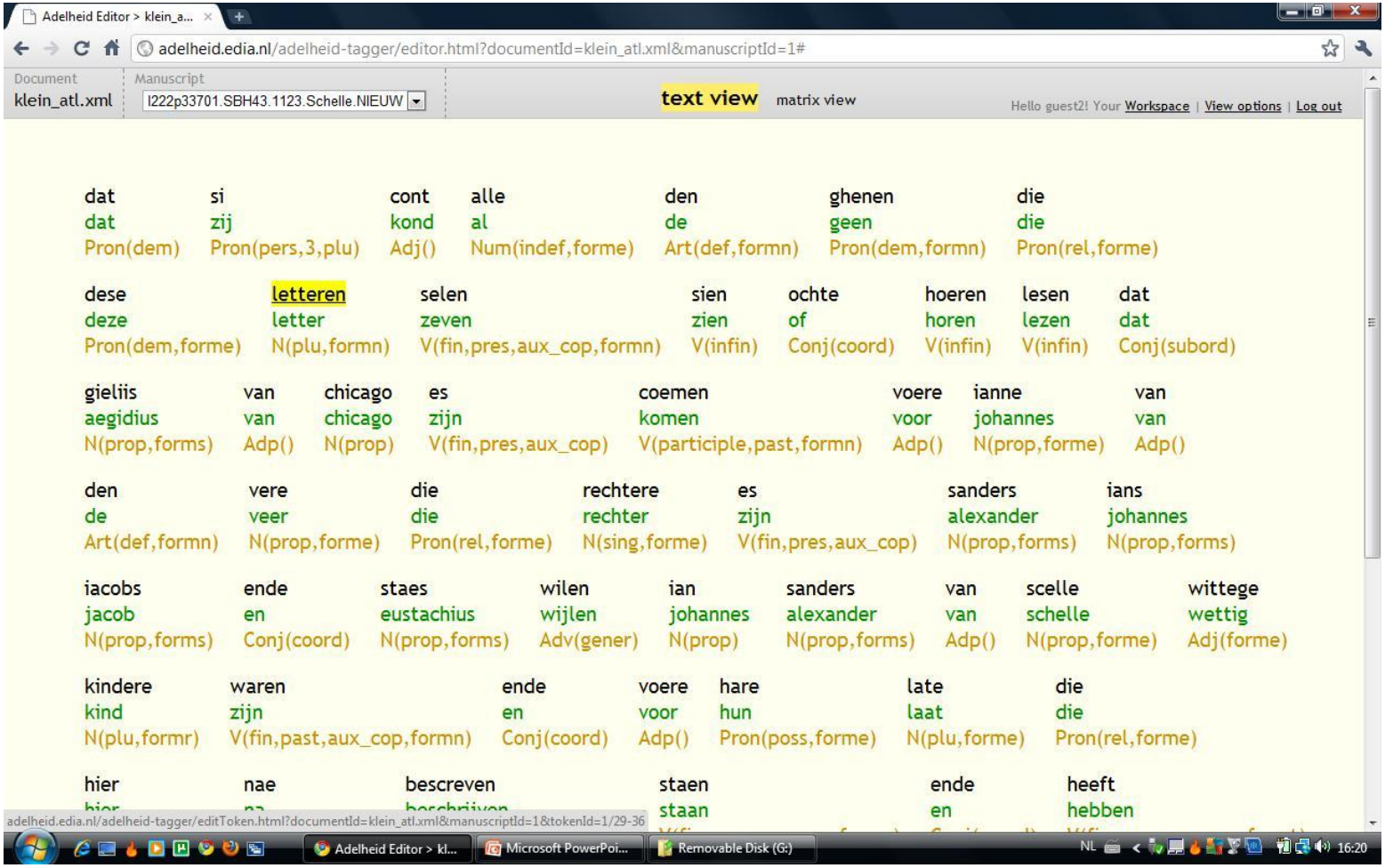


- Selecting manuscripts for processing



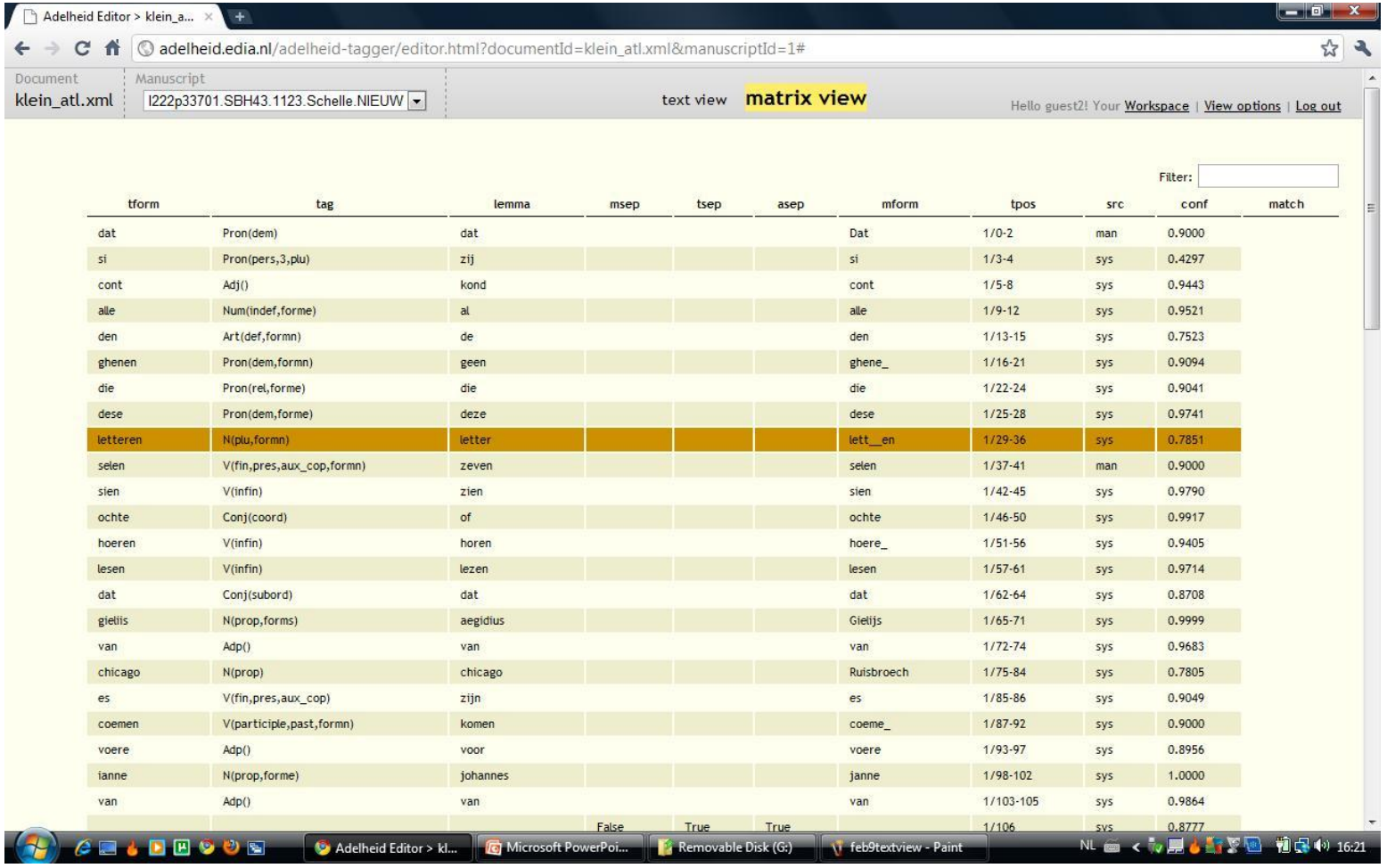
Annotation tool: Functionality

- Seeing tokens, tags and lemmas: Text View



Annotation tool: Functionality

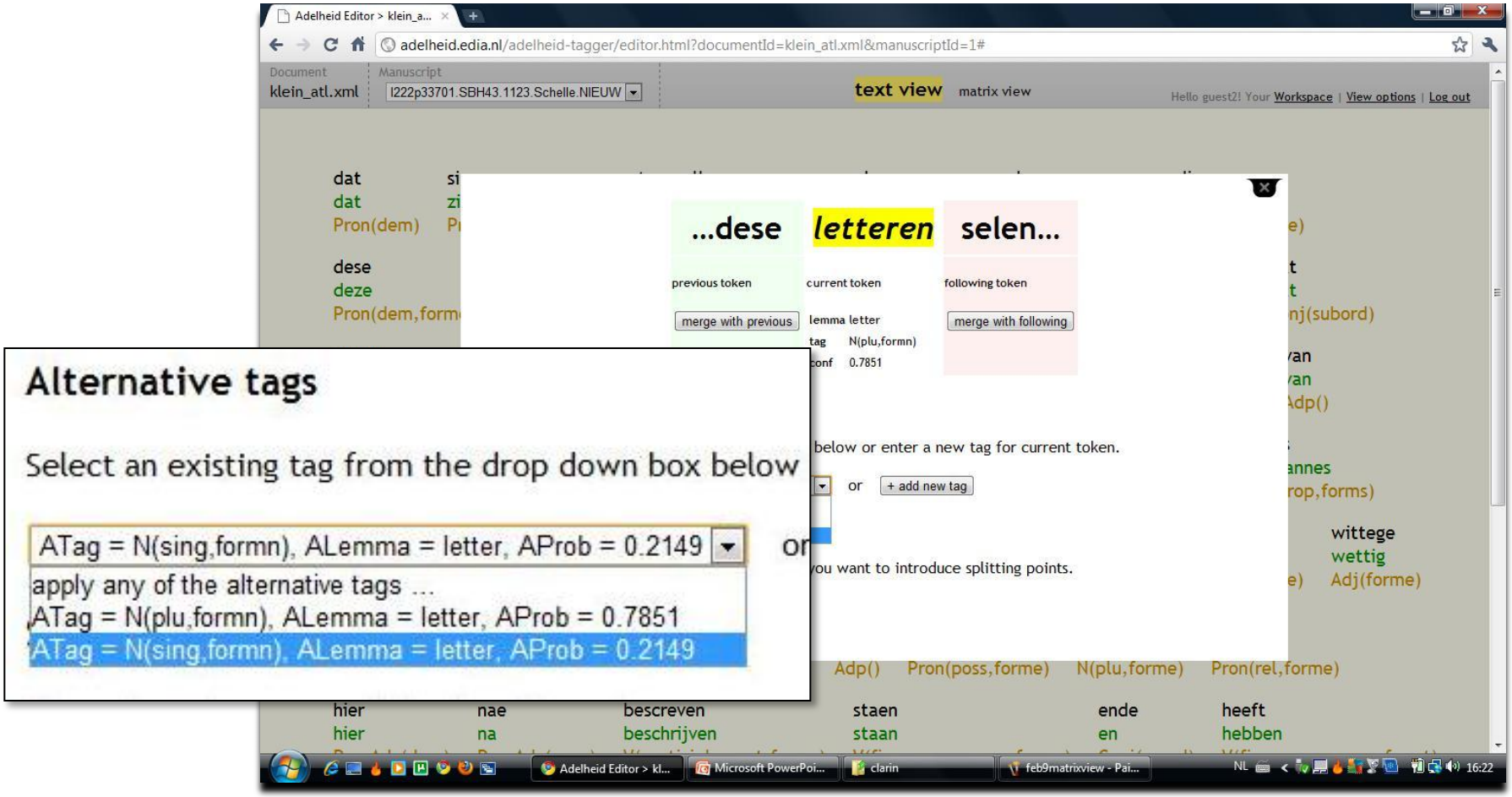
- Seeing tokens, tags and lemmas: Matrix View



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

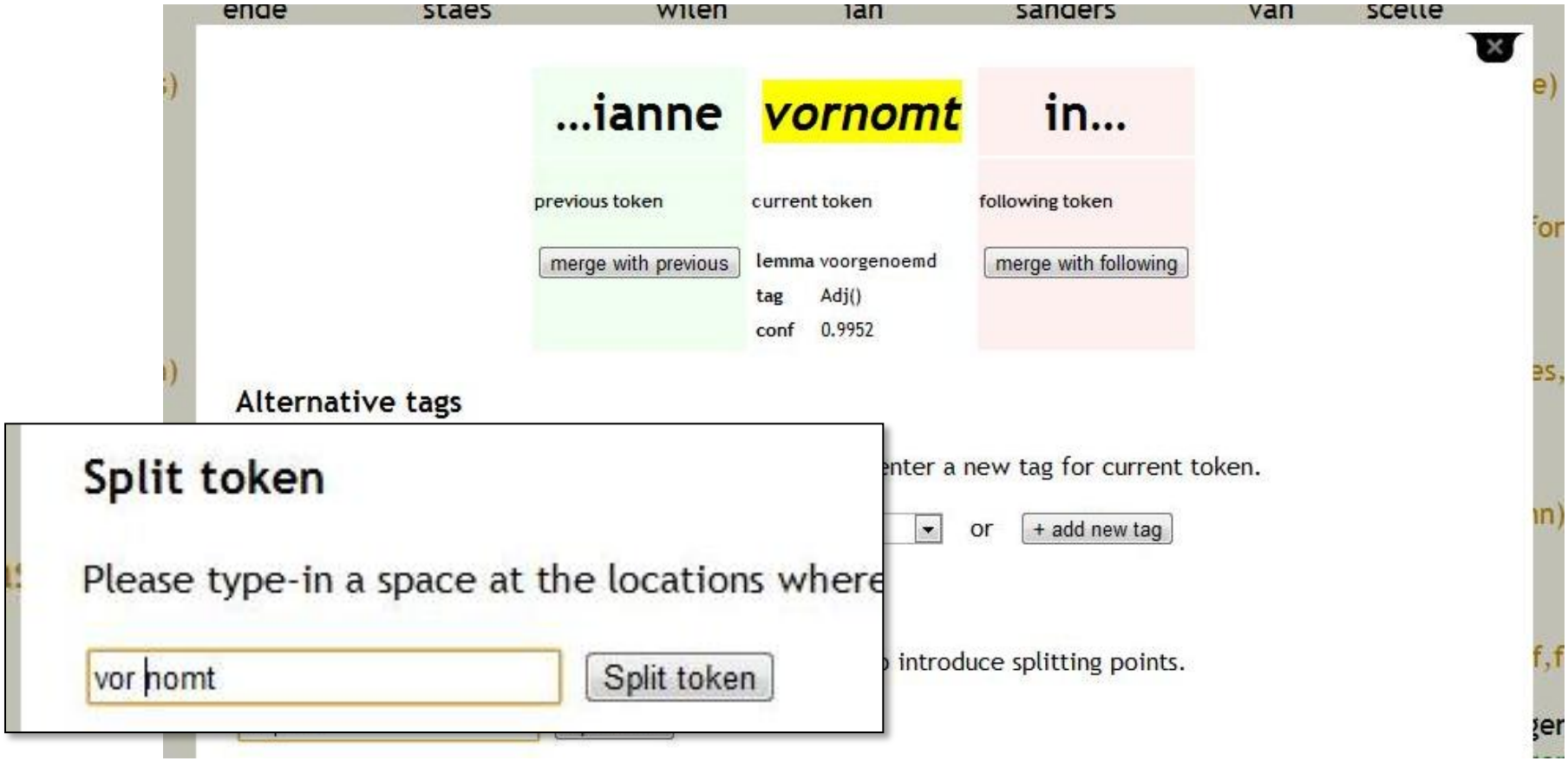
- Entering annotation not suggested by system

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

Questions?

Major and Minor Experiences

- From experimental to available is a lot of work
  - Reimplementing
  - Speeding up
  - Making available via web (Clam very nice!)
  - Modernizing interfaces (XML)
  - Documenting
  - Maintaining

Major and Minor Experiences

- External parties can help
  - More expertise, in our case on Java/XML
  - But can be expensive
  - And who will do maintenance?
- Using modern methods is vital
  - For accessibility, in our case XML
  - But also has its problems
    - Need to learn properly
    - Special problem: character entities vs Unicode vs DTD