Nederlab pilot project *Language Dynamics*

Marijn Schraagen
Marjo van Koppen
Feike Dietz

Utrecht University

June 6, 2018

Universiteit Utrecht

Nederlab pilot project

- Corpus linguistics for 17th century Dutch
- Full manual annotation on POS and sociolinguistic variables
  - automatic POS annotation not sufficiently accurate
  - automatic sociolinguistic annotation not available
- Corpus: letters of P.C. Hooft from DBNL
- Incorporate into Nederlab
  - use existing search interface, CQP, visualizations
  - available for other Nederlab/CLARIN users
- Possible extension: create treebank to investigate more complex linguistic phenomena

Nederlab integration

- Add `lemma`, `pos`, `feat` tags to existing FoLiA <w> nodes
- ```
  <w xml:id="hoof001hwva02_01.TEI.2.text.body.div.p.28035.s.2.w.10" class="WORD">
    <t>vinde</t>
    <pos class="N(soort,ev,basis,onz,stan)" confidence="0.3" head="N" textclass="nederlab-orig">
    <pos class="WW(pv,tgw,+lex,ev,1,conj,+forme)" confidence="0" head="WW" textclass="gustave">
  </w>
  ```
- Search interface (mock-up):

De briefwisseling van P.C. Hooft. Deel 1
datering: 1976
auteur: **P.C. Hooft (Amsterdam (Noord-Holland), 1581-Den Haag (Zuid-Holland), 1647) H.W. van Tricht (Arnhem (Gelderland), 1897-Velp (Gelderland), 1982)**
genre: fictie, non-fictie, proza, egodocumenten, brieven
collectie: DBNL
aantal hits: 1

	dat	ick	mij	in	staet	vinde	van	niet	te	konnen
Annotatie (standaard):	dat VNW aanw pron stan vol 3o ev	ick ADJ vrij basis zonder	mij VNW pr pron obl vol 1 ev	in SPEC vreemd	staet SPEC vreemd	vinde SPEC vreemd	van VZ init	niet BW	te VZ init	konnen WW pv tgw mv
Annotatie (alternatief): *meer informatie*		ik VNW pers pron nomin vol 1			staat N soort ev basis zijd stan	vinden WW pv tgw ev				
  ```

Integration issues

- Separate corpus or additional annotation layer
  - Limited search results because of small corpus
  - Possible confusion over tagset extensions
- Representing token alignment in FoLiA and search results

Aenden		Advocaet	van	Hollandt	...	verschej-	den	onwaerdicheden	zijn	toegedreven
Aen	den	Advocaet	van	Hollandt	...	verschejden		onwaerdicheden	zijn	toegedreven

- How to integrate sociolinguistic annotation?
- Data provided to Nederlab, but not yet available online

Part-of-Speech features

- Features added to CGN tagset to study 17th century Dutch
  - particle: een' masker **te** doen draeghen **om** wit **te** gaêren
  - nominative (position): dat zij$_{\textbf{nom}}$ ijemandt$_{\textbf{nonnom}}$ gebrujke
  - suffix forms: 11en$_{\textbf{form-n}}$ Aprilis$_{\textbf{form-latin}}$ laestleden
  - lexical/nonlexical, imperative verbs: evenwel laet$_{\textbf{nonlex,imp}}$ mij voorstaen
  - negation clitic: ick **en** ben geen schrijver
  - ...
- Not indexed by Nederlab, not valid in CGN, unclear for external users

Tokenization in FoLiA: split

- Annotation guidelines: split morphological compounds and syntactic compounds in two tokens
  - *uit+de*: heb ick **wter** hechtenisse ontslaeghen
  - *zal men*: zoo **zalmen** moeten rekenen hoe veele tienduimde steenen in een vloer van 30 voeten lang, 12 breedt
- Representation using non-standard FoLiA
- ```
  <w xml:id="p.3.s.15.w.30" class="WORD">
    <t>zalmen</t>
    <t textclass="combination-token">zal</t>
    <lemma class="zullen"/>
    <pos class="V(fin,+nonlex,sg,3,pres)" head="V"/>
    <t_2 textclass="combination-token">men</t_2>
    <lemma_2 class="men"/>
    <pos_2 class="PRON(indef,+nom)" head="PRON"/>
  </w>
  ```


  ```

Tokenization in FoLiA: merge

- Annotation guidelines: combine compound elements, hyphenated tokens, fixed expressions
  - op dat → opdat, verschej- den → verschejden, U E. → UE
- Representation using non-standard FoLiA
- ```
<w xml:id="p.28035.s.5.w.52" class="WORD">
  <t>op</t>
  <!-- lemma, features, POS, etc.-->
  <alt xml:id="p.28035.s.5.w.52.merge.1" src="gustave-merge">
    <merge src="p.28035.s.5.w.52" dest="p.28035.s.5.w.53"/>
    <merge src="p.28035.s.5.w.53" dest="p.28035.s.5.w.53"/>
  </alt>
</w>
<w xml:id="p.28035.s.5.w.53" class="WORD">
  <t>dat</t>
  <!-- lemma, features, POS, etc.-->
  <alt xml:id="p.28035.s.5.w.53.merge.1" src="gustave-merge">
    <merge src="p.28035.s.5.w.52" dest="p.28035.s.5.w.53"/>
    <merge src="p.28035.s.5.w.53" dest="p.28035.s.5.w.53"/>
  </alt>
</w>
```


```

Mê Joffre

De prujmen beginnen al teffens op een bodt te rijpen, en te roepen Tesseltje, Tesseltjes mondtje. Etlijke deuntjes van Belusar en andre roepen daer tegen aen, Tesseltje, Tesseltjes keeltje; daer zij gejrne van gezongen waeren, ende wenschten wel dat U.E. Joffre Francisca te hulpe meêbraght. Wat ik haer zeg, Tesseltje suft; Tesseltjen heeft pen nocht inkt om een briefken te beantwoorden; zij neemen 't niet aen, ende willen dat ik U.E. ujt den droom wekke. Op, op dan.

10 Rozemondt, hoor je speelen nocht zingen?

Wij verwachten U.E. op 't spoedighste, met U.E. dochter, ende Joffre Duart met haer' E. man: maer een briefken voor ujt, om wat gissings te moghen maeken. Ondertussen zullen wij in den windt zien, ende happen nae den geenen die van Alkmaer komt en snuffen oft hij nae U.E. adem riekt. Godt beboede U.E. op de *zejze*, en eeuwlijk in genaede, met alle die haer lief zijn; gelijk van heelen, heeten harte wenscht,

Mê Joffre

U.E.

verplichte, dienstwste

P C Hóóft.

557 / 1219 1 Selecteer

wonder vruchtbare akker des verstands . zo ghij wakker verhoet datter de ikker gheen onkruijd meer in en zaaije (die

Vorige Volgende Gebruik pijltjestoetsen

Combineer *datterde* Splits < > *dat | ter*

	huidig	*controle*	
lemma	dat+er	dat+er	modern alternatief
pos	VNW	N ADJ WW BW VNW LID TW VZ VG SPEC LET	pos/features onduidelijk
features	betr	pers aanw onbep +negonb vrag bez refl betr 1 2 3 ev mv +nom +nonnom +forme +formn +formr +forms	

-

pos	BW	N ADJ WW BW VNW LID TW VZ VG SPEC LET
features	+pers	+aanw +gener +onbep +vrag +pers +vz +neg +negcl +betr +comp +super +prtcl

Universiteit Utrecht

Manual vs. automatic annotation

type	*different*	*equal*	*ratio*
lemma (Frog default)	27,179	70,560	0.28
lemma (Frog modernized)	18,052	53,648	0.25
main POS	23,365	82,830	0.22
features (only Frog/CGN default)	46,725	67,273	0.41
features (extended)	140,888	67,273	0.68

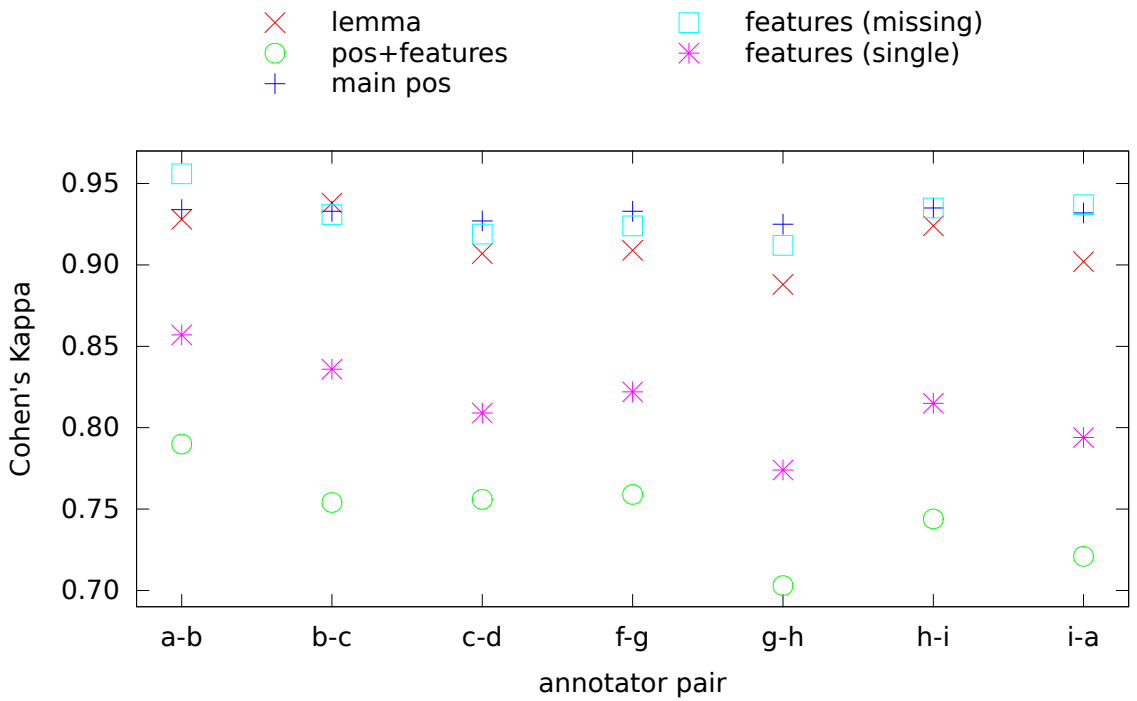
ratio: different / (different+equal)

Application: re-evaluation of sociolinguistic categories

Original framework	
Goal	express thanks, compliment, excuse, ask a favour, ask information, ask advice, admonish, inform, remember, persuade, order, allow, invite
Topic	business, literature, domestic affairs, love, death, news, religion/ethics
Revised framework	
Goal	prompt for action, honour, help, inform, keeping contact, ask for reply
Topic	political work, literary work, current events, social circle

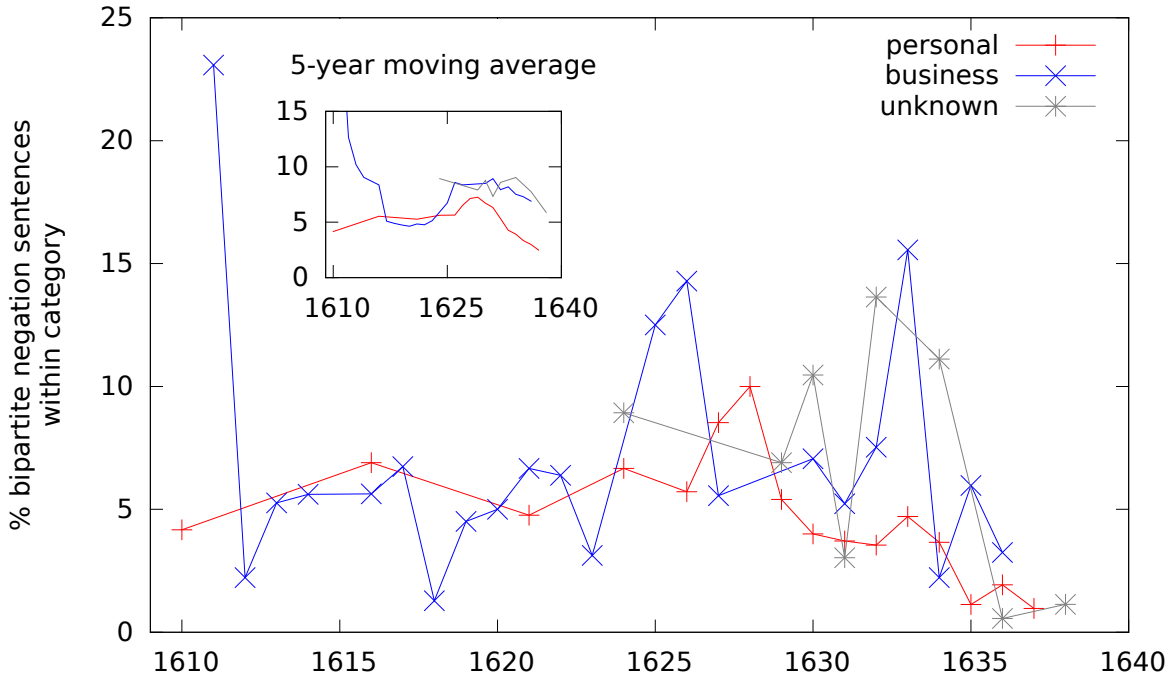
Application in methodology: inter-annotator agreement

token	*annotation 1*	*annotation 2*	*description*
fraejicheden	fraaiheid	fraaiigheid	lemma difference
gedicht	N(nonnom,sg)	V(lex,pp)	past participle vs. noun
ijet	PRON(indef,3,sg,nonnom)	PRON(indef,nonnom)	missing features
kan	V(simple,pres,nonlex,sg,3)	V(simple,pres,nonlex,sg,1)	1/5 features different
etc	SPEC(unclear)	SPEC(foreign)	ambiguous feature



Universiteit Utrecht

Application: analysis of bipartite negation



	single	%	bipartite	%	total
subclause	237	81	75	19	312
subclause	69	83	14	17	83
V1	34	100	0	0	34
V1	4	100	0	0	4

Universiteit Utrecht

Application: verb classification

- Research topic: *have-doubling* in D.V. Coornhert
  - wyder verspreyt ende vermeerdert **heeft ghehad** dan te voren
- How many occurrences? How many verbs in total?
- Train a verb classifier on annotated snippets from P.C. Hooft
  - Long Short Term Memory network using word embeddings
  - dat ick beter meen te **verstaen** als Duitsch weet te schrijven

```
Using TensorFlow backend.
training: 30254 testing: 5338
Found 38398 word vectors.
Train on 30254 samples, validate on 5338 samples
Epoch 1/5
30254/30254 [==============================] - 106s 3ms/step - loss: 0.5453 - acc: 0.7158 - val_loss: 0.4394 - val_acc: 0.7855
Epoch 2/5
30254/30254 [==============================] - 107s 4ms/step - loss: 0.2795 - acc: 0.8830 - val_loss: 0.1830 - val_acc: 0.9311
Epoch 3/5
30254/30254 [==============================] - 104s 3ms/step - loss: 0.1109 - acc: 0.9617 - val_loss: 0.1898 - val_acc: 0.9399
Epoch 4/5
30254/30254 [==============================] - 104s 3ms/step - loss: 0.0655 - acc: 0.9797 - val_loss: 0.2058 - val_acc: 0.9365
Epoch 5/5
30254/30254 [==============================] - 115s 4ms/step - loss: 0.0453 - acc: 0.9859 - val_loss: 0.2356 - val_acc: 0.9416
```