Nederlab

progress and challenges in linguistic enrichment of historical Dutch texts

Katrien Depuydt, Maarten van Gompel, Jesse de Does, Hennie Brugman, Gosse Bouma

Radboud University

Meertens instituut

university of groningen

/instituut voor de Nederlandse taal/

Outline

- Project and corpus
- The current tagging pipeline
  - “Modernization” procedure
  - Linguistic enrichment
  - Metadata
- Evaluation
  - Principles
  - Preparation of ground truth
- Conclusions and future work

Nederlab: Project and corpus

The project

- “NWO groot”, from 1-1-2013 until mid 2018
- Research environment aimed primarily at historians, linguists and literary scholars
- Bring together existing digital text collections
- Diachronic corpus: from year ~800 until present
- Uniform, curated and enriched metadata and texts

http://www.nederlab.nl

Corpus: dataproviders

Koninklijke Bibliotheek

Huygens-ING

Meertensinstituut

Instituut voor de Nederlandse Taal

Taalunie

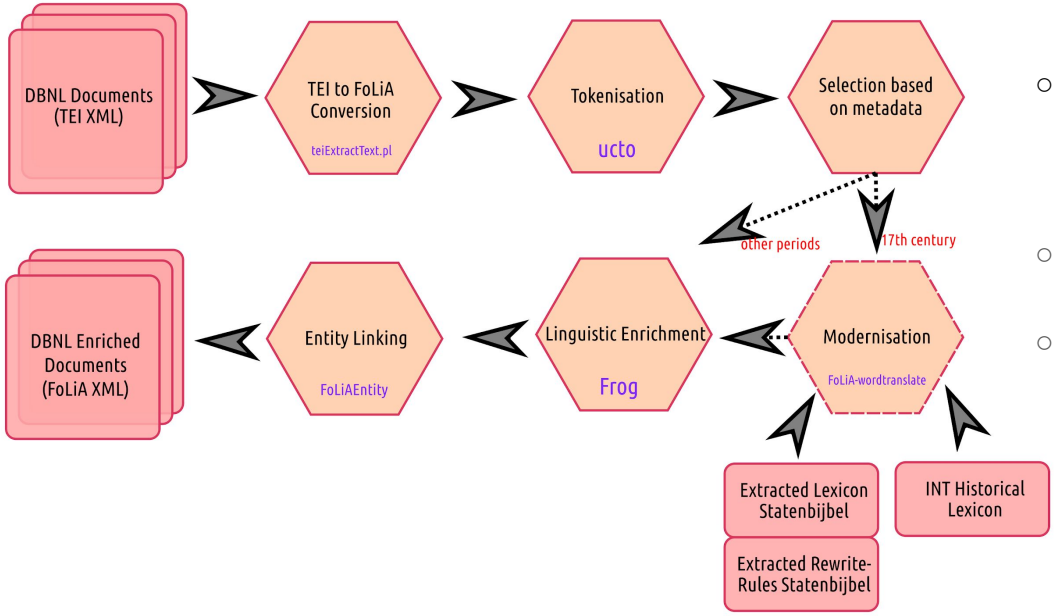
Corpus collections

DBNL XML
KB-newspapers tot 1700-1899
KB-EDBO
Political Mashup
NTU SoNaR
KB-Newspapers 1900-1940
Briefwisseling van Anthonie Heinsius
Dagboeken en aantekeningen van Willem Hendrik de Beaufort
Acta der Particuliere synoden van Zuid-Holland
Notulen van de vergaderingen der Staten van Holland 1620-1640 door N. Stellingwerff en S. Schot
Pieter van Dam's Beschrijvinge van de Oostindische Compagnie
Dagboek Willem de Clercq 1811-1844
Huygens-ING varia (eLaborate)
Clusius Correspondentie
Epistolarium
Corpus Van Reenen-Mulder
Dagboeken van P.J.M. Aalberse
Eindhoven corpus
Van Gogh Nederlandse brieven
Sailing Letters

Linguistic annotation

- Processing depends on
  - language period
  - availability of gold standard material
- Research: Modernisation before linguistic annotation with Frog
- Evaluation
- PoS and modern Dutch equivalent

Current tagging pipeline



PICCL/dbnl.nf pipeline

It is not feasible to produce training data for all varieties in the corpus

- Requires lots of training data
- Sensitive to language variation

So far: two pathways:

- Historical text → "Modernized" text → Modern tagger/lemmatizer (Frog)
- Contemporary only (Frog)

Use **period metadata** to determine what path to follow

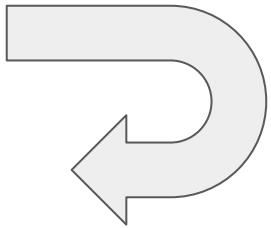
Corpus delivered in **FoLiA** XML (huge!!), to be indexed for actual use in the Nederlab portal using MTAS' Indexer

Implemented as part of PICCL (https://github.com/LanguageMachines/PICCL), powered by Nextflow (https://nextflow.io)

“Modernization”

What does this actually mean??

- “Translation” to modern Dutch (MT)
  - wyf → vrouw
  - wyven → vrouwen
- “Transverbation”
  - wyf → wijf
  - wyven → wijven
- “Modern lemmatization”
  - wyf → wijf N(ev)
  - wyven → wijf N(mv)



Modernisation: Current procedure

- Modernisation of 17th century Dutch and 18th century Dutch
- Word-by-word ‘translation’ trained on:
  - Extracted lexicon from Statenbijbels (parallel corpus 1637 vs 1888)
  - Extracted substitution rules
  - Word-by-word translations are inherently limited !!
    - Can’t deal with contractions, split words, etc..
- Tjong Kim Sang (2016); *Improving Part-of-Speech Tagging of Historical Text by First Translating to Modern Text*
- Reimplemented in FoLiA-wordtranslate (part of https://github.com/LanguageMachines/foliautils)

Linguistic Enrichment

- **Frog:** Part-of-Speech tagging, Lemmatisation & Named Entity Recognition
  - Supervised machine learning (memory-based learning)
  - Trained on large corpora of contemporary dutch
  - Part-of-Speech: CGN tagset
  - https://languagemachines.github.io/frog
- Lemma linking to INT Historical Lexicon
- Entity linking with **FoliaEntity**
  - Links found entities to external resources (e.g DBpedia)
  - https://github.com/ErwinKomen/FoliaEntity

Gentse Spelen (1539)

http://www.dbnl.org/tekst/_gen001gent01_01/

	PoS	Lemma
Frog as is:	37.6%	31%
With modernization:	52.7%	53.7%
Retrained Frog (on Letters as Loot)	54.3%	45.3%

(choose the right training material!)

Metadata

J. V. VONDELS

LUCIFER.

TREURSPEL.

PRÆCIPITEMQUE IMMANI TURBINE ADEGIT.

t'AMSTERDAM,
Voor ABRAHAM de WEES, Boeckverkooper op den Middeldam, in 't Nieuwe Testament, in 't jaer 1654.

1654

DE WERKEN

VAN

J. VAN DEN VONDEL

UITGEGEVEN DOOR

M^R. J. VAN LENNEP

Herzien en bijgewerkt door J. H. W. UNGER

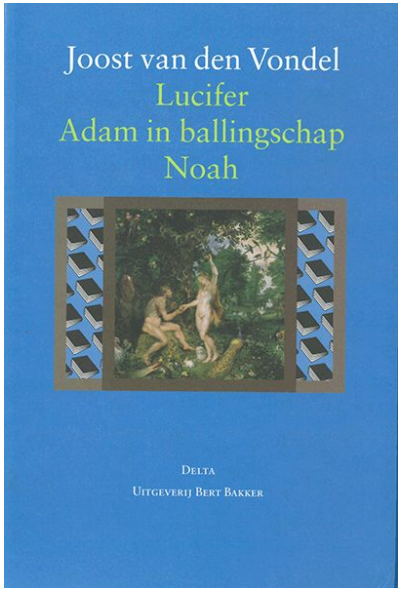
1654–1657

Lucifer — Inwydinge van 't Stadthuis t' Amsterdam — Salmoneus — Koning Davids Harpzangen

LEIDEN — A. W. SIJTHOFF

ANTWERPEN. — DE NEDERLANDSCHE BOEKHANDEL.

ca. 1891



2004

Metadata (2)

- Book vs. Text
- Date witness vs. date text
- Editorial matter v.s. original text
- Several texts with several authors, dates, etc. in one publication

- Searching
- Linguistic analysis
- Tagging strategy!

Evaluation tagset (1)

- CGN tagset is starting point
- Full CGN tagset (i.e. with features) too detailed to be feasible
- Problems in applying it to historical Dutch
  - Principle tag + lemma ⇒ (almost) determines word form cannot be maintained historically
  - (likewise for CG/CRM): tagging inflection instead of case/number/mood/tense etc does not benefit (diachronic) analysis
  - Mood cannot be lumped with tense historically (PVTIJD = tegenwoordig, verleden, imperatief, conjunctief.)
  - “dial” for all non-standard-modern words is unsatisfactory

Evaluation tagset (2)

- Just main part of speech is not enough
  - Mapping problems (lexical vs. functional tagging)
    - extra features: enable mapping POS / differences in lemmatisation
      - WW(vd,prenom) → ADJ, lemma: *gebroken*
      - WW(vd,vrij) → VRB, lemma: *breken*
  - Too much detail lost

Evaluation tagset (3)

- Intermediate option
  - Keep “position” information
  - Keep pronoun types and some other important distinctions
- Taking to account mappability to other tagsets
  - Functional tagging of main part of speech (as in e.g. CHN, Corpus Gysseling, Brieven als buit)
  - Universal dependencies

Evaluation tagset (4)

adj_nom
adj_postnom
adj_prenom
adj_vrij
adj_vrij_dim

bw

let

lid

n
n_dim

spec
spec_afgebr
spec_afk
spec_deeleigen
spec_vreemd
spec_symb
spec_meta

tsw

tw_hoofd_nom
tw_hoofd_nom_dim
tw_hoofd_prenom
tw_hoofd_vrij
tw_rang
tw_rang_nom
tw_rang_prenom

vg_neven
vg_onder

vnw_aanw_adv
vnw_aanw_nom
vnw_aanw_prenom
vnw_aanw_vrij
vnw_betr
vnw_bez_nom
vnw_bez_prenom
vnw_excl
vnw_onbep_adv
vnw_onbep_nom
vnw_onbep_nom_dim
vnw_onbep_prenom
vnw_onbep_vrij

vnw_pers
vnw_recip
vnw_refl
vnw_vb_adv
vnw_vb_nom
vnw_vb_prenom
vnw_vb_vrij

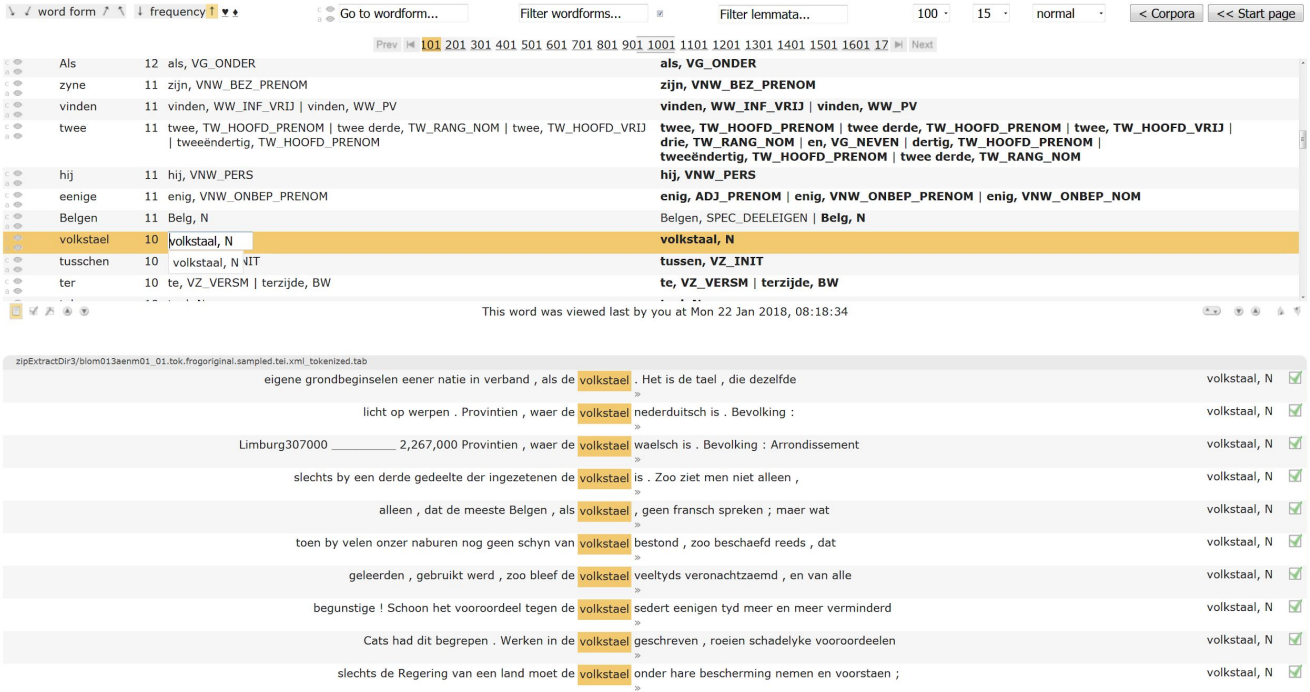
vz_fin
vz_init
vz_versm

ww_inf_nom
ww_inf_vrij
ww_od_nom
ww_od_postnom
ww_od_prenom
ww_od_vrij
ww_pv
ww_vd_nom
ww_vd_postnom
ww_vd_prenom
ww_vd_vrij

Preparation of ground truth

- At least 10K tokens per century
- Sets for 18th and 19th century

Tool: CoBaLT



https://github.com/INL/COBALT

Radboud University

Meertens Instituut

university of groningen

/instituut voor de Nederlandse taal/

Evaluation

```
Toch     S       (BW, BW, true, true)    (toch, toch, true)
heb      S       (WW_PV, WW_PV, true, true)      (hebben, hebben, true)
ik       S       (VNW_PERS, VNW_PERS, true, true)        (ik, ik, true)
ook      S       (BW, BW, true, true)    (ook, ook, true)
hier     S       (BW, VNW_AANW_ADV, false, false)        (hier, hier, true)
de       S       (LID, LID, true, true)  (de, de, true)
leelyke S        (ADJ_PRENOM, ADJ_PRENOM, true, true)    (lelijk, leelyke, false)
i-j      S       (SPEC_VREEMD, N, false, false)  (i-j, i-j, true)
die      S       (VNW_BETR, VNW_BETR, true, true)        (die, die, true)
door     S       (VZ_INIT, VZ_INIT, true, true)  (door, door, true)
sommigen         S       (VNW_ONBEP_NOM, VNW_ONBEP_NOM, true, true)      (sommige, sommig, false)
als      S       (VG_ONDER, VG_ONDER, true, true)        (als, als, true)
y-klank S        (N, N, true, true)      (ij-klank, y-klank, false)
gebruikt         S       (WW_VD_VRIJ, WW_VD_VRIJ, true, true)    (gebruiken, gebruiken, true)
wordt    S       (WW_PV, WW_PV, true, true)      (worden, worden, true)
voor     M       (BW, VZ_INIT, false, false)     (voorgoed, voor, false)
goed     M       (BW, ADJ_PRENOM, false, false)  (voorgoed, goed, false)
congé    S       (N, N, true, true)      (congé, congé, true)
gegeven S        (WW_VD_VRIJ, WW_VD_VRIJ, true, true)    (geven, geven, true)
Tant     M       (TSW, SPEC_VREEMD, false, false)        (tant pis, Tant, false)
pis      M       (TSW, N, false, false)  (tant pis, pis, false)
voor     S       (VZ_INIT, VZ_INIT, true, true)  (voor, voor, true)
de       S       (LID, LID, true, true)  (de, de, true)
Hilaridessen     S       (SPEC_DEELEIGEN, SPEC_DEELEIGEN, true, true)    (Hilarides, Hilaridessen, false)
die      S       (VNW_BETR, VNW_BETR, true, true)        (die, die, true)
er       M       (BW, VNW_AANW_ADV, false, false)        (erom, er, false)
om       M       (BW, VZ_INIT, false, false)     (erom, om, false)
```

Preliminary results

Set	size	Main PoS	PoS + features in evaluation tagset	Lemma
18e eeuw, met moderniseren	8594	87.07	76.06	84.54
zonder multiwords/clitics	7810	91.34	79.92	93.17
18e eeuw, zonder moderniseren	8594	80.50	71.72	73.61
zonder multiwords/clitics	7810	84.58	75.53	81.12
19e eeuw	13404	84.91	78.53	79.57
zonder multiwords/clitics	12722	87.07	80.62	84.66

Tag confusion: 18th century

reference	tagger	count
N	SPEC	102
VG	VZ	63
WW	N	53
BW	ADJ	43
ADJ	N	38
N	WW	28
BW	VNW	27
N	ADJ	26
VNW	ADJ	24
VNW	N	20

	ADJ	BW	LID	N	SPEC	TSW	TW	VG	VNW	VZ	WW
ADJ	545	14	0	38	16	0	0	0	3	1	20
BW	43	455	0	12	0	0	0	7	27	13	0
LID	0	0	768	1	3	0	0	1	4	0	1
N	26	7	5	1304	102	0	1	0	7	5	28
SPEC	2	0	0	4	22	2	0	0	0	1	8
TSW	0	0	0	0	0	0	0	0	0	0	0
TW	3	0	9	0	0	0	110	0	0	0	0
VG	9	9	0	5	8	0	0	622	15	63	3
VNW	24	9	10	20	1	1	0	3	990	0	4
VZ	0	3	0	1	0	0	0	1	0	859	2
WW	8	0	0	53	3	0	0	0	4	1	1431

Tag confusion: 19th century

reference	tagger	count
WW	N	111
VNW	N	106
BW	N	105
ADJ	N	100
N	SPEC	94
VNW	SPEC	89
N	WW	86
BW	VNW	67
ADJ	WW	59
VNW	ADJ	53

	ADJ	BW	LET	LID	N	SPEC	TSW	TW	VG	VNW	VZ	WW
ADJ	864	21	0	0	100	33	0	6	0	2	0	59
BW	31	724	0	0	105	29	0	0	10	67	15	16
LET	0	0	29	0	0	0	0	0	0	0	0	0
LID	27	3	0	1387	15	6	0	0	0	27	0	4
N	31	11	0	0	2224	94	0	2	0	1	0	86
SPEC	5	0	0	8	23	94	0	0	1	0	0	2
TSW	0	1	0	0	8	3	8	0	0	0	0	1
TW	1	0	0	6	1	9	0	104	0	0	0	0
VG	15	26	0	0	32	5	0	0	743	46	49	2
VNW	53	10	0	10	106	89	1	2	12	1189	0	21
VZ	3	4	0	0	18	33	0	0	1	0	1508	25
WW	21	2	0	0	111	13	0	2	0	1	1	1794

Frequent lemmatization errors

18th century

reference	*tagger*	*count*
welk	die	15
welk	wie	15
dezelve	die	12
schoon	mooi	11
haar	haaar	10
malkander	elkaar	9
dezelve	derzelver	9
meer	veel	9
wijze	wijs	7
doch	dog	7
omtrent	ongeveer	7
nootmuskaat	notemuscaat	6
dezelve	dezelven	6
gans	heel	6

19th century

reference	*tagger*	*count*
zijn	zyn	67
zo	zoo	45
hij	hy	32
meer	veel	32
taal	tael	31
zij	zy	31
bij	by	30
een	eene	26
Frans	fransch	26
aan	aen	24
mijn	myn	21
wij	wy	16
enig	eenig	14
daar	daer	12
mij	my	12
tussen	tusschen	11
Belg	Belgen	11
zijn	zyne	11

Lemma accuracy (influence of PoS)

18th century

ADJ	0.8901098901098901
BW	0.9317773788150808
LID	0.9922879177377892
N	0.8888888888888888
SPEC	0.5384615384615384
TW	0.7704918032786885
VG	0.9809264305177112
VNW	0.9048964218455744
VZ	0.9930715935334873
WW	0.9433333333333334

19th century

ADJ	0.6027649769585254
BW	0.8054162487462387
LET	1.0
LID	0.9639210347174949
N	0.7848101265822784
SPEC	0.8796992481203008
TSW	1.0
TW	0.8760330578512396
VG	0.9477124183006536
VNW	0.8218352310783658
VZ	0.9535175879396985
WW	0.8688946015424165

Lemma error types / interpretation differences

Subtleties

Sellery ?sellerie ?selderij

Remnants of the MT approach:

beest → dier
gezicht → visioen
vel → huid

But mostly unrecognized or wrongly resolved variants

Conclusions

- Modernization works in 18th century
  - But it should be possible to improve on current modernization by optimising the modernization lexicon
  - and by consistenly choosing for “transverbation” to obtain correct lemmata
- Normalisation of some form is needed for ca.1800-1954
  - But different training data for modernization required
- Enriching thousands of documents is computationally very expensive

Future work

- Evaluation sets for other language periods
- Training Frog with existing gold standards
  - Brieven als buit > Sailing letters
  - Gentse spelen > 16th Century material
  - CG and CRM > Middle Dutch and 16th Century material
- Evaluation measures more detailed
- Related projects
  - Nederlab small research projects (Hooft letters, Statenvertaling)
  - Adelheid/Midas/…
- CLARIAH-plus tasks: improved infrastructure for historical Dutch

17e eeuw: gedichtje _aar004aard01_01

PoS Accuracy: 0.8669201520912547 (max=0.944233206590621)
Lemma Accuracy: 0.8719898605830165 (max=0.9416983523447402)
http://www.dbnl.org/tekst/_aar004aard01_01/_aar004aard01_01_0001.php (789 tokens)

	ADJ	BW	LET	LID	N	POS	SPEC	TSW	TW	VG	VNW	VZ	WW
ADJ	24	1	0	0	2	0	0	0	0	0	0	0	4
BW	0	26	0	1	2	0	0	0	0	1	1	0	1
LET	0	0	159	0	0	0	1	0	0	0	0	0	0
LID	0	0	0	45	0	0	0	0	0	0	0	0	0
N	1	2	0	0	91	0	1	2	1	0	1	0	10
POS	0	0	0	0	0	1	0	0	0	0	0	0	0
SPEC	2	0	0	1	17	0	29	3	1	0	0	1	4
TSW	0	0	0	0	2	0	0	0	0	0	0	0	0
TW	0	0	0	0	0	0	0	0	0	0	0	0	1
VG	0	0	0	0	0	0	0	0	0	30	0	0	0
VNW	0	1	0	1	0	0	0	0	0	0	99	0	0
VZ	0	0	0	0	0	0	0	0	1	0	0	79	0
WW	3	0	0	0	4	0	0	0	0	0	0	0	101