Searching and analyzing large annotated text collections in Nederlab

Menzo Windhouwer
Hennie Brugman

menzo.windhouwer@meertens.knaw.nl

Introduction to nederlab

- Started: 2013 – ends: 2017
- Meertens Institute, Huygens ING, Institute for Dutch Lexicology, Radboud University Nijmegen- CLS
- Aims:
  - Detect and analyze historical changes in digitized Dutch and Flemish texts
  - History, literature, culture, linguistics
  - Bring ‘full text production’ together
  - User-friendly and tool-enriched research portal for scholars
  - Covers 800 until present
  - Most important metadata: time, place, author, text type
  - Enrichment of data by team and by scholarly users
  - Focus on data quality by including an editorial staff

Nederlab status

- Target: approximately 20 collections
  - Order of magnitude: tens of billions of annotated words
  - Including KB newspapers and some regional newspapers
- Major update of research portal expected: october, 2016
- Then available through Nederlab
  - 7 collections, including KB newspapers until 1900
  - 15,7 million titles (from articles up to books)
  - 150k persons

Collection workflow

1. Arrangement with collection provider
2. Quality Assessment
3. Mapping
4. Scripting and processing
5. Thesaurus linking
6. Manual curation
7. Automatic spelling correction/normalization
8. Add modern Dutch
9. Add annotation layers
10. Indexing and search
11. Make available to end users

7. Spelling correction/normalization

- TiCCL – Text Induced Corpus Cleanup (Reynaert, 2010)
- Improved to better deal with historical texts
- However
  - Many old OCR texts are of very bad quality
  - TiCCL improves on this, but quality stays mediocre at best
  - TiCCL works better for more recent texts

8. Add modern Dutch

- Approaches to apply language tools on historical text varieties
  - Adapt the tools to the language: time consuming, requires training data
  - Adapt the language to the tools
    - CLIN shared task: http://ifarm.nl/clin2017st/

9. Add annotation layers

Available:

- Postcorrected text (ticcl)
- Lemma, part-of-speech with sub features (frog)
- Entities (also multi-token) (frog)
- Sentence, div, paragraph, head

Planned/in progress (not –yet– for newspapers):

- Translation to modern word forms
- Entity linking
- Links to historical lexicon (INL)
- Speakers
- Language used
- Syntax, dependency structures
- Use case specific annotations

10. Indexing and search

- Current implementation: Multi-Tier Annotation Search (MTAS)
- Based on Lucene and SOLR
- Scalable, maintainable, parallel search
- 'Broker' middleware layer mixes in other services
  - Historical lexical query expansion
  - Joins
  - Later: semantic query expansion
- Fully configurable FoLiA parser and indexer

nederlab

samenstelling

- onzelfstandige titels ✓
- zelfstandige titels ✓
- koepeltitels ✓
- doublures uitsluiten

beperk tot genre(s)

reset zoek

- ⊞ ☐ fictie
- ⊞ ☐ non-fictie
- ⊞ ☐ periodieken
- ☐ bloemlezing
- ☐ verzameld werk
- ☐ verzamelhandschrift
- ☐ hertaling

beperk tot collectie(s)

DBNL
SoNaR

beschikbaarheid tekst

tekst beschikbaar
automatisch verbeterde tekst beschikbaar

woordvarianten

varianten inbegrepen

mijn nederlab zoeken over nederlab help uitloggen

↘ zoeken in inhoud

↘ zoeken in titelgegevens

↘ zoeken in auteursgegevens

achternaam auteur

selecteer rol +

☐ man ☐ vrouw ☐ onbekend

geboortejaar sterfjaar

↔ + ↔ +

☐ alleen exact bekende leefjaren

geboorteland

selecteer

geboorteprovincie

selecteer

geboorteplaats

sterfplaats

zoek reset ?

Corpus Query Language

pos is VNW

OR

AND

feat.getal is ev

OR

AND

feat.persoon is 1

OR

pos is WW

OR

[pos="VNW"&feat.getal="ev"&feat.persoon="1"][pos="WW"]

Voorbeelden
'koe'
twee adjectieven
+ 'geit'
1e pers. enk. +
werkwoord

zoek reset ?

Rodebeuk

collectie: **SoNaR**

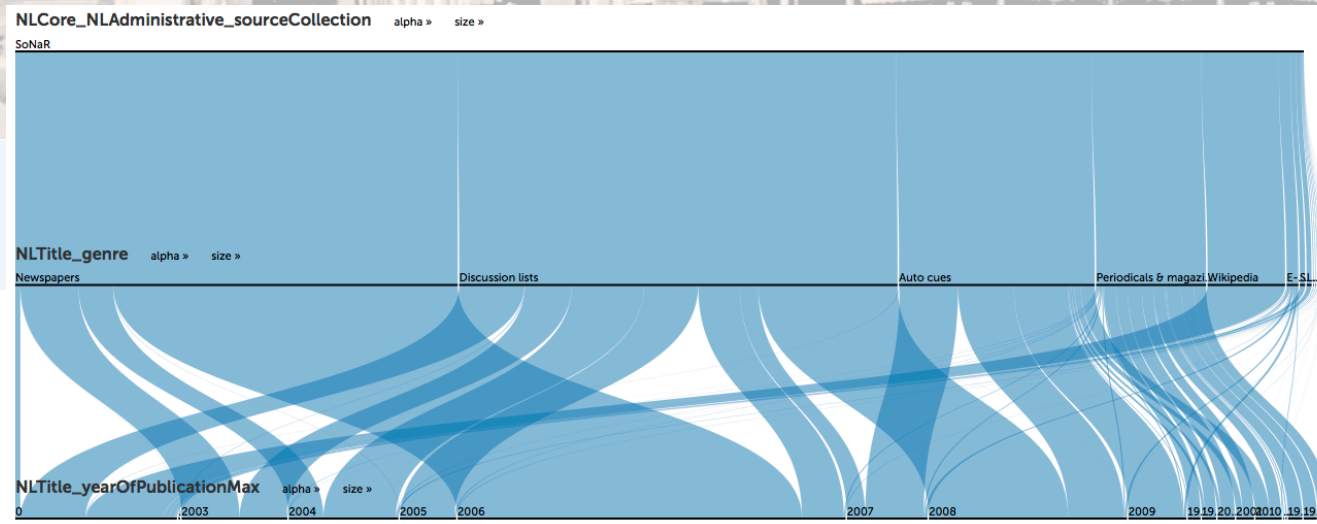
aantal hits: 3

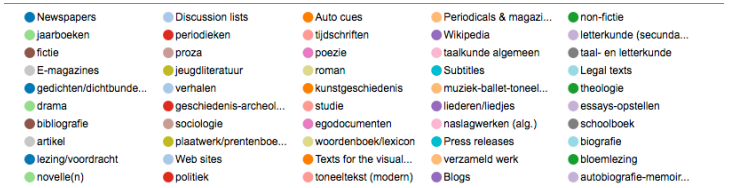
woord:	**geniet**	**van**	**mijn**	**verhaal**	**.**	**Ik**	**wist**	**niet**	**dat**	**Knex**	**is**	**overgenomen**
lemma:	genieten	van	mijn	verhaal	.	ik	weten	niet	dat	Knex	zijn	overnemen
pos:	WW	VZ	VNW	N	LET	VNW	WW	BW	VG	SPEC	WW	WW
kenmerken:	tgw pv ev	init	agr prenom zonder stan vol 1 ev bez det	onz basis stan soort ev		nomin vol 1 ev pers pron	verl pv ev		onder	deeleigen	tgw pv ev	vrij zonder vd

woord:	**daar**	**mee**	**te**	**maken**	**.**	**Ik**	**was**	**dan**	**dus**	**echt**	**meer**	**gecharmeerd**
lemma:	daar	mee	te	maken	.	ik	zijn	dan	dus	echt	veel	charmeren
pos:	VNW	VZ	VZ	WW	LET	VNW	WW	BW	BW	ADJ	VNW	WW
kenmerken:	obl vol 3o getal aanw adv-pron	fin	init	vrij zonder inf		nomin vol 1 ev pers pron	verl pv ev			basis vrij zonder	comp vrij zonder stan onbep grad	vrij zonder vd

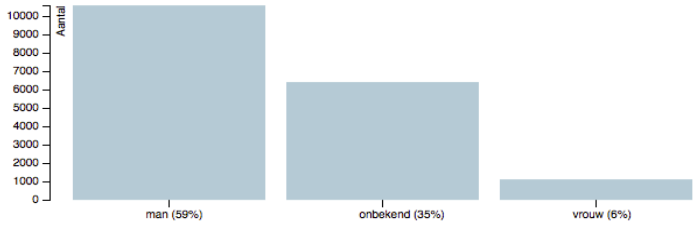
woord:	**"**	**oude**	**"**	**Knex**	**.**	**Ik**	**ben**	**wel**	**benieuwd**	**of**	**er**	**nog**
lemma:	"	oud	"	Knex	.	ik	zijn	wel	benieuwd	of	er	nog
pos:	LET	ADJ	LET	SPEC	LET	VNW	WW	BW	ADJ	VG	VNW	BW
kenmerken:		basis prenom met-e stan		deeleigen		nomin vol 1 ev pers pron	tgw pv ev		basis vrij zonder	onder	stan red 3 getal aanw adv-pron	

nederlab

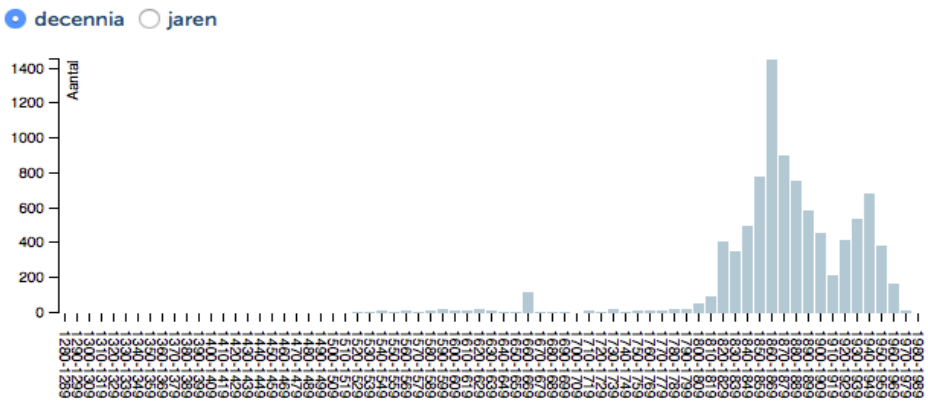


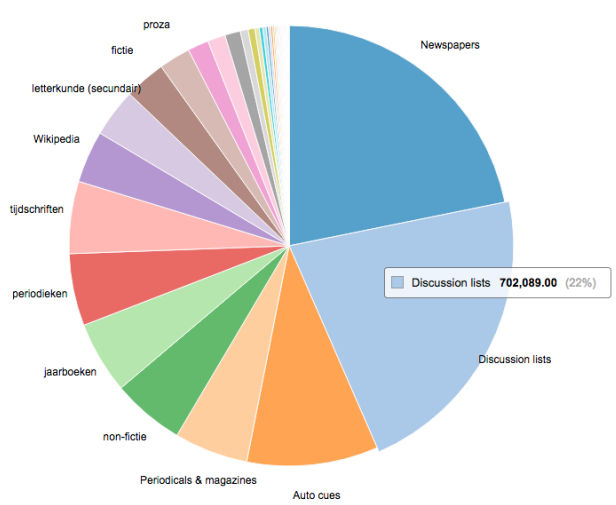


auteurs: mannen vs. vrouwen



geboortejaren auteurs





nederlab

type woorden top 100 min. lengte 6 begint met: ge eindigt op: heid regexp: ge[a-z]*heid reset toon

<< < 1 2 3 > >>

1.	gelegenheid	483
2.	gerechtigheid	231
3.	gezondheid	205
4.	gehoorzaamheid	114
5.	gelijkheid	89
6.	geleerdheid	74
7.	genegenheid	66
8.	gekheid	65
9.	geestelijkheid	45
10.	gezindheid	41
11.	geaardheid	31
12.	gesteldheid	31
13.	gezelligheid	31
14.	geestigheid	28
15.	gerustheid	28
16.	gereedheid	26
17.	gemeenzaamheid	25
18.	geldigheid	22
19.	gevoeligheid	19
20.	geschiktheid	16

type lemma's top 10

1.	gelegenheid	542
2.	gerechtigheid	231
3.	gezondheid	206
4.	gehoorzaamheid	115
5.	gelijkheid	89
6.	geleerdheid	74
7.	genegenheid	71
8.	gekheid	67
9.	geestelijkheid	45
10.	gezindheid	43
11.	gesteldheid	32
12.	geaardheid	31
13.	geestigheid	30
14.	gezelligheid	30
15.	gerustheid	28
16.	gereedheid	26
17.	gemeenzaamheid	25
18.	geldigheid	22
19.	gevoeligheid	21
20.	geschiktheid	16

type woordsoorten

1.	LET	1.256.452
2.	N	1.090.304
3.	WW	908.442
4.	SPEC	777.906
5.	VNW	639.232
6.	VZ	588.908
7.	LID	459.247
8.	ADJ	382.463
9.	BW	305.760
10.	VG	268.717
11.	TW	171.173
12.	TSW	6.816

type entities

1.	per	145.499
2.	loc	89.827
3.	pro	31.213
4.	misc	24.153
5.	org	17.893
6.	eve	1.305

Statistieken

98 hits, gevonden in **76** documenten
voor CQL query: **[pos="ADJ"]{2,2}[t_lc="geit"]**

matchende documenten

maximum aantal woorden: **2201607**
minimum aantal woorden: **12**
gemiddeld aantal woorden: **90202.89**
mediaan aantal woorden: **8656.5**
totaal aantal woorden: **6855420**

hit statistieken

maximum aantal hits per document: **5**
minimum aantal hits per document: **1**
gemiddeld aantal hits per document: **1.29**
mediaan aantal hits: **1**
totaal aantal hits: **98**

Conclusions

- Nederlab will be very large, with many different annotation layers, with different structures
- So far, indexes scale well and can be efficiently maintained (tested up to 1.5 billion word forms)
- Most of search and analysis requirements are met
- We support different forms of results: lists, grouped results, statistics.
  - Good input for different types of end user research tools
- Nederlab is extendible, both for collections and for tools