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
<table>
<thead>
<tr>
<th>sammenstelling</th>
<th>mijn nederlab</th>
<th>zoeken</th>
<th>over nederlab</th>
<th>help</th>
<th>uitloggen</th>
</tr>
</thead>
<tbody>
<tr>
<td>onzelfstandige titels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>zelfstandige titels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>koepeltitels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>doubles uitsluiten</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>beperk tot genre(s)</th>
<th>zoek in inhoud</th>
</tr>
</thead>
<tbody>
<tr>
<td>fictie</td>
<td>selecteer rol</td>
</tr>
<tr>
<td>non-fictie</td>
<td>+</td>
</tr>
<tr>
<td>periodieken</td>
<td></td>
</tr>
<tr>
<td>bloemlezing</td>
<td></td>
</tr>
<tr>
<td>verzameld werk</td>
<td></td>
</tr>
<tr>
<td>verzamelhandschrift</td>
<td></td>
</tr>
<tr>
<td>hertaling</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>beperk tot collectie(s)</th>
<th>zoek in titelgegevens</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBNL</td>
<td></td>
</tr>
<tr>
<td>SoNaR</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>beschikbaarheid tekst</th>
<th>zoek in auteursgegevens</th>
</tr>
</thead>
<tbody>
<tr>
<td>tekst beschikbaar</td>
<td></td>
</tr>
<tr>
<td>automatisch verbeterde tekst beschikbaar</td>
<td></td>
</tr>
</tbody>
</table>

| woordvarianten | |
|----------------||
| varianten ingebegrepen | |
Corpus Query Language

AND

pos - is VNW

feat.getal - is ev

feat.persoon - is 1

AND

pos - is WW

Voorkbeelden
'
koë'
twee adjectieven
+'geit'
1e pers. enk. + werkwoord

zoek reset ?
Rodebeuk

collectie: SoNaR

woord: geniet
lemma: genieten
pos: VV
kenmerken: genieten

woord: daar
lemma: daar
pos: VV
kenmerken: daar

woord: "oude"
lemma: oud
pos: ADJ
kenmerken: oud

woord: wel
lemma: wel
pos: VV
kenmerken: wel
auteurs: mannen vs. vrouwen

geboortejaren auteurs

- decennia
- jaren
**Statistieken**

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

<table>
<thead>
<tr>
<th>matchende documenten</th>
<th>hit statistieken</th>
</tr>
</thead>
<tbody>
<tr>
<td>maximum aantal woorden: 2201607</td>
<td>maximum aantal hits per document: 5</td>
</tr>
<tr>
<td>minimum aantal woorden: 12</td>
<td>minimum aantal hits per document: 1</td>
</tr>
<tr>
<td>gemiddeld aantal woorden: 90202.89</td>
<td>gemiddeld aantal hits per document: 1.29</td>
</tr>
<tr>
<td>mediaan aantal woorden: 8656.5</td>
<td>mediaan aantal hits: 1</td>
</tr>
<tr>
<td>totaal aantal woorden: 6855420</td>
<td>totaal aantal hits: 98</td>
</tr>
</tbody>
</table>
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