Experiences & requirements: Nederlab

Hennie Brugman
hennie.brugman@meertens.knaw.nl

Meertens Institute

MTAS workshop, September 25-26, 2017

Introduction to Nederlab

- Started: 2013 – ends: mid-2018
- Meertens Institute, Huygens ING, Dutch Language Institute, Radboud University Nijmegen-CLS
- Some facts:
  - Bring 'full text production' together
  - User-friendly and
    scholars
  - History, literatur
  - Diachronic: cove
  - Most important
  - Enrichment of data by team and by scholarly users
  - Funding: NWO, KNAW, CLARIN-NL, CLARIAH

Requirement: scalability

 # collections: 8 (+3), aim: 20-30 by mid-2018
 # documents: 15 million
 # tokens: 10 billion
 # annotations: 36 billion

Nederlab main objectives

- Detect the beginning of innovation in language and culture (concepts, word forms, patterns
- Determine the spread of change
- Detect connections and networks
- Find similarities and differences between (sets of) texts

Nederlab main objectives

- Detect the beginning of innovation in language and culture (concepts, word forms, patterns
- Determine the spread of change
- Detect connections and networks
- Find similarities and differences between (sets of) texts

nederlab

FoLiA XML

```
  <entities xml:id="hoof001hwva03_01.TEI.2.text.body.div.note.8223.s.1.entities.1">
    <entity xml:id="hoof001hwva03_01.TEI.2.text.body.div.note.8223.s.1.entities.1.entity.1" class="misc" confidence="0.790142" textclass="contemporary">
      <wref id="hoof001hwva03_01.TEI.2.text.body.div.note.8223.s.1.w.3" t="T."/>
    </entity>
    <entity xml:id="hoof001hwva03_01.TEI.2.text.body.div.note.8223.s.1.entities.1.entity.2" class="loc" confidence="0.950087" textclass="contemporary">
      <wref id="hoof001hwva03_01.TEI.2.text.body.div.note.8223.s.1.w.6" t="Mengelwerken"/>
    </entity>
  </entities>
 </s>
</note>
<p xml:id="hoof001hwva03_01.TEI.2.text.body.div.p.8225">
 <t> Mijn Heer en Broeder,</t>
 <s xml:id="hoof001hwva03_01.TEI.2.text.body.div.p.8225.s.1">
  <t>Mijn Heer en Broeder,</t>
  <w xml:id="hoof001hwva03_01.TEI.2.text.body.div.p.8225.s.1.w.1" class="WORD">
   <t>Mijn</t>
   <t class="contemporary">Mijn</
   <lemma class="WNT:M039335"
   <lemma class="mijn" set="https
   <metric class="modernisationso
   <pos class="VNW(bez,det,stan,v
    <feat class="bez" subset="vwty
    <feat class="det" subset="pdty
    <feat class="stan" subset="naa
    <feat class="vol" subset="statu
    <feat class="1" subset="persoo
    <feat class="ev" subset="getal"
    <feat class="prenom" subset="
    <feat class="zonder" subset="b
    <feat class="agr" subset="npag
   </pos>
   <lemma class="mijn" set="http://ilk.uvt.nl/folia/sets/frog-mblem-nl" textclass="contemporary"/>
  </w>
```

Requirements:

- *search on combination of metadata, text and annotations*
- *Corpus Query Language (CQL)*
- *search over (combinations of) any of the mentioned annotation structures and types*
- *joins*

Nederlab main objectives

CQL builder

nederlab

Nederlab main objectives

Even mijn neus laten zien.

originele tekst < > ? Omschakelen naar kwic-regel modus Verberg kwic-verrijkingen 1

DE **DRONGO** *PROFITEERT*

Hoe de drongo's profiteren van de voedingsstrategieën van andere dieren .

Drongo's zijn insectenetende Afrikaanse vogels met een opvallende gevorkte staart , die zich gemakkelijk aansluiten bij gemengde groepen van insecteneters . Tot voor kort werd aangenomen dat ze daarin vooral voedsel stalen , dat andere soorten al bemachtigd hadden . Kleptoparasitisme , heet dat . Maar deze veronderstelling werd niet hard gemaakt door observaties van ornitholoog Marc Herremans van het Koninklijk Museum voor Midden-Afrika in Tervuren . Herremans beschreef in het vakblad Bird Behavior zijn conclusies terzake .

De	drongo	bleek
lemma: de	lemma: drongo	lemma: blijk
pos: LID	pos: N	pos: WW
feat.naamval: stan	feat.ntype: soort	feat.wvorm:
feat.npagr: rest	feat.getal: ev	feat.pvtijd: v
feat.lwtype: bep	feat.graad: basis	feat.pvagr: e
	feat.genus: zijd	
	feat.naamval: stan	

feat.persoon: 3
feat.status: red

Requirements:

- *results as keyword in context*
- *retrieve pages of annotated document text*

aan te sluiten bij andere dieren , meer bepaald bij vee , bij wild zoals olifanten , of bij een breed gamma aan vogels die vooral op de grond zelf naar voedsel zochten , zoals struisvogels en wevers . Die jagen in dat proces allemaal insecten op , waarvan de **drongo** kan profiteren . Meestal was de **drongo** de enige vogel van het gezelschap die niet op de grond naar voedsel zocht .

Het actief stelen van insecten gebeurde relatief zelden . Toch viel het wel eens voor , vooral in omstandigheden waarin heel weinig insecten te vinden waren . Interessant daarbij was dat drongo's afleidingsmanoeuvres gebruikten : ze werden vanuit hun homen verondersteld de

masc

Nederlab main objectives

- Detect the beginning of innovation in language and culture (concepts, word forms, patterns
- Determine the spread of change
- Detect connections and networks
- Find similarities and differences between (sets of) texts

Nederlab main objectives

- Determine the spread of change
  - Spread over time
  - Spread over places
  - Spread over authors
  - Change of meaning, over time

Requirements:
- *distributions over sets of (metadata) values → facets*
- *distributions over numeric values → facet ranges*
- *distributions over multiple dimensions simultaneously → multifacets*

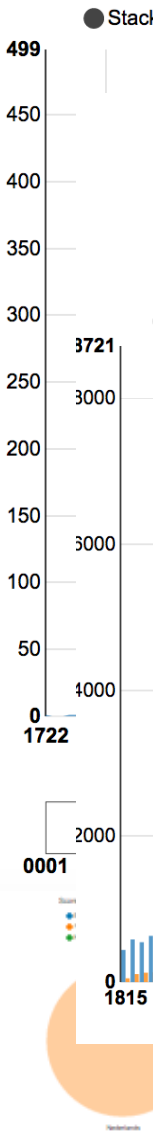
Requirement:
- *inspect word context at different times → CQL search + grouping of results*

Broncollectie

Datering benaderd

Metadata over de tijd

[lemma="aardig"][pos="N"]



38 hits, gevonden in **10** document
voor CQL query: **[lemma="aardig**

query: **[lemma="aardig"][pos="N**

1. **aardige puntreden** **3** hits
2. **aardige quinkslag** **3** hits
3. **aardige verstanden** **2** hits
4. **Aardig diertje** **1** hit (
5. **Aardig vraagstuk** **1** hit (
6. **Aardige beschrijving** **1** hit (
7. **Aardige gelikenissen** **1** hit (
8. **Aardige spreuk** **1** hit (
9. **Aardige plaats** **1** hit (
10. **Aardige print** **1** hit (
11. **Aardige puntreden** **1** hit (
12. **Aardige redeneering** **1** hit (
13. **Aardige redenstrijt** **1** hit (
14. **Aardige zinspreuk** **1** hit (
15. **aardig deuntje** **1** hit (

5.682 hits, gevonden in **5.042** docum
voor CQL query: **[lemma="aardig"][p**

query: **[lemma="aardig"][pos="N"]** (

1. **aardige mensen** **211** hits
2. **aardige man** **209** hits
3. **aardige jongen** **121** hits
4. **aardig stukje** **89** hits
5. **aardige dingen** **88** hits
6. **aardige duit** **81** hits
7. **aardige vent** **81** hits
8. **aardig idee** **76** hits
9. **aardig bedrag** **73** hits
10. **aardig mondje** **68** hits
11. **aardig beeld** **61** hits
12. **aardige stuiver** **61** hits
13. **aardig woordje** **59** hits
14. **aardige som** **51** hits
15. **aardige cent** **47** hits

Nederlab main objectives

- Detect the beginning of innovation in language and culture (concepts, word forms, patterns
- Determine the spread of change
- Detect connections and networks
- Find similarities and differences between (sets of) texts

Nederlab main objectives

- [groepeer op woord toon

uitgangspunt

een lemma

lemma

varen

woordsoort: (maakt niet uit)

collocatie

alleen de woordsoort

woordsoort: N

+ - - -

bereik

zoek naar collocaties binnen:

woorden voor en na

bereik:

+ 4 3 2 1 0 0 1 2 3 4 +

3 links | rechts 3

reset zoek collocaties

De documenten waarin naar collocaties wordt gezocht:

collectie: SoNaR
metadata van bron: tekst beschikbaar

Zoek naar woordsoort N binnen de context van 3 woorden vóór en 3 woorden ná [varen]

Totaal aantal gevonden collocaties: **11026** in **7003** documenten

download csv

CQL zoekvraag

Frequentie Woord 1 2 3 4 5 6 7 8 9 > >>

#	Woord	Frequentie	Documenten	
1	koers	879	852	
2	boot	331	309	
3	schip	281	265	
4	vlag	176	135	
5	uur	137	126	
6	dag	133	132	
7	jaar	132	129	
8	water	113	104	
9	zee	97	96	
10	d	80	80	
11	e	80	80	
12	r	80	80	
13	plan	76	76	
14	haven	71	67	
15	idee	63	63	

Nederlab main objectives

- Detect the beginning of innovation in language and culture (concepts, word forms, patterns
- Determine the spread of change
- Detect connections and networks
- Find similarities and differences between (sets of) texts

Nederlab main objectives

Statistieken

10.500 hits, gevonden in **8.890** documen
voor CQL query: **[lemma="er"][][lemma=**

matchende documenten

maximum aantal woorden: **2.201.607**
minimum aantal woorden: **17**
gemiddeld aantal woorden: **46.463,31**
totaal aantal woorden: **413.058.859**

type woorden top 100

✓ woorden
lemma's

1 (4.550%)

Statistieken

49 hits gevonden in document
voor CQL query: **[t_lc="oorlog"]**

totaal aantal woorden: **64.549**
aantal entities: **1.801**
aantal paragrafen: **619**
aantal zinnen: **3.278**
aantal divs: **107**
aantal headings: **107**

Requirements:

- *calculate statist*
- *determine char*
- *frequency lists for word form, lemma, part of spe*
  - *basis for e.g. stylistics*

15. **met** **2.747.909** (0,665%)
16. **voor** **2.597.831** (0,629%)
17. **aan** **2.255.337** (0,546%)
18. **hij** **2.171.201** (0,526%)
19. **er** **2.000.070** (0,484%)
20. **als** **1.965.724** (0,476%)

Thank you.