Een baseline voor het parsen van historische tekst

Marijn Schraagen
Digital Humanities Lab
Universiteit Utrecht
M.P.Schraagen@uu.nl

Aanpak

- Gebruik parsers voor modern Nederlands

- Verbeter de resultaten met aanpassingen in de spelling en grammatica van de brontekst

Parsers

- Alpino
  - Handmatige grammaticale regels (335) en lexicon
  - Head-Driven Phrase Structure Grammar
- Frog
  - Statistische machine learning
  - Automatisch afgeleide beslisbomen
  - Trainingsdata nodig

Alpino

- Lexicon met features
  - *tafel: non-derived, sg, count, de, lexical, non-temp/refl, noun, …*
- Onbekende woorden passen niet goed in dit formalisme

  o.a. *Van der Beek et al., The Alpino Dependency Treebank, 2002*

Frog

- Training: 10,979,827 woorden
  - Corpus Gesproken Nederlands
  - ILK, D-Coi, Eindhoven Corpus
- Onbekende woorden: features
  - Prefix (2), suffix (3)
  - hoofdletter, koppelteken (-), cijfers
  - contextwoorden, voorgaande POS-tags

Van den Bosch et al., An efficient memory-based morphosyntactic tagger and parser for Dutch (2007)

Demonstratie

- Bron: correspondentie Hugo de Groot, 17e eeuw
- Corpus Geleerde Brieven
- Selectie: <interp type="language" value="nl"/>
- 27-08-1611: *Het verdryet my, dat wy soo langsaeme aen 't werck comen, 't welck my zeer te onpasse comt.*

Baseline-aanpassingen 1

- Fonetische herschrijfregels
- `ae → aa`
  `gh → g`
  `ck → k`
  `lick → lijk`
  etc.
- Vgl. Loes Braun, *Information Retrieval from Dutch Historical Corpora,* 2002
- ~50 regels

Baseline-aanpassingen 2

- Lexicale herschrijfregels
- Hugo de Groot-tekst gesorteerd op woordfrequentie, top 200

1.	52158	de	
2.	47424	van	
3.	26268	te	
4.	25749	ende	→ en
11.	12115	ick	→ ik *(niet nodig, fonetische herschrijfregel)*
18.	7249	sijn	→ zijn
23.	6673	sal	→ zal
28.	5805	soo	→ zo
...			

- Beperkt aantal aanpassingen, groot bereik

Baseline-aanpassingen 3

- Afstand tot modern lexicon
- Woordenlijst Opentaal.org
- Voor elk onbekend woord: vergelijk Levenshtein edit distance voor alle bekende woorden
- Niet altijd goede resultaten

Baseline-aanpassingen 3

- Afstand (langsaame,langzaam)=2
- Ook: (langsaame,langsdam)=2

- Kies alternatief op basis van frequentie, beschikbaar in Opentaal woordenlijst
- *langzaam*: 26759, *langsdam*: 24

Baseline-aanpassingen 3

- Lange rekentijd: vergelijk ieder woord met 300,000 moderne woorden
- Efficiënter: begin met woorden van eigen lengte, sla andere lengtes over die geen beter resultaat meer kunnen opleveren
- (langsaame,aangename)=3, bekijk geen woorden met lengte <6 of >12
- Binnen de huidige lengte: bekijk verschillen in prefix
- (langsaame,Hebreeuws) → (lang,Hebr), altijd >3

Mogelijkheden voor onderzoek

- Trainingsmateriaal verzamelen
  - Trainen parser / lexicon
  - Trainen omzetting historisch → modern
- Grammar induction
  - Supervised / unsupervised

Mogelijkheden voor onderzoek

- Domeinkennis
  - Lexica / gestructureerde knowledge bases voor het onderwerp van de tekst
  - Mapping naar Latijn/Frans/...
    - *admitteeren, refusie, disputatien*
- Grammaticale omzetting (?)
  - Ick en weet U L. niet nieus te schriven, alsoo ick niemant en hebbe die my wat tuis brent.
  - Ik en weet U L. niet niets te schrijven, omdat ik niemant en heb die mij wat tuis brent. (fonetisch, top 200 lexicaal, edit distance)
  - Ik kan u niets nieuws schrijven, omdat ik niemand heb die mij wat thuis brengt.
- Afhankelijk van taal- of letterkundig onderzoeksonderwerp