Language Dynamics in the Dutch Golden Age

linguistic and socio-cultural aspects of intra-author variation

Feike Dietz, Marjo van Koppen

Jorik van Engeland, Cora van de Poppe, Marijn Schraagen

Universiteit Utrecht

Project overview

- 16th and 17th century Dutch: linguistic variation and change
- This project: **intra-author variation**, i.e., variation within individual, influential language users
- Three main approaches:
  1. Theoretical linguistics
  2. Literary and socio-cultural analysis
  3. Computational linguistics
- **Data:** P.C. Hooft, Stevin, Coornhert, Vondel, ...
- **Method:**
  1. Corpus selection, cleaning, annotation
  2. *Close reading* for linguistic and literary analysis
  3. Extracting morphological features
     - negation, case, infinitival clauses, ...
     - use, improve and/or create tools for automatic analysis
  4. Identifying sociological features
     - location, genre, audience, gender, ...

1 Automatic and manual enrichment

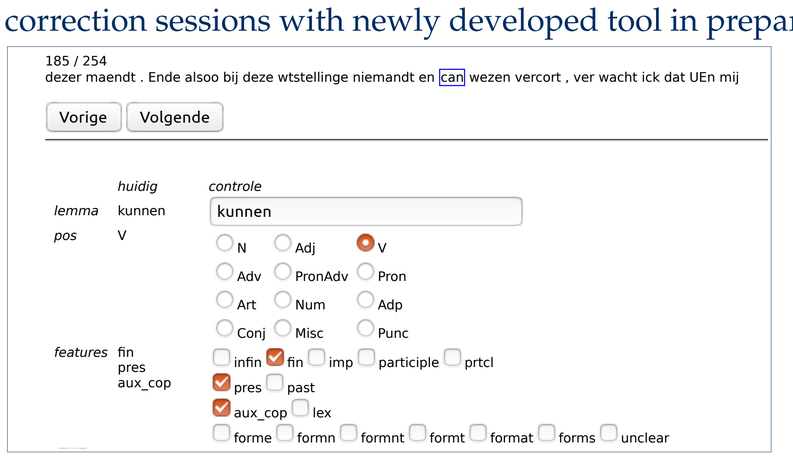
Letter of P.C. Hooft [1] to the mayors of Muiden, June 18, 1609, asking to postpone the election for guard commanders. *En* is used both as conjuction and negation.

> [...] dat UE. de keur **en** bevestinge der bevelhebberen over de schutterie gelieven sal wt te stellen **ende** op te houden tot op Sondach over acht daeghen werdende den achtentwintichsten dezer maendt. **Ende** alsoo bij deze wtstellinge niemandt **en** can wezen vercort [...]
> *that you please postpone the choice and confirmation of the commanders of the guard and hold off until Sunday in eight days, being the 28th of this month. And also with this delay nobody will be opposed*

- Enrichment using the Adelheid tagger for Middle Dutch [2]:

word	lemma	POS
dat	dat	Art(def)
UE	vee	N(sing,forme)
.	.	Punc(period)
de	te	Adp()
keur	eekhoorn	N(prop)
en	in	Adp()
bevestinge	bezating	N(sing,forme)
[...]		
Ende	en	Conj(coord)
alsoo	alzo	Adv(gener)
bij	bij	Adp()
deze	deze	Art(def,forme)
wtstellinge	stalling	N(plu,forme)
niemandt	???	N(sing)
en	en	Adv(neg)
can	kunnen	V(fin,pres,aux_cop)
wezen	zijn	V(infin)
vercort	???	N(sing)

- Automatic tagging provides useful results
- Still many errors, manual correction desired
- Annotation correction sessions with newly developed tool in preparation



2 Automatic modernization

- Modernization of spelling and grammar allows use of tools for modern Dutch
- *Note:* some features (e.g., double negation and case marking) are lost after modernization
- Automatic conversion is possible using parallel text to train algorithms and construct a translation lexicon
- Relatively large parallel text available in diachronic translations of the Bible

> 1637: Verlost my, o Godt : want de wateren zijn gekomen tot aen de ziele.
> 1888: Verlos mij, o God ! want de wateren zijn gekomen tot aan de ziel.
> *Save me, O God; for the waters are come in unto my soul.*

- Statistical Machine Translation using *Moses* [3]
- Alternative: start from scratch using various rule-based and machine learning-based approaches
  - Construct 1-to-1 translation lexicon using sentences of equal length
  - Perform alignment to handle sentences of unequal length
  - Compile a set of manual modernization rules (e.g., strip case markers)
  - Construct many-to-1 translation lexicon using aligned sentences
  - Use POS-information for already modernized words to choose the right alternative for historical words
    - ◇ *haer* + V → *hen*
    - ◇ *haer* + N → *hun*
  - Compile rules to address punctuation differences
- Results of both approaches are comparably accurate
  - BLEU score for evaluation of machine translation [4]
  - Moses: 0.61631 (after post-processing: 0.63867)
  - From scratch: 0.62715
- Combination of approaches not straightforward
- Best so far: Moses + manual rules (0.64418)

References

[1] Hendrik van Tricht. *De briefwisseling van Pieter Corneliszoon Hooft*. Tjeenk Willink / Noorduijn, 1976.

[2] Hans van Halteren and Margit Rem. Dealing with orthographic variation in a tagger-lemmatizer for fourteenth century Dutch charters. *Language Resources and Evaluation*, 47(4):1233–1259, 2013.

[3] Philipp Köhn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics, 2007.

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

Acknowledgements

NWO Netherlands Organisation for Scientific Research

This work is financed by the Netherlands Organisation for Scientific Research (NWO), grant 360-78-020.