# Linguistic and Sociolinguistic Annotation of 17th Century Dutch Letters

Marijn Schraagen, Feike Dietz, Marjo van Koppen

Department of Languages, Literature and Communication, Utrecht University, The Netherlands

## Summary

- **Domain:** letters by Dutch author and politician P.C. Hooft (1581–1647)
- **Annotation:** lemma, POS, correspondent metadata, text content classification, text segmentation
- **Comparison:** large differences between automatic and manual tagging
- **Inter-annotator agreement:** strict vs. permissive agreement measures
- **Availability:** integration with CLARIN infrastructure
- **Application:** historical (socio-)linguistics, e.g., negation constructs

## Domain

- 17th century: linguistic developments in Dutch
  - Changes in vocabulary, spelling, case marking, negation, verb constructs, ...
- Limited availability of manually annotated data and automatic tools
- Manual annotation task: letters of P.C. Hooft
- Total corpus: 1300 documents, 300k tokens
- POS annotations provided for 333 letters (108k tokens) from 1600-1638
- Additional sociolinguistic categorization of meta-information and content
- Document example: fragment from a letter to the mayors of Muiden, June 18, 1609, asking to postpone the election for guard commanders.

> 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*

## Annotation

- Lemmatization
- Part-of-Speech tagging using Spoken Dutch Corpus (CGN) tagset
  - Main tags with features (e.g., number, gender, tense, case)
  - Example: *can* → `V(present,nonlexical,singular,1st person,simple)`
  - Features added for 17th century Dutch (e.g., case marking, negation)
- Sociolinguistic annotation
  - Document level and correspondent level annotations
  - Categorization revised due to low agreement scores, re-annotation currently in progress

| | |
|---|---|
| *Document characteristics* | |
| Type | business/personal, regular/appendix, individual/group correspondent |
| Goal | express thanks, compliment, excuse, ask a favour, ask information, ask advice, admonish, inform, remember, persuade, order, allow, invite |
| Topic | business, literature, domestic affairs, love, death, news, religion & ethics |
| *Revised document characteristics* | |
| Goal | prompt for action, honour, help, inform, keeping contact, ask for reply |
| Topic | political work, literary work, current events, social circle |
| *Correspondent characteristics (individual correspondents only)* | |
| birth/death date, gender, occupation, literary author, relation to P.C. Hooft | |
| *Letter segmentation* | |
| Initial greeting, opening (optional), narrative, closing (optional), final greeting | |

## Comparison with automatic tagging

- Automatic tagger/parser for Dutch using CGN tagset: Frog
- Modernization layer using look-up in historical dictionary
- Substantial differences compared to manual annotation
  - Note: several features not present in CGN tagset

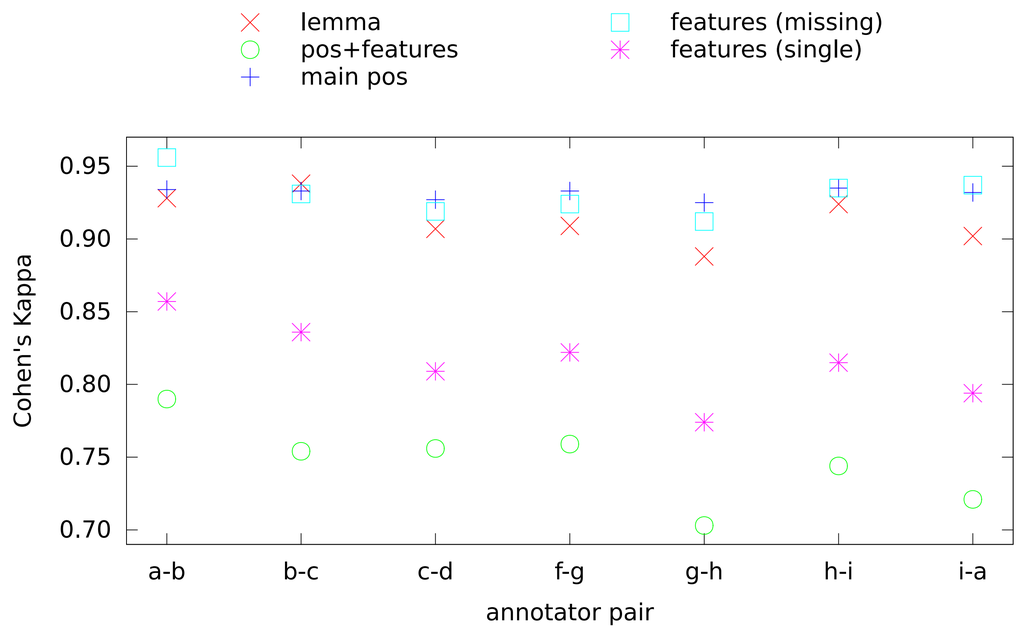
| *type* | *different* | *equal* | *ratio* |
|---|---|---|---|
| lemma (Frog default) | 27,179 | 70,560 | 0.28 |
| lemma (Frog modernized) | 18,052 | 53,648 | 0.25 |
| main POS | 23,365 | 82,830 | 0.22 |
| features (only Frog/CGN default) | 46,725 | 67,273 | 0.41 |
| features (extended) | 140,888 | 67,273 | 0.68 |

*ratio: different / (different+equal)*

- Current annotations potentially useful for improving automatic tagging

## Inter-annotator agreement

- In general: binary agreement → a POS tag is the same or not
- Widely used tagsets: features not always explicit, low granularity
- **For a high granularity tagset, partial agreement may be more informative than binary agreement**
- Various possible measures
  - `V(present,nonlexical,singular,1st person,simple)` vs. `V(present,singular,3rd person,simple)`
  - Binary/full agreement on POS and features: 0.0
  - Agreement on main POS: 1.0
  - Agreement on single features: 3/5=0.6
  - Ignore disagreement due to missing features (here: *lexical*/*nonlexical*): 3/4=0.75
- Current corpus: eight annotators, 1000 words also tagged by second annotator




- Full agreement on POS and features: **too strict**
- Agreement on lemma, main POS, missing features: **too permissive**
- Agreement on single features: **more balanced**
- Sociolinguistic features: reasonable agreement on segmentation (~0.8)
- Document characteristics: low agreement, categorization has been revised

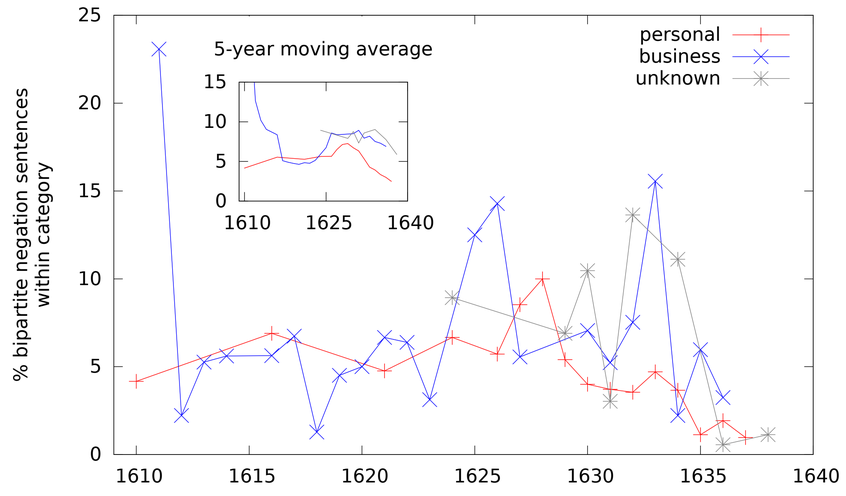
## Availability

- Annotations will be available through CLARIN-federated login in the Nederlab online research environment (`http://www.nederlab.nl`)
- Separate layer on top of automatic annotations
- Alignment to account for tokenization differences

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Aenden | | Advocaet | van | Hollandt | ... | verschej- | den | onwaerdicheden | zijn | toegedreven |
| Aen | den | Advocaet | van | Hollandt | ... | verschejden | | onwaerdicheden | zijn | toegedreven |
| *To* | *the* | *attorney* | *of* | *Holland* | ... | *several* | | *untruths* | *are* | *added* |

## Application

- Bipartite negation (adverb + clitic) in decline in 17th century Dutch
- Sociolinguistic hypothesis: decline slower and/or later in business letters




## Acknowledgments

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

Netherlands Organisation for Scientific Research

Universiteit Utrecht