Universiteit Utrecht

Provenance Language Dynamics in the Dutch Golden Age

Marjo van Koppen

DANS, Den Haag, 2018-09-03

Meertens Instituut

Overview

- Input data
- Editing data
- Availability of the data
- Provenance for researchers

Meertens Instituut

Input Data

- Dbnl-web version of 'The letters of P.C. Hooft'. https://dbnl.org/tekst/hoof001hwva02_01/colofon.php
- FoLiA-data files from Nederlab: no version number, just a date (19-09-2017).
- Separated actual letters from notes, foreword etcetera.

Meertens Instituut

Editing data

- Tokenization and Tagging with *Adelheid* (versie 1.0).
- *Adelheid* was bound to give a lot of mistakes.
- We used it as a tool to preproces the data for manual correction (maar veel fouten en onvolledig).

Editing data

(Semi) Manual enrichment: pos-tagging and socio-linguistic information

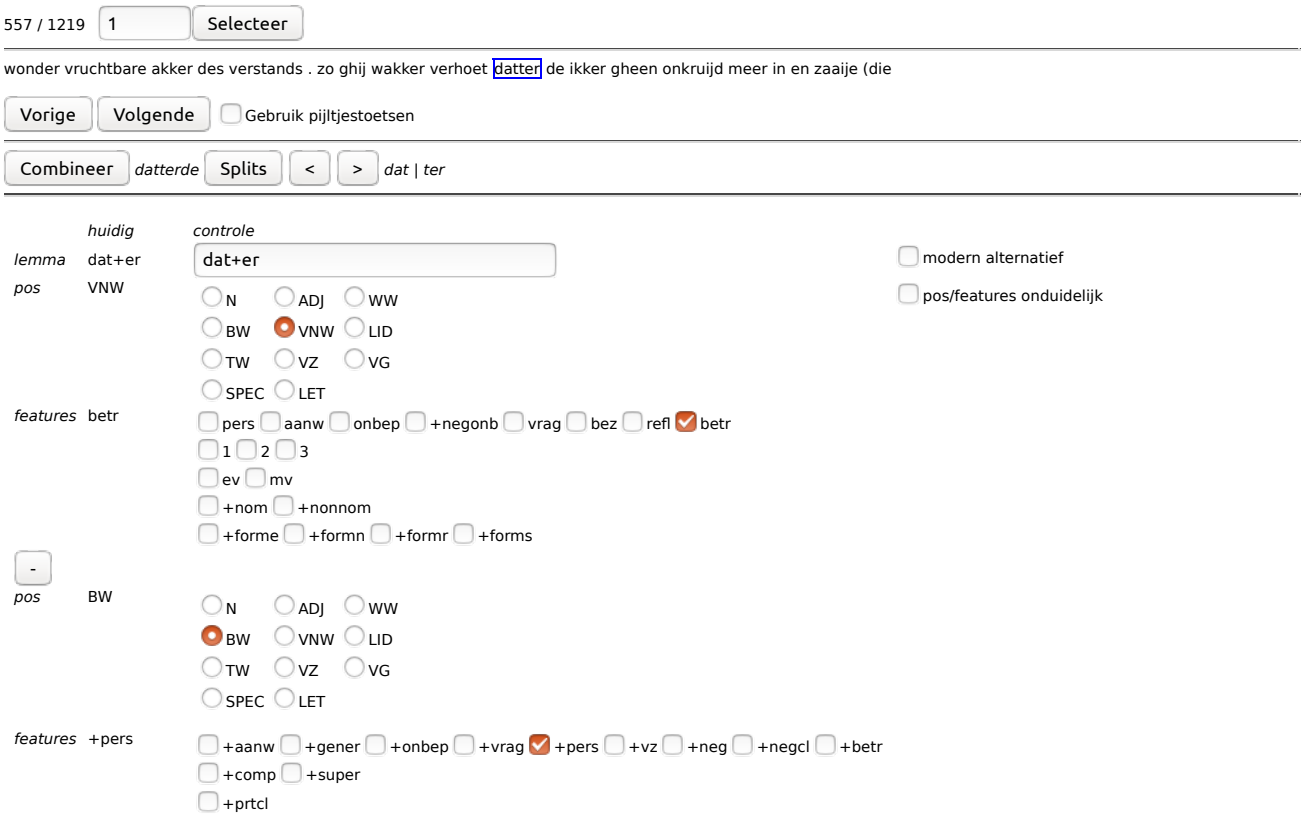
By 10 (R)MAstudents

Document characteristics	
Category	business, personal
Type	regular, appendix
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
Correspondent characteristics	
Group	name
Individual	name, birth/death date, gender, occupation, literary author, relation to P.C. Hooft
Letter segmentation	
Introductory greeting, opening (optional), narrative, closing (optional), final greeting	

Table 1: Sociolinguistic annotation set.

Meertens Instituut

Editing data



POS-tagging data: Gustave tool

Availability of the corpus

- The corpus, including the manual of the tagset and the way in which the sociolinguistics enrichment is defined will become part of Nederlab
- Problem: the letters of P.C. Hooft are from an edited volume (part of DBNL) which is copyrighted.

Meertens Instituut

Provenance for researchers:

- Metadata about the source file (preprocessing):
  - which version of the text (which print? Date of original tekst? Author? Editor? Etc.)
- Information of the input text into the tools:
  - Editorial matter separated or not?
  - Notes separated or not?

Meertens Instituut

Universiteit Utrecht

Provenance for researchers:

- Information about the enrichment
  - What information has been added?
  - How has that information been added? Manually or automatically?
    - If automatically: with which tool?
    - If manually: according to what protocol?
  - How is that information defined?

Meertens Instituut

Universiteit Utrecht

Provenance for researchers:

- Information about the query:
  - What was the query used to search for the data?
  - Are (all or part of) the data manually checked or not?

Meertens Instituut

Thanks for Your Attention!

Meertens Instituut