Provenance
Language Dynamics
in
the Dutch Golden Age

Marjo van Koppen
DANS, Den Haag, 2018-09-03
Overview

• Input data
• Editing data
• Availability of the data
• Provenance for researchers
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.
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).
(Semi) Manual enrichment: pos-tagging and socio-linguistic information

By 10 (R)MA students

<table>
<thead>
<tr>
<th>Document characteristics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
<td>business, personal</td>
</tr>
<tr>
<td>Type</td>
<td>regular, appendix</td>
</tr>
<tr>
<td>Goal</td>
<td>express thanks, compliment, excuse, ask a favour, ask information, ask advice, admonish, inform, remember, persuade, order, allow, invite</td>
</tr>
<tr>
<td>Topic</td>
<td>business, literature, domestic affairs, love, death, news, religion/ethics</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correspondent characteristics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>name</td>
</tr>
<tr>
<td>Individual</td>
<td>name, birth/death date, gender, occupation, literary author, relation to P.C. Hooft</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Letter segmentation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Introductory greeting, opening (optional), narrative, closing (optional), final greeting</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Sociolinguistic annotation set.
wonder vruchtbare akker des verstands. zo ghij wakker verhoet [datter] de ikker gheen onkruijd meer in en zaaije (die

<table>
<thead>
<tr>
<th>Selecteer</th>
<th>Vorige</th>
<th>Volgende</th>
<th>□ Gebruik pijltjestoetsen</th>
<th>Combineer</th>
<th>datterde</th>
<th>Splits</th>
<th>&lt;</th>
<th>&gt;</th>
<th>dat</th>
<th>ter</th>
</tr>
</thead>
<tbody>
<tr>
<td>huidig</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lemma</td>
<td>dat+er</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pos</td>
<td>VNW</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>controle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pos</td>
<td>BW</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>features</td>
<td>betr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>features</td>
<td>+pers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

□ N  □ ADJ  □ WW
□ BW □ VNW □ LID
□ TW  □ VZ  □ VG
□ SPEC □ LET
□ pers □ aanw □ onbep □ +negonb □ vrag □ bez □ refl □ retr
□ 1 □ 2 □ 3
□ ev □ mv
□ +nom □ +nonnom
□ +forme □ +formn □ +formr □ +forms
□ +aanw □ +gener □ +onbep □ +vrag □ +pers □ +vz □ +neg □ +negcl □ +betr
□ +comp □ +super
□ +prtcl

□ modern alternatief
□ pos/features onduidelijk
wonder vruchtbare akker des verstands. zo ghij wakker verhoet [dat+er] de ikker gheen onkruijd meer in en zaaije (die...
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.
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?
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?
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?
Thanks for Your Attention!