Here it becomes clear that cluster 1 prefers thrillers and police novels, while cluster 2 has a less-focussed interest in family, writing and the country. It is worthwhile to repeat that these clusters of content words result from clustering reviewers on the basis of function words.

Conclusion

Taken together, the correlations and the exploratory analysis show that there is a relation between the function words that people use and their preferences for books. This relation still holds at the level of part-of-speech tags. This clearly shows that the word usage that helps tell authors apart is to some extent related to artistic preference. A possible explanation would be that the reviewers unconsciously imitate the books they read in their use of function words. That seems unlikely, among other reasons because the effect is also visible when we just look at the reviews in a single genre (second and third column of table 1). The more likely explanation is that function word usage is at least in part determined by artistic preference and related personality characteristics. The 'fingerprint' metaphor that is often used in this context, with its suggestion of an essentially random identifier, unlikely to be related to artistic preference, must therefore be considered as inappropriate.

Literature

Boot, P. (2014). *Dimensions of literary appreciation. Word use and ratings on a book discussion site*. Digital Humanities 2014. Retrieved from http://dharchive.org/paper/DH2014/Paper-825.xml

Boot, P., Zijlstra, H., & Geenen, R. (2017, in press). The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics, 6*(1).

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing, 17*(3), 267-287.

Cantador, I., Fernández-Tobías, I., Bellogín, A., Kosinski, M., & Stillwell, D. (2013). *Relating Personality Types with User Preferences in Multiple Entertainment Domains.* Proceedings of the 1st Workshop on Emotions and Personality in Personalized Services (EMPIRE 2013), at the 21st Conference on User Modeling, Adaptation and Personalization (UMAP 2013).

Gabrielatos, C., & Marchi, A. (2011). Keyness: Matching metrics to definitions. *Theoretical-methodological challenges in corpus approaches to discourse studies-and some ways of addressing them*.

Noecker, J., Ryan, M., & Juola, P. (2013). Psychological profiling through textual analysis. *Literary and Linguistic Computing, 28*(3), 382-387.

Székely, G. J., & Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis, 117*, 193-213.

3. Corpus enrichment for 17th century Dutch: a pilot study

Feike Dietz[1], Marjo van Koppen[2], Irene Kramer[1] and Marijn Schraagen[2]
[1]Institute for Cultural Inquiry, [2]Utrecht Institute of Linguistics OTS
Utrecht University

1 Introduction

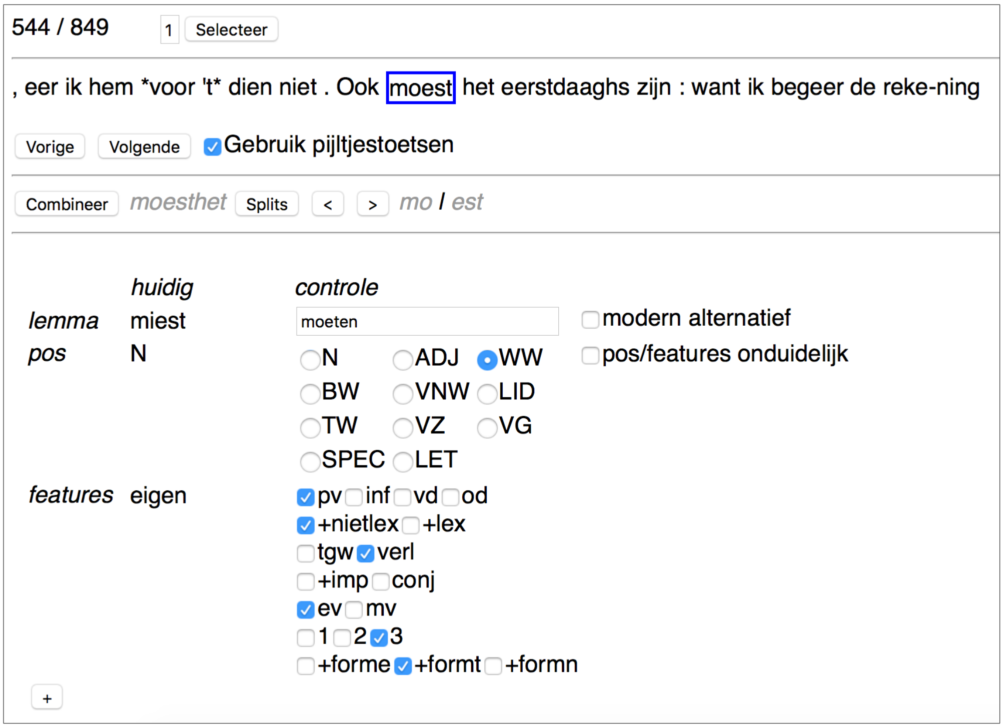
The Dutch language in the 17th century was a mixture of fading linguistic properties from the preceding language phase, Middle Dutch, and upcoming new ways to construct words and sentences. Within these language dynamics we observe a type of language variation that has rarely been addressed before: variation within individual language users (intra-author variation). The aim of the current project is to describe and analyse in detail the linguistic and literary/rhetorical contexts in which intra-author variation occurs. As a prerequisite, the data needs to be annotated linguistically, using part of speech (POS) information and (morpho-) syntactic structure, and sociolinguistically, describing various factors that influence language use.

In a pilot project we restrict our research to the letters of the famous Dutch author and politician P.C. Hooft, written between 1600 and 1638. This collection is relatively large (approximately 800 letters, ~300.000 words) and contains sociolinguistic variation in type of correspondent and type of letter. The corpus can be used, i.a., to study the loss of negative concord in Dutch, which is observed in Hooft's letters from this period (Paardekooper, 2016).

As a starting point for obtaining POS tags, the Adelheid tagger for Middle Dutch (van Halteren and Rem, 2013) is used. Because the tagger is trained on Middle Dutch, the results are not highly accurate for 17th century texts. Therefore, a correction procedure for POS-tags and lemmas is performed by human annotators. Additionally, the annotators provide the necessary sociolinguistic information about letters and correspondents. When annotation is completed, a detailed and systematic analysis of linguistic phenomena will become feasible.

2 Approach

The source data is available in a diplomatic edition (Van Tricht, 1976). We use this edition after separating Hoofts original seventeenth century texts from the metadata (page numbers, foot notes, annotations).



544 / 849 1 Selecteer

, eer ik hem *voor 't* dien niet . Ook moest het eerstdaaghs zijn : want ik begeer de reke-ning

☑ Gebruik pijltjestoetsen

moesthet

mo / est

	huidig	*controle*	
lemma	miest	moeten	☐ modern alternatief
pos	N	○ N ○ ADJ ● WW ○ BW ○ VNW ○ LID ○ TW ○ VZ ○ VG ○ SPEC ○ LET	☐ pos/features onduidelijk
features	eigen	☑ pv ☐ inf ☐ vd ☐ od ☑ +nietlex ☐ +lex ☐ tgw ☑ verl ☐ +imp ☐ conj ☑ ev ☐ mv ☐ 1 ☐ 2 ☑ 3 ☐ +forme ☑ +formt ☐ +formn	

Figure 1: Example of the newly developed annotation tool

2.1 Part-of-Speech tagging

A collaboration with the Nederlab project (Brugman et al., 2016) is established to increase availability of the enriched corpus, by including the POS tagging and sociolinguistic metadata in the Nederlab research infrastructure. The integration necessitates conversion of the CRM tagset used by Adelheid to the CGN tagset used by Nederlab. Additionally, the tagging needs to be represented into the FoLiA

XML format for linguistic annotation (van Gompel and Reynaert, 2013). The CRM tagset is more extensive than CGN, notably in the use of surface form features such as form-e (words ending in -e). Surface form features are related to case marking, which is an important aspect in the study of linguistic variation in 17th century Dutch. Therefore, we decided to keep these features in the mapping to CGN tags (see Figure 1).

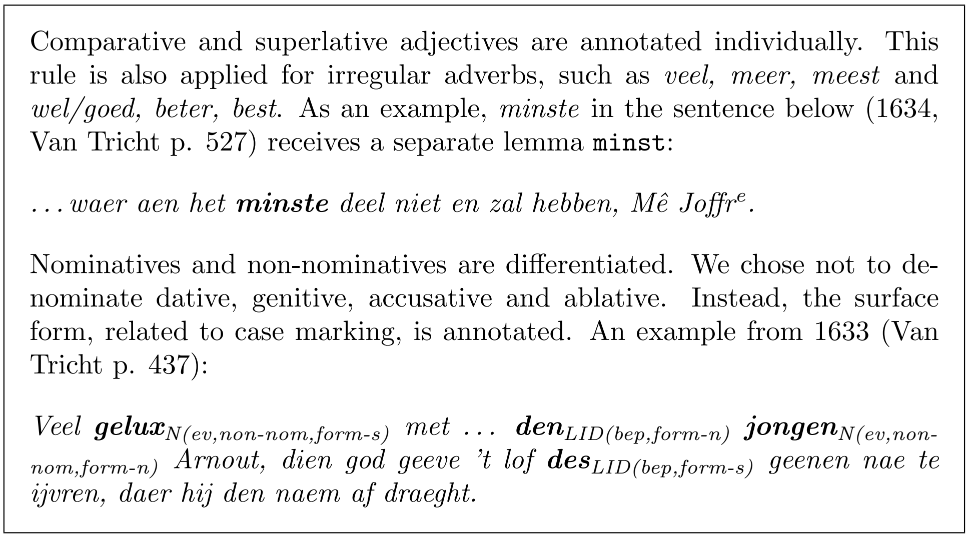
2.2 Sociolinguistic tagging

A key hypothesis in intra-author variation is the influence of sociological factors on linguistic choices. To evaluate this hypothesis systematically, all letters are being annotated with the following information:

- Goal: express thanks, ask advice, recommend, invite
- Topic: politics, religion, personal affairs, administration
- For individual correspondents:
  - name, gender, year of birth and death
  - status of correspondent as literary author
  - relation to Hooft: family members, literary friends, politicians, etc.
- For group correspondents:
  - name
  - domain: government, financial or legal institutions, civil associations
- Letter structure: greeting, introduction, narratio, closing formulas

2.3 Annotation process

A tool has been developed (see Figure 1) to perform POS and sociolinguistic annotation in an efficient way. A pool of annotators is available for the task, which will perform partly overlapping annotations to allow for agreement measurements. The annotation process is currently ongoing. A protocol has been developed to guide the post-correction process (see Figure 2 for examples).



Comparative and superlative adjectives are annotated individually. This rule is also applied for irregular adverbs, such as *veel, meer, meest* and *wel/goed, beter, best.* As an example, *minste* in the sentence below (1634, Van Tricht p. 527) receives a separate lemma `minst`:

. . . waer aen het ***minste*** *deel niet en zal hebben, Mê Joffr*^{e}.

Nominatives and non-nominatives are differentiated. We chose not to denominate dative, genitive, accusative and ablative. Instead, the surface form, related to case marking, is annotated. An example from 1633 (Van Tricht p. 437):

Veel ***gelux***$_{N(ev,non\text{-}nom,form\text{-}s)}$ *met . . .* ***den***$_{LID(bep,form\text{-}n)}$ ***jongen***$_{N(ev,non\text{-}nom,form\text{-}n)}$ *Arnout, dien god geeve 't lof* ***des***$_{LID(bep,form\text{-}s)}$ *geenen nae te ijvren, daer hij den naem af draeght.*

Figure 2: Annotation guideline examples

3 Analysis

In related work (Kramer, 2016) the use of negation by Hooft has been studied manually. Kramer shows that Hooft uses mostly single negation in different syntactical environments (subclauses, inversion, main clauses, local negation, V1 (verb-initial) sentences). Additionally, the negation particle *niet* can be used as alternative for the noun *nothing*. Furthermore, Hooft uses bipartite negation in almost all syntactical environments as well (all except in V1). In Kramer's research, not one environment seemed to particularly ask for the use of bipartite negation. This research, however, encompassed only 107 letters. The fully annotated corpus will allow a more quantitative analysis, as well as a larger range and higher level of detail of linguistic phenomena.

Nobels and Rutten (2014) note the influence of gender and social class on negation (p. 41): ‘while single negation spread from the north to the south, it also turned into a social variant, as the upper ranks in society and male letter writers seemed to be quicker to pick up on the incoming variant than the lower ranks and female letter writers’. Nobels and Rutten (2014) also note (p. 43) that traditions in letter writing affect linguistic development: ‘fixed formulae were memorized as a whole (or copied) by writers from any social background. These fixed formulae occur in certain parts of the letters, mostly in the beginning and the ending’. With the current annotation effort, this type of observations can be studied systematically.

References

Brugman, H., Reynaert, M., van der Sijs, N., van Stipriaan, R., Tjong Kim Sang, E., and van den Bosch, A. (2016). Nederlab: Towards a single portal and research environment for diachronic Dutch text corpora. In Proceedings of LREC 2016.

van Gompel, M. and Reynaert, M. (2013). Folia: A practical xml format for linguistic annotation-a descriptive and comparative study. Computational Linguistics in the Netherlands Journal, 3:63–81.

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

Kramer, I. (2016). Variatie in negatie, een syntactisch en retorische analyse van het gebruik van enkele en tweeledige negatie in de brieven van P.C. Hooft van 1633 tot 1638 aan Joost Baek en Tesselschade Roemersdochter Visser. BA thesis, Universiteit Utrecht.

Nobels, J. and Rutten, G. (2014). Language norms and language use in seventeenth-century Dutch: negation and the genitive. In Rutten, G., editor, Norms and usage in language history, 1600-1900. A sociolinguistic and comparative perspective., pages 21–48. John Benjamins Publishing Company.

Paardekooper, P. (2016). Bloei en ondergang van onbeperkt ne/en, vooral dat bij niet-woorden. Neerlandistiek.nl.

van Tricht, H. (1976). De briefwisseling van Pieter Corneliszoon Hooft. Tjeenk Willink / Noorduijn.