Het gewicht van de auteur

Stylometrische auteursherkenning in Middelnederlandse literatuur

Deze publicatie heeft aantoonbaar een peer-reviewprocedure doorlopen die beantwoordt aan de internationale wetenschappelijke standaarden.

GPRC
Guaranteed
Peer Reviewed
Content
boekenvak.be/gprc

Dit werk gaat terug op een doctoraal proefschrift dat op 23 mei 2012 werd verdedigd aan de de departementen Taal- en Letterkunde van de Universiteit Antwerpen. De leden van de jury waren:

- Prof. dr. Frank Willaert (promotor, Universiteit Antwerpen)
- Prof. dr. Walter Daelemans (promotor, Universiteit Antwerpen)
- Prof. dr. Dominiek Sandra (voorzitter, Universiteit Antwerpen)
- Prof. dr. Veerle Fraeters (secretaris, Universiteit Antwerpen)
- Dr. Bart Besamusca (Universiteit Utrecht)
- Prof. dr. Herman Brinkman (Universiteit van Amsterdam)
- Prof. dr. Antal van den Bosch (Radboud Universiteit Nijmegen)
- Prof. dr. Thom Mertens (Universiteit Antwerpen)

Het onderzoek werd uitgevoerd aan volgende onderzoeksgroepen van de Universiteit Antwerpen: het Instituut voor de Studie van de Letterkunde in de Nederlanden en het CLiPS Computational linguistics and psycholinguistics research center.

Universiteit Antwerpen
ISLN Instituut voor de Studie van de Letterkunde in de Nederlanden
CLiPS

Het gewicht van de auteur

Stylometrische auteursherkenning in Middelnederlandse literatuur

Mike Kestemont

Gent, Koninklijke Academie voor
Nederlandse Taal- en Letterkunde, 2013

Dit boek is gedrukt op zuurvrij en permanent papier dat beantwoordt aan de norm ISO 9706.

Omslagontwerp: Ae Jin Huys

Omslagillustratie: Nicolas Leus

Vormgeving: Mike Kestemont, met behulp van de LaTeX-bundel CLASSICTHESIS door © André Miede, *A Classic Thesis Style. An Homage to the Elements of Typographic Style* (2010), http://code.google.com/p/classicthesis/.

Alle rechten voorbehouden. Niets uit deze uitgave mag worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand, of openbaar gemaakt, in enige vorm of op enige wijze, hetzij elektronisch, mechanisch, door fotokopieën, opnamen, of op enige andere manier, zonder voorafgaande schriftelijke toestemming van de auteur.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, by photo-copying, recording, or otherwise, without the prior written permission of the author.

Uitgegeven met de steun van de Universitaire Stichting van België, de Vlaamse Overheid en het Fonds voor Wetenschappelijk Onderzoek Vlaanderen.

Fonds Wetenschappelijk Onderzoek
Research Foundation – Flanders

Met steun van de Vlaamse overheid

Deze studie werd in 2012 bekroond met de Frans van Cauwelaertprijs, een fondsprijs van de Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten.

© 2013 Mike Kestemont & Koninklijke Academie voor Nederlandse Taal- en Letterkunde

ISSN 1378-6563
ISBN 978-90-72474-91-1
D/2013/0228/3

Studies op het gebied van de oudere Nederlandse letterkunde 5

Singulari gaudio, et lassate mentis unico solamini,
ille cuius vita sine te mors est: quid amplius
quam seipsum quantum corpore et anima valet.

"Aan de uitzonderlijke vreugde en de enige troost
voor een vermoeide geest, hij voor wie het leven
zonder jou de dood is: wat meer dan hijzelf,
in zoverre hij sterk is in lichaam en ziel."

Naar de *Epistolae duorum amantium*
in handschrift Troyes, Bibliothèque
municipale, MS 1452, fols. 159r-167v.

Dit werk is in liefde opgedragen
aan Hannelore Vanrobaeys.

DANKWOORD

Ik had dit proefschrift niet kunnen schrijven zonder de jarenlange steun van mijn ouders, Hilde Van Wateghem en René Kestemont. Zij hebben steevast blijk gegeven van een blind vertrouwen in mij, zelfs wanneer mijn leven de gekste wendingen nam. Hun steun was altijd onvoorwaardelijk en ik hoop dat zij beseffen hoe dankbaar ik hen daarvoor ben. Samen met mijn ouders, mijn grootvader Robert Van Wateghem en Paula Kestemont-De Meyst ben ik nog dagelijks mijn overleden grootouders indachtig, Anna Nolst, Christiane Liefooghe en Wilfried Kestemont. Dat zij het verschijnen van dit werk niet meer hebben mogen meemaken, neemt niet weg dat zij ons in gedachten nabij blijven. Even erkentelijk ben ik mijn lieve zussen, Britta en Lina, voor hun morele steun tijdens de afgelopen jaren en mijn zwagers, Rien en Bram, voor hun technische bijstand bij veel gelegenheden.

Vanzelfsprekend had dit werk nooit tot stand kunnen komen zonder mijn tandem van promotoren, Frank Willaert en Walter Daelemans. Nu ik kan terugblikken op mijn promotietraject, besef ik eens te meer dat ik nergens ter wereld mijn onderzoek beter had kunnen uitvoeren dan onder hun vakkundige en complementaire begeleiding. Walter en Frank behoren tot de internationale top van hun respectievelijke vakgebieden, maar naast hun grote onderzoeksbekwaamheid bewonder ik hen vooral om hun wetenschappelijke attitude. Zij geven blijk van een haast monastieke toewijding en loyauteit ten aanzien van de wetenschap. Ik blijf bewondering hebben voor de verfijnde stijl en het grote altruïsme waarmee Frank Willaert door het leven gaat. Als ik van Frank één zaak zal onthouden, is het zijn hoofse *savoir vivre* en de verzoenende wijze waarop hij de filologie bedrijft. Van Walter

Daelemans zal mij vooral zijn pragmatische vooruitgangsgeloof bijblijven. Walter belichaamt in mijn ogen Galileo Galilei's optimistische wetenschapsideaal dat stelt dat men eerst het meetbare moet meten en vervolgens het onmeetbare meetbaar moet maken.

In de medioneerlandistiek ben ik weinigen zoveel dank verschuldigd als Remco Sleiderink, *een minen vrient van Bruesele*. Het was Remco die mij in mijn studentendagen aan de KUB de liefde voor de filologie heeft bijgebracht en mij steeds met een onnavolgbaar enthousiasme heeft gesteund in mijn onderzoek. Ik waardeer ten zeerste onze avondvullende gesprekken op de Brusselse Vismarkt, die inmiddels een gewoonte zijn geworden die ik niet graag zou opgeven. Remco, *merci*. Evenzeer enthousiasmerend waren de colleges van Jef Janssens, een andere sympathieke Brusselaar aan wie ik veel dank ben verschuldigd. Ik beschouw het als een privilege dat ik ook na mijn studies nog geregeld een beroep heb kunnen doen op Jefs eruditie en gastvrijheid.

Ik heb goede herinneringen aan mijn Antwerpse werkomgeving, in het bijzonder de jonge onderzoeksgroep van medioneerlandici die Frank Willaert de laatste jaren heeft uitgebouwd. Binnen deze onderzoekshabitat heb ik meer dan eens een aangename samenwerking met collega's kunnen aangaan. Bedankt, Alexia, Bram, Elisabeth en Patricia, voor jullie grote collegialiteit. Zeker mijn kantoorgenote Elisabeth heeft mij meer dan eens praktisch uit de nood geholpen. Andere leden van de wekelijkse *labmeeting* liggen mij evenzeer nauw aan het hart: Daniël, Daniëlle, Kees, Markus, Thom en Veerle. Ook deze Ruusbroecgenoten wil ik danken voor de vruchtbare samenwerking en de vele stimulerende discussies.

Mijn bijzondere dank gaat ook uit naar de CLiPS'ers van de Lange Winkelstraat die mij van in het begin en zonder de minste aarzeling hebben opgenomen in hun hechte onderzoeksteam. Zij hebben mij door de jaren heen veel meer bijgeleerd dan ik hier kan weergeven. Dank jullie wel, Claudia, Frederik, Guy, Hanne, Inge, Iris, Janneke, Kathy, Lieve, Renate, Roser, Sarah, Tom en Vincent. Ook via de collega's van de *Atila*-gemeenschap heb ik de wereld van de computationele linguïstiek in de afgelopen jaren leren kennen als een warme omgeving, waarin innovatie, precisie en vriendschap op eigenzinnige wijze hand in hand gaan.

Ook veel andere collega's aan de UA hebben gemaakt dat mijn onderzoek de afgelopen jaren meer aanvoelde als een hobby dan als arbeid. Deze collega's zijn stuk voor stuk ongelooflijke persoonlijkheden die mijn blik op het leven – maar ook op de Scheldestad en haar inwoners – ingrijpend beïnvloed hebben. Dank dat ik jullie deur te pas en te onpas heb mogen platlopen: Adelheid, Astrid, Bram, Ciska, Dirk, Geert, Jasper, Lien, Maartje, de 'Mechelen-mensen', Peter, Pim, Reinhild, Rita, Sara V., Sarah B., Sarah V.H., ... Ook verschillende buitenlandse vakgenoten hebben mij herhaaldelijk met raad en daad bijgestaan: Antal van den Bosch, André Bouwman, Bart Besamusca, Dirk Schoenaers,

Godfried Croenen, Herman Brinkman, Johan Oosterman, Jos Biemans, Karina van Dalen-Oskam, Marjolein Hogenbirk, Renée Gabriël, Robert Stein, Willem Kuiper, ... Antal, Bart en Herman wil ik trouwens net als Thom en Veerle hartelijk bedanken omdat zij zich als juryleden hebben geëngageerd om mijn proefschrift na te lezen. Ik sta bij hen in het krijt. Ook Dominiek Sandra, de voorzitter van mijn doctoraatsjury, wil ik danken voor zijn niet-aflatende enthousiasme en de talloze keren dat zijn deur voor mij openstond.

Ik durf verschillende van mijn collega's inmiddels tot mijn vriendenkring rekenen. Dat geldt in het bijzonder voor Kim Luyckx en Thomas Crombez: het is wat mij betreft een voorrecht geweest om samen met hen het project van de Digitale Geesteswetenschappen gestalte te mogen geven aan de UA. Bijzondere vermelding verdienen de vele collega's die mij belangeloos hun digitale edities en andere data voor dit onderzoek ter beschikking hebben gesteld. Zonder deze tastbare hulp had ik dit onderzoeksproject nooit tot een goed einde kunnen brengen. Bart Besamusca, David Johnson, Dirk Geirnaert, Frank Brandsma, Geert Claassens, Herman Brinkman, Karina van Dalen-Oskam, Marjolein Hogenbirk, Reindert van Eekelen, Willem Kuiper, ...: ik dank jullie uit de grond van mijn hart voor jullie vitale bijdrage aan dit onderzoek. Ook verschillende medewerkers van het *Instituut voor Nederlandse Lexicologie* (Dirk Geirnaert, Jesse de Does, Katrien Depuydt, ...) ben ik meer dan erkentelijk. Ook de mensen van het onderzoekssecretariaat van mijn faculteit wil ik hartelijk danken voor de administratieve bijstand.

Het onderzoek waarvan de resultaten in deze studie worden beschreven, kon ik uitvoeren tijdens een mandaat als predoctoraal aspirant van het Fonds voor Wetenschappelijk Onderzoek Vlaanderen. Het FWO is voor mij steeds als een correcte en ondersteunende werkgever opgetreden en ik ben het Fonds dan ook zeer erkentelijk voor dit wetenschappelijk mecenaat. Ik kan tevens dankbaar melden dat deze publicatie is uitgegeven met de financiële steun van de Universitaire Stichting van België. Ook de vast secretaris, Willy Vandeweghe, en de uitgeverij van de Koninklijke Academie voor Nederlandse Taal- en Letterkunde wil ik danken voor het uitbrengen van dit boek. Een bijzonder woord van dank gaat uit naar Edward Vanhoutte voor de goede zorgen voor mijn manuscript.

Verschillende hoofdstukken in deze studie gaan terug op artikelen die reeds eerder werden gepubliceerd in vaktijdschriften, die hebben toegestaan dat dit werk in de onderhavige studie verwerkt zou worden. Ik wil de redacties, medewerkers en anonieme beoordelaars van volgende tijdschriften en hun uitgeverijen dan ook hartelijk danken: *Belgisch Tijdschrift voor Filologie en Geschiedenis*, *Digital Philology: A Journal of Medieval Cultures* (John Hopkins University Press), *Journal of Quantitative Linguistics* (Taylor & Francis), *Queeste: Tijdschrift over middeleeuwse letterkunde in de Nederlanden* (Uitgeverij Verloren) en *Tijd-*

schrift voor Nederlandse Taal- en Letterkunde (Uitgeverij Verloren). Een bijzonder woord van dank gaat uit naar Anja van Leusden (Uitgeverij Verloren), met wie het steeds prettig samenwerken is.

Tot slot wil ik ook de volgende bibliotheken danken, omdat zij toestemming hebben gegeven voor de reproductie van enkele afbeeldingen van preciosa uit hun collecties: de Stedelijke Musea Kortrijk (Sylvie De Coster), de Universiteitsbibliotheek van Gent (Hendrik Defoort), de Universiteitsbibliotheek van Leiden (André Bouwman), de Universiteitsbibliotheek van Leuven (An Smets) en de bibliotheek van het Fürst zu Bentheimsche Domänenkammer te Steinfurt (Lieselotte Meyeringh).

Anderlecht/Antwerpen, 12 september 2012

INHOUDSOPGAVE

I DE AANLOOP 1

1 VELE BOUKE 3
- 1.1 De dood van de auteur 3
  - 1.1.1 Een schrale oogst 4
  - 1.1.2 De blinde vlek van het auteurschap 6
  - 1.1.3 Probleemstelling 8
- 1.2 Methodologie 12
  - 1.2.1 Kwantitatief versus computationeel 12
  - 1.2.2 Transdisciplinariteit 14

2 DE HAND VAN DE MEESTER 17
- 2.1 Niet-traditionele auteursherkenning 17
  - 2.1.1 Wat voorafging 17
  - 2.1.2 Het functiewoord 19
  - 2.1.3 Het onbewuste van de schrijver 21
  - 2.1.4 Lettermodellen 23
  - 2.1.5 Van functiewoord naar functor 24
- 2.2 De medioneerlandistiek 27
  - 2.2.1 De eerste zwaluwen 28
  - 2.2.2 Delta voor Walewein 30
  - 2.2.3 Tien geboden 33

3 DE JACHT OP DE KOPIIST 39
- 3.1 Een dozijn Rijmbijbels 39
- 3.2 Van tekst tot tabel 43
  - 3.2.1 Steekproeven 43
  - 3.2.2 Knikkeren met woorden 45
  - 3.2.3 Kolomselectie 51
- 3.3 Principal Components Analysis 54
- 3.4 Rijmbijbel-analyses 59
- 3.5 Kopiistenherkenning? 65

II HET NOODRIJM 69

4 DE SPIEGEL 71
- 4.1 Het rijm als reddingsboei 71
  - 4.1.1 Een atelier te Damme 72
  - 4.1.2 Lemmatisering 75
  - 4.1.3 De Wetten van Zipf 78
- 4.2 Cijferen voor gevorderden 84
  - 4.2.1 Tekstclassificatie 84
  - 4.2.2 Geheugen-gebaseerd leren 87
  - 4.2.3 Spiegel-experimenten 88
- 4.3 Conclusie 93

5 VELTHEM ET AL. 97

5.1 De Voortzetting 97
5.1.1 Velthem 97
5.1.2 Een verwarrende matrix 101
5.2 Het vierde boek 104
5.2.1 De Guldensporenslag 104
5.2.2 Opnieuw de dichtste buur 106
5.2.3 Goede buren en valse vrienden 108
5.2.4 De Vijfde Partie 111
5.2.5 Een palimpsest? 114
5.2.6 De Vlaamse brontekst 121
5.3 Conclusie 125
6 MAERLANT IN MEERVOUD 127
6.1 Een oeuvre van formaat 127
6.2 Spreiding 130
6.2.1 Inverse Document Frequency 136
6.2.2 De variatie-coëfficient 139
6.3 Maerlants genre-driehoek 141
6.4 Het genre-gevaar 146

III ONGEBAANDE WEGEN 151
7 BOENDALES SCHOOL 153
7.1 Jan van Boendale 153
7.1.1 Jan, gheheten Clerk 153
7.1.2 Boendale erbij 157
7.2 De klasse van Boendale 160
7.2.1 Naïève Delta? 162
7.2.2 Delta-experimenten 167
7.3 De didactische teksten 170
7.4 Het Complex van de Yeesten 179
7.4.1 De ontstaansgeschiedenis van de Brabantse yeesten 180
7.4.2 De toeschrijving van de Yeesten 183
7.4.3 Rondom de Yeesten 187
7.4.4 Analyses 190
7.4.5 Nicolaas de clerk? 194
7.5 Conclusie 197
8 EEN WOUT SONDER GENADE 199
8.1 Codex-Burgsteinfurt 200
8.2 Het boek van Lancelot 211
8.2.1 Centroïdes 219
8.2.2 De kernteksten 223
8.2.3 De interpolaties 229
8.2.4 De wereld buiten de compilatie 235
8.3 Conclusie 240
9 KONING EN KEIZER 243
9.1 Coninc ende keyser mede 243

9.1.1 Twee matières 243
9.1.2 Eindjes aan elkaar knopen 245
9.1.3 Clusters 250
9.1.4 Rangsom 255
9.2 De zaak-Heeroma 259
9.3 Eerherstel voor Icarus? 264
9.3.1 Reminiscenties 264
9.3.2 Ohrenphilologie? 268
9.4 Conclusie 272

IV TOT SLOT 273
10 BESLUIT 275
10.1 Wat is onderzocht? 275
10.1.1 Hoogfrequente rijmwoorden 275
10.1.2 Boendale 279
10.1.3 De Artur- en Karelepiek 281
10.2 Wat is (nog) niet onderzocht? 285
11 APPENDIX 289
11.1 Annotatiepraktijk 289
11.1.1 Voorbewerking 289
11.1.2 Tokenisering 290
11.1.3 Lemmatisering 293

BIBLIOGRAFIE 299

LIJST VAN FIGUREN

Figuur 1 Leuven, Universiteitsbibliotheek, Ms. 1488, olim: Fragm. 598 5
Figuur 2 Delta-experiment op de *Walewein* (Van Dalen-Oskam & Van Zundert 2007) 37
Figuur 3 Detail van fol. 152v in het *Walewein*-handschrift Leiden, Universiteitsbibliotheek, Ltk. 195 38
Figuur 4 PCA van 6 parallelle samples uit de *Rijmbijbel*-handschriften B en D 60
Figuur 5 PCA van 6 parallelle samples uit de *Rijmbijbel*-handschriften G en E 62
Figuur 6 PCA van 6 parallelle samples uit de *Rijmbijbel*-handschriften A en D 63
Figuur 7 De vijftig meest frequente rijmwoorden in Utenbroekes Tweede Partie 81
Figuur 8 De vijftig meest frequente rijmwoorden in Maerlants Derde Partie 82
Figuur 9 Grafische voorstelling van enkele experimenten uit Tab. 13 92
Figuur 10 Openingsfolium van het Vijfde Partie-handschrift (Leiden, Universiteitsbibliotheek, bpl e 14) 99
Figuur 11 Naïeve voorstelling van classificatie volgens het *nearest neighbour*-model 107
Figuur 12 *Leave one out*-experimenten met Velthems *P4* vs. Utenbroekes *P2* en Maerlants *P3* 109
Figuur 13 Naïeve voorstelling van fictieve situaties binnen een *nearest neighbours*-redenering 111
Figuur 14 Waardes van de $\delta\delta$ voor leave-one-out-experimenten van Velthems Vierde Partie vs. Utenbroekes Tweede Partie 112
Figuur 15 Waardes van de $\delta\delta$ voor leave-one-out-experimenten van Velthems Vierde Partie vs. Maerlants Derde Partie 113
Figuur 16 Waardes van $\delta\delta$ voor *windows* uit *P5*, gegeven de controle-tekst *P2* 115
Figuur 17 Waardes van $\delta\delta$ voor *windows* uit *P5*, gegeven de controle-tekst *P3* 116
Figuur 18 Fol. 71v-72r in het Heber-Serrure-handschrift (Gent, Universiteitsbibliotheek, hs. 1476) 118
Figuur 19 Bovendeel van de 3e kolom op fol. 49r in het Vijfde Partie-handschrift (Leiden, Universiteitsbibliotheek, bpl e 14) 122

Figuur 20 Auteursafbeelding, vermoedelijk van Jacob van Maerlant (Den Haag, Koninklijke Bibliotheek, ka xx) 128
Figuur 21 De 50 meest frequente rijmwoorden uit Maerlants oeuvre 133
Figuur 22 Grafiek van de rijmwoorden uit Maerlants oeuvre 134
Figuur 23 Fig. 22 na een logaritmische omzetting 135
Figuur 24 Rijmwoord-analyse van Maerlants oeuvre op basis van de *Inverse-Document Frequency* 137
Figuur 25 Rijmwoord-analyse van Maerlants oeuvre op basis van de *Inverse-Document Frequency* (gelimiteerde samples) 138
Figuur 26 Rijmwoord-analyse van Maerlants oeuvre op basis van de variatie-coëfficient (gelimiteerde samples) 141
Figuur 27 Rijmwoord-analyse van Maerlants oeuvre op basis van de varietie-coëfficient (gelimiteerde samples) 142
Figuur 28 Correspondentie-analyse van 68 samples (2.156 rijmwoorden) uit Maerlants oeuvre 144
Figuur 29 Correspondentie-analyse van 68 samples (2.156 rijmwoorden) uit Maerlants oeuvre (voorgesteld aan de hand van genre) 145
Figuur 30 Statistische test (*Tukeys Honestly Significant Difference*) van de genre-spreiding in Fig. 29 147
Figuur 31 Ladingen van rijmwoorden voor het experiment in Fig. 29 148
Figuur 32 Grafiek met de variatie-coëfficient van rijmwoorden in samples van 2.000 rijmwoorden uit de *Spiegel historiael* (Utenbroeke, Maerlant, Velthem) en Boendales *JT* en *LS* 159
Figuur 33 Pca van samples (3.000 rijmwoorden) uit *LS* en *JT* (Boendale), *P3* (Maerlant) en *P4P5* (Velthem) 160
Figuur 34 Pca van samples (3.000 rijmwoorden) uit *LS* en *JT* (Boendale), *P3* (Maerlant) en *P2* (Utenbroeke) 161
Figuur 35 Pca van samples (3.000 rijmwoorden) uit *LS* en *JT* (Boendale), *P2* (Utenbroeke) en *P4P5* (Velthem) 162
Figuur 36 Pca van samples (3000 rijmwoorden) uit *P2* (Utenbroeke), *P3* (Maerlant) en *P4P5* (Velthem) 163
Figuur 37 *Leave-one-out*-experimenten (3d-weergave) op samples uit *P3* en *P4P5* 168

Figuur 38 *Leave-one-out*-experimenten (3D-weergave) op samples uit *P2*, *P3* en *P4P5* 169
Figuur 39 *Leave-one-out*-experimenten (3D-weergave) op samples uit *P2*, *P3*, *P4P5*, *LS* en *JT* 170
Figuur 40 PCA van samples (2.000 rijmwoorden) uit *LS* en *JT* (zeker Boendale), *Mel*, *Doc*, *Wrak* ('Antwerpse School') en Maerlant (*P3*) als controleauteur 175
Figuur 41 PCA van samples (2.000 rijmwoorden) uit *LS* en *JT* (zeker Boendale), *Mel*, *Doc* en *Wrak* ('Antwerpse School') en Utenbroeke (*P2*) als controleauteur 176
Figuur 42 PCA van samples (2.000 rijmwoorden) uit *LS* en *JT* (zeker Boendale), *Mel*, *Doc* en *Wrak* ('Antwerpse School') en Velthem (*P4P5*) als controleauteur 177
Figuur 43 PCA van samples uit *LS* en *JT* (zeker Boendale) en *Mel*, *Doc* en *Wrak* ('Antwerpse School') 178
Figuur 44 *Leave-one-out*-experimenten (3D-weergave) op samples uit *P2*, *P3*, *P4P5*, *LS*, *JT*, *Mel*, *Doc* en *Wrak* 179
Figuur 45 CA van alle hier bestudeerde teksten uit de 'Antwerpse School', waaronder ook de historiografische (samples van 1.000 rijmwoorden) 191
Figuur 46 CA van alle hier bestudeerde teksten uit de 'Antwerpse School', waaronder ook de historiografische (samples van 2.000 rijmwoorden) 192
Figuur 47 CA van de teksten uit de 'Antwerpse School', waaronder ook de historiografische en Maerlants *P3* als controlemateriaal 193
Figuur 48 CA van de zuiver historiografische teksten uit de 'Antwerpse School' 194
Figuur 49 CA met ladingen van het historiografische materiaal uit de 'Antwerpse School', met uitzondering van *BY1-3* 196
Figuur 50 Fol. 62v-63r van codex BURGSTEINFURT, FÜRST ZU BENTHEIMISCHE SCHLOSSBIBLIOTHEK, B37 201
Figuur 51 Bovenkant van de tweede kolom op fol. 62v van het Burgsteinfurtse handschrift 202
Figuur 52 Delta-experiment met Maerlants *P3* als referentiemateriaal voor het Burgsteinfurtse handschrift 206
Figuur 53 Delta-experiment met Velthems *P4P5* als referentiemateriaal voor het Burgsteinfurtse handschrift 207
Figuur 54 CA van samples (3.000 rijmwoorden) uit *P3*, *MerlM*, *P4P5* en *MerlV* 208

Figuur 55 Dezelfde analyse als in Fig. 54 maar met de 'ladingen' van de rijmwoorden 210
Figuur 56 Delta-experiment met Velthems *P4P5* als referentiemateriaal voor het Burgsteinfurtse handschrift; Maerlants *P3* wordt gebruikt als controlecentroïde 212
Figuur 57 Laatste beschreven folium uit de Haagse *Lancelotcompilatie* (Den Haag, Koninklijke Bibliotheek, 129 a 10) 216
Figuur 58 Colofon van de Haagse *Lancelotcompilatie* (Den Haag, Koninklijke Bibliotheek, 129 a 10) 217
Figuur 59 Delta-experiment met Velthems *MerlV* als referentiemateriaal voor de *Lancelotcompilatie* 221
Figuur 60 Dubbele Delta-experiment met Velthems *MerlV* en *P4P5* als referentiemateriaal voor de *Lancelotcompilatie* 222
Figuur 61 Ca van samples (3.000 rijmwoorden) uit de kerndelen van de *Lancelotcompilatie* 225
Figuur 62 Ca van samples (3.000 rijmwoorden) uit de kerndelen van de *Lancelotcompilatie*, met slechts de 30 meest frequente rijmwoorden 226
Figuur 63 Ca van samples (3.000 rijmwoorden) uit de kerndelen van de *Lancelotcompilatie* en Velthems *MerlV*, met slechts de 30 meest frequente rijmwoorden 229
Figuur 64 Ca van samples (856 rijmwoorden) uit alle teksten in de *Lancelotcompilatie* 233
Figuur 65 Ca van samples (856 rijmwoorden) uit alle teksten in de *Lancelotcompilatie*, met uitsluiting van de kerndelen maar met *MerlV* 234
Figuur 66 Ca van samples (856 rijmwoorden) uit de geïnterpoleerde romans in de *Lancelotcompilatie* 235
Figuur 67 Ca alle teksten (samples van 1.000 rijmwoorden) uit de *Lancelotcompilatie*, alsook de Arturteksten die buiten dit handschrift zijn overgeleverd 239
Figuur 68 Ca van samples (1.000 rijmwoorden) uit alle geïnterpoleerde teksten uit de *Lancelotcompilatie*, alsook beide *Walewein*-delen 240
Figuur 69 Ca van samples (1.000 rijmwoorden) uit Vostaerts bijdrage aan de *Walewein*, de *Roman van Moriaen*, *Lantsloot vander Haghedochte* en de *Ferguut*. 241
Figuur 70 Titelbladzijde uit Incunabel A van *Karel ende Elegast* 249
Figuur 71 Cluster-analyse van samples uit de Middelnederlandse Artur- en Karelepiek 253

Figuur 72 Cluster-analyse van samples (ca. 1500 rijmwoorden) uit Mor, Hagh, Elegast, Ferg, WalA, WalB, VlaAiol en Merl 261
Figuur 73 Ca van kleine samples (ca. 700 rijmwoorden) uit Mor, Hagh, Elegast, Ferg en WalB 262
Figuur 74 Ca van de 50 vaakst voorkomende rijmwoorden in Ferg, WalB, Hagh, Mor en Elegast (samples ca. 1500 rijmwoorden), met weergave van de 'ladingen' van rijmwoorden 263

LIJST VAN TABELLEN

Tabel 1 Overzicht van de 15 parallelle handschriften in het *Rijmbijbel*-corpus 41
Tabel 2 Illustratie van de variatie in het *Rijmbijbel*-corpus 42
Tabel 3 De eerste 15 parallelle verzen uit de *Rijmbijbel*-handschriften A en D 44
Tabel 4 Drie samples (5 verzen) uit de parallelle *Rijmbijbel*-handschriften A en D 45
Tabel 5 De unigram-frequenties in de eerste 10 samples (5 regels) uit de parallelle *Rijmbijbel*-handschriften A en D 50
Tabel 6 Relatieve bigram-frequenties in de eerste 10 samples (5 regels) in de parallelle *Rijmbijbel*-handschriften A en D 55
Tabel 7 Overzicht van een reeks PCA-experimenten, telkens van 3 samples uit twee *Rijmbijbel*-handschriften 66
Tabel 8 Eerste 10 gelemmatiseerde rijmwoorden uit de Tweede Partie 79
Tabel 9 Eerste 10 gelemmatiseerde rijmwoorden uit de Derde Partie 80
Tabel 10 Fictieve dataset voor *spam filtering* via tekstclassificatie 86
Tabel 11 Illustratie van het *nearest neighbour*-algoritme voor de Tweede Partie en Derde Partie 89
Tabel 12 *Leave one out*-experiment op boekniveau voor de Tweede en Derde Partie 91
Tabel 13 Leave one out-experiment voor samples uit de Tweede en Derde Partie (variabele parameters) 96
Tabel 14 *Leave one out*-experimenten op samples uit de Tweede, Derde en Vijfde Partie 101
Tabel 15 *Confusion matrices* voor *leave one out*-experimenten op 45 samples uit de Tweede, Derde en Vijfde Partie 102

Tabel 16 *Confusion matrices* voor *leave one out*-experimenten op de boeken van de Tweede, Derde en Vijfde Partie 103
Tabel 17 Overzicht van Maerlants oeuvre in deze studie 131
Tabel 18 Overzicht van Maerlants oeuvre in deze studie (tellingen) 132
Tabel 19 Illustratie van de variatie-coëfficient aan de hand van een fictief corpus 140
Tabel 20 Overzicht van de gebruikte teksten uit de Karelepiek (deel 1) 247
Tabel 21 Overzicht van de gebruikte teksten uit de Karelepiek (deel 2) 248
Tabel 22 Illustratie van de *Wilcoxon ranks sums*-proef (Kilgariff 2001) 255
Tabel 23 *Ranksums*-analyse (deel 1, Karelepiek) van de 1.000 meest frequente rijmwoorden in samples van de Middelnederlandse Artur- en Karelepiek 257
Tabel 24 *Ranksums*-analyse (deel 2, Arturepiek) van de 1.000 meest frequente rijmwoorden in samples van de Middelnederlandse Artur- en Karelepiek 258
Tabel 25 Voorbeeld van trainingsvectoren voor de context-gevoelige tokenisering van Middelnederlandse oppervlaktevormen 292

LIJST VAN TECHNISCHE AFKORTINGEN

CA	*Correspondence analysis*, correspondentie-analyse
CV	*Coefficient of variation*, variatie-coëfficient
δ	*Burrows's Delta*
$\delta\delta$	*Double Delta*, dubbele delta
HSD	*Tukey's Honestly Significant Difference*
IDF	*Inverse document frequency*
μ	*Mean*, gemiddelde
PC	*Principal component*, hoofdcomponent
PCA	*Principal components analysis*; hoofdcomponenten-analyse
ss	*Step size*
σ	*Standard deviation*, standaarddeviatie
ws	*Window size*

Deel I

DE AANLOOP

In deze studie doe ik verslag van een onderzoek naar stylometrische auteursherkenning in Middelnederlandse literaire teksten. In dit eerste deel zal ik in drie hoofdstukken de aanloop naar het eigenlijke onderzoek voorstellen. Het eerste hoofdstuk benadrukt het belang van de studie van auteurschap in de medioneerlandistiek en blikt vooruit op de implicaties van de transdisciplinaire methodologie die ik zal hanteren. In het tweede hoofdstuk stel ik het vakgebied voor waaraan ik veel van mijn methodes zal ontlenen (*non-traditional authorship attribution*). Verder beschrijf ik de stand van zaken in het computationele onderzoek naar auteurschap in de medioneerlandistiek. In het derde hoofdstuk laat ik aan de hand van een aantal handschriften van de *Rijmbijbel* zien hoe men teksten stylometrisch kan bestuderen. Vervolgens wordt de techniek *Principal components analysis* geïntroduceerd en toegepast op parallelle handschriften van de *Rijmbijbel*.

1 VELE BOUKE

1.1 DE DOOD VAN DE AUTEUR

Voor de uitvinding van de boekdrukkunst werden middeleeuwse teksten met de hand gekopieerd, bijvoorbeeld in perkamenten of papieren handschriften.[1] Bij het kopiëren kon er echter heel wat fout gaan. Vaak durfden kopiisten teksten grondig aan te passen. Met onschuldige ingrepen raakten zij aan de spelling of het dialect van een tekst, maar ook meer gewichtige 'emendaties' kwamen voor. Het schoolvoorbeeld van dat laatste biedt *Van den vos Reynaerde*, een bekend dertiende-eeuws dierenepos dat blijkens de proloog door ene Willem werd gedicht, die buiten deze *Reynaerde* ook een *Madocke* maakte. Althans, dat is de lezing van het Dyckse en Brusselse handschrift van de tekst. In de proloog van het Comburgse afschrift van de *Reynaert* heeft een vijftiende-eeuwse kopiist het woord *Madocke* weggekrast en vervangen door de slordig geschreven woordgroep *vele bouke* (Lagast 2010).

Liters inkt zijn reeds gevloeid over deze merkwaardige aanpassing.[2] Soms neemt men aan dat de Comburg-corrector nog nooit gehoord had van de mysterieuze *Madocke* en de titelvermelding daarom door het weinigzeggende *vele bouke* verving. Of misschien was de *Madocke* ondertussen een verboden tekst geworden en werd de titel daarom uitgeradeerd. Een andere, meer prozaïsche mogelijkheid is dat de corrector weet had van andere werken van Willems hand. Willem zou dan niet enkel *Reynaert* en *Madocke* maar, integendeel, veel meer *bouke* hebben geschreven. Dan rijst de prikkelende vraag welke *bouke* dit geweest kunnen zijn. We kunnen zelfs niet uitsluiten dat een van die werken schuilgaat achter een van de talloze anonieme gedichten die bewaard zijn. Zouden bekende werken als *Karel ende Elegast* of de *Beatrijs* eveneens van Willems hand kunnen zijn? Dat lijkt even onwaarschijnlijk als fascinerend.

medioneerlandistiek

Toch moet de medioneerlandistiek, de studie van Middelnederlandse (literaire) teksten, dergelijke auteursvragen jammer genoeg uit de weg gaan bij gebrek aan betrouwbare technieken om het auteurschap van werken vast te stellen op basis van tekst-interne kenmerken, zoals stijl (Van Driel 2007, 163-164). Nu zou men kunnen berusten in deze situatie en lijdzaam de vondst van nieuwe handschriften afwach-

1 Deze sectie steunt in belangrijke mate op delen van eerder gepubliceerde artikels (Kestemont 2011c, d).

2 Zie Lagast (2010) voor een recente stand van zaken. Letterlijk leest het handschrift-Brussel: *Willam die madock maecte / Dair hi dicke om waecte* en het Dyckse handschrift: *Willam, die madocke makede, / Daer hi dicke omme wakede*. Een facsimile van het Comburgse handschrift bieden Janssens et al. (1998).

auteurschap, anonieme werken

stylometrie, de kwantitatieve studie van (literaire) schrijfstijl

ten. Mogelijk zullen deze meer licht werpen op het auteurschap van de vele anonieme werken in de Middelnederlandse letterkunde. In deze studie wordt echter een andere weg ingeslagen. Hieronder doe ik verslag van een onderzoek naar stijlgebaseerde auteursherkenning in Middelnederlandse epische teksten.[3] Daarbij maak ik gebruik van moderne computationele technieken uit de stylometrie, de kwantitatieve studie van (literaire) schrijfstijl.

Dergelijke technieken zijn in het internationale onderzoek vooral toegepast op meer hedendaagse en vooral Engelstalige teksten. Daarom vereist de toepassing van stylometrische methodes op Middelnederlandse poëzie heel wat toelichting. In dit eerste hoofdstuk beschrijf ik eerst kort de overleveringstoestand van Middelnederlandse literatuur en licht ik bondig toe hoe deze (in combinatie met andere factoren) het zicht belemmert op het auteurschap in de Middelnederlandse epiek. Vervolgens zal ik kort ingaan op het belang van auteurschap voor een beter begrip van de Middelnederlandse letterkunde als geheel. Zo hoop ik de focus van dit onderzoek te rechtvaardigen. Omwille van de praktische invalshoek van deze studie heb ik de theoretische inbedding van mijn werkzaamheden opzettelijk erg beknopt gehouden. Op die manier kan ik snel overgaan tot de eigenlijke kern van dit onderzoek.

1.1.1 *Een schrale oogst*

In 2010 publiceerde Willem Kuiper op het online platform *Neder-L* een column, getiteld 'Een digitale knekelput'. Via Ingrid Biesheuvel was hij op het spoor gekomen van een nog onbekend fragment van een middeleeuws handschrift. De snipper was ontdekt in de universiteitsbibliotheek van Leuven (Ms. 1488, OLIM: FRAGM. 598) en zou de restanten van een Middelnederlandse ridderroman bevatten. Toen Kuiper enkele uren na Biesheuvels signalement een digitale afbeelding (zie Fig. 1, p. 5) van de vondst bemachtigde, bleek het echter geen ontdekking 'van formaat'. Slechts de bovenrand van een enkel perkamenten blad was bewaard, waarop een luttele acht Middelnederlandse verzen waren te lezen...

De Leuvense snipper behoort tot het trieste type boekrestanten dat wel vaker opduikt. Verschillende middeleeuwse boeken vonden in latere tijden geen lezerspubliek meer en al snel hechtte men meer waarde aan hun stevige perkament dan aan de (onbegrijpelijk geworden) teksten erin (Geirnaert 2000, 86-90). Afgedankte folia werden door ambtenaren dan ook dankbaar gerecycleerd tot rekeningomsla-

3 In navolging van Berteloot (2000, 26) reken ik tot het Middelnederlands de historische variëteiten van het Nederlands die 'gesproken en geschreven werden tussen het eind van de twaalfde en het begin van de zestiende eeuw'. Aardrijkskundig laat Berteloot het Middelnederlandse taalgebied samenvallen met 'het huidige Nederlandse taalgebied, met inbegrip van Frans-Vlaanderen en de Nederrijn maar met uitzondering van de streek ten oosten van de IJssel'.

Figuur 1: De (verso?)zijde van Leuven, Universiteitsbibliotheek, Ms. 1488, olim: Fragm. 598. De tekst waarvan hier (zoals op de achterzijde) een luttele 4 verzen zijn bewaard, is verder onbekend. Waarschijnlijk dateert dit tweekolommige handschrift uit het laatste kwart van de veertiende eeuw. Mogelijk werd het door een Vlaamse kopiist afgeschreven (Kuiper 2010). In een diplomatische weergave leest de afgebeelde zijde: Salic bliuen ongehangen / En*de* gine sult mi ni*et* ontga*n*gen / [...] Des hadden si wond*er* groot / Entie coni*n*gi*n*ne al bloot / [...]'. In de zomer van 2011 werden de handschriften uit de Leuvense universiteitsbibliotheek opnieuw genummerd, waardoor het fragment nu de signatuur Ms. 1488 draagt, i.p.v. Fragm. 598 zoals in de column van Kuiper (2010) nog het geval was.

gen. Ook boekbinders versneden ze genadeloos tot reepjes die dan ter versteviging in nieuwe boekbanden werden verwerkt (Kienhorst 1988, III-IV). De meer creatieve zielen knutselden er zelfs bladwijzers of doosjes van.

De beperkte omvang van het Leuvense fragment is dan ook verre van uniek. Kuiper had zelfs in zekere zin nog geluk, want in Fragment 598 zijn hele verzen overgeleverd. Enkele jaren geleden werden Oxfordse fragmenten van een Middelnederlandse ridderroman uitgegeven die verticaal werden versneden, zodat slechts halve tekstkolommen bewaard zijn (Kienhorst 1998). De overlevering van dergelijke fragmenten is voor literair-historici bijzonder frustrerend. Enerzijds laten zij toe om een unieke blik te werpen op een tekst die wij zonder dergelijke fragmenten nooit gekend zouden hebben. Anderzijds herinneren zij op pijnlijke wijze aan de rest van het boek dat niet is overgeleverd.

Vaak zijn deze boekrestanten te fragmentarisch om tot nieuwe literair-historische inzichten te leiden. Toegegeven, de hierboven beschreven voorbeelden zijn erg extreem, maar in het algemeen speelt de gebrekkige bronnenoverlevering de studie van de Middelnederlandse letterkunde lelijk parten. Obbema merkte reeds op dat onze wereldlijke letterkunde slechts een negental volledig bewaarde Middelnederlandse handschriften telt (Kienhorst 1988, XII, n. 1). Voor de rest is men aangewezen op een disparate verzameling fragmenten die alle tezamen nog geen volledige verhuisdoos in beslag nemen (Kienhorst 1999, 49).

gebrekkige overlevering van materiële bronnen

1.1.2 *De blinde vlek van het auteurschap*

Afgezien van het feit dat deze fragmenten (Geirnaert 2000) zich verzetten tegen een tekst-interne analyse, zijn ze ook moeilijk te situeren. Waar, wanneer, waarom, voor wie en door wie een tekst werd geschreven, zijn vragen waarop de medioneerlandistiek niet zelden het antwoord schuldig moet blijven (Van Oostrom 2006, 233). In het bijzonder over het auteurschap van teksten is men slecht ingelicht, want slechts een klein aantal auteursnamen is bekend, en een nog kleiner aantal valt daadwerkelijk met een (al dan niet overgeleverde) tekst te verbinden (Van Driel 2007, 159). De anonimiteit van de Middelnederlandse letterkunde hangt voor een groot deel samen met materieel bronnenverlies. Auteurs gaven weleens meer informatie over de ontstaanscontext van een werk in pro- of epilogen (Sonnemans 1995), maar dergelijke passages zijn net als de rest van de teksten zelden bewaard in fragmenten.

proloog, epiloog

Bovendien zijn de overgeleverde teksten zelden 'originelen' en vormen zij in de regel afschriften van een legger die op zijn beurt weer op andere afschriften teruggaat (Sonnemans 1992, 324). Middeleeuwse autografen – het auteursexemplaar van een tekst – of directe afschrif-

autograaf, het auteursexemplaar van een tekst

ten daarvan zijn in de medioneerlandistiek uitzonderlijk (Houthuys 2009, 51ff). Latere kopiisten van teksten waren bovendien niet altijd meer geïnteresseerd in de herkomst van een tekst (vgl. Van Hemelryck 2007). Men veronderstelt dat zij deze in hun ogen minder relevante informatie vaak hebben weggelaten om een afschrift beter geschikt te maken voor een nieuwe omgeving. Veel Middelnederlandse teksten schijnen daarom letterlijk 'kop- en staartloos' overgeleverd (Janssens 1995, 94-95).

Toch biedt een volledig overgeleverd Middelnederlands handschrift – zelfs een autograaf (Houthuys 2009) – geen garantie op bijvoorbeeld een auteursvermelding. De individualiteit van een schrijver was in de middeleeuwen over het algemeen minder van tel. In kloosters zorgde het ideaal van de *humilitas* ervoor dat schrijvende geestelijken niet noodzakelijk hun naam aan een werk wilden verbinden, maar ook in de buitenwereld maakten wereldlijke schrijvers zich – voor zover wij weten – eigenlijk maar uitzonderlijk bekend in hun teksten (Van Hemelryck 2007, 276). Een kanttekening bij die ogenschijnlijke bescheidenheid is dat middeleeuwse literatuur meestal kleinschalig was (Van Oostrom 1994, 6). Een auteur schreef in de regel voor intimi en anticipeerde op een publiek dat hem (of haar) persoonlijk kende. Technologieën voor een wijde verspreiding van geschreven teksten, zoals de drukpers, bestonden nog niet. Bijgevolg hielden auteurs niet altijd rekening met een brede receptie van hun teksten. Een auteursvermelding in een tekst – die de dichter misschien zelf voor zijn publiek zou voordragen – was in een dergelijke context sowieso minder zinvol.

humilitas-ideaal

Als een tekst toch elders of later werd gerecipieerd, zien we anderzijds dat kopiisten soms de nood voelden om informatie toe te voegen in plaats van weg te halen. Toen een vijftiende-eeuwse afschrijver de dertiende-eeuwse *Rijmkroniek van Woeringen* in handen kreeg, voegde die zelf een proloog toe aan de tekst (Sonnemans 1992, 334). In die introductie verwerkte deze – zelf anoniem gebleven – kopiist een opvallend gedetailleerde toeschrijving: volgens de kopiist werd de rijmkroniek gedicht door broeder Jan van Heelu, die men ook Jan van Leeuwen pleegde te noemen. Lodewijk van Velthem had in het begin van de veertiende eeuw grote lappen tekst uit de *Rijmkroniek van Woeringen* in zijn eigen werk ingebouwd, maar vermeldt Heelu nergens bij naam (Mantingh 2009). Hoewel Velthem niet bepaald bekend staat om zijn secure bronvermeldingen, is het goed mogelijk dat Heelu's kroniek tot in de vijftiende eeuw gewoon anoniem heeft gecirculeerd.

toeschrijving, attributie

In het geval van Heelu twijfelt men zelden aan de juistheid van de veel jongere attributie (Sleiderink 2003, 87). Toch moeten er ook gevallen zijn geweest waarin kopiisten anonieme werken ten onrechte aan bekende auteurs hebben toegeschreven. In het geval van een schijnattributie kan het vergroten van de autoriteit van een werk immers een motief zijn geweest. Zelfs indien een toeschrijving in een

schijnattributie

middeleeuwse bron is overgeleverd, mogen we die dus niet blindelings aannemen (vgl. Van Mierlo 1994). Een gelieerd probleem is het hergebruik van teksten. Zeker middeleeuwse auteurs grepen graag terug naar bestaande bronnen die ze al dan niet woordelijk in hun eigen teksten verwerkten. Dat gebeurde niet altijd met een duidelijke bronvermelding. De eerder vermelde Lodewijk van Velthem bijvoorbeeld, integreerde met de regelmaat van de klok passages van andere auteurs in zijn eigen oeuvre, vaak zonder de lezer expliciet van de ontlening op de hoogte te stellen (Besamusca 2009). Verschillende teksten kunnen dus een gelaagd en meervoudig auteurschap vertonen, hoewel dat zeker bij een oppervlakkige lectuur niet noodzakelijk meteen duidelijk wordt. Het auteurschap van werken is nog in veel opzichten een blinde vlek in onze kennis over Middelnederlandse literatuur.

1.1.3 *Probleemstelling*

Al bij al kan men zich de vraag stellen of het eigenlijk wel problematisch is dat wij zo weinig weten over het auteurschap van Middelnederlandse teksten. Roland Barthes, bijvoorbeeld, heeft in een uiterst invloedrijk essay *La mort de l'auteur* (Barthes 1968) de auteur dood verklaard. In poëtische bewoordingen beargumenteerde Barthes dat eens een auteur een tekst aan een tekstdrager had toevertrouwd, de auteur die tekst ook onvermijdelijk had losgelaten. De tekst werd vanaf dat moment losgekoppeld van elke intentie die de auteur ermee kon gehad hebben en ging een eigen leven leiden. Amper een jaar later zou Michel Foucault zijn landgenoot in die opvatting bijtreden. In zijn beroemde essay *Qu'est-ce qu'un auteur?* citeerde Foucault (1969) – niet toevallig zonder verwijzing naar de oorspronkelijke auteur – Samuel Beckett: *Qu'importe qui parle, quelqu'un a dit qu'importe qui parle*. Oftewel: wat maakt het uit wie spreekt?

La mort de l'auteur

Qu'est-ce qu'un auteur?

Deze filosofische overlijdensakte van de auteur heeft zijn effect op de literatuurwetenschap niet gemist. Barthes' visie werkte duidelijk bevrijdend: de laatste decennia van de vorige eeuw ademden een haast vrolijk *dédain* voor de romantische notie van de auteur als ontzagwekkend figuur (Claassen 2009, 12-19). De tekst en vooral de lezer traden daarentegen uit de schaduw. De vrijheid van interpretatie – los van elke auteursintentie – kreeg opnieuw de ademruimte die zij lang had moeten missen.

Toch was niet iedereen overtuigd en zeker de laatste jaren schijnt men schoorvoetend terug te komen op de postmodernistische auteursopvatting die door mensen als Barthes en Foucault werd verdedigd (Claassen 2009, 19-27). Harold Love, bijvoorbeeld, auteur van de indrukwekkende monografie *Attributing authorship: An introduction*, toonde zich reeds in 2002 bijzonder kritisch en meende dat de

postmodernistische auteursopvatting

geruchten over de dood van de auteur sterk overdreven waren. Niet zonder ironische ondertoon merkte hij het volgende op:

> Yet this erasure of the author has remained paradoxical, not least because it was powered by the creativity of a number of profoundly individualistic thinkers and writers. There was never any doubt as to where royalty cheques for Barthes, Foucault, Lacan, Kristeva and Derrida ought to be sent: they reasserted heroic authorship even in their questioning it (Love 2002, 7).

mediëvistiek

In de mediëvistiek hebben de denkbeelden van Barthes en Foucault overigens nooit diepgaand wortel geschoten.[4] Enerzijds is dat vreemd omdat zij precies daar heel bevrijdend hadden kunnen werken. Gezien de beperkte informatie over middeleeuws auteurschap, hadden mediëvisten de spreekwoordelijke dood van de auteur kunnen bejubelen. Als de auteur inderdaad niet langer van tel was in de literatuurwetenschap, verloste dat onderzoekers immers van de last en plicht om zich nog druk te maken over het auteurschap van werken.[5] Anderzijds moet men de essays van Barthes en Foucault begrijpen als een reactie op het overdreven 'biografisme' waaraan auteurs wel eens werden blootgesteld in de periode voorafgaand aan de *sixties* (vgl. Greene 2006, 220). Door het schrijnende gebrek aan biografische informatie heeft dit euvel zich echter veel minder prominent gemanifesteerd in de mediëvistiek, zodat men daar waarschijnlijk niet het gevoel had dat er *überhaupt* op iets moest gereageerd worden.

biografisme

auteursintentie
poetica

Misschien verbaast het dus niet dat precies in de periode dat men elders nauwelijks nog aandacht besteedde aan auteur(sintentie)s, een groot aantal publicaties over oeuvres en auteurspoetica's verschenen in de medioneerlandistiek – net zoals in andere taalgebieden (Greene 2006, 218). Bekende voorbeelden zijn het gelauwerde *Maerlants wereld* over het oeuvre van Jacob van Maerlant (Van Oostrom 2006) of de Griffioen-vertaling van het *Boec vander Wraken* waarin Wim van Anrooij de poetica van Jan van Boendale verklaart vanuit de levensloop van deze dichter (Van Anrooij 1994b, 1995). Ook in de

4 Een uitstekend overzicht van de problematiek biedt Greene (2006) waar zij het belang van *La mort de l'auteur* voor de (hedendaagse) mediëvistiek sterk relativeert, maar wel de invloed bespreekt van o.m. Barthes op P. Zumthor en van Foucault op B. Cerquiglini. Vgl. ook Van Hemelryck (2007, 274).

5 Vgl. bijvoorbeeld Gerritsen (1975) die pleitte voor een literatuurgeschiedenis in de vorm van een schetskaart die zich radicaal verwijdert van de focus op auteursteksten en de beter dateerbare handschriften centraal plaatst (Caers & Kestemont 2011, 1-3). In het verlengde hiervan kan de zogeheten *New* of *Material Philology* (Nichols 1997) beschouwd worden als een variatie op *La mort de l'auteur*, omdat de nadruk daar komt te liggen op de handschriftelijke receptie van teksten en niet de oorspronkelijke auteursteksten – vgl. ook Foucaults invloed op B. Cerquiglini (Greene 2006, 217-218). Daarbij zij echter opgemerkt dat onderzoekers in dit vakgebied ertoe neigen kopiisten een soort auteursstatus toe te kennen, zodat er soms gewoon een nieuw auteurstype in het leven wordt geroepen.

slothoofdstukken van Joost van Driels recente studie over de stijl van Middelnederlandse epische poëzie blijkt van Barthes' visie nauwelijks sprake: dé auteur wordt resoluut naar voren geschoven als drijvende kracht achter de stilistische diversiteit van deze literatuur (Van Driel 2007).

Veel andere recente studies binnen de medioneerlandistiek geven evenzeer blijk van een niet-aflatende fascinatie voor auteurs als de 'Bijbelvertaler van 1360' (De Bundel 2009), de schrijvende zusters uit Jericho (Stoop 2009) of Lodewijk van Velthem (Besamusca et al. 2009a). Het mooiste voorbeeld is wellicht de studie van Astrid Houthuys waarin zij de unieke autograaf van Wein van Cotthem analyseerde en zo tot een haarscherp beeld kwam van de werk- en schrijfwijze van deze historiograaf (Houthuys 2009). De medioneerlandistiek lijkt daarom in te stemmen met de romaniste V. Greene die zei:

> I do not believe that any of us can claim to be in a position in which we could honestly affirm that "who is speaking" makes no difference (Greene 2006, 221).

De grote aandacht voor auteurschap in de hedendaagse studie van Middelnederlandse letterkunde lijkt tot op zekere hoogte samen te hangen met het klimaat dat de laatste decennia in het vak heerste. Tijdens het NLCM-project onder leiding van Frits van Oostrom focuste een groot aantal studies op de 'historisch-functionele' benadering van teksten (Janssens 2000; Fraeters 2004): men trachtte te reconstrueren welke rol teksten vervulden in hun oorspronkelijke context en welke reacties zij konden uitlokken bij hun geïntendeerd publiek. De auteur en auteursintenties namen vaak een centrale plaats in omdat veel studies de complexe relaties tussen mecenas en schrijver belichtten. Precies in dit type onderzoek groeide dus de bezorgdheid om onze gebrekkige kennis inzake het auteurschap van Middelnederlandse teksten (Fraeters 2004, 300).

historisch functionalisme

geïntendeerd publiek

mecenas

intertextualiteit

In onderzoek naar intertextualiteit, bijvoorbeeld, vormen anonieme teksten een groot probleem. In ridderromans treft men geregeld passages aan die woordelijk aan een andere tekst zijn ontleend of er stilistisch sterk aan herinneren. Vaak blijft onduidelijk welke conclusies men uit dergelijke verbanden mag afleiden (Willaert 1995, 261). Verwoordden auteurs via dergelijke wenken een subtiele literaire stellingname ten aanzien van andere werken (Janssens 1998a)? Of werkten de dichters in een gelijkaardig hofmilieu of dezelfde streek en resulteerden deze overeenkomsten uit een gemeenschappelijk poëtisch idioom (Besamusca 1993, 27-29)? Mensen als Heeroma gingen echter veel verder en herkenden in dergelijke verbanden geen gewone knipoogjes maar zwaarwichtige indicaties voor een identiek auteurschap (Heeroma 1973). Een beter inzicht in het auteurschap van teksten zou vele discussies over de waarschijnlijkheid van intertekstuele netwerken ongetwijfeld helpen beslechten.

intertekstuele netwerken

Een ander twistpunt is de stilistische flexibiliteit van middeleeuwse auteurs. Van Driel (2007) claimde zoals gezegd dat de Middelnederlands auteur de belangrijkste verklarende factor is voor de stilistische variatie in epische teksten, maar betrok wel geen enkel oeuvre in zijn onderzoek (Kestemont 2007, 179-180). Van Driel (2007, 167) betwijfelt bijvoorbeeld de veronderstelling 'dat elke dichter een unieke, persoonlijke stijl heeft die in de loop van zijn werkzame leven constant en homogeen zou blijven'. Hij verwijst wel naar het oeuvre van Maerlant, dat een zekere evolutie zou kennen (vgl. De Vries & Verwijs 1982, XXI-XXII), maar heeft zichzelf niet de mogelijkheid gegund om deze opvatting verder te onderzoeken. Dat is jammer want uiteindelijk is maar weinig bekend over de interne samenhang van oeuvres of de schommelingen in stijl die een middeleeuws dichter zich kon veroorloven.

oeuvre

Nog veel belangrijker voor de interpretatie van literaire werken zijn het profiel en de achtergrond van een auteur – of zelfs zijn of haar sekse (Lie 2001). Als niets vaststaat over het auteurschap van een tekst, is het bijzonder moeilijk om de ideologische standpunten die een tekst inneemt, te kaderen. Historici doen bijvoorbeeld graag een beroep op Middelnederlandse rijmkronieken om historische gebeurtenissen beter te begrijpen (Stein 1991, bv. 194). Voor anonieme teksten is dit echter een hachelijke onderneming. Of een auteur tot de stedelijke elite dan wel de clerus of een hofmilieu behoorde, is vanzelfsprekend cruciale informatie om tot een verantwoorde analyse van de ideologie van een auteur te komen (Kinable 1991). Natuurlijk kan men ook zonder auteursnaam een profiel van de auteur van een anonieme tekst trachten op te stellen, maar een correcte toeschrijving van de tekst aan een specifiek individu zou meer dan welkom zijn.

Met dat laatste raken we een interessant punt: bij een onderzoek naar auteursidentiteit is een auteurs*naam* vaak van ondergeschikt belang. De toeschrijving van een tekst is meestal geen einddoel en wordt slechts aangegrepen om andere literair-historische inzichten te formuleren, die het belang van de attributie zelf overstijgen. Iedereen zou bijvoorbeeld wat graag de nietszeggende voornaam *Willem* inruilen voor een kopie – zelfs een fragment – van de mysterieuze *Madocke*. De naam van een auteur lijkt in sommige gevallen dus ondergeschikt aan het belang van zijn teksten.

naam vs. identiteit van een auteur

Toch kan een auteursnaam helpen om werken beter te begrijpen. Recent nog werd op interessante gronden beargumenteerd dat *Willem die Madocke maecte* identiek zou zijn aan Willem van Boudelo, een dertiende-eeuwse klerk die onder meer werkte voor de gravin van Vlaanderen (Van Daele 2005). Deze Willem van Boudelo werkte in de onmiddellijke omgeving van Diederik van Assenede, de bij naam gekende auteur van de oosterse liefdesroman *Floris ende Blancefloer* en eveneens klerk van de Vlaamse gravin (Kuiper 1998, 176). De Boudelo-attributie werpt zo mogelijk een interessant licht op de relatie tussen

beide – onderling erg verschillende – werken, aangezien zij in dat geval in hetzelfde milieu tot stand kwamen. Vermoedelijk werden deze werken dan door dezelfde mensen gerecipieerd (en gesmaakt?). Attributies kunnen daarom een niet te versmaden rol spelen in de analyse van literaire milieus en netwerken.

Tot slot kan men opmerken dat auteurschap een merkwaardige aantrekkingskracht uitoefent op het grotere publiek (vgl. Greene 2006, 218). Een zekere fascinatie voor de mens-achter-het-werk is onze tijd zeker niet vreemd. Verscheidene vormen van auteurschap staan de dag van vandaag opnieuw volop in de belangstelling (Vaessens 2009). Daarbij kan men wijzen op de recente gevallen van plagiaat (door een Duitse minister en een Leuvense hoogleraar) die breed in de pers werden uitgesmeerd. Dat illustreert overigens hoe ver de hedendaagse, voyeuristische samenleving – 'Piraterij is een misdrijf' – verwijderd is van *La mort de l'auteur* (Claassen 2009, 27).

plagiaat

In de muziek-industrie staat de notie auteurschap anderzijds onder druk: 'gepirateerde' muziek-opnames zijn vrijwel straffeloos illegaal te downloaden op het internet, zodat nog maar een fractie van de wettelijk verplichte auteursvergoedingen bij de muzikanten terecht komt. Toch zijn mensen meer dan ooit bereid om goed geld te betalen voor live performances op allerhande festivals, hetgeen suggereert dat de status van de individuele artiest verre van tanende is. Ook op het internet vormt auteurschap voer voor een intens maatschappelijk debat. Hoe schatten we de waarde in van – om slechts één voorbeeld te noemen – een anoniem Wikipedia-artikel? In deze tijd van haast complexloos *copy-pasten* zaaien informatie en gedachtegoed stuurloos uit over het internet. In die context groeit begrijpelijkerwijs de bekommernis om de oorsprong, *casu quo* het auteurschap van teksten. Deze (weinig exhaustieve) lijst voorbeelden dient slechts ter illustratie van het feit dat een wetenschappelijk onderzoek naar (om het even welk type) auteurschap de dag van vandaag een zekere maatschappelijke relevantie heeft.

maatschappelijk debat

1.2 METHODOLOGIE

1.2.1 *Kwantitatief versus computationeel*

Hieronder zal ik gebruik maken van een aantal onderzoeksmethodes uit het veld van de *non-traditional authorship attribution*. In het volgende hoofdstuk zal ik grondig ingaan op de inzichten die ten grondslag liggen aan dit paradigma. Nu mag reeds benadrukt worden dat de methodes uit deze discipline zich onderscheiden van de gangbare onderzoekspraktijk in de medioneerlandistiek, omdat zij sterk 'kwantitatief' en 'computationeel' zijn. Het is raadzaam om in dit inleidend hoofdstuk al even stil te staan bij de implicaties hiervan.

non-traditional authorship attribution

In theorie is het belangrijk om beide aspecten – kwantitatief versus computationeel – goed te onderscheiden. Men kan immers een kwantitatief onderzoek uitvoeren, zonder bij het 'tellen' een beroep te doen op de computer als rekeninstrument.[6] Bovendien is lang niet al het computationele onderzoek kwantitatief. Steeds vaker zetten onderzoekers computers in voor het verzamelen of raadplegen van onderzoeksmateriaal (bijvoorbeeld via zoekmachines), hoewel zij verder geen kwantitatieve analyse van dat materiaal beogen (zie het historisch overzicht hieronder).

kwantitatief vs. computationeel onderzoek

Gezien de opmars van het gebruik van computers in de geesteswetenschappen – de zogenaamde *computational turn* – ziet het ernaar uit dat steeds meer onderzoek tenminste deels computationeel zal worden uitgevoerd (Berry 2011). Dat is echter nu reeds veel sterker het geval voor kwantitatief onderzoek, zodat de termen 'kwantitatief' en 'computationeel' in de praktijk vaak verward worden. Wat betreft computationeel onderzoek, zij in het bijzonder verwezen naar het initiatief van de *Digital Humanities* (Schreibman et al. 2004), het sterk internationale onderzoeksveld waarin men nagaat hoe digitale hulpmiddelen, zoals computers, kunnen worden ingezet in de geesteswetenschappen.

computational turn

Digital Humanities, digitale geesteswetenschappen

Hoewel het gebruik van de computer al lang geen nieuwigheid meer is in de medioneerlandistiek, blijft kwantitatief onderzoek wel ongebruikelijk.[7]. De studie van de Middelnederlandse letterkunde is tot op heden vooral 'kwalitatief' aangepakt, waarbij onderzoeksresultaten werden verkregen op basis van een gegevensinterpretatie zonder doorgedreven telwerk. Vanzelfsprekend is het traditionele onderscheid tussen kwantitatief en kwalitatief onderzoek tot op zekere hoogte artificieel en moet men rekening houden met een continuüm (Hinskens & Van Dalen-Oskam 2007, 4). Zelfs in erg traditioneel 'kwalitatief' onderzoek ziet men vaak hoe onderzoekers impliciet tellen bij het afwegen van hypotheses. Bovendien kan zelfs de meest fanatieke kwantitatieve onderzoeker niet om een 'kwalitatieve' duiding van zijn telwerk heen. Het onderscheid hangt mogelijk deels samen met de retorische presentatie van een onderzoek en de mate waarin de kwantificering wordt gethematiseerd of geëxpliciteerd. Hoewel ik mijn kwantitatieve aanpak hieronder sterk zal expliciteren, vormt ook dit aspect van mijn werk eigenlijk geen fundamentele nieuwigheid. De echte vernieuwing van dit onderzoek ligt in de *wijze van kwantificering* en specifieker het transdisciplinaire karakter van die kwantificering.

traditioneel, 'kwalitatief' onderzoek

6 Hinskens & Van Dalen-Oskam (2007, 3) verstaan onder kwantitatief onderzoek 'onderzoek waarbij tellen en het verder (veelal statistisch) bewerken van de daaruit resulterende getallen een belangrijke rol speelt'.

7 Een historisch overzicht bied ik in het volgende hoofdstuk maar hier zij reeds verwezen naar de overzichten bij Van Dalen-Oskam (2007, 2011); Kestemont (2012b)

1.2.2 *Transdisciplinariteit*

De studie van Middelnederlandse literatuur kent nog geen uitgebreide traditie op het gebied van expliciet kwantificerend onderzoek (Van Dalen-Oskam 2007, 13). Net zoals in andere taalgebieden schijnt de historische taalkunde overigens verder te staan op dit gebied (Hinskens & Van Dalen-Oskam 2007, 1). Evenals in andere disciplines binnen de geesteswetenschappen zijn vakken als statistiek trouwens zelden opgenomen in neerlandistische curricula. Dat biedt reeds een praktische verklaring waarom veel literair-historici geen vaardigheden bezitten op het gebied van kwantitatief onderzoek.

Ook computer-vaardigheden worden anno 2012 meestal nog maar oppervlakkig onderwezen, hetgeen niet helpt aangezien het makkelijker is om kwantitatief onderzoek uit te voeren met een computer. Daarenboven is de computer als onderzoeksinstrument nog maar enkele decennia beschikbaar voor individuele onderzoekers, zodat er zich misschien gewoon nog geen kwantitatieve onderzoekstraditie heeft kunnen ontwikkelen. De ware reden voor de geringe populariteit van kwantitatief onderzoek naar Middelnederlandse literatuur is natuurlijk dieperliggend.

computer als onderzoeksinstrument

Kwantitatief onderzoek veronderstelt doorgaans de beschikbaarheid van 'data' (Hinskens & Van Dalen-Oskam 2007, 6-7). Bij voorkeur zijn die data, ten eerste, omvangrijk en, ten tweede, betrouwbaar. In het domein van *Machine Learning*, bijvoorbeeld, bestaat het adagium *There's no data, like more data* (Moore 2001). Bovendien beschikt men daar graag over een *ground truth* of een *gold standard*: een deel van de onderzoeksdata zijn van niet-twijfelachtige herkomst en kunnen gebruikt worden als veilige uitvalsbasis om andere data – waarvan bepaalde aspecten wel onzeker zijn – te onderzoeken (Manning et al. 2008, 152). Op voorhand is al duidelijk dat data uit de middeleeuwse letterkunde vaak niet aan deze criteria zullen voldoen: niet heel veel teksten zijn overgeleverd, en als zij al zijn overgeleverd, is dat vaak op fragmentarische wijze. Bovendien kan de status van de overgeleverde teksten erg onduidelijk zijn (Hinskens & Van Dalen-Oskam 2007, 7), zodat *gold standards*, bijvoorbeeld inzake auteurschap, zeldzaam zijn.

data

ground truth, gold standard

Tevens veronderstelt kwantitatief onderzoek 'telbare' data: bepaalde aspecten van de data (bijvoorbeeld teksten) moeten op een zinvolle manier numeriek kunnen worden voorgesteld, zodat deze meetpunten kwantitatief kunnen worden verwerkt (Hoover 2007). Hier wringt in de kwantitatieve letterkunde niet zelden het schoentje: zoals in andere geesteswetenschappen worden in de literatuurgeschiedschrijving veel concepten onderzocht die niet rechtstreeks 'telbaar' zijn (Hinskens & Van Dalen-Oskam 2007, 3). De subjectieve leesbeleving van een tekst laat zich bijvoorbeeld niet eenvoudig in cijfers vatten. In de linguïstiek is het meestal makkelijker om telbare taalaspecten uit

numerieke, 'telbare' representatie van data

teksten of corpora te isoleren, zodat het kwantitatief onderzoek hier sneller is doorgebroken (Hinskens & Van Dalen-Oskam 2007, 1).

Wie aan kwantitatief letterkundig onderzoek wil doen, kan dus niet teruggrijpen naar het eigen vakgebied. Bij gebrek aan een 'inheemse' kwantitatieve traditie is men daarom haast verplicht om beproefde methodes te importeren uit wetenschapsgebieden die een rijkere traditie op het gebied van kwantificering kennen. Wanneer een onderzoeker dergelijke methodes aan een ander onderzoeksgebied ontleent, spreekt men doorgaans van 'transdisciplinair' onderzoek: de methodes uit een ander vakgebied worden geëxploiteerd voor het beantwoorden van een onderzoeksvraag uit het eigen vakgebied.[8]

transdisciplinair onderzoek

Dit proefschrift geniet het zeldzame voordeel dat verschillende van de deelonderzoeken die hierin beschreven worden reeds zijn beoordeeld voor andere publicaties. Uit de leesrapporten van deze anonieme *peer reviewers* durf ik concluderen dat vooral het transdisciplinaire karakter van mijn onderzoek als vernieuwend maar ook problematisch werd ervaren. Het kwantitatieve aspect van het onderzoek lokte op zich geen kritische reacties uit, maar wel het feit dat gekwantificeerd werd op een wijze die vreemd was aan het eigen vakgebied.

Als ik abstractie maak van de kritiek op de zuivere fouten die mijn werk vaak nog bevatte, bleek het transdisciplinaire karakter van mijn onderzoek vaak dezelfde types reacties teweeg te brengen. Een vijftal – erg gegronde – bezwaren kwam geregeld terug in de leesrapporten. Daarbij werd de transdisciplinaire methode door beoordelaars vaak ervaren als:

ONAANGEPAST – de methode is ontwikkeld voor het beantwoorden van exogene onderzoeksvragen die niet relevant zijn in het eigen vakgebied. De methode houdt onvoldoende rekening met de eigenheid van het traditionele vakgebied en de specifieke gevoeligheden ervan;

ONBEGRIJPELIJK – de methode is onbekend en wekt argwaan omdat zij op onaangepaste wijze werd toegelicht. De methode is ofwel zo oppervlakkig, ofwel zo technisch toegelicht, dat men zich niet aangesproken voelt. De beoordelaar voelt zich niet in staat om het onderzoek te beoordelen en laat het links liggen;

ONBEWEZEN – de methode wordt meteen toegepast op een echte casus, zonder dat werd aangetoond dat ze werkt (of tenminste zou kunnen werken) via een relevante test-casus;

ONVOLDOENDE INGEBED – de transdisciplinaire onderzoeker schijnt de traditionele werkwijzen uit het eigen vakgebied te 'verlooche-

8 Als er in het 'donor-vakgebied' geen vooruitgang wordt geboekt als gevolg van het transdisciplinaire onderzoek, mag men in theorie niet van 'interdisciplinariteit' spreken – hoewel dat in de praktijk vaak gebeurt. In een transdisciplinair onderzoek is de ontleende methodologie louter een onderzoeksinstrument en zelf geen onderzoeksobject waarvan de werking wordt bestudeerd (Fraeters 2004, 305, n. 25).

nen', hoewel die kunnen bogen op een veel langere traditie. De methode wordt niet voorgesteld als een aanvulling op oudere methodes maar als een surrogaat voor methodes die helemaal niet aan vervanging toe zijn;

ZELFINGENOMEN – de transdisciplinaire onderzoeker is niet kritisch genoeg voor het 'onfeilbare' karakter van de methode en veegt onterecht methodologische minpunten onder de mat. De methode wordt voorgesteld als had zij louter een meerwaarde *omdat* zij transdisciplinair is.

Vanzelfsprekend wil ik rekening houden met deze bezwaren en zoveel mogelijk vermijden dat deze studie dergelijke reacties uitlokt. Dit werk is immers expliciet gericht op een publiek van letterkundige (medio)neerlandici en de voorkennis die wordt verondersteld op het gebied van techniek en linguïstiek is minimaal gehouden. Relevante technische termen worden hieronder ter benadrukking in de kantlijn opgenomen, als zij verderop nog belangrijk zullen zijn voor een begrip van de tekst. Ik zal bovendien veel aandacht besteden aan de inbedding van de gehanteerde onderzoekstechnieken, zodat zowel hun meerwaarde als minpunten ten aanzien van traditionele onderzoeksmethodes duidelijk worden. Van meet af aan zullen methodes getest worden via relevante test-casussen. Gevallen waarin de methodes geen sluitende onderzoeksresultaten (kunnen) leveren, zullen duidelijk worden aangegeven.

stylometrie als 'controle' of 'correctie'?

Een laatste en cruciaal voornemen betreft het volgende. Stylometrisch onderzoek is kwantitatief onderzoek en wordt daarom geregeld voorgesteld als een 'controle' of zelfs 'correctie' van traditioneel, kwalitatief onderzoek. De stylometrie staat echter nog maar in de kinderschoenen en kent een veel minder lange traditie dan de kwalitatieve filologie. In veel gevallen is het daarom nog raadzaam om, juist omgekeerd, met kwalitatieve onderzoeksmethodes het kwantitatieve onderzoeksresultaat te evalueren. In deze studie wil ik dan ook geen hiërcharchie veronderstellen tussen beide onderzoekstypes. Veeleer wil ik verkennen hoe beide benaderingen elkaar kunnen aanvullen.

2 DE HAND VAN DE MEESTER

In deze studie zal ik een onderzoek beschrijven naar het auteurschap van Middelnederlandse werken op basis van de kwantitatieve methodes van een paradigma dat bekend staat als *non-traditional authorship attribution*. Dit tweede hoofdstuk biedt een inleiding op de relatief jonge discipline van de 'niet-traditionele' auteursherkenning zoals die vandaag de dag wordt beoefend binnen de kwantitatieve stijlleer. Het belang van auteurschap voor de medioneerlandistiek is in het vorige hoofdstuk reeds toegelicht. De relevantie van kwantitatief onderzoek hiernaar is echter nog niet aan bod gekomen en vormt het onderwerp van de eerste sectie (§ 2.1) hieronder. Ik zal daarom de historische ontwikkeling van dit vakgebied tijdens de tweede helft van de twintigste eeuw schetsen. In de laatste sectie (§ 2.2) bespreek ik de opkomst van het computationele onderzoek naar auteurschap binnen de medioneerlandistiek.

2.1 NIET-TRADITIONELE AUTEURSHERKENNING

2.1.1 *Wat voorafging*

Niet-traditionele auteursherkenning

stijlgebaseerde auteursherkenning

'Niet-traditionele auteursherkenning' wordt als term vaak gebruikt om het kwantitatieve onderzoek dat binnen dit paradigma wordt verricht, af te zetten tegen traditioneel (veelal ouder) onderzoek naar auteursherkenning. Men doelt hier meestal specifiek op stijlgebaseerde auteursherkenning. Het uitgangspunt van de moderne auteursherkenning is namelijk de hypothese dat men louter op basis van tekst-interne eigenschappen uitspraken kan doen over het auteurschap van een (bv. anonieme) tekst (Stamatatos 2009, 538). Binnen de niet-traditionele auteursherkenning gebruikt men daarvoor kwantitatieve onderzoekstechnieken die zich richten op stilistische en linguïstische aspecten van een tekst.

Traditioneel onderzoek naar auteursherkenning kent vanzelfsprekend een oude en boeiende voorgeschiedenis in de filologie die ver in de tijd terugreikt (Love 2002). Hoewel het onmogelijk is om al dit onderzoek hier te bespreken of diepgaand te karakteriseren, is er toch een opvallend aspect dat veel traditionele studies gemeen hadden: de afwezigheid van een consensus over de methodologie die men voor auteursherkenning diende te hanteren. In voorgaand onderzoek werden namelijk zulke uiteenlopende werkwijzen gehanteerd dat studies onderling nauwelijks te vergelijken waren. Bij gebrek aan een methodologische consensus was het bovendien moeilijk om de kwaliteit van

een onderzoek naar de stijl van een auteur te evalueren (Stamatatos 2009, 538).

Toch vertonen vroege stilistische studies van auteurschap ook wel gelijkenissen. Veel studies stelden namelijk het subjectieve oordeel van een onderzoeker centraal en waren zo *expert-based* (Luyckx 2010, 13-14). Onderzoekers steunden veelal op hun individuele expertise die men – soms neerbuigend – heeft aangeduid met termen als *Gelehrtenintuition*, *connoisseurship*, *Sprachgefühl* of *Ohrenphilologie*. Vaak wendden onderzoekers zich tot een soort 'checklist': men trachtte uit het oeuvre van een gekende auteur stilistische kenmerken te extraheren, waarvan men op basis van eerdere leeservaringen oordeelde dat deze stilistica veel minder voorkwamen in het werk van andere auteurs. Voor de toeschrijving van een anoniem of mogelijk foutief toegeschreven werk kon men vervolgens een beroep doen op de checklist. Als een tekst met een onbekende of betwiste signatuur vaak gebruik maakte van de stilistica in de checklist van een auteur kon dat de aanleiding vormen om de tekst aan dezelfde auteur toe te schrijven. *In se* is er natuurlijk niets fout met deze werkwijze en zoals hieronder nog zal blijken, is de niet-traditionele aanpak hier niet altijd erg verschillend van.

expert-based

Het grote probleem was echter dat het aanleggen van een dergelijke checklist vaak nogal wazig bleef en slecht gedocumenteerd werd. Deze lijsten werden immers met de hand vervaardigd en waren daarom nogal beperkt in lengte. Verschillende, soms nauwelijks vergelijkbare stilistische kenmerken werden vrijelijk met elkaar gecombineerd (bv. syntactische vs. lexicale karakteristieken). Bovendien bleef de samenstelling van de lijst vaak een uiterst subjectieve aangelegenheid: waarom bepaalde kenmerken wel en andere helemaal niet in de lijst werden opgenomen, bleef in de regel onverantwoord.

Nog meer onrustwekkend was het feit dat de meeste checklists gedomineerd werden door opvallende stilistische kenmerken – bijvoorbeeld een ongewone woordkeuze – die zo weinig frequent waren dat zij automatisch de aandacht van de onderzoeker trokken. Een eerste probleem daarmee betrof de spreiding: precies omdat deze fenomenen laagfrequent waren, waren zij zelfs in het oeuvre van de bewuste auteur niet altijd even goed gespreid. Als deze ongewone kenmerken niet voorkwamen in een andere – mogelijk kortere – tekst, bleef het onduidelijk of dit een betrouwbaar argument leverde tégen een toeschrijving aan de bewuste auteur. Bovendien waren dergelijke, laagfrequente kenmerken niet zelden gebonden aan een specifiek onderwerp of genre. Zo bleef het meestal onzeker of men ze als argument kon aanvoeren voor een attributie over de grenzen van onderwerpen en genres heen.

Het grootste probleem was echter dat deze kenmerken niet alleen opvallend waren voor moderne onderzoekers, maar dat in het verleden ook moeten zijn geweest in de ogen van vervalsers of imitators. Zulke opvallende tekstkenmerken moeten voor imitators immers de

eerste zijn geweest om na te bootsen. In het geval dat de stijl van een succesvolle auteur navolging had gevonden, leidde de focus op laagfrequente woorden dus niet noodzakelijk tot een betrouwbaar onderscheid tussen de oorspronkelijke kunstenaar en zijn eventuele epigonen. Globaal gesproken was de traditionele onderzoekspraktijk weinig expliciet op methodologisch vlak. Zij valt met andere woorden te contrasteren met de algemene criteria die men vandaag de dag aan wetenschappelijk onderzoek stelt, zoals transparantie en repliceerbaarheid.

2.1.2 *Het functiewoord*

In de niet-traditionele auteursherkenning heeft men van in het begin voor een andere methodologie gekozen dan in het traditionele onderzoek. Cruciaal was daarbij de vraag welke aspecten van een tekst zich nu eigenlijk het best leenden voor het vellen van een gefundeerd oordeel over het auteurschap van een werk. Tot voor kort werd aangenomen dat een onderzoek naar auteurschap het best werd uitgevoerd op woord-niveau. In een overzichtspaper uit 1994 claimde Holmes nog dat 'to date, no stylometrist has managed to establish a methodology which is better able to capture the style of a text than that based on lexical items' (Holmes 1994, 87).

Holmes biedt met dit citaat een samenvatting van een onderzoekslijn die in de zestiger jaren van de vorige eeuw werd ingezet door de Amerikaanse statistici Mosteller & Wallace (1964), de 'niet-traditionele auteursherkenning' oftewel *non-traditional authorship attribution* (Holmes 1994, 1998). Mosteller en Wallace richtten zich in hun invloedrijke onderzoek op de *Federalist papers*, een verzameling Engelstalige pamflettistische artikelen die aan het einde van de achttiende eeuw werden gepubliceerd onder het pseudoniem 'Publius'. Naar alle waarschijnlijkheid gaat achter dit pseudoniem een groep auteurs schuil die tot de goedkeuring van de Amerikaanse grondwet wilden aanzetten. Het werk van Mosteller en Wallace wordt vandaag de dag in het bijzonder geprezen omwille van een specifieke nieuwigheid die zij introduceerden: de nadruk op de frequenties van *functiewoorden* in het onderzoek naar auteursherkenning.

non-traditional authorship attribution

In de hedendaagse linguïstiek maakt men doorgaans een onderscheid tussen twee types woorden (Morrow 1986, 423). De zogenaamde inhoudswoorden vormen een open woordklasse, die veelal woordsoorten omvat als zelfstandige naamwoorden, adjectieven of werkwoorden (Clark & Clark 1977). De verzameling inhoudswoorden is erg groot – er zijn bijvoorbeeld veel verschillende naamwoorden – en bovendien makkelijk uitbreidbaar – want iedere dag wordt er wel een nieuw naamwoord gecreëerd. De categorie van de functiewoorden is daarentegen veel beperkter: deze gesloten woordklasse bevat woordsoorten als lidwoorden, voornaamwoorden of voegwoorden. Een taal telt bij-

inhoudswoorden

functiewoorden

gevolg veel minder functiewoorden dan inhoudswoorden, onder meer omdat de verzameling functiewoorden moeilijk uitbreidbaar is – het is bijvoorbeeld moeilijk om een nieuw voorzetsel uit te vinden.

Inhoudswoorden dragen veel betekenis in zich en zijn bovendien ook betekenisvol als zij in isolatie worden gebruikt (bv. 'kat'). Functiewoorden zijn daarentegen veel sterker gegrammaticaliseerd en dragen veel minder betekenis in zich, zeker als ze alleen worden gebruikt (bv. 'de'). Hoewel de verzameling functiewoorden in een taal doorgaans veel kleiner is dan de verzameling inhoudswoorden, zien we dat de meeste functiewoorden veel vaker gebruikt worden dan de meeste inhoudswoorden (Zipf 1949). Nog anders gesteld, zien we dat minder dan 0,04% van ons volledige vocabulaire goed is voor meer dan de helft van de woorden die we dagelijks gebruiken (Chung & Pennebaker 2007, 347).

gegrammaticaliseerd

Het gebruik van functiewoorden kent veel voordelen in onderzoek naar auteursstijl (Binongo 2003, 11):

- alle auteurs die in dezelfde taal en periode schrijven, maken noodzakelijkerwijs gebruik van dezelfde functiewoorden. Zij vormen daarom een betrouwbare basis voor een vergelijking van schrijfstijlen;
- omdat zij zo frequent zijn, leveren zij veel meetpunten of 'observaties' op voor een tekst, hetgeen aantrekkelijk is vanuit statistisch oogpunt;
- het gebruik van functiewoorden wordt amper beïnvloed door het onderwerp of het genre van een tekst: het gebruik van de voorzetsels 'op' of 'van' hangt nauwelijks samen met het onderwerp van een tekst.

Alle verschillen en gelijkenissen tussen teksten wat betreft de frequenties van functiewoorden kunnen daarmee veel makkelijker in verband gebracht worden met auteursstijl dan met andere factoren, zoals onderwerpsgerelateerde stilistica. Het onderliggende idee is natuurlijk een beetje contradictorisch: we gaan op zoek naar verschillen tussen teksten die worden gereduceerd tot een verzameling kenmerken waarin zij eigenlijk helemaal niet zouden mogen verschillen (Juola 2006, 264-265). Niettemin hebben verschillende empirische studies aangetoond dat analyses die worden beperkt tot de meest functionele woorden van een taal betrouwbare indicaties leveren inzake het auteurschap van teksten.[1]

De meerwaarde van functiewoorden is te illustreren aan de hand van een parallel in de schilderkunst, meer bepaald in de theorie van Giovanni Morelli (1816-1891). Ook veel schilderijen zijn anoniem overgeleverd, wat heeft geleid tot bloeiend attributie-onderzoek in de

Giovanni Morelli

1 Specifiek hierover is Argamon & Levitan (2005), maar zie ook de overzichten in Stamatatos (2009, 540-541) of Koppel et al. (2009, 11-12).

kunstgeschiedenis (Wollheim 1972, 177ff). Bijvoorbeeld in het geval van de Italiaanse schilders uit het *Quattrocento* bleek voor Morelli dat de attributie van een werk aan een bepaalde meester niet kon gebeuren aan de hand van de inhoud van een schilderij. Of Christus met vier dan wel drie kruisnagels werd afgebeeld, is goed zichtbaar en daarom makkelijk te imiteren en onderhevig aan processen als beïnvloeding en schoolvorming.

Het scheen beter uit te wijken naar minder opvallende, eerder formele aspecten. Morelli claimde dat de hand van de meester het meest betrouwbaar werd herkend in hoogfrequente maar op het eerste gezicht weinig bijzondere picturale elementen als handen, oren en voeten. Haast ieder kruisigingstafereel verbeeldt immers mensen met handen en voeten, zodat die een betrouwbare basis voor een vergelijking vormen. Een ander voordeel is dat het voorkomen van handen en voeten niet gebonden is aan de inhoud van een schilderij want zowel een kruisigingstafereel als een annunciatie bevatten deze elementen, wat ook in deze kunsttak de vergelijking mogelijk maakt over de grenzen van onderwerpen en genres heen. Ook voor attributies in de schilderkunst wordt dus opvallend genoeg gebruikt gemaakt van 'functiewoorden'.[2]

2.1.3 *Het onbewuste van de schrijver*

De meerwaarde van functiewoorden kan dus op verschillende manieren worden betoogd. Een bijkomend argument dat vaak in de secundaire literatuur wordt vermeld, is hier echter nog niet besproken. Het gebruik van functiewoorden zou niet bewust door een auteur beregeld of gecontroleerd worden tijdens het schrijfproces.[3] Dat zou inderdaad helpen verklaren waarom functiewoorden zich mogelijk gedragen als een 'constante' (*author invariant*) in de oeuvres van auteurs (Koppel et al. 2009, 11). Bovendien is deze verklaring cruciaal vanuit een methodologisch perspectief: als deze onoplettendheid voor functiewoorden geldt voor alle auteurs, geldt deze ook voor vervalsers en imitators. Vanuit deze optiek zouden functiewoorden dus bestand kunnen zijn tegen stilistische imitatie. Hoewel deze assumptie vanzelfsprekend kan lijken, wordt zij in stylometrische publicaties in de regel amper onderbouwd met verwijzingen naar ander onderzoek.[4] Hieronder wil ik ruimte vrijmaken om kort enkele interessante inzichten te bespreken.

author invariant

Beschouw volgende Engelstalige zin en tel hoe vaak de letter 'f' erin voorkomt:

2 Merk op dat deze interpretatie afwijkt van die van Ziolkowski (2004, 185–186), die Morelli nog met de praktijk van 'scientific connoisseurship' en *Sprachgefühl* associeert maar voorbijgaat aan de nieuwigheden die Morelli introduceerde.

3 Dit wordt o.m. vermeld bij Stamatatos (2009, 540), Binongo (2003, 11), Argamon & Levitan (2005) en Peng et al. (2003, 267).

4 Een uitzondering vormt Koppel et al. (2009, 11), met een – hoewel summiere – verwijzing naar Chung & Pennebaker (2007).

Finished files are the result
of years of scientific study
combined with the experience
of many years.

Gewoonlijk zullen de meeste mensen slechts 4 of 5 attestaties van de letter 'f' in deze zin opmerken (Schindler 1978). De meeste lezers gaan voorbij aan de voorkomens van de letter in het voorzetsel 'of'. Dit komt overeen met ander onderzoek dat uitwijst dat lezers meer moeilijkheden hebben met het opsporen van spelfouten in functiewoorden dan inhoudswoorden (Drewnowki & Healy 1978). Een vergelijkbaar effect is geassocieerd met frases als 'Paris in the the spring' (Aronoff & Fudeman 2005, 40-41). Leesexperimenten tonen aan dat veel mensen zich bij een eerste lezing niet bewust zullen zijn van de verdubbeling van het lidwoord 'the'. Lezers merken dergelijke fouten niet op omdat zij sneller voorbijgaan aan lidwoorden en dergelijke functiewoorden minder bewust verwerken. Een dergelijk leeseffect kan immers niet worden waargenomen bij frases als 'Paris in the spring spring', waarin een inhoudswoord foutief wordt verdubbeld.

De onoplettendheid waarmee lezers functiewoorden verwerken, zou vergelijkbaar kunnen zijn met de houding van schrijvers ten opzichte van functiewoorden. Paradoxaal genoeg hoeft deze onoplettende attitude echter niet te impliceren dat functiewoorden onbelangrijk zijn in talige communicatie (Talmy 2000, 21ff). Beschouw de volgende passage:

> Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

Hoewel de letters in deze passage willekeurig lijken dooreen gehaspeld, blijkt de tekst nog makkelijk leesbaar (Rawlinson 1976). Zoals de passage zelf speels vermeldt, is het wel noodzakelijk dat de eerste en laatste letter van een woord niet van plaats veranderen – en dit zijn niet de enige voorwaarden die van kracht zijn.[5] Het is van cruciaal belang dat deze beperking ervoor zorgt dat – de meestal relatief korte – functiewoorden niet fundamenteel van uitzicht veranderen (McCusker et al. 1981). Enkel het intacte uitzicht van deze functiewoordjes in dergelijke passages zou volgens cognitieve wetenschappers al in

5 Deze paragraaf ontleent veel informatie aan een interessante website die wordt onderhouden door M. Davis (Cambridge): `http://www.mrc-cbu.cam.ac.uk/people/matt.davis/Cmabrigde/`. Ik dank B. Vandekerckhove (Antwerpen) voor de verwijzing naar deze site. Merk op dat verschillende (ook Nederlandstalige) varianten van de *Cmabrigde*- en *Finished files*-passages een tijdlang op het internet hebben gecirculeerd.

belangrijke mate bijdragen aan de leesbaarheid ervan. In dat opzicht is het ook interessant om de volgende twee zinnen te vergelijken (Aronoff & Fudeman 2005, 40-41):

> (a) Likewise, if we took the function words out of speech, it would be hard to figure out what was going on.
> (b) Took function words speech hard figure going on.

Functiewoorden zijn duidelijk cruciaal voor de wijze waarop wij informatie structureren in onze talige communicatie (Morrow 1986; Talmy 2000). Toch schijnen leesonderzoekers ervan overtuigd dat zij in natuurlijke taal niet op dezelfde wijze als inhoudswoorden de aandacht op zichzelf vestigen.

2.1.4 *Lettermodellen*

Laten we even terugkeren naar het citaat van Holmes (1994, 87): 'to date, no stylometrist has managed to establish a methodology which is better able to capture the style of a text than that based on lexical items'. In 1994 waren andere types stijlmarkeerders (*style markers*) nooit in staat geweest om functiewoorden in vergelijkende experimenten te verslaan. Combinaties van functiewoorden met andere – bijvoorbeeld syntactische – stijlkenmerken leverden soms wel betere resultaten op dan wanneer men enkel functiewoorden gebruikte, maar deze bijkomende kenmerken versloegen in isolatie nooit de functiewoorden (Van Halteren et al. 2005).[6] De populariteit van de functiewoorden kwam echter onder druk te staan in de jaren negentig van de vorige eeuw toen het gebruik van de zogenaamde *character* n-*grams* opkwam (Kjell 1994).

style markers

character n-grams

Deze wijze van tekstrepresentatie werd oorspronkelijk gebruikt in de Information Retrieval voor de automatische herkenning van talen. In plaats van een stuk tekst in woorden te verdelen, wordt een tekst in deze benadering opgehakt in segmenten van opeenvolgende, gedeeltelijk overlappende reeksen van n letters. Een 'first-order model' zal enkel unigrammen beschouwen (n=1), een 'second-order model' bigrammen (n=2), enzovoorts. Onderstaande voorbeelden illustreren hoe de woorden 'unigram', 'bigram' en 'trigram' onder de overeenkomstige lettermodellen zouden worden voorgesteld. Merk op dat de underscore (_) in deze representatie witruimte (*whitespace*) voorstelt.

unigram
bigram
trigram
whitespace

UNIGRAM – _, u, n, i, g, r, a, m, _ (eerste orde)

BIGRAM – _b, bi, ig, gr, ra, am, m_ (tweede orde)

6 Een technische noot: het is fascinerend dat geavanceerde methodes voor *feature selection* niet altijd selectie-methodes konden verslaan die gewoon frequentie-criteria hanteerden en zo hoogfrequente functiewoorden selecteerden. Zie bv. Argamon & Levitan (2005) of Stamatatos (2009, 545).

TRIGRAM – _tr, tri, rig, igr, gra, ram, am_ (derde orde)

Sinds hun introductie (Kjell 1994) zijn dergelijke modellen de meest krachtige representatie van auteursstijl gebleken in de stylometrie (Juola 2006, § 6.3). Vervolgonderzoek toonde tevens aan dat dit effect zo goed als onafhankelijk was van de onderzochte taal en de gebruikte attributie-algoritmes.[7] Het gebruik van lettermodellen in auteursherkenning is sindsdien enkel in populariteit toegenomen, hoewel zij vooral veel gebruikt worden in de meer technische literatuur waar de techniek oorspronkelijk vandaan kwam. In deze studies speelt 'performantie' namelijk een grote rol en zijn onderzoekers meer geïnteresseerd in de accuraatheid van methodes dan in een theoretische verklaring voor de werking ervan (Luyckx 2010, § 2.2). Mogelijk werden er daarom tot dusver amper geloofwaardige pogingen ondernomen om te verklaren waarom deze lettermodellen zo goed werken.

lettermodellen

performantie

Een populaire stelling is dat *n*-grammen 'een beetje van alles' schijnen te capteren, omdat zij zowel gevoelig zijn voor de vorm als de inhoud van een tekst.[8] Men moet echter opmerken dat een dergelijke 'verklaring' weinig meer doet dan de vraagstelling herhalen, immers: waarom werkt het dan? Koppel et al. (2009, 13) hebben in dat opzicht trouwens terecht reserves uitgedrukt met betrekking tot lettermodellen: precies omdat zij eveneens gevoelig kunnen zijn voor inhoudsgerelateerde aspecten (bijvoorbeeld woordstammen), is hun rol in auteursherkenning onduidelijk.

Het moet meteen worden opgemerkt dat de kracht van lettermodellen kan geassocieerd zijn met een prozaïsche, informatietheoretische verklaring die teruggaat op de eenheid van stilistische vergelijking. Functiewoorden zijn vanuit kwantitatief oogpunt interessant omdat zij vaak optreden en dus veel 'meetmomenten' mogelijk maken in teksten. Hoe meer meetmomenten, hoe betrouwbaarder de representatie van een tekst. In het geval van lettermodellen is dit nog sterker het geval want een tekst telt nu eenmaal meer character *n*-grams dan woorden (Stamatatos 2009, 542). Het woord 'bigram' levert bijvoorbeeld slechts één meetmoment op woordniveau, maar 7 op het bigram-niveau. Aangezien het aantal meetmomenten binnen teksten dus groter is bij lettermodellen dan bij modellen op het woordniveau, biedt dit alvast een gedeeltelijke verklaring voor de superioriteit van lettermodellen. Toch ligt de verklaring mogelijk nog dieper.

2.1.5 *Van functiewoord naar functor*

Recent verscheen het artikel 'Deeper Delta across genres and languages: do we really need the most frequent words?' van Rybicki &

7 Dit effect wordt erkend door Peng et al. (2003), Stamatatos (2009, 541-542) en Koppel et al. (2009, 12-13).

8 Die stelling wordt geopperd bij Houvardas & Stamatatos (2006, 78), Koppel et al. (2009, 12) en Stamatatos (2009, 541).

Eder (2011). In dit artikel stellen deze Poolse onderzoekers de taalonafhankelijkheid ter discussie van *Burrows's Delta* (Burrows 2002), een invloedrijk attributie-algoritme dat hieronder nog in detail aan bod zal komen (§ 7.2). De interne werking van Delta is op dit moment minder relevant dan het feit dat deze techniek werkt met frequentietellingen op het woordniveau. Door de meest frequente woorden uit een corpus te selecteren, extraheert de techniek de frequenties van functiewoorden en gebruikt ze die vervolgens in wiskundige berekeningen. Voortbouwend op een suggestie van Juola (2006, 269), trokken Rybicki & Eder (2011, 315) de aanname in twijfel 'that the use of methods relying on the most frequent words in a corpus should work just as well in other languages as it does in English'.

Burrows's Delta

In hun experimenten betrokken deze onderzoekers een zorgvuldig opgebouwd corpus met teksten in 7 verschillende talen (Engels, Pools, Frans, Latijn, Duits, Hongaars en Italiaans). Hun onderzoeksresultaten tonen aan dat de aanname van Delta's taalonafhankelijkheid waarschijnlijk fout is:

taalonafhankelijkheid

> while Delta is still the most successful method of authorship attribution based on word frequencies, its success is not independent of the language of the texts studied. This has not been noticed so far for the simple reason that Delta studies have been done, in a great majority, on English-language prose. [...] The relatively poorer results for Latin and Polish – both highly inflected in comparison with English and German – suggests the degree of inflection as a possible factor. This would make sense in that the top strata of word frequency lists for languages with low inflection contain more uniform words, especially function words; as a result, the most frequent words in languages such as English are relatively more frequent than the most frequent words in agglutinative languages such as Latin (Rybicki & Eder 2011, 319-320).

Hun punt van kritiek is evident: de beperking tot functiewoorden in stylometrisch onderzoek is suboptimaal voor talen die minder gebruik maken van functiewoorden. Zij suggereren bovendien terecht dat dit inzicht slechts relatief recent werd verkregen, omdat de meerderheid van het stylometrisch onderzoek totnogtoe werd uitgevoerd op Engelstalige teksten.

Engels is een schoolvoorbeeld van een taal die geen hevig gebruik maakt van casus-uitgangen (of andere vormen van inflectie) voor het uitdrukken van grammaticale informatie (Sapir 1921, hoofdstuk VI). Functionele linguïstische informatie wordt in dergelijke talen uitgedrukt via het gebruik van – meestal korte – functiewoorden (bv. '*with* a sword'). Structurele informatie wordt dus meestal vormgegeven via minimale betekeniseenheden of 'grammaticale morfemen' die als een

inflectie

grammaticale morfemen

klein maar zelfstandig woord zijn gerealiseerd (Morrow 1986). Engels kan zo gecontrasteerd worden met een andere historische *lingua franca*. Latijn is een schoolvoorbeeld van een taal waarin grammaticale informatie op een heel andere manier wordt uitgedrukt, namelijk via affixen. Affixen zijn net als functiewoorden grammaticale morfemen; zij worden echter niet gerealiseerd als afzonderlijke woorden maar worden gehecht aan inhoudswoorden. In het Latijnse woord *ensi* (ablatief enkelvoud: 'met een zwaard') is het affix *-i* als een gebonden morfeem aan een stam gehecht en vervult het zo een gelijkaardige rol als het voorzetsel 'with' in de Engelstalige constructie 'with a sword'.

affix

gebonden morfeem

Het is daarom evident dat de Delta-methode minder goed werkt voor synthetische talen als het Latijn of het Pools: de methode voert immers metingen uit op het woordniveau en is blind voor alle andere processen die zich afspelen onder het woordniveau. Al hetgeen hierboven werd gezegd over functiewoorden, schijnt slechts toepasbaar op functiewoorden als 'grammaticale *morfemen*' en niet als 'grammaticale *woorden*'. Voor een taal als het Engels is het een gelukkig toeval dat men grammaticale morfemen makkelijk kan bereiken door een tekst in woorden te verdelen. Voor talen die gebruik maken van een andere linguïstische logica zullen metingen op het woordniveau vaak niet toereikend zijn: men moet daar immers een morfologische analyse doorvoeren om de relevante grammaticale morfemen los te maken van hun stammen.

synthetische taal

Vanaf het volgende hoofdstuk zal ik daarom niet langer spreken van 'functiewoorden', zoals gebruikelijk in het internationale onderzoek, omdat deze misleidende nomenclatuur te sterk voortkomt uit de sterke positie van het onderzoek naar Engelstalige teksten in de stylometrie. In plaats daarvan zal ik gebruik maken van de term 'functor' uit het psycholinguïstisch onderzoek – waar de term overigens equivalent is aan het begrip 'grammaticaal morfeem' (Kwon 2005, 1-2). Brown definieerde het concept 'functor' als volgt:

functor

> forms that do not, in any simple way, make reference. They mark grammatical structures and carry subtle modulatory meanings. The word classes or parts of speech involved (inflections, auxiliary verbs, articles, prepositions, and conjunctions) all have few members and do not readily admit new members (Brown 1973, 75).

Afgezien van het feit dat de term 'functor' iets aangenamer klinkt dan de term 'grammaticaal morfeem', heeft de term 'functor' een duidelijke meerwaarde. Ten eerste, zijn alle functie*woorden* volgens Brown's definitie ook functoren. De term functor verandert de traditionele focus van de stylometrie dus niet maar breidt deze wel uit naar alle andere grammaticale morfemen die niet noodzakelijk als afzonderlijke woorden zijn gerealiseerd. Merk op dat alle hierboven vermelde voordelen – zoals inhoudsonafhankelijkheid, minder bewust gebruik en hoge frequentie – ook toepasbaar zijn op functoren.

Een ander voordeel is dat het internationaal onderzoek vaak de nadruk legt op de taalonafhankelijkheid van technieken. In 2005 werd bijvoorbeeld de zogenaamde *Stylome Hypothesis* geformuleerd (Van Halteren et al. 2005). Deze ambitieuze hypothese stelt dat het op een dag mogelijk moet zijn om het 'styloom' of individueel stijleigen van iedere auteur te karakteriseren, en wel op zulk een manier dat men een individueel styloom kan onderscheiden van het styloom van iedere andere auteur op aarde. Deze – vanzelfsprekend onbewezen – hypothese maakt geen onderscheid tussen auteurs wat betreft leeftijd, regio, tijd maar ook taal. Als het blootleggen van stylomen gebruik zal maken van grammaticale morfemen, schijnt het aangewezen om deze laatsten aan te duiden met de term functor aangezien deze definitie veel sterker onafhankelijk is van de taal van een auteur (Talmy 2000, 21ff).

Styloom-hypothese

2.2 DE MEDIONEERLANDISTIEK

Hierboven is toegelicht waarom men in de internationale stylometrie veel aandacht besteedt aan het gebruik van functiewoorden en andere grammaticale morfemen bij het karakteriseren van een individuele schrijfstijl. Bovendien is uitgelegd hoe deze nadruk op functoren historisch is gegroeid in het internationale auteursonderzoek. Tot voor kort waren dergelijke inzichten veelal onbekend bij onderzoekers van Middelnederlandse literatuur.[9] In deze sectie wil ik daarom kort de geschiedenis bespreken van het jongste onderzoek naar stilistische auteursherkenning in Middelnederlandse teksten. Ik beperk mij daarbij tot onderzoek dat expliciet kwantitatief en/of computationeel van aard was, omdat een exhaustief overzicht van het onderzoek naar auteursattributie in het algemeen nodeloos ver zou leiden.[10]

Het is aan het begin van dit hoofdstuk raadzaam op te merken dat men uitgerekend in de kwantitatief geïnspireerde stylometrie in zekere zin een relatief wazig gedefinieerd stijlbegrip hanteert. Men baseert zich bijvoorbeeld vaak op tellingen, onder andere van individuele

9 Merk op dat de mediëvistiek in andere taalgebieden niet heel veel beter was ingelicht. Stylometrisch mediëvistisch onderzoek is bijzonder schaars – althans, als men het onderzoek naar computationele stemmatologie niet tot de stylometrie rekent (consulteer daarvoor de excellente studie in Salemans 2000). Vermeldenswaardige uitzonderingen in het internationale onderzoek (met dank aan K. van Dalen-Oskam) zijn: García & Martín (2007) en Treschow et al. (2009) voor Oud-Engels, Dimpel (2004, 2006) voor Middelhoogduitse teksten en de casus van Veldeke, Allen (1974) voor het *Chanson de Roland* en Schmidt (1993) en Thaisen (2010) voor Middelengelse teksten. Het Latijn kent een iets uitgebreidere traditie; zie bv. Ziolkowski (2004) en Stella (2008) maar zeker ook het kritische werk van Rudman (1998a, b). Omdat deze studies erg disparaat zijn, wijd ik er geen aparte bespreking aan. Onbesproken blijft het interessante maar nog ongepubliceerde onderzoek van Hans van Halteren en Margit Rem. Van Rem (2005) is wel reeds een kort artikel verschenen dat onrechtstreeks handelt over literaire auteursherkenning. Binnenkort verschijnt ook een mooie studie van Stapel (2012), die nu reeds online raadpleegbaar is.

10 Dit overzicht is grotendeels gebaseerd op Kestemont (2012b).

woordjes, die lastig zijn te verbinden aan de complexe fenomenen die men in de traditionele stilistiek bestudeert – een schoolvoorbeeld daarvan zijn de stijlfiguren uit de klassieke retorica. Een anonieme beoordelaar van deze studie stelde daarom terecht dat men in de stylometrie lijkt te veronderstellen 'dat alles wat maar in een tekst zichtbaar of meetbaar is een weerspiegeling is van de stijl van een auteur'. Een dergelijk, nogal *impliciet* stijlbegrip valt te contrasteren met de wijze waarop de stilistiek traditioneel is beoefend in de Nederlandse letterkunde.

In een belangrijk overzichtsartikel over stijlonderzoek binnen de medioneerlandistiek spreekt Joost van Driel over stijl als een lastig definieerbaar begrip dat hijzelf opvat 'als een notie die betrekking heeft op het taalgebruik van een auteur, als de wijze waarop hij zich uitdrukt in taal' (Van Driel 2003, 98). 'Formele aspecten', zo stelt hij, 'zoals de vormgeving van woorden, verzen of dialogen, zijn bij uitstek het terrein van stijlonderzoek'. Belangrijk is dat dergelijk onderzoek zich meestal toelegt op een aantal stilistische categorieën die op voorhand expliciet worden afgebakend, zoals metriek, epitheta, publieksaansporingen, vormgeving van dialogen etc. In zijn eigen proefschrift onderzocht Van Driel (2007) enkele jaren later bijvoorbeeld in opeenvolgende hoofdstukken rijmtechniek, informatiedichtheid, repetitieve stijl, conversaties en 'aanschouwelijkheid' in Middelnederlandse epische teksten. Dergelijk onderzoek maakt op voorhand expliciet duidelijk op welke specifieke stijlverschijnselen het is gericht. Zoals hieronder ook nog ruimschoots zal blijken, onderzoekt men in de hedendaagse stylometrie verschijnselen die minder duidelijk zijn te verbinden met de traditionele focuspunten in stilistisch onderzoek.

2.2.1 *De eerste zwaluwen*

Eén van de eerste toepassingen van de inzichten uit de internationale stylometrie vinden we in een relatief vroeg artikel in *Dutch crossing* van Saskia Murk Jansen (1988). In deze bijdrage beschrijft zij een experiment aangaande het auteurschap van de zogenaamde *Mengeldichten*, een serie van 29 gedichten die voorkomt in een aantal handschriften met werk van de Brabantse mystica Hadewijch (De Baere 1993). Traditioneel herkent men in deze verzameling drie reeksen van gedichten (1-16, 17-24, 25-29). Slechts de eerste 16 *Mengeldichten* worden traditioneel aan Hadewijch toegeschreven, omdat de daaropvolgende gedichten op verschillende niveaus sterk afwijken van de rest van Hadewijchs oeuvre.

Principal components analysis

Murk Jansen was benieuwd of de stylometrie deze bevindingen zou onderschrijven en voerde een *Principal components analysis* uit van de *Mengeldichten*. Deze techniek uit de multivariate statistiek komt hieronder nog ruimschoots aan bod en wordt vandaag de dag vaak gebruikt in kwantitatieve auteursstudies. Murk Jansen verwees expli-

ciet naar het werk van Mosteller en Wallace en beperkte haar analyses tot de relatieve frequenties van een dertiental Middelnederlandse functiewoorden. Haar resultaten leverden geen eenduidig resultaat op, maar zij dacht dat het wel interessant zou zijn om een dergelijke, kwantitatieve onderzoekslijn verder te verkennen.[11]

Hoewel Murk Jansens onderzoek niet meteen navolging heeft gevonden, ontstond er in de jaren tachtig en negentig van de vorige eeuw wel een trend waarbij verschillende onderzoekers in toenemende mate de mogelijkheden verkenden van kwantificerend stilistisch onderzoek.[12] Initieel werd er echter maar zelden statistiek toegepast en de computationele hulpmiddelen die her en der werden ingezet, bleven meestal bescheiden. Van den Berg (1983) promoveerde bijvoorbeeld op een kwantitatief onderzoek naar de geografische en diachrone variatie van onder meer enjambementen, epitheta en andere stilistica. Van den Bergs invloedrijke werk was niet gebaseerd op geavanceerde kwantificering, hoewel de onderzoeker occasioneel wel statistische tests uitvoerde om de relevantie van zijn resultaten te controleren.

Gelijkaardig was het promotie-onderzoek door Kuiper (1989) van de *Ferguut*. Kuiper beargumenteerde dat deze ridderroman door twee verschillende auteurs was geschreven en baseerde zich daarvoor op de grote verschillen tussen het eerste en tweede deel inzake de frequenties van diverse taal- en stijlkundige karakteristieken. Hij maakte daarbij geregeld gebruik van statistische toetsen. Nieuw in Kuipers aanpak was dat hij gedeeltelijk steunde op een computationele verwerking van de tekst, bijvoorbeeld voor het automatisch aanleggen van woordenschatlijsten. Een vergelijkbare strategie werd aangewend door Besamusca (1991) in zijn studie van de *Lancelotcompilatie*, een grote collage van Arturromans waarin een anonieme compilator verschillende, reeds bestaande teksten in elkaar heeft gevlochten (vgl. § 8.2). Ook Besamusca gebruikte de computer bij het aanleggen van woordenschatlijsten voor het stilistisch onderzoek naar de identiteit van deze compilator.

Cd-rom Middelnederlands

De computationele studie van middeleeuws auteurschap nam pas een hoge vlucht met het verschijnen van de *Cd-rom Middelnederlands* in 1998 die – naast het volledige *Middelnederlandsch woordenboek* – een haast exhaustieve verzameling bevatte van de belangrijkste Middel-

11 Een gelijkaardig kwantitatief (hoewel niet-computationeel) onderzoek werd uitgevoerd door Reynaert (1974). De Baere (1993) biedt een kritische analyse van het stylometrisch onderzoek in Murk Jansens latere proefschrift (Murk Jansen 1991), waarin zij concludeert dat de tweede reeks *Mengeldichten* mogelijk zijn geschreven door Hadewijch – of toch iemand uit de onmiddellijke omgeving van de mystica. De Baere (1993) uitte echter gegronde twijfels bij de door Murk Jansen gevolgde onderzoeksmethode en pleitte voor doorgedreven onderzoek.

12 Frank Willaert wees mij erop dat het onderzoek van Murk Jansen mogelijk weinig weerklank heeft gevonden omdat het onderwerp van haar bijdrage op dat moment reeds als een *non-issue* werd ervaren. Het auteurschap van de tweede en derde reeks *Mengeldichten* werd – zeker sinds de editie door Van Mierlo (1952) – immers nog maar zelden aan Hadewijch werd toegekend.

nederlandse teksten. Het is wellicht moeilijk om de impact van dit schijfje te overschatten, want sinds 1998 heeft de *Cd-rom* misschien wel het belangrijkste vertrekpunt gevormd voor de computationele studie van de Middelnederlandse letterkunde. Het belangrijkste voordeel was dat men teksten niet alleen in digitale vorm van de *Cd-rom* kon 'oogsten', maar dat deze teksten ook volledig doorzoekbaar waren. Vooral deze geavanceerde zoekfunctionaliteit is vrijwel onmiddellijk door onderzoekers uitgebuit die geïnteresseerd waren in auteursherkenning.

In 2001 heeft Westgeest (2001) de zogenaamde Leidse *Lapidarium*-fragmenten toegeschreven aan Jacob van Maerlant. Westgeest maakte in zijn onderzoek volop gebruik van de *Cd-rom* en was de eerste om de doorzoekbare teksten als een soort controle-corpus te gebruiken. Hij betoogde dat de stilistische eigenheden van de *Lapidarium*-fragmenten relatief sterk verwant waren met andere Maerlant-teksten maar nagenoeg afwezig of toch veel minder frequent waren in teksten van andere auteurs.

Een ander bekend artikel verscheen in 2002. Daarin onderzocht Reynaert (2002) de stijl van de teksten binnen de 'Antwerpse School': een coherente verzameling didactische, ethische en historiografische geschriften die alle tot stand kwamen in de eerste helft van de veertiende eeuw te Antwerpen (vgl. § 7). Slechts een enkele auteur uit deze vermeende 'School' kent men bij naam: Jan van Boendale. Reynaert beargumenteerde dat veel meer – zoniet alle – Antwerpse werken mogen toegeschreven worden aan de Antwerpse schepenklerk Boendale. Reynaert herkende in de 'Antwerpse School' dus een wetenschappelijke schijnconstructie, die in werkelijkheid gewoon samenviel met een enkel oeuvre, namelijk dat van Boendale. In zijn onderzoek had Reynaert de *Cd-rom* gebruikt op nagenoeg dezelfde wijze als Westgeest: hij beweerde dat de teksten in het Antwerpse corpus veel kenmerken gemeen hadden die zij niet deelden met andere teksten op de *Cd-rom*. Hogenbirk (2009b) publiceerde recent nog een onderzoek waarin zij Besamusca's eerder vermelde studie van de *Lancelot*-compilator (Besamusca 1991) aanvulde op basis van een methodologie die heel sterk gelijkt op die van Westgeest en Reynaert.

2.2.2 *Delta voor Walewein*

Het lijkt kenschetsend voor deze vroege toepassingen dat men wel enthousiast verkende hoe computationeel onderzoek kon worden ingezet voor auteursonderzoek maar dat men erg onzeker was over de te volgen methodes. Die onzekerheid lijkt in grote mate samen te hangen met het feit dat men in deze periode nog opvallend slecht was ingelicht over het internationale onderzoek op het gebied van de sty-

lometrie.[13] Daar had men intussen veel vooruitgang geboekt – hoewel aanvankelijk slechts een enkeling op de hoogte leek van het onderzoek van bijvoorbeeld Mosteller en Wallace. Het was pas wanneer de medioneerlandistiek in het eerste decennium van de eenentwintigste eeuw nauwer aansluiting zocht bij de internationale *Digital humanities* dat het computationele onderzoek naar auteursherkenning echt van de grond kwam.

Het eerste voorbeeld van deze ontwikkeling vinden we in een korte maar interessante bijdrage van Croenen (2005), een belangrijke protagonist in computer-geassisteerd onderzoek aan de universiteit van Liverpool. In dit artikel behandelde Croenen de eerder geopperde mogelijkheid van een dubbel auteurschap in de *Grimbergsche oorlog*, een gefictionaliseerde rijmtekst over een vroegmiddeleeuwse vete tussen het Huis van Grimbergen en de Brabantse hertogen. Croenen verdeelde de tekst in opeenvolgende steekproeven van 500 verzen. Vervolgens voerde hij een statistische cluster-analyse uit van deze samples. Hoewel zijn methodologie slechts kort wordt toegelicht, boden zijn resultaten ondersteuning voor de hypothese dat een tweede dichter de auteur die de tekst was begonnen, heeft afgelost. De cluster-analyse maakte eveneens aannemelijk dat de tweede auteur de proloog van de tekst later heeft toegevoegd.

De meest verreikende bijdrage aan de Middelnederlandse stylometrie tot dusver is van Van Dalen-Oskam & Van Zundert (2007) en is te vinden in *Literary and Linguistic Computing*, een tijdschrift dat het onbetwiste vlaggenschip vormt van de internationale stylometrie.[14] In dit artikel onderzochten zij de *Roman van Walewein*, een Middelnederlandse ridderroman over de neef van koning Artur, Walewein (zie § 8.2.4). De verteller van de tekst vermeldt in de epiloog dat ene *Penninc* de tekst was begonnen maar niet heeft afgemaakt: de *Walewein* (ca. 11.000 verzen in totaal) zou zijn vervolledigd door een tweede auteur, *Pieter Vostaert*, die 'ongeveer' 3.300 verzen aan Penncs aandeel zou hebben toegevoegd. Verscheidene onderzoekers hebben gespeculeerd over de exacte plaats in de tekst waar Vostaert zijn voorganger zou hebben afgelost. Men beschikte echter nog niet over doorslaggevend bewijs voor een van de vele hypotheses die werden geopperd. Van Dalen-Oskam en Van Zundert wilden daarom gebruik maken van kwantitatieve auteursherkenning om de exacte plaats van de auteursovergang in de *Walewein* vast te stellen.

Literary and Linguistic Computing

Van Dalen-Oskam en Van Zundert pasten daarom 'Burrows's Delta' toe op de *Walewein*. Deze invloedrijke Delta-metriek werd oorspronke-

Burrows's Delta

13 Illustratief is dat Burgers (1999, 83, n. 83) nog stellig beweerde dat kwantitatieve auteursherkenning niet bestond, terwijl op dat moment wel reeds enkele belangrijke overzichtspublicaties van het vakgebied waren verschenen.

14 Recent verscheen ook een (deels stylometrisch) onderzoek naar het auteurschap van de *Beatrijs* (Draux 2009). Ik bespreek deze publicatie verder niet in dit overzicht omdat ik in een eerdere recensie heb beweerd dat dit onderzoek zowel los staat van de recente internationale stylometrie als van de ontwikkelingen in de medioneerlandistiek zelf

lijk voorgesteld door J. Burrows en komt hieronder nog uitgebreid aan bod: het is een manier om de afstand in schrijfstijl tussen twee stukken tekst te berekenen. Simpel gesteld: Delta vergelijkt twee samples met tekst op basis van de relatieve frequenties erin van een kleine verzameling (bv. 50) hoogfrequente woordjes. Delta berekent daarna een dissimilariteitsscore (vandaar: 'Delta') door de frequentie-verschillen tussen beide samples statistisch te vergelijken (vgl. § 7.2). Een tekst van onbekende signatuur kan vervolgens aan een gekende auteur worden toegeschreven door de afstand van de anonieme tekst te berekenen tot de teksten van een aantal relevante controle-auteurs. Het auteurschap van de betwiste tekst kan worden toegekend aan de auteur ten aanzien van wiens andere teksten de anonieme tekst het dichtste staat. Empirische testen tonen aan dat Delta vaak betrouwbare resultaten oplevert, terwijl het een intuïtieve en eenvoudige methode is.

windowing procedure

windows

crossfade

Voor de *Walewein* ontwikkelden Van Dalen-Oskam en Van Zundert een raamwerk-procedure of *windowing procedure*. Ze isoleerden het verdachte gebied uit de tekst waar de auteursovergang kon hebben plaatsgevonden – omstreeks vers 8.000 – en verdeelden dit gebied in opeenvolgende steekproeven of *windows* van een gelijke lengte. Ze extraheerden ook schrijfsamples uit Pennincs begin en Vostaerts einde – respectievelijk lang vóór en lang ná het gebied met de vermeende auteursovergang. Vervolgens gingen zij door het verdachte gebied en berekenden de Delta tussen, enerzijds, elk van de verdachte 'ramen' en, anderzijds, de stukken die zeker werden geschreven door Penninc en Vostaert. Zij hoopten dat de reeks Delta-waarden die hieruit voortkwam op een specifiek punt snel zou inboeten aan stilistische gelijkenis met Pennincs werk en steeds sterker op Vostaerts stijl zouden gaan gelijken. Als een dergelijke *crossfade* zou optreden, lag het natuurlijk voor de hand dat de plaats daarvan zou overeenkomen met de auteursovergang in de *Walewein*.

De reeks Delta-waarden die zij verkregen voor verschillende experimenten, vertoonde inderdaad de verhoopte *crossfade*. Zij beargumenteerden overtuigend dat de auteursovergang zich voordoet omstreeks vers 7.880 van de tekst. Niettemin bleek hun onderzoek een schoolvoorbeeld van serendipiteit. Verschillende van de visualisaties van de Delta-waardenreeks bevatten een merkwaardige piek (c.q. stijlovergang) omstreeks vers 6.000 in de tekst (zie Fig. 2, p. 37). Vreemd genoeg bevatte de tekst zelf op dat moment geen duidelijk narratieve breuk die deze piek zou kunnen verklaren – hoewel die piek in veel van hun curves nog prominenter was dan het gebied met de auteursovergang.

Al snel achterhaalden de onderzoekers dat deze piek overeenkwam met een kopiistenovergang. Voor hun onderzoek maakten de Nederlandse onderzoekers gebruik van een digitale editie van het enige

(Kestemont 2011b). Verder is ook de recente dissertatie van De Bundel (2009) relevant (de handelseditie hiervan is in voorbereiding).

volledige handschrift waarin de *Walewein* is overgeleverd (LEIDEN, UNIVERSITEITSBIBLIOTHEEK, LTK. 195). Eerder had paleografisch onderzoek al aangetoond dat twee verschillende kopiisten een hand hebben gehad in de vervaardiging van het *Walewein*-afschrift in deze codex. De eerste kopieerde slechts de eerste 5.783 regels van de tekst, terwijl een tweede kopiist de rest voor zijn rekening nam (zie Fig. 3, p. 38). De tweede kopiist voegde een epiloogje aan de tekst toe waarin hij zijn werkzaamheden dateerde in 1350. Van Dalen-Oskam en Van Zundert moesten daarom vaststellen dat Delta, indien toegepast op middeleeuwse teksten, niet alleen gevoelig bleek voor auteursstijl maar ook – en misschien nog meer – voor de stijl van de (laatste?) kopiist van de tekst. Een cruciaal inzicht dat zij met dit onderzoek bereikten, was dat de karakteristieken van kopiisten kunnen interfereren met auteurskenmerken in de computationele stijlanalyse van middeleeuwse teksten.

kopiisteninterferentie

2.2.3 *Tien geboden*

Het onderzoek van Van Dalen-Oskam & Van Zundert (2007) legde belangrijke theoretische inzichten bloot en veroorzaakte een breekpunt in de computationele medioneerlandistiek. Blijkbaar waren attributietechnieken, veelal ontwikkeld voor veel modernere teksten, niet *out of the box* geschikt voor de auteursanalyse van middeleeuwse werken. Zij moesten namelijk een belangrijke hindernis zien te neutraliseren: de kopiist. In het eerste hoofdstuk werd al aangestipt dat middeleeuwse teksten voor de uitvinding van de boekdrukkunst individueel werden overgeschreven door kopiisten. Het ontbreken van een standaardspelling en -taal in deze periode leidde tot een grote variatie tussen handschriften (Kestemont et al. 2010, 287-288). Van sommige teksten zijn meerdere, parallelle handschriften door verschillende kopiisten overgeleverd. Deze tonen aan hoe makkelijk kopiisten hun voorbeeldexemplaar konden aanpassen aan hun eigen wensen (Kestemont & van Dalen-Oskam 2009, 121-122).

Een belangrijk evenement voor deze problematiek was de oratie die Brinkman (2010) heeft uitgesproken bij zijn aanstelling als hoogleraar aan de universiteit van Amsterdam.[15] Deze benoeming werd overigens financieel mogelijk gemaakt door het Huygens-instituut: die institutionele vermelding is niet zonder belang omdat ook Van Dalen-Oskam en Van Zundert verbonden zijn aan dit instituut. Het Huygens ING heeft zich de laatste jaren duidelijk gepositioneerd als een toonaange-

Huygens-instituut

15 Deze rede werd uitgesproken op 19 maart 2010 te Amsterdam en droeg de titel: 'Als de nachtuil. Auteurschap en overlevering van middeleeuwse teksten'. Ik dank Brinkman voor het beschikbaar stellen van zijn tekst en de bijhorende presentatie. Een herwerkte versie van deze rede zal in de nabije toekomst verschijnen bij de Vossiuspers, een imprint van University Press Amsterdam.

vende speler wat betreft computationele filologie in de Lage Landen.[16] Brinkman had tot dan nog geen opvallende initiatieven ondernomen inzake kwantitatieve filologie maar verraste het publiek met een uitstekende presentatie over de kwantitatieve studie van het auteurschap van Middelnederlandse teksten met computationele hulpmiddelen.

Brinkman illustreerde hoe verschillende technieken uit de stylometrie traditionele onderzoeksresultaten over het auteurschap van teksten konden bevestigen. Niettemin bleef Brinkman voorzichtig want de stylometrische studie van Middelnederlandse teksten gaat niet over rozen. Ten eerste benadrukte hij terecht dat op attributie-technieken sowieso een aantal beperkingen rusten: ook moderne teksten moeten bijvoorbeeld lang genoeg zijn om ze op een overtuigende manier te kunnen toeschrijven aan een auteur. Bovendien, stelde Brinkman, moet de middeleeuwse filologie een groot aantal bijkomende hindernissen zien te overwinnen. Veel problemen, zoals het schoolvoorbeeld van de orthografische variatie, zijn eigen aan middeleeuwse teksten en moeten door kwantitatieve onderzoekers op gepaste wijze worden omzeild.

Tien Geboden van Brinkman

Voor de stylometrische studie van middeleeuws auteurschap heeft Brinkman 'Tien Geboden' geformuleerd: als en slechts als een combinatie van middeleeuwse teksten aan deze tien criteria beantwoordden, leek het volgens Brinkman aanvaardbaar om hun auteurschap te onderwerpen aan een stylometrische studie. Een vrije parafrase daarvan is hieronder weergegeven; merk op dat enkele geboden stylometrische inzichten repliceren die in dit hoofdstuk reeds uitvoerig belicht zijn. Ik zal op dit moment nog niet heel expliciet of gedetailleerd ingaan op deze Geboden, die in zekere zin kunnen beschouwd worden als een mijns inziens betrouwbare *status quaestionis* wat betreft de stylometrische studie van auteurschap in de medioneerlandistiek. Ik zal in de rest van deze studie echter geregeld mijn onderzoeksresultaten terugkoppelen naar Brinkmans richtlijnen, omdat zij het methodologische achterdoek vormen waartegen veel van mijn onderzoeksresultaten moeten gesitueerd worden. Ook in het besluit van deze studie zal ik van deze Geboden gebruik maken om kritisch het resultaat van mijn onderzoek te evalueren.

Eerste Gebod: Eenheid van kopiist

Op dit moment schijnt vooral het Eerste Gebod – DE EENHEID VAN KOPIIST – wel reeds bijzonder belangwekkend (vgl. ook Kestemont & van Dalen-Oskam 2009), niet het minst omdat het voor de studie van auteurschap vanzelfsprekend slecht nieuws inhoudt. De auteursexemplaren van teksten zijn immers zelden overgeleverd en in de regel moet men werken met veel jongere, mogelijk onbetrouwbare afschriften van de autografen. In de doorgeefketting van kopiisten die tot een dergelijk afschrift heeft geleid, kan iedere opeenvolgende afschrijver

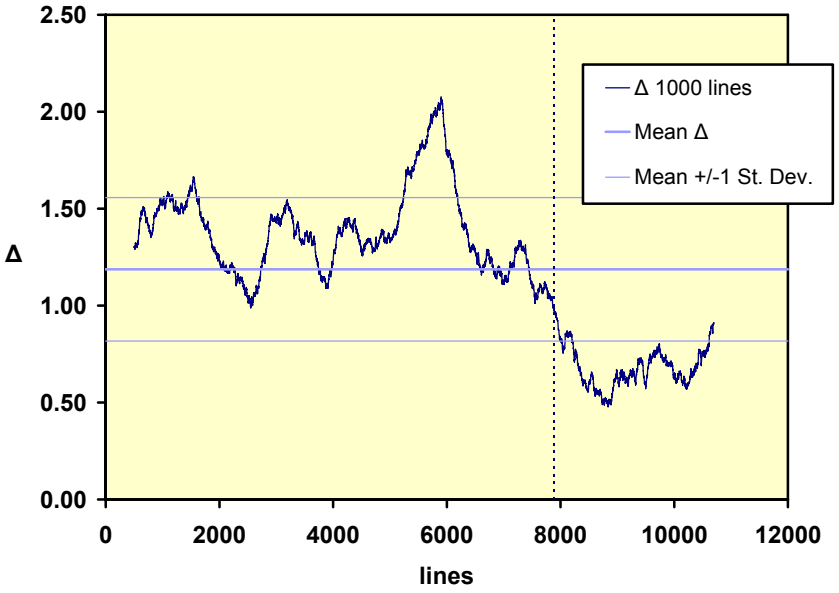
16 Zie http://www.huygens.knaw.nl/. Na een recente fusie met het vroegere Instituut voor Nederlandse Geschiedenis maakt het oorspronkelijke Huygens-instituut nu deel uit van het 'Huygens ING', een onderzoeksinstituut in Den Haag dat wordt gefinancierd door de Koninklijke Nederlandse Akademie van Wetenschappen.

vrijelijk aanpassingen hebben doorgevoerd, zodat de auteursstijl met iedere nieuwe kopie het gevaar liep om verder te verbleken. Aangezien slechts weinig Middelnederlandse teksten in een afschrift van dezelfde kopiist zijn overgeleverd (Van Oostrom 1994, 6), zijn de meeste werken niet rechtstreeks vergelijkbaar voor een stylometrische auteursanalyse.

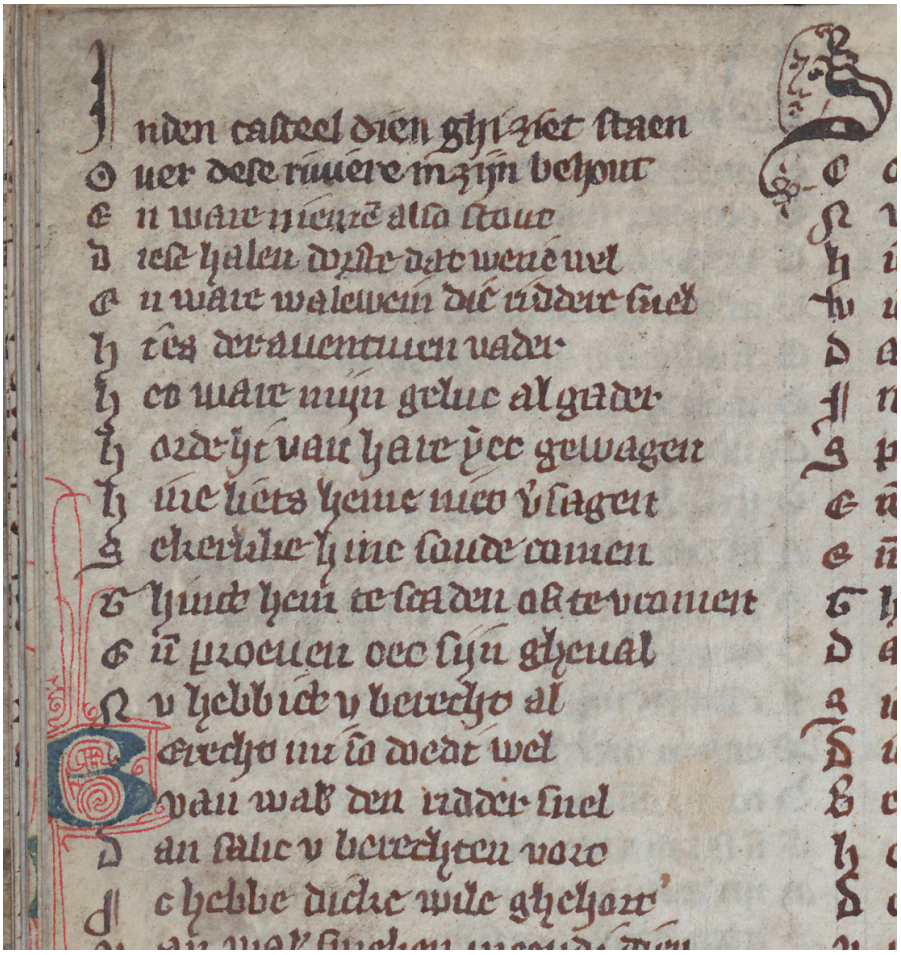
Het Eerste Gebod brengt dus een allesbehalve hoopvolle boodschap voor stylometrici in de medioneerlandistiek die geïnteresseerd zijn in het auteurschap van teksten. Voor de codicologie en paleografie daarentegen roept het Eerste Gebod nieuwe en vooral spannende onderzoeksvragen in het leven. Als de karakteristieken van kopiisten auteursanalyses hinderen, kan men zich immers de vraag stellen hoe stabiel deze karakteristieken dan precies zijn. Zouden deze bijvoorbeeld kenmerkend genoeg zijn om middeleeuwse kopiisten te herkennen, analoog aan de wijze waarop men in de stylometrie auteurs tracht te herkennen? In dat geval kunnen we de intrigerende mogelijkheid opperen dat men heel binnenkort misschien wel de hand van een kopiist kan herkennen in een afschrift, zonder dat afschrift ooit echt gezien te hebben. Stylometrische analyses beperken zich immers tot louter linguïstische gegevens. Over die problematiek handelt het volgende hoofdstuk.

De 'Tien Geboden'
voor stylometrische auteursherkenning
in middeleeuwse teksten (Brinkman 2010)

1. Eenheid van kopiist – Slechts teksten die zijn overgeleverd in een afschrift van dezelfde kopiist zijn rechtstreeks vergelijkbaar wat betreft auteursstijl;
2. Inhoudsonafhankelijkheid – Analyses mogen slechts worden uitgevoerd op basis van woordfrequenties die onafhankelijk zijn van specifieke inhoud, genre, discours en vertelperspectief;
3. Tekstlengte – Teksten moeten lang genoeg zijn (ca. 2000 woorden) voor een betrouwbare stylometrische analyse;
4. Sample-grootte – Als samples van teksten worden vergeleken, wordt de grootte van deze samples best constant gehouden;
5. Eenheid van genre – Teksten moeten tot hetzelfde of toch een vergelijkbaar genre behoren;
6. Eenheid van ontstaansdatum – Teksten moeten eenzelfde of toch een vergelijkbare ontstaansdatum hebben;
7. Eenheid van ontstaansregio – Teksten moeten in eenzelfde of toch een vergelijkbare regio zijn geschreven;
8. Minimale afstand tot de autograaf – Slechts teksten komen in aanmerking die niet te ver staan van de oorspronkelijke auteurstekst;
9. Tertium comparationis – Twee teksten kunnen slechts zinnig vergeleken worden als er ook relevante 'derden' in het onderzoek worden betrokken;
10. Samples – Als twee teksten worden vergeleken is het aangewezen om deze in samples van gelijke grootte te verdelen, zodat men een zicht krijgt op de stabiliteit van het resultaat van de analyses.



Figuur 2: Grafiek resulterend uit een Delta-experiment op de *Walewein* door Van Dalen-Oskam & Van Zundert (2007, 538, Fig. 19 – met dank aan de auteurs). De merkwaardige piek in de Delta-curve omstreeks vers 5.784 lijkt geassocieerd met de kopiisten-overgang op deze plaats in het handschrift (vgl. Fig. 3, p. 38). Verder is ook het auteursverschil zichtbaar tussen Penninc en Vostaert want aan weerszijden van de verticale stippellijn (ca. vers 7.880) vertoont de curve respectievelijk hogere en lagere waardes.



Figuur 3: Fotografische reproductie van het begin van de eerste kolom op fol. 152v in handschrift Leiden, Universiteitsbibliotheek, Ltk. 195. Dit handschrift is het enige volledige overgeleverde afschrift van de *Walewein* en lag aan de basis van de experimenten van Van Dalen-Oskam & Van Zundert (2007). De eerste twee verzen in deze afbeelding zijn nog van de eerste kopiist; de rest van het afschrift is geschreven door een tweede hand (die zijn werk volgens het colofon in 1350 beëindigde).

3 DE JACHT OP DE KOPIIST

Hoewel deze studie vanaf het volgende hoofdstuk volledig gewijd is aan auteursherkenning in de Middelnederlandse epiek, zal ik mij in dit hoofdstuk toeleggen op kopiistenherkenning in Middelnederlandse teksten. Dit hoofdstuk moet vooral illustreren hoe stylometrische technieken die werden ontwikkeld voor moderne teksten erg gevoelig zijn voor de (in de eerste plaats orthografische) variatie in middeleeuwse teksten. Het is belangrijk om hier lang genoeg bij stil te staan, omdat deze kopiistenproblematiek bepalend zal zijn voor de manier waarop auteursherkenning in de rest van deze studie zal worden benaderd.

De casus waarmee ik zal werken, is die van de parallelle handschriften van de Middelnederlandse *Rijmbijbel*. Aangezien de gebruikte methode voor deze kopiistenanalyse hieronder volledig wordt ontleend aan het onderzoek naar auteursherkenning in moderne teksten, laat deze kopiistencasus toe om in detail een stylometrische methode te introduceren die ook in de volgende hoofdstukken nog vaak zal gebruikt worden. Eerst wordt een korte inleiding gegeven op het gebruikte *Rijmbijbel*-corpus (§ 3.1). Aan de hand daarvan zal ik ingaan op de types variatie die we aantreffen in Middelnederlandse handschriften. Vervolgens behandel ik enkele technische aspecten. Eerst zal ik een manier bespreken waarop teksten in een numeriek formaat kunnen worden voorgesteld met het oog op een stylometrische analyse (§ 3.2).

Vervolgens zal ik een techniek voor auteursherkenning introduceren die in het internationale onderzoek al vaak werd toegepast (*Principal components analysis*, § 3.3). Met behulp van deze techniek zullen tot slot enkele parallelle handschriften uit de Middelnederlandse *Rijmbijbel*-traditie worden geanalyseerd (§ 3.4). Daaruit zal blijken dat technieken die werden ontwikkeld voor auteursherkenning in moderne teksten zonder noemenswaardige aanpassingen kunnen toegepast worden op kopiistenherkenning in middeleeuwse handschriften (§ 3.5).

3.1 EEN DOZIJN RIJMBIJBELS

Zoals ik in het vorige hoofdstuk heb toegelicht, zijn technieken voor auteursherkenning in moderne teksten waarschijnlijk niet zonder meer toepasbaar op middeleeuwse teksten. Dergelijke technieken lijken vooral niet opgewassen tegen de grote tekstuele variatie die we aantreffen in middeleeuwse handschriften. Het lijkt aangewezen om dit te illustreren met een relevant voorbeeld. Daarvoor biedt de *Rijmbijbel* van Jacob van Maerlant een interessante casus. Maerlant

beëindigde deze omvangrijke Middelnederlandse rijmtekst in 1271 (Van Dalen-Oskam 1997, 25). De tekst vormt een omvangrijke, volkstalige bewerking van de Latijnse *Historia scolastica* door Petrus Comestor, een invloedrijk werk over de Bijbelse geschiedenis.[1] Ondanks zijn aanzienlijke lengte moet de tekst al snel in verschillende afschriften in omloop zijn geraakt (Postma 1991). Het oudste handschrift C (Brussel, Koninklijke bibliotheek, hs. 15.001) waarin de tekst bijna 35.000 verzen telt, moet nog tijdens Maerlants leven zijn vervaardigd, misschien zelfs in de onmiddellijke nabijheid van de schrijver. Naast talrijke fragmenten zijn 15 min of meer complete handschriften overgeleverd, wat voor een Middelnederlandse rijmtekst uitzonderlijk veel is. Deze afschriften bestrijken bovendien een lange periode die tot in de vroegmoderne tijd doorloopt.

synoptisch corpus

Karina van Dalen-Oskam heeft een interessant synoptisch corpus aangelegd met vijf passages over vrouwelijke hoofdrolspeelsters in Maerlants oorspronkelijke tekst. Uit het Oude Testament selecteerde zij de passages over Eva (A) en Debora (D), uit apocriefe bronnen de Judith-passage (J), uit het Nieuwe Testament een passage over Martha, Maria en Maria Magdalena (M) en uit de Joodse oorlog een passage over Maria van Transjordanië (T). Van elk van deze passages heeft zij diplomatische transcripties gemaakt, waarin de oorspronkelijke spelling van het handschrift nauwgezet werd aangehouden.[2] Niet alle handschriften bevatten Maerlants volledige tekst (zie Tab. 1, p. 41) en de complete passages variëren in lengte tussen de 215 en 240 regels. Tezamen bevat dit corpus net geen 70 passages, goed voor 87.617 tokens.[3]

token

De tekstuele variatie die we aantreffen in dit corpus mag representatief heten voor de overleveringstoestand van de Middelnederlandse letterkunde. Reeds een oppervlakkige vergelijking van de varianten in deze parallelle afschriften is inzichtelijk wat betreft de vrijheden die kopiisten zich konden veroorloven. Tab. 2 (p. 42) biedt een weergave van de eerste regel uit de *Eva*-passage in enkele handschriften. De belangrijkste types kopiistenvariatie in Middelnederlandse handschrif-

1 Voor Maerlants oeuvre in het algemeen, zie Van Oostrom (1996). Belangrijke studies over de *Rijmbijbel* vindt men bij Van Dalen-Oskam (1997) en Van Moolenbroek & Mulder (1991).

2 Ik wil Karina van Dalen-Oskam hartelijk danken voor het beschikbaar stellen van dit materiaal. Bij het transcriberen kreeg zij hulp van Willem Kuiper. Verder heeft zij de tekst geannoteerd met lemmata en woordsoorten maar van die annotatielaag zal ik hier verder geen gebruik maken. In een Engelstalige publicatie van Van Dalen-Oskam (2012) worden ook deze lagen stylometrisch geanalyseerd. De handschriften C en I zal ik bij de eigenlijke analyses niet betrekken, omdat deze afschriften mogelijk van meerdere kopiisten zijn en als geheel dus onzuiver zouden kunnen zijn.

3 Onder tokens versta ik de woorden zoals zij voorkomen in de tekst oftewel letterreeksen die in de spelling van het handschrift aan weerszijden door witruimte zijn begrensd. Middeleeuwse tokens vallen niet noodzakelijk samen met onze opvatting van een woord omdat het gebruik van spaties toen nog niet strikt beregeld was (vgl. *hoe dat* vs. *hoedat*).

Siglum	Bewaarplaats	Datum	Afwezig/Incompleet
A	Berlijn	ca. 1331	*Geen*
B	Brussel	ca. 1300	*Geen*
C	Brussel	ca. 1285	*Geen*
D	Den Haag	medio 14e eeuw	*Geen*
E	Den Haag	ca. 1400	*Geen*
F	Den Haag	ca. 1400	*Geen*
G	Groningen	ca. 1339	*Geen*
H	Leiden	ca. 1465	T
I	Brussel	ca. 1450	M, T
J	Leiden	ca. 1451	*Geen*
K	Londen	1370–1385	*Geen*
L	Londen	1393	*Geen*
M	Den Haag	ca. 1330	*Geen*
N	Den Haag	1453	*Geen*
O	Den Haag	ca. 1475	E(42), D(139), M, T

Tabel 1: Overzicht van de 15 parallelle handschriften in het *Rijmbijbel*-corpus. Naast de sigla, huidige signatuur en de oorspronkelijke tijd van vervaardiging (bij benadering) wordt met een lettersymbool (zie hoofdtekst) vermeld welke van de vijf passages in een manuscript ontbreken of onvolledig zijn (in dat geval volgt tussen ronde haken het aantal aanwezige regels).

ten kunnen aan de hand van deze tabel besproken worden (Kestemont et al. 2010, 287-288):

ALLOGRAFIE – bepaalde letters (grafemen) of lettergroepen kunnen onderling worden ingewisseld zonder enig betekenisverschil, zoals v – u, uu – vv – uv – vu – w, u – v, i – j (vgl. *uren* vs. *vren*). Vergelijkbaar is het inwisselen van c – k;

HOOFDLETTERGEBRUIK – het gebruik van kapitalen is niet streng beregeld (vgl. *Tier* vs. *TIere*);

DIALECT – middeleeuwse spelling is nog in hoge mate fonologisch; verschillende spellingsverschillen gaan tenminste terug op dialectale (uitspraak)verschillen (vgl. *Tier* vs. *Tiere*);

CLITICA – middeleeuwse spelling bootst soms gesproken taal na en verbindt woorden (al dan niet met assimilatie-verschijnselen) die wij nu als zelfstandige woorden zouden spellen (vgl. *ende ter* vs. *enter*);

SPELLINGGEWOONTE – verschillende (supra-)regionale spellingsystemen waren in omloop zodat bepaalde verschijnselen in verschil-

lende regio's anders worden gespeld, zoals de verlenging van klinkers in open lettergrepen (vgl. *veren* vs. *vren*).

Merk echter op dat de variatie tussen deze handschriften niet beperkt blijft tot relatief oppervlakkige variaties, zoals bijvoorbeeld in dialect, morfologie of orthografie. In handschrift G lijkt ten aanzien van de andere versies, bijvoorbeeld zowel een volledig woord toegevoegd (bv. *Tot-* in *Tottien*) als weggelaten (bv. *seluer*). Dergelijke (veelal lexicale) verschillen worden door onderzoekers in de stemmatologie uitgebuit om de onderlinge afhankelijkheidsrelaties tussen afschriften vast te stellen (vgl. Salemans 2000). Hieronder zullen echter vooral de meer oppervlakkige variatieverschijnselen worden belicht. Verder moet opgemerkt worden dat ik geen expliciete aandacht zal besteden aan de onderlinge afhankelijkheid van de 15 *Rijmbijbel*-handschriften maar ze zal benaderen als een synchroon gegeven, waarbij er geen superioriteit of toch anterioriteit voor een van de overgeleverde tekstdragers wordt geclaimd.

stemmatologie

Kopiist	Regel
D	Ter stont ende ter seluer vren
E	Tier stont ende ter seluer vren
F	TIere stont enter seluer vren
G	Tottien stonden en ter uren
H	TEn stonden ende ter seluer vren
I	Tjerst stont ende tier veren
J	Tyer stont ende tier seluer vren
N	TJer stont tier seluer vre

Tabel 2: Illustratie van de variatie in het *Rijmbijbel*-corpus: de eerste regel uit de *Eva*-passage voor de parallelle handschriften D, E, F, G, H, I, J en N.

Aan dit corpus kunnen vanzelfsprekend verschillende interessante onderzoeksvragen worden gesteld, maar in dit hoofdstuk beperk ik mij tot de volgende kwestie. Aangezien deze afschriften alle teruggaan op dezelfde auteurstekst, zou men naïef kunnen veronderstellen dat stylometrische technieken voor auteursherkenning in principe geen verschil zouden mogen maken tussen de stijl in deze afschriften. Anders gesteld: een stylometrische analyse zou geen noemenswaardige verschillen tussen afschriften aan het licht mogen brengen. Hieronder zal ik een analyse uitvoeren waarin de parallelle passages uit verschillende handschriften stylometrisch worden vergeleken. Ter illustratie wordt bij deze analyse Brinkmans Eerste Gebod dus brutaal geschonden.

3.2 VAN TEKST TOT TABEL

Zoals hierboven reeds aangehaald, moet men teksten op een zinvolle wijze numeriek voorstellen om ze geschikt te maken voor een kwantitatieve analyse (Hinskens & Van Dalen-Oskam 2007; Hoover 2007). In deze studie zal ik vooral gebruik maken van één specifieke tekstrepresentatie die momenteel wijdverspreid is in studiegebieden die te maken hebben met de kwantitatieve verwerking van natuurlijke taal (zoals de corpus- en computerlinguïstiek, Machine Learning of Information Retrieval). Deze tekstrepresentatie wordt soms aangeduid met de naam *term-frequency vector space model* of gewoon *vector space model*, hoewel lang niet alle onderzoekers deze wijze van tekstvoorstelling expliciet zo benoemen. In tegenstelling tot wat deze benaming suggereert, is deze manier van tekstvoorstelling in essentie relatief eenvoudig.

(term-frequency) vector space model

3.2.1 *Steekproeven*

Een eerste stap in veel stylometrisch onderzoek betreft 'sampling', het proces waarbij we het beschikbare materiaal in kleinere steekproeven verdelen. Dat laat toe om rekening te houden met de interne variatie die mogelijk in teksten aanwezig is. Stel dat men een kwantitatieve vergelijking wil uitvoeren tussen twee fictieve teksten, A en B, die elk exact 2.000 woorden tellen. Gewoonlijk zal men dergelijke tekstdocumenten eerst samplen.[4] Omdat we de interne variatie (bv. in woordgebruik) die in deze teksten mogelijk optreedt niet willen negeren (cf. Brinkmans 10e Gebod), verdelen we A en B in niet-overlappende, opeenvolgende steekproeven of *samples* van een gelijke grootte. Als we uitgaan van een sample-grootte van 200 woorden, levert dat voor A en B elk 10 samples op, zodat onze totale dataset op dat moment bestaat uit 20 samples. Als we een index gebruiken om deze samples van een unieke naam te voorzien, omvat de oorspronkelijke tekst A de volgende samples: $\{A_1, A_2, A_3, \ldots, A_8, A_9, A_{10}\}$. Geheel analoog bevat B de volgende samples: $\{B_1, B_2, B_3, \ldots, B_8, B_9, B_{10}\}$. Het maximum aan samples dat we per tekst verkrijgen (in dit geval 10), is gelijk aan de eenvoudige deling van het totaal aantal woorden in de tekst door de gewenste sample-grootte, die eveneens wordt uitgedrukt in woordlengte (hier 2.000/200=10).

sampling

samples

sample-grootte

Laat ons deze sampling-procedure alvast illustreren aan de hand van twee willekeurige handschriften uit de *Rijmbijbel*-traditie. Handschriften A en D, bijvoorbeeld, bevatten beide de vijf passages uit het corpus en zijn dus goed vergelijkbaar. Deze handschriften worden eerst manueel synoptisch naast elkaar geplaatst ('gealigneerd'). Twee verzen uit beide handschriften met een duidelijk overeenkomstige lexicale inhoud worden op dezelfde hoogte geplaatst. Regels die de

4 Stamatatos (2009, 548) spreekt in dit geval van 'instance-based approaches'.

alignering

handschriften volgens de manuele alignering van deze teksten niet gemeenschappelijk hebben, zijn wel interessant voor de stemmatologie, maar laten zich niet rechtstreeks vergelijken in een stylometrisch onderzoek. De eerste 15 verzen uit beide handschriften zien er na deze alignering uit als in Tab. 3 (p. 44). In een oogopslag wordt reeds duidelijk dat de handschriften zich van gelijkwaardige, maar uiterlijk verschillende orthografische alternatieven bedienen. Vaak gaat het om zuiver allografische alternatieven (bv. *i* vs. *j* of *u* vs. *v*) maar ook tussen andere letterreeksen (bv. *g* vs. *gh*) wordt in de handschriften gealterneerd.

Handschrift A	**Handschrift** D
Nota Tier stont ende ter seluer uren	Ter stont ende ter seluer vren
So was van allen creaturen	was van allen creaturen
Dat uerstaet van allen beesten	Dat verstaet van allen beesten
Tserpent lustechst na der Jeesten	serpent lustech na der ieesten
Ende harde scalc ooc van sinne	Ende herde scalc oec van sinne
Want die duuel waser Jnne	want die duuel waser inne
bedi dat lucifer die scone	bedi dat lucifer die scone
Was gheworpen uten trone	was geworpen uten trone
hadde hijs nijt in alre wijs	had hijs nijt in alre wijs
Dat adam was int paradijs	dat adam was int paradijs
Ende pensde hoe dat hine uut dade steken	Ende pensde hoe dat hine vt dade steken
Dade hi hem tghebod gods breken	dadi hem tgebod gods breken
So moeste hiere ute pensde hi dan	so moeste hire vte pensdi dan
Maer dat hi meer ontsach den man	maer dat hi meer ontsach den man
Dies ghinc hi eerst toten wiue	dies ginc hi erst toten wiue

Tabel 3: De eerste 15 parallelle verzen uit het *Rijmbijbel*-corpus voor de handschriften A en D. Het is duidelijk dat beide handschriften zich van gelijkwaardige, maar uiterlijk verschillende orthografische alternatieven bedienen.

Laat ons de tekst in beide handschriften nu in samples verdelen. In dit voorbeeld opteren we voor een artificieel kleine sample-grootte van vijf regels. Als we ons beperken tot de 1.106 parallelle regels in beide handschriften, bekomen we in totaal 221 samples of 221 groepjes van telkens 5 opeenvolgende verzen in elk handschrift. Om de vergelijkbare grootte van de samples te garanderen, laten we het laatste parallelle vers vallen om een mooie verdeling van de tekst over de steekproeven te verkrijgen (1106-1= 1105 en 1105/5 = 221). In totaal levert dat voor deze handschriften tweemaal 221 samples op (442 in totaal) met indices van 1 tot 221. Analoog aan de notatie hierboven, levert dat voor handschrift A de reeks $\{A_1, A_2, A_3, \ldots, A_{219}, A_{220}, A_{221}\}$ op en voor handschrift D de reeks $\{D_1, D_2, D_3, \ldots, D_{219}, D_{220}, D_{221}\}$. Merk op dat de indices in dit geval een parallellie uitdrukken: de vijf regels in het willekeurige sample A_{176} zullen immers parallel zijn aan

de vijf verzen in sample D_{176}. In deze specifieke opzet is de gebruikte index dus betekenisvol (zie Tab. 4, p. 45). Indien men echter werkt met niet-parallelle teksten, zoals gebruikelijk in auteursherkenning, duidt de index gewoon de relatieve positie aan van de steekproef ten opzichte van de andere samples in de tekst.

Handschrift A	**Handschrift D**
Sample A_1	Sample D_1
Nota Tier stont ende ter seluer uren	Ter stont ende ter seluer vren
So was van allen creaturen	was van allen creaturen
Dat uerstaet van allen beesten	Dat verstaet van allen beesten
Tserpent lustechst na der Jeesten	serpent lustech na der ieesten
Ende harde scalc ooc van sinne	Ende herde scalc oec van sinne
Sample A_2	Sample D_2
Want die duuel waser Jnne	want die duuel waser inne
bedi dat lucifer die scone	bedi dat lucifer die scone
Was gheworpen uten trone	was geworpen uten trone
hadde hijs nijt in alre wijs	had hijs nijt in alre wijs
Dat adam was int paradijs	dat adam was int paradijs
Sample A_3	Sample D_3
Ende pensde hoe dat hine uut dade steken	Ende pensde hoe dat hine vt dade steken
Dade hi hem tghebod gods breken	dadi hem tgebod gods breken
So moeste hiere ute pensde hi dan	so moeste hire vte pensdi dan
Maer dat hi meer ontsach den man	maer dat hi meer ontsach den man
Dies ghinc hi eerst toten wiue	dies ginc hi erst toten wiue

Tabel 4: Een voorbeeld van sampling. De parallelle handschriften A en D hebben 1106 parallelle verzen die worden verdeeld in steekproefjes van 5 opeenvolgende verzen. De eerste drie samples van elk handschrift worden afgebeeld.

3.2.2 *Knikkeren met woorden*

Als deze sampling uit de weg is, kan men werk maken van de eigenlijke numerieke voorstelling van de samples. In de stylometrie maakt men graag gebruik van metingen op het niveau van individuele woorden (Holmes 1994, 87, vgl. § 2.1.2 hierboven). Het is handig dat men in de meeste moderne, westerse talen makkelijk de woorden uit een tekst kan halen (Stamatatos 2009, 540). Men splitst een tekst uit in letterreeksen die geen spaties bevatten en eventueel gebruikt men een *tokeniser* om de moeilijkere gevallen (bv. afkortingen) op te lossen.[5] Afhankelijk van het doel van een toepassing, kan men ook nog leestekens verwijderen en bepaalde hoofdletters (bv. aan het begin

tokeniser

5 Een *tokeniser* is een software-toepassing die lopende tekst omzet naar een reeks 'tokens', bijvoorbeeld door de interpunctie van belendende woorden te scheiden of

tokenisering

van een zin) omzetten in kleine letters. Na een dergelijke 'tokenisering' is het voor machines kinderspel om een exhaustieve woordenlijst uit een tekst te distilleren (Manning et al. 2008, 22-23).

Een dergelijke woordenlijst kan men vervolgens gebruiken voor metingen in de tekst op woordniveau. Voor sommige toepassingen kan men louter geïnteresseerd zijn in de aan- of afwezigheid van een woord in een document: verschillende zoekmachines voor online databases zullen bijvoorbeeld enkel bijhouden ('indexeren') of een specifiek woord al dan niet voorkomt in een bepaalde entry in de database (bv. Manning et al. 2008, 5). Op dat moment volstaat het om met een zogenaamde booleaanse waarde (*True* of *False*) te werken. In de stylometrie is men echter vaak gediend met meer fijnmazige metingen en zal men de absolute frequentie van een woord bepalen door het globale aantal voorkomens van dit woord in een document te tellen (Stamatatos 2009, 540-541).

absolute frequentie

Als men samples van een gelijke grootte vergelijkt, zijn dergelijke absolute woordtellingen rechtstreeks vergelijkbaar. Vaak worden dergelijke tellingen echter genormaliseerd tot relatieve frequenties, omdat men dan tot op zekere hoogte ook het woordgebruik tussen teksten met een ongelijke lengte kan vergelijken.[6] De relatieve frequentie van een woord bekomt men in dat geval door de absolute frequentie van een woord in het document te delen door het totale aantal woorden erin. Een woord dat 163 keer voorkomt in een document dat in totaal 2.378 woorden telt heeft bijgevolg een relatieve frequentie van $\approx 0,0685$ in dat document.

relatieve frequentie

Dergelijke frequenties kunnen echter ook geëxtraheerd worden op het niveau van lettergroepen in plaats van op woordniveau. In de stylometrie werkt men immers ook graag met *character n-grams*, want de frequenties van dergelijke lettergroepen zijn verbonden aan de functoren in een tekst. Ook voor onze kopiistenanalyse lijken metingen op het niveau van letterreeksen interessanter dan die op het woordniveau. Spellingvariatie, bijvoorbeeld, laat zich beter onderzoeken op basis van eenheden die kleiner zijn dan het woord. Een woordvergelijking zal wel opmerken dat *tghebod* anders is gespeld dan *tgebod*, maar specifieker zijn we geïnteresseerd in het conflict tussen de reeks bigrammen *tg*, *gh*, *he* in *tghebod* versus het corresponderende duo *tg*, *ge* in *tgebod*.

Onderstaande opsomming biedt een concreet voorbeeld van een dergelijke voorstelling onder een *n*-grammodel. Sample D_2 wordt hierin opgedeeld in zowel uni-, bi- en trigrammen op letterniveau. Merk op dat ik alle hoofdletters automatisch omzet naar kleine letters en dat alle leestekens of editoriale aantekeningen worden verwijderd. De witruimte aan weerszijden van tokens wordt weergegeven als

---

woorden die bij elkaar horen te verbinden (bijv. partikels in het geval van scheidbare werkwoorden).

6 Zie voor woordfrequenties het inleidende hoofdstuk van Baayen (2002), maar merk op dat zelfs relatieve woordfrequenties niet volledig vergelijkbaar zijn als de lengte van de vergeleken documenten té verschillend is.

een grote letter 'S'. De *n*-grammen worden onderscheiden door een komma. Het einde van een vers wordt als een gewone spatie ('S') gemarkeerd.

OORSPRONKELIJK – want die duuel waser inne bedi dat lucifer die scone was geworpen uten trone had hijs nijt in alre wijs dat adam was int paradijs

UNIGRAMMEN – S, w, a, n, t, S, d, i, e, S, d, u, u, e, l, S, w, a, s, e, r, S, i, n, n, e, S, b, e, d, i, S, d, a, t, S, l, u, c, i, f, e, r, S, d, i, e, S, s, c, o, n, e, S, w, a, s, S, g, e, w, o, r, p, e, n, S, u, t, e, n, S, t, r, o, n, e, S, h, a, d, S, h, i, j, s, S, n, i, j, t, S, i, n, S, a, l, r, e, S, w, i, j, s, S, d, a, t, S, a, d, a, m, S, w, a, s, S, i, n, t, S, p, a, r, a, d, i, j, s, S

BIGRAMMEN – Sw, wa, an, nt, tS, Sd, di, ie, eS, Sd, du, uu, ue, el, lS, Sw, wa, as, se, er, rS, Si, in, nn, ne, eS, Sb, be, ed, di, iS, Sd, da, at, tS, Sl, lu, uc, ci, if, fe, er, rS, Sd, di, ie, eS, Ss, sc, co, on, ne, eS, Sw, wa, as, sS, Sg, ge, ew, wo, or, rp, pe, en, nS, Su, ut, te, en, nS, St, tr, ro, on, ne, eS, Sh, ha, ad, dS, Sh, hi, ij, js, sS, Sn, ni, ij, jt, tS, Si, in, nS, Sa, al, lr, re, eS, Sw, wi, ij, js, sS, Sd, da, at, tS, Sa, ad, da, am, mS, Sw, wa, as, sS, Si, in, nt, tS, Sp, pa, ar, ra, ad, di, ij, js, sS

TRIGRAMMEN – Swa, wan, ant, ntS, tSd, Sdi, die, ieS, eSd, Sdu, duu, uue, uel, elS, lSw, Swa, was, ase, ser, erS, rSi, Sin, inn, nne, neS, eSb, Sbe, bed, edi, diS, iSd, Sda, dat, atS, tSl, Slu, luc, uci, cif, ife, fer, erS, rSd, Sdi, die, ieS, eSs, Ssc, sco, con, one, neS, eSw, Swa, was, asS, sSg, Sge, gew, ewo, wor, orp, rpe, pen, enS, nSu, Sut, ute, ten, enS, nSt, Str, tro, ron, one, neS, eSh, Sha, had, adS, dSh, Shi, hij, ijs, jsS, sSn, Sni, nij, ijt, jtS, tSi, Sin, inS, nSa, Sal, alr, lre, reS, eSw, Swi, wij, ijs, jsS, sSd, Sda, dat, atS, tSa, Sad, ada, dam, amS, mSw, Swa, was, asS, sSi, Sin, int, ntS, tSp, Spa, par, ara, rad, adi, dij, ijs, jsS, sS

Vanzelfsprekend is een dergelijke representatie nog maar amper betekenisvol voor een mens, laat staan vlot leesbaar. Dat wordt nog erger wanneer we aan het tellen slaan en het oorspronkelijke sample reduceren tot een 'woordenboek' of 'index' waarin we voor ieder *n*-gram de absolute of relatieve frequentie weergeven. Hieronder geef ik als voorbeeld de absolute en relatieve frequenties (weergegeven tot op vier cijfers na de komma) voor de bigrammen in sample D_2. Het bigram 'Sd' (een *d*- aan het begin van een woord) treedt 5 keer op in dit mini-sample; de relatieve frequentie van het bigram in dit mini-sample is dus $\approx 0,0384$, want in totaal telt dit sample 130 bigrammen (5/130).

ABSOLUTE FREQUENTIES – **Sa**(2), **Sb**(1), **Sd**(5), **Sg**(1), **Sh**(2), **Si**(3), **Sl**(1), **Sn**(1), **Sp**(1), **Ss**(1), **St**(1), **Su**(1), **Sw**(5), **ad**(3), **al**(1), **am**(1), **an**(1), **ar**(1), **as**(3), **at**(2), **be**(1), **ci**(1), **co**(1), **dS**(1), **da**(3), **di**(4),

du(1), **eS**(6), **ed**(1), **el**(1), **en**(2), **er**(2), **ew**(1), **fe**(1), **ge**(1), **ha**(1), **hi**(1), **iS**(1), **ie**(2), **if**(1), **ij**(4), **in**(3), **js**(3), **jt**(1), **lS**(1), **lr**(1), **lu**(1), **mS**(1), **nS**(3), **ne**(3), **ni**(1), **nn**(1), **nt**(2), **on**(2), **or**(1), **pa**(1), **pe**(1), **rS**(2), **ra**(1), **re**(1), **ro**(1), **rp**(1), **sS**(5), **sc**(1), **se**(1), **tS**(5), **te**(1), **tr**(1), **uc**(1), **ue**(1), **ut**(1), **uu**(1), **wa**(4), **wi**(1), **wo**(1)

RELATIEVE FREQUENTIES – **Sa**(0,0153), **Sb**(0,0076), **Sd**(0,0384), **Sg**(0,0076), **Sh**(0,0153), **Si**(0,0230), **Sl**(0,0076), **Sn**(0,0076), **Sp**(0,0076), **Ss**(0,0076), **St**(0,0076), **Su**(0,0076), **Sw**(0,0384), **ad**(0,0230), **al**(0,0076), **am**(0,0076), **an**(0,0076), **ar**(0,0076), **as**(0,0230), **at**(0,0153), **be**(0,0076), **ci**(0,0076), **co**(0,0076), **dS**(0,0076), **da**(0,0230), **di**(0,0307), **du**(0,0076), **eS**(0,0461), **ed**(0,0076), **el**(0,0076), **en**(0,0153), **er**(0,0153), **ew**(0,0076), **fe**(0,0076), **ge**(0,0076), **ha**(0,0076), **hi**(0,0076), **iS**(0,0076), **ie**(0,0153), **if**(0,0076), **ij**(0,0307), **in**(0,0230), **js**(0,0230), **jt**(0,0076), **lS**(0,0076), **lr**(0,0076), **lu**(0,0076), **mS**(0,0076), **nS**(0,0230), **ne**(0,0230), **ni**(0,0076), **nn**(0,0076), **nt**(0,0153), **on**(0,0153), **or**(0,0076), **pa**(0,0076), **pe**(0,0076), **rS**(0,0153), **ra**(0,0076), **re**(0,0076), **ro**(0,0076), **rp**(0,0076), **sS**(0,0384), **sc**(0,0076), **se**(0,0076), **tS**(0,0384), **te**(0,0076), **tr**(0,0076), **uc**(0,0076), **ue**(0,0076), **ut**(0,0076), **uu**(0,0076), **wa**(0,0307), **wi**(0,0076), **wo**(0,0076)

De volgende stap is om deze tellingen in een globale tabel om te zetten. Meestal is een dergelijke tabel in essentie twee-dimensioneel, in die zin dat een reeks kolommen met een reeks rijen wordt verbonden zoals in standaard *spreadsheet*-applicaties.[7] De geanalyseerde documenten (bijv. de samples uit de *Rijmbijbel*-handschriften A en D) worden in deze tabel voorgesteld op een afzonderlijke, horizontale rij. In ons kopiistenvoorbeeld zal de tabel dus 442 rijen bevatten die elk een sample van 5 verzen voorstellen: 221 voor A en 221 voor D. Voor ieder uniek woord of *n*-gram dat in deze teksten voorkwam, wordt een aparte, verticale kolom in de tabel voorzien. De individuele cellen in de tabel worden bevolkt met de frequentie van elk woord of *n*-gram (kolom) in elk sample (rij).

Natuurlijk kan een dergelijke tabel een enorme omvang aannemen, zeker indien veel en grote documenten worden voorgesteld. Vooral het aantal kolommen kan snel oplopen, aangezien elk verschillend woord of *n*-gram een eigen kolom wordt toegewezen, zelfs al komt het misschien maar één keer voor in de sample-collectie (een zogenaamd hapax legomenon). Niettemin blijven de eenvoudige basisprincipes voor de opbouw van de tabel ook in deze gevallen ongewijzigd.[8]

hapax legomenon

Het is vanzelfsprekend moeilijk (en ook weinig interessant) om dergelijk grote tabellen hier nog visueel weer te geven. Om toch een

7 Verschillende formaten voor dergelijke 'tabellen' zijn in omloop; veel gebruikt is het 'Attribute-Relation File Format' van de *Weka*-software (Hall et al. 2009).

8 Dat de teksten op rijen worden voorgesteld en de woorden of *n*-grammen in kolommen is overigens louter een willekeurige aanname: de representatie kan ook perfect worden omgedraaid worden zodat teksten in kolommen terechtkomen en woorden of *n*-grammen in rijen.

idee te geven van de aanzienlijke complexiteit van een dergelijke tekstvoorstelling geef ik hieronder Tab. 5 (p. 50) weer: deze bevat de relatieve unigram-frequenties in de eerste 10 samples (van 5 regels) in de parallelle *Rijmbijbel*-handschriften A en D. De kapitaal 'S' stelt nog steeds het optreden van witruimte tussen twee tokens voor. De relatieve frequenties werden voor de leesbaarheid genormaliseerd naar een promille waarbij cijfers na de komma niet zijn weergegeven. Aangezien deze tabel al betrekkelijk onoverzichtelijk is, kan men zich voorstellen hoe complex een dergelijke weergave zou zijn voor bijvoorbeeld álle, niet-afgeronde frequenties in álle samples en álle trigrammen.

Waarom heeft men deze eenvoudige tekstrepresentatie nu eigenlijk bedacht met de ingewikkelde naam *term-frequency vector space model* (Manning et al. 2008, 289-313)? Woorden worden in de Information Retrieval vaak aangeduid als *termen* (cf. 'zoektermen': Manning et al. 2008, 22). Aangezien daar vaak gewerkt wordt met de frequenties van deze woorden, spreekt men doorgaans van een *term-frequency* (afgekort TF). De afzonderlijke rijen die in onze tabel de samples voorstellen, noemt men in Information Retrieval en Machine Learning vaak een 'vector'. In wezen betekent de naam '(feature) vector' in deze context niet meer dan dat er een geordende reeks getallen wordt gebruikt om de eigenschappen of *features* van een bepaald object voor te stellen (Sebastiani 2002, 10). In ons huidige *Rijmbijbel*-voorbeeld zijn de features eenvoudigweg de frequenties van de *n*-grammen waarmee wij de samples numeriek voorstellen.[9]

term-frequency vector space model

feature vector

Tezamen stellen de vectoren in onze tabel een hoog-dimensionele ruimte voor: iedere kolom kan immers opgevat worden als een soort geometrische dimensie waarin een tekst een positie wordt toegewezen, op basis van de frequentie erin van het woord dat de kolom voorstelt. Als alle woorden in de tekstcollectie een kolom krijgen toegewezen, vormt deze representatie dus een hoog-dimensionele ruimte (Argamon 2008, 146-147). Vandaar dat men over deze tekstrepresentatie spreekt als een 'space model'. Als men teksten voorstelt als vectoren die uit enkelvoudige woordfrequenties bestaan, spreekt men trouwens ook wel eens over een *bag-of-words model* (Sebastiani 2002, 10). De teksten worden immers benaderd als een 'zak knikkers' waarin de oorspronkelijke positie of volgorde van de woorden in de tekst niet langer van tel is. Men houdt zo slechts rekening met de globale frequentie van woorden in het document (Stamatatos 2009, 541).[10]

hoog-dimensionele ruimte

bag-of-words model

9 In de statistiek zal men de rijen met de vectoren in onze tabel overigens vaak aanduiden met het woord 'observatie', terwijl de kolommen zullen worden beschouwd als de eigenschappen (*properties, variables, attributes*) van deze observaties. Voor een veelgeprezen linguïstische inleiding tot statistiek, zie Baayen (2008). De termen *observation – vector* zijn daarom in veel studies inwisselbaar, net zoals *feature – property – variable – attribute*.

10 Als men op het niveau van woorden met *n*-grammen van een hogere orde werkt – bv. met trigrammen van drie opeenvolgende woorden – zal woordvolgorde natuurlijk wel een zekere rol spelen in de tekstrepresentatie.

	S	**e**	**n**	**d**	**a**	**t**	**i**	**s**	**r**	**o**	**h**	**l**	**m**	**c**	**u**	**g**	**w**	**b**	**v**	**p**	**j**	**f**	**k**	**y**	**q**	**x**	**z**
A_1	180	167	105	31	80	93	12	74	62	31	12	43	0	31	31	0	6	6	18	6	6	0	0	0	0	0	0
A_2	194	104	74	74	89	52	74	52	44	22	22	22	7	14	29	7	44	7	0	14	37	7	0	0	0	0	0
A_3	203	174	69	81	46	63	52	52	29	40	63	0	29	11	23	17	5	11	0	11	0	0	11	0	0	0	0
A_4	189	183	94	44	69	75	12	44	44	31	56	12	18	31	6	12	18	6	6	18	6	6	6	0	0	0	0
A_5	176	156	58	39	52	58	65	65	52	32	39	26	13	39	13	13	39	6	13	13	13	13	0	0	0	0	0
A_6	200	193	90	70	64	64	32	19	12	45	45	19	19	0	25	32	6	12	25	0	0	12	0	0	0	0	6
A_7	192	153	76	70	57	38	76	57	57	70	19	6	6	6	44	6	19	19	6	6	0	0	0	0	0	0	6
A_8	185	197	53	53	35	65	47	35	53	59	47	23	23	23	11	17	17	5	0	11	5	17	0	0	0	0	0
A_9	178	178	46	46	46	66	39	39	46	66	39	26	0	13	33	39	39	6	6	6	6	13	6	6	0	0	0
A_{10}	185	158	99	52	59	46	33	52	26	72	46	0	33	59	6	6	0	26	6	0	0	0	13	0	0	0	13
D_1	181	194	107	33	73	80	13	67	67	13	13	46	0	33	20	0	6	6	33	6	0	0	0	0	0	0	0
D_2	198	99	76	68	91	53	83	53	45	22	15	22	7	15	30	7	45	7	0	15	30	7	0	0	0	0	0
D_3	204	161	74	86	49	68	55	55	31	43	43	0	31	12	6	18	6	12	12	12	0	0	12	0	0	0	0
D_4	192	178	99	46	72	79	19	46	46	19	33	13	19	33	6	13	19	6	13	19	6	6	6	0	0	0	0
D_5	173	160	60	40	53	60	66	66	53	33	20	26	13	40	33	13	40	6	6	13	13	6	0	0	0	0	0
D_6	215	180	90	62	69	76	27	27	13	41	27	20	20	0	41	34	6	13	27	0	0	0	0	0	0	0	0
D_7	194	147	80	80	60	40	93	60	60	40	13	6	6	6	40	6	20	20	13	6	0	0	0	0	0	0	0
D_8	201	213	54	48	36	67	48	42	54	30	30	24	24	24	12	18	18	6	18	12	0	12	0	0	0	0	0
D_9	186	179	48	48	55	68	48	41	48	55	20	27	0	13	13	41	41	6	20	6	6	6	6	6	0	0	0
D_{10}	191	157	102	47	61	54	54	75	27	41	34	0	34	54	6	6	0	27	6	0	0	0	13	0	0	0	0

Tabel 5: Voorstelling van de relatieve unigram-frequenties in de eerste 10 samples (van 5 regels) uit de parallelle *Rijmbijbel*-handschriften A en D. De kapitaal 'S' stelt het optreden van witruimte tussen twee tokens voor. De relatieve frequenties werden genormaliseerd naar een promille waarbij cijfers na de komma (louter omwille van de leesbaarheid) niet zijn weergegeven. De tabel illustreert de grote complexiteit en onoverzichtelijkheid van een tekstvoorstelling onder een *term-frequency vector space model*.

3.2.3 *Kolomselectie*

Als men een groep teksten voorstelt onder een *term-frequency vector space model*, bekomt men dus een tekstrepresentatie in de vorm van een tabel die bijzonder omvangrijk kan worden. We kennen aanvankelijk immers aan ieder geteld item een aparte kolom toe, onafhankelijk van de relevantie van dat item. Een tabel kan zo een groot aantal minder relevante kolommen tellen (vgl. Alpaydin 2004, 33-34). Dat is, ten eerste, niet efficiënt want zo kan een computationele analyse veel rekenkracht verspillen aan het verwerken van de numerieke gegevens in irrelevante kolommen. Ten tweede, is deze tekstvoorstelling ook niet noodzakelijk effectief: als een woord slechts één keer werd aangetroffen in de gehele document-collectie, zal de corresponderende kolom met uitzondering van één cel alleen nullen bevatten. Vanzelfsprekend zijn dergelijke, laagfrequente woorden weinigzeggend voor de vergelijking van teksten. Bovendien vereisen veel statistische tests dat een rij getallen niet al te veel nulwaardes telt (bv. Kilgariff 2001, 246–247), zodat zij niet gediend zijn met hapax-kolommen.

Het is dus gebruikelijk om minder relevante kolommen uit een tabel te verwijderen, alvorens de eigenlijke analyse te beginnen. Dit proces wordt doorgaans aangeduid als *feature selection* (Guyon & Elisseeff 2003). Het doel is om enkel de relevante kolommen in een tabel bij te houden om zowel de efficiëntie als effectiviteit van de daaropvolgende analyse te maximaliseren. De simpelste methode is het gebruik van een *frequency cut-off filter*: men verwijdert de kolommen met woorden waarvan de globale frequentie in alle teksten niet boven een bepaalde drempelwaarde uitkomt (bv. Kilgariff 2001, 246–247).

feature selection

frequency cut-off filter

Er bestaat echter ook een hele resem meer geavanceerde selectiemethodes die hiervoor kunnen gebruikt worden. Op dit moment kan men al rekening houden met het eigenlijke probleem dat men probeert op te lossen. Als men in een analyse een onderscheid tracht te maken tussen de stijlen van verschillende auteurs, zal men bijvoorbeeld slechts die woorden bijhouden waarvan het gebruik de grootste verschillen tussen de betrokken auteurs laat zien. Een bekende metriek is bijvoorbeeld *Information Gain* (Quinlan 1986), die probeert vast te stellen wat de 'informativiteit' van een bepaald woord in de tabel is, gegeven het auteurschap van een tekst.

Hoewel een geavanceerde kolomselectie in veel studies tot een positief resultaat heeft geleid, moet meteen worden opgemerkt dat dit voor auteursherkenning niet altijd het geval is.[11] Zo kan een selectietechniek opmerken dat het gebruik van een woord markante verschillen vertoont tussen de teksten van twee auteurs, maar eraan voorbijgaan dat dit woord sterk gerelateerd is aan een bepaald onderwerp in plaats van aan de algemene schrijfstijl van een auteur.

11 Luyckx (2010, 19-20) verwijst bv. naar Argamon & Levitan (2005) en Stamatatos (2009, 545).

Net zoals dat gevaar bestond in traditionele auteursherkenning kan een analyse daarom een artificieel groot gewicht toekennen aan een specifiek inhoudswoord. Een tekst over een bepaald onderwerp van auteur A kan dan verkeerdelijk worden toegeschreven aan auteur B, omdat die een tekst heeft geschreven over hetzelfde onderwerp (vgl. Craig 2004). Dit probleem staat bekend als *overfitting* (Alpaydin 2004, 33-34): men kent te veel belang toe aan stilistische kenmerken die niet goed 'schalen' of uitbreidbaar zijn naar andere teksten (Luyckx 2010, 2).

overfitting

In auteursherkenning maakt men daarom vaak gebruik van een veel eenvoudigere selectie-techniek die men in het Engels kan aanduiden als *frequency-based feature selection*.[12] Als men werkt met tellingen op woordniveau, maakt men daarvoor eerst een woordenlijst aan waarin men de absolute frequentie bijhoudt van alle woorden in de gehele document-collectie (of een relevant controlecorpus, vgl. Burrows 2002, 268). Men kan hier spreken van een cumulatieve frequentie omdat deze teruggaat op álle samples. Vervolgens worden de woorden in deze lijst gerangschikt op basis van hun cumulatieve frequentie. De woorden die globaal gesproken het meest frequent zijn, komen in deze ranking bovenaan terecht. Vervolgens selecteert men op basis van een drempelwaarde of *threshold* de meest frequente woorden in de gehele tabel (Stamatatos 2009, 541). Alle andere woordkolommen worden verder genegeerd. Welke drempelwaardes men hiervoor hanteert, varieert sterk tussen studies onderling: zowel de 30, 300 als 1.000 meest frequente woorden uit corpora zijn reeds voorgesteld (Stamatatos 2009, 540-541). In het algemeen wordt het aantal geselecteerde items echter klein gehouden. Dat heeft te maken met het volgende.

frequency-based feature selection

cumulatieve frequentie

ranking

threshold

Wie de woorden uit een tekstcorpus extraheert en deze vervolgens rangschikt op basis van cumulatieve frequentie, stelt vast dat de woorden bovenin de rangschikking veelal functiewoorden zijn.[13] Zoals toegelicht in het vorige hoofdstuk hebben dergelijke functoren een erg vage betekenis die niet gebonden is aan een specifiek onderwerp. Omdat zij zo vaak en in de meest diverse contexten optreden, vormen zij de meest frequente woorden in een natuurlijke taal. Stylometrici zijn precies omwille van deze eigenschap erg geïnteresseerd in de frequenties van functoren. Omdat zij zo goed gespreid zijn over teksten, lenen zij zich goed voor een vergelijkende stijlanalyse over de grenzen van diverse tekstsoorten heen. Zij 'schalen' met andere woorden goed tussen teksten.

Daarom kan men in stylometrische studies vaak volstaan met een *frequency-based feature selection* (Luyckx 2010, 19-20). Men selecteert slechts de meest frequente woorden uit het onderzochte corpus omdat deze gewoonlijk samenvallen met de functoren in een taal (Stamatatos

12 Het overzicht bij Luyckx (2010, 19) verwijst hiervoor naar het onderzoek van Koppel et al. (2006) en Houvardas & Stamatatos (2006).

13 Stamatatos (2009, 541). Bijzonder relevant in deze context zijn de wetten van Zipf (1949). Zie Baayen (2002) voor woordfrequenties in het algemeen.

2009, 544-545). Met deze werkwijze zet men bijgevolg tot op zekere hoogte het euvel van *overfitting* buiten spel. Wel is het belangrijk dat bijvoorbeeld de aandelen van verschillende auteurs in een analyse min of meer gelijk zijn: als de teksten van één auteur over-vertegenwoordigd zijn, kan dat leiden tot een frequentielijst die het werk van deze auteur ten onrechte bevoordeeld. Om diezelfde reden is het ook belangrijk om slechts te werken met de kruin van de rangschikking: hoe meer laagfrequente woorden men betrekt, hoe groter de kans dat men ook sterker inhoudsgerelateerde woorden toelaat.

Alleszins kent deze aanpak nog een laatste belangrijk voordeel: indien men zich beperkt tot een woordselectie op basis van grammaticale criteria moet men voor Engelstalige teksten bijvoorbeeld het woord 'notwithstanding' toelaten. Dit voegwoord is in theorie een functor maar zal amper voorkomen in veel minder formele teksten, waar men de voorkeur zal geven aan een alternatief als 'however'. Het gebruik van het woord 'notwithstanding' zou in dat opzicht meer indicatief zijn voor het genre van de tekst dan voor de schrijfstijl van een auteur (vgl. Stamatatos 2009, 544-545). Wie zich voor de woordselectie echter baseert op zuivere frequentiecriteria zal dergelijke woorden niet meenemen als zij niet tot de meest frequente functoren in het onderzochte corpus behoren.

Laat ons deze frequentiegebaseerde kolomselectie toepassen op onze parallelle *Rijmbijbel*-handschriften A en D. Voor deze handschriften werken we in ons voorbeeld met een representatie op basis van *character n-grams* maar vanzelfsprekend is de redenering hier volledig analoog. We kunnen immers aannemen dat de meest frequente *n*-grammen in deze teksten vooral zullen samenhangen met het gebruik van functoren. Deze hoogfrequente *n*-grammen zullen bijgevolg erg goed gespreid zijn in de geanalyseerde samples en niet samenhangen met specifieke passages.

Hieronder volgt een lijst met de cumulatieve, absolute frequentie van 100 bigrammen uit de samples van A en D. Deze bigrammen zijn de honderd meest frequente in de bewuste samples en werden zo gerangschikt dat de meest frequente bigrammen zich bovenin in de lijst bevinden. Merk op dat de meest frequente bigrammen inderdaad gelieerd kunnen worden aan Middelnederlandse functoren. De hoog gerangschikte bigrammen *en* (2e positie), *nd* (17e positie) en *de* (5e positie) kunnen bijvoorbeeld worden teruggevonden in het nevenschikkend voegwoord *ende*, terwijl het optreden van een *d-* aan het begin van een woord (het bigram *Sd*, 4e positie) in verschillende aanwijzende voornaamwoorden *de, dat, die, ...* of het lidwoord *den* voorkomt. Het optreden van een *-e* of *-n* in de *Auslaut* van woorden (*eS*, 1e positie of *nS*, 3e positie) is dan weer typisch voor de uitgang van veel substantieven in het Middelnederlands. Al deze functorgerelateerde bigrammen bekleden de hoogste posities in deze rangschikking, hetgeen onze verwachtingen bevestigt.

1 – **eS** (3613), 2 – **en** (2725), 3 – **nS** (2413), 4 – **Sd** (2078), 5 – **de** (1988), 6 – **tS** (1763), 7 – **er** (1501), 8 – **Ss** (1407), 9 – **Sh** (1072), 10 – **ie** (986), 11 – **da** (900), 12 – **rS** (852), 13 – **Se** (817), 14 – **te** (793), 15 – **re** (754), 16 – **he** (741), 17 – **nd** (737), 18 – **an** (734), 19 – **Sw** (725), 20 – **di** (705), 21 – **sS** (695), 22 – **Sm** (669), 23 – **et** (660), 24 – **iS** (657), 25 – **ae** (634), 26 – **me** (620), 27 – **Sg** (605), 28 – **Sv** (586), 29 – **ne** (586), 30 – **at** (578), 31 – **Sa** (557), 32 – **se** (553), 33 – **oe** (534), 34 – **St** (489), 35 – **in** (474), 36 – **Sb** (459), 37 – **al** (457), 38 – **So** (447), 39 – **ar** (447), 40 – **ed** (443), 41 – **es** (438), 42 – **on** (437), 43 – **hi** (427), 44 – **Si** (420), 45 – **si** (412), 46 – **Sn** (399), 47 – **wa** (388), 48 – **ch** (383), 49 – **st** (380), 50 – **el** (379)

Als we de voorstelling van onze samples nu beperken tot de frequentie van de vaakst voorkomende bigrammen, bekomen we een representatie die veel compacter is dan de oorspronkelijke tabel. Tab. 6 (p. 55) is dezelfde als diegene die hierboven de frequentie-vectoren weergaf voor de eerste van onze 10 samples van A en D voor unigram-frequenties. Hier wordt echter gewerkt met een representatie onder een bigram-model. Doordat deze representatie via een eenvoudige kolomselectie is teruggedrongen tot de relatieve frequenties van de 15 vaakst voorkomende bigrammen in de samples, is deze veel compacter geworden en dus makkelijker te hanteren. Merk trouwens op hoe relatief weinig nulwaardes in deze cellen voorkomen: zelfs in onze artificieel kleine samples van vijf regels komen de meeste van onze 15 geselecteerde bigrammen tenminste één keer voor. Dat illustreert waarom dergelijke lettergroepen zo interessant zijn voor het stylometrisch vergelijken van teksten. Het is een dergelijke, gereduceerde tekstvoorstelling waarmee wij in de volgende sectie aan de slag zullen gaan.

3.3 PRINCIPAL COMPONENTS ANALYSIS

In de vorige sectie is toegelicht hoe men teksten of samples daaruit kan voorstellen onder een *term-frequency vector space model*. Een tekst-collectie wordt daarbij omgezet in een soort tabel waarin ieder sample als een rij wordt voorgesteld en individuele items, zoals woordtellingen, in kolommen. De cellen in een dergelijke tabel stellen de relatieve frequenties voor van de items in de samples. *Feature selection* kan worden toegepast om de tabel minder omvangrijk te maken. In auteursherkenning zal men zich bijvoorbeeld vaak beperken tot een analyse van functoren die cumulatief gesproken het meest frequent zijn in deze teksten. Op basis van een dergelijke representatie kan men nu een echte kwantitatieve analyse uitvoeren.

	eS	en	nS	Sd	de	tS	er	Ss	Sh	ie	da	rS	Se	te	re
A_1	0,025	0,0562	0,0562	0,0125	0,025	0,0312	0,0375	0,0312	0,0062	0,0062	0,0062	0,025	0,0125	0,025	0,0187
A_2	0,0526	0,0150	0,0225	0,0375	0,0075	0,0375	0,0150	0,0075	0,0150	0,0150	0,0225	0,0150	0,0	0,0075	0,0075
A_3	0,0643	0,0409	0,0350	0,0409	0,0350	0,0233	0,0233	0,0116	0,0467	0,0116	0,0292	0,0116	0,0116	0,0233	0,0116
A_4	0,0509	0,0445	0,0318	0,0191	0,0254	0,0445	0,0254	0,0191	0,0191	0,0	0,0127	0,0191	0,0063	0,0191	0,0127
A_5	0,0526	0,0263	0,0065	0,0263	0,0	0,0328	0,0263	0,0197	0,0197	0,0263	0,0065	0,0	0,0131	0,0131	0,0131
A_6	0,0649	0,0584	0,0454	0,0324	0,0194	0,0324	0,0064	0,0129	0,0194	0,0129	0,0194	0,0064	0,0129	0,0259	0,0064
A_7	0,0774	0,0516	0,0258	0,0258	0,0451	0,0193	0,0064	0,0451	0,0193	0,0193	0,0064	0,0	0,0064	0,0	0,0258
A_8	0,0843	0,0240	0,0060	0,0240	0,0361	0,0361	0,0301	0,0180	0,0301	0,0180	0,0060	0,0120	0,0120	0,0240	0,0120
A_9	0,0533	0,04	0,0266	0,0066	0,0333	0,04	0,0333	0,0266	0,0066	0,0133	0,0	0,0	0,0066	0,0133	0,0066
A_{10}	0,0733	0,0666	0,06	0,0266	0,0266	0,0266	0,0066	0,04	0,0066	0,02	0,0133	0,0	0,02	0,0133	0,0066
D_1	0,0270	0,0608	0,0608	0,0135	0,0270	0,0270	0,0472	0,0337	0,0067	0,0067	0,0067	0,0270	0,0135	0,0337	0,0202
D_2	0,0461	0,0153	0,0230	0,0384	0,0	0,0384	0,0153	0,0076	0,0153	0,0153	0,0230	0,0153	0,0	0,0076	0,0076
D_3	0,0562	0,0437	0,0375	0,0437	0,025	0,025	0,0187	0,0125	0,0375	0,0062	0,0312	0,0125	0,0125	0,025	0,0125
D_4	0,0466	0,0466	0,0333	0,02	0,0266	0,04	0,0266	0,02	0,0133	0,0	0,0133	0,02	0,0066	0,02	0,0133
D_5	0,0536	0,0268	0,0067	0,0268	0,0	0,0402	0,0268	0,0201	0,0134	0,0268	0,0067	0,0	0,0134	0,0067	0,0134
D_6	0,0559	0,0629	0,0559	0,0349	0,0209	0,0419	0,0069	0,0209	0,0279	0,0069	0,0209	0,0069	0,0139	0,0279	0,0069
D_7	0,0472	0,0540	0,0337	0,0270	0,0405	0,0270	0,0135	0,0405	0,0135	0,0202	0,0067	0,0067	0,0135	0,0	0,0202
D_8	0,0613	0,0306	0,0122	0,0184	0,0368	0,0429	0,0306	0,0184	0,0245	0,0122	0,0061	0,0122	0,0306	0,0184	0,0122
D_9	0,0555	0,0416	0,0277	0,0069	0,0347	0,0486	0,0208	0,0277	0,0069	0,0069	0,0	0,0	0,0069	0,0138	0,0069
D_{10}	0,0551	0,0689	0,0620	0,0275	0,0275	0,0275	0,0068	0,0551	0,0068	0,0206	0,0137	0,0	0,0206	0,0137	0,0068

Tabel 6: Voorstelling van de relatieve bigram-frequenties in de eerste 10 samples (van 5 regels) in de parallelle *Rijmbijbel*-handschriften A en D. De oorspronkelijke voorstelling met álle frequenties is gereduceerd tot een representatie die slechts de relatieve frequenties bijhoudt van de 15 bigrammen die cumulatief gesproken het meest frequent zijn in alle samples. (De relatieve frequenties zijn weergegeven tot op 4 cijfers na de komma.)

In deze sectie introduceer ik bondig *Principal components analysis* (PCA), ook ook hoofdcomponenten-analyse genoemd.[14] Deze techniek kent reeds een lange voorgeschiedenis in de multivariate statistiek maar wordt de jongste decennia ook vaak toegepast binnen studies naar auteursherkenning. Gegeven het geïntendeerd publiek van deze studie zal ik vooral focussen op het effect van PCA en niet op de vrij complexe werking ervan. Wie meer wil te weten komen over de mathematische principes waarop PCA stoelt, kan te rade gaan bij een aantal inleidende studies die expliciet gericht zijn op een niet-geïnitieerd publiek van filologen (Binongo & Smith 1999; Binongo 2003). Omdat deze bespreking onvermijdelijk nogal abstract zal blijven, wordt de techniek in de volgende sectie toegepast op een relevante casus die het beoogde effect van een PCA onmiddellijk duidelijker zal maken.

Als de oorspronkelijke samples zijn herleid tot numerieke vectoren in onze tabel, staan we daarmee nog niet zo heel veel verder. Zelfs als men besluit om het aantal bestudeerde functoren terug te dringen tot 15 woordfrequenties, is onze tabel te omvangrijk om bestudeerd te worden (vgl. Tab. 6, p. 55). Een rechtstreekse visualisatie van de tabel is bovendien moeilijk, aangezien de teksten zouden moeten geprojecteerd worden in niet minder dan 15 verschillende dimensies. PCA is een van de technieken die statistici gebruiken voor 'dimensiereductie'. De bedoeling is namelijk om tot een bevattelijke 'samenvatting' te komen van de complexe informatie in onze tabel. Nadat de oorspronkelijke tabel reeds is herleid tot een relatief klein aantal functorfrequenties, zal PCA de overgebleven informatie nóg verder reduceren. Vanzelfsprekend is het wel de bedoeling om tot een betrouwbare samenvatting te komen die zoveel mogelijk recht doet aan de oorspronkelijke verhoudingen tussen de samples in de tabel.

Technieken voor dimensiereductie zijn overigens niet de enige methodes die voor de visualisatie van vectoren gebruikt kunnen worden. Clustering (zie § 9.1.3) of Correspondentie-analyse (zie § 6.3) zijn andere aantrekkelijke technieken die hiervoor kunnen dienen. Omdat deze studie tot op zekere hoogte een inleidend karakter heeft, zullen hieronder ter illustratie verschillende kwantitatieve methodes naast elkaar gebruikt worden. Ik zal daarbij zelden expliciet de voorkeur geven aan een specifieke techniek, want vaak leveren verschillende kwantitatieve methodes een vergelijkbaar eindresultaat op. Als hieronder tussen verschillende technieken geschakeld wordt, is dat dus niet omdat deze technieken heel verschillende resultaten opleveren, maar veeleer om het brede scala aan kwantitatieve analysetechnieken te illustreren.

14 De basis voor de techniek werd gelegd door Pearson (1901). Een (relatief technische) monografie over het onderwerp biedt Dunteman (1989). Leesbare introducties bieden Binongo & Smith (1999) en Binongo (2003). Uitstekend is ook de *tutorial* van Schlens (2009) – zie zeker het inzichtelijke voorbeeld met de veer en de camera's.

Belangrijk voor een goed begrip van PCA is het inzicht dat de woordfrequenties in teksten onderling gerelateerd zijn (Argamon 2008, 140-145). Als een bepaald woord in een tekst relatief meer frequent is, betekent dit per definitie dat andere woorden in die tekst relatief minder frequent zullen zijn. Dat betekent in statistische termen dat woordfrequenties 'gecorreleerd' zijn: er bestaat een verband of correlatie tussen het optreden van woorden. Het is erg belangrijk dat PCA – in tegenstelling tot andere technieken zoals Delta – deze onderlinge relaties tussen woorden expliciet tracht in kaart te brengen (Koppel et al. 2009, 10).

gecorreleerde woordfrequenties

Een dergelijke afhankelijkheidsrelatie is het duidelijkst bij woorden die alternatieven voor elkaar vormen. In het Nederlands treden het aanwijzend lidwoord 'de' en het aanwijzend voornaamwoord 'die' vaak in dezelfde contexten op, bijvoorbeeld in de parallelle frases 'de kat' vs. 'die kat'. Belangrijk is dat 'de' en 'die' elkaar onderling uitsluiten in deze frases, want '*de die kat' en '*die de kat' zijn ongrammaticaal. In deze context vormen 'de' en 'die' dus structurele synoniemen, waartussen onvermijdelijk een keuze moet worden gemaakt. Als 'die' in een tekst opvallend frequent is, is het bijgevolg goed mogelijk dat 'de' heel wat minder frequent zal zijn. Opnieuw is voor *n*-grammen een volledig analoge redenering op te bouwen. Als een kopiist graag gebruik maakt van de letters 'v' en 'i' zullen de corresponderende allografen 'u' en 'j' per definitie minder frequent zijn in een tekst. Ook op het niveau van bigrammen moeten dergelijke effecten waarneembaar zijn: als een kopiist zich opvallend graag bedient van de letterreeks 'ae', kan het haast niet anders dan dat de frequentie van de alternatieve schrijfwijzes 'aa' of 'ai' in diezelfde tekst is teruggedrongen.

Een dergelijke correlatie tussen twee variabelen zou men eigenlijk ook moeten samenvatten in één enkele, abstracte variabele. Die variabele zou dan een abstracte betekenis hebben, iets in de zin van *een hoge frequentie van 'de', maar een lage frequentie van 'die'*. Die nieuwe variabele kunnen we dan 'samengesteld' of *compound* noemen, omdat die eigenlijk de informatie uit meerdere variabelen bevat. Als we een dergelijke samenvatting herhalen voor alle items in onze tabel zullen we op een gegeven moment een aantal *compound variables* bekomen die onderling niet meer gecorreleerd zijn, zodat de samenhang tussen de oorspronkelijke variabelen is herleid tot een abstracte samenvatting. Een dergelijke abstracte variabele wordt een *principal component* of hoofdcomponent (PC) genoemd. Het gaat om een soort 'oer-variabele' die men heeft gedistilleerd uit de oorspronkelijke frequenties.

compound variable, samengestelde variabele

principal component, hoofdcomponent

Cruciaal is niet alleen dat deze *principal components* ongecorreleerd zijn, maar vooral dat het aantal PCs meestal veel kleiner is dan het aantal oorspronkelijke variabelen. Dat is voor stylometrici interessant. Als de oorspronkelijke variabelen zonder veel informatie-verlies kunnen worden herleid tot een veel kleiner aantal variabelen, wordt

de informatie in onze onoverzichtelijke frequentie-tabel veel bevattelijker. Hoeveel componenten uit een PCA zullen resulteren, valt op voorhand moeilijk te voorspellen. Aan elk van de componenten kan wel een waarde worden toegekend die uitdrukt hoe 'informatief' de component is ten aanzien van de oorspronkelijke data in de tabel. Hoe getrouwer de samenvatting in een component, hoe hoger deze waarde. Componenten worden dan meestal gerangschikt op basis van deze getrouwheidswaarde. Meestal wordt aangenomen dat slechts PCS die meer dan 5% procent van de oorspronkelijke variatie verklaren, relevant zijn.

Hét grootste voordeel van PCA is echter dat het uiteindelijk een 'plaatje' oplevert. De stilistische informatie in onze frequentie-tabel was oorspronkelijk niet bepaald aanschouwelijk. De *compound variables* die resulteren uit een PCA kunnen echter beschouwd worden als geometrische dimensies, waarin aan de oorspronkelijke teksten een positie wordt toegekend op basis van hun woordfrequenties. Voor een visualisatie wordt dan meestal gebruikt gemaakt van 'strooidiagrammen' of *scatterplots*. In deze diagrammen worden teksten als 'stippen' uitgezet in een twee-dimensioneel vlak, gewoonlijk op basis van hun positie in de twee meest informatieve 'dimensies' die uit de PCA resulteerden. Dergelijke strooidiagrammen zijn handig om het resultaat van een PCA te interpreteren. Als stippen in deze diagrammen dicht op elkaar worden afgebeeld, betekent dit dat de onderliggende samples belangrijke stilistische gelijkenissen vertonen. Als stippen daarentegen ver uit elkaar liggen, impliceert dit een grotere stilistische afstand tussen de onderliggende samples.

strooidiagram, scatterplots

De spreiding van teksten in dergelijke strooidiagrammen kan men vervolgens aanwenden om tot een interpretatie van het resultaat te komen. In stilistische studies is bijvoorbeeld opgemerkt dat een component soms de diachronie van de teksten in een oeuvre weet te vatten (Juola 2007). Bij de romans van Jane Austen bleek dan weer dat PCA de spreekbeurten van verschillende personages van elkaar wist te onderscheiden (Burrows 1987). In studies naar auteursherkenning is echter opgemerkt dat de spreiding van teksten in de eerste twee (soms drie) PCS vaak samenhangen met de factor auteurschap (Binongo & Smith 1999). Dat is zeker het geval als men een PCA uitvoert op teksten die zijn voorgesteld aan de hand van vectoren met een beperkte set functorfrequenties (Stamatatos et al. 2000, 475). Teksten van dezelfde auteur vormen in strooidiagrammen vaak hechte clusters van stippen, terwijl werken van verschillende auteurs relatief ver van elkaar worden weergegeven. Daarom vormen PCA en andere technieken voor dimensiereductie – zoals *Correspondence analysis* of *Multi-dimensional scaling* – een belangrijk hulpmiddel in veel recente studies naar stylometrische auteursherkenning.

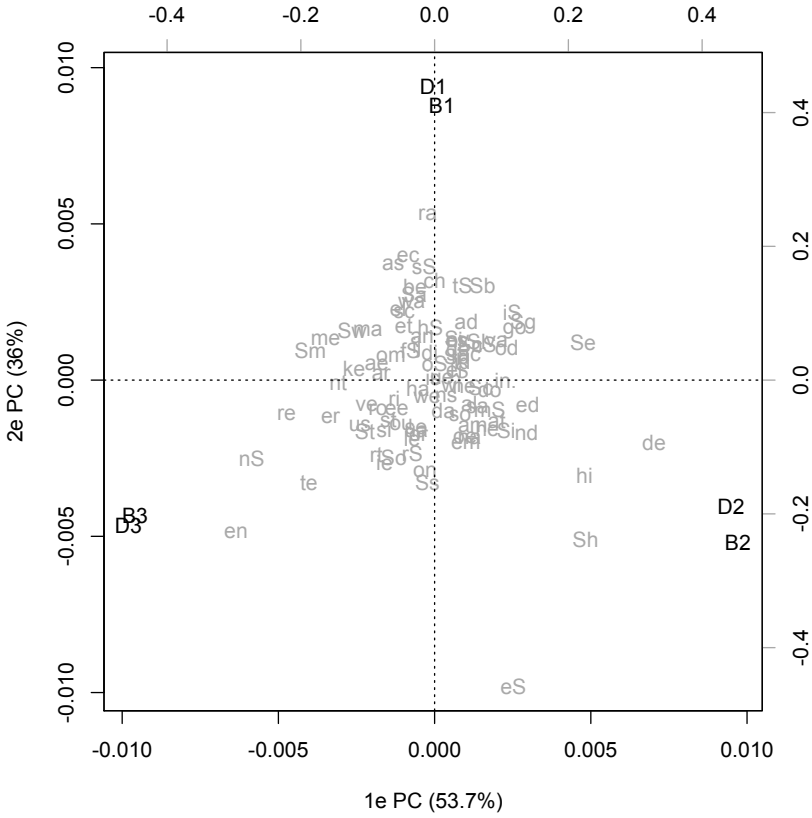
3.4 RIJMBIJBEL-ANALYSES

Laat ons nu een PCA uitvoeren van enkele parallelle handschriften van de *Rijmbijbel*. Zoals gezegd zal het resultaat van een PCA die werd uitgevoerd op de functorfrequenties van moderne teksten, vaak het auteurschap van teksten blootleggen. Wij verkeren op dit moment nog in de naïeve veronderstelling dat men een dergelijk effect mogelijk ook voor middeleeuwse teksten zal kunnen vaststellen. Het *Rijmbijbel*-corpus waarmee wij werken, bevat een aantal parallelle handschriften die in principe allemaal teruggaan op dezelfde laatdertiende-eeuwse auteurstekst van Jacob van Maerlant. We laten even in het midden dat sommige handschriften relatief oud en oorspronkelijk, of dat een bepaalde groep handschriften van elkaar of een gemeenschappelijke legger kan zijn afgeschreven. Van heel veel Middelnederlandse teksten is vaak slechts één afschrift overgeleverd, zodat men bij gebrek aan parallelle versies vaak wel *moet* aannemen dat het bewuste afschrift vrij oorspronkelijk is. Aangezien de grondtekst van deze Maerlant-afschriften 'van dezelfde auteur' zijn, zouden we dus kunnen veronderstellen dat een PCA geen onderlinge verschillen zal bemerken tussen de stijl van de parallelle handschriften.

Laat ons twee willekeurige handschriften beschouwen, bijvoorbeeld de codices B en D, die 1.102 parallelle regels bevatten. Hierboven hebben we ter illustratie gewerkt met een artificieel kleine sample-grootte van 5 regels. Voor een echte analyse is het echter aangewezen om met grotere steekproeven te werken, waaruit makkelijker betrouwbare tendensen zijn te distilleren. Als we ieder handschrift in drie samples van 367 regels verdelen, hoeven we slechts één parallelle regel te laten vallen en bekomen we in totaal 6 parallelle samples voor deze analyses: B_1, B_2, B_3 en D_1, D_2, D_3. Vervolgens zetten we deze samples om naar een frequentie-tabel en selecteren we de 100 meest frequente bigrammen in deze teksten. Als we een PCA uitvoeren op de resulterende 100-kolommige tabel met bigram-frequenties, verkrijgen we het strooidiagram dat hieronder is afgebeeld (Fig. 4, p. 60).

De donkerzwarte lettergroepen in deze figuur stellen de positie voor van de gelijknamige samples (de donkerzwarte letters D2 corresponderen bv. met het sample D_2). Deze zijn zo in het strooidiagram gespreid dat zij de structuur in onze frequentie-tabel weergeven, zoals die naar voor komt uit de eerste twee PCs. De horizontale spreiding in het diagram gaat terug op de belangrijkste of eerste component (x-as, '1e PC'); de verticale positie van de samples gaat terug op de tweede belangrijkste component PC (y-as, '2e PC'). (De lichtgrijze symbolen negeren we voorlopig nog.) Aan de beschrijving van de assen in dit vlak is een percentage toegevoegd: dit is de eerder besproken 'getrouwheidswaarde' die uitdrukt hoeveel van de oorspronkelijke informatie door de bewuste component word gevrijwaard. Onze eerste dimensie bewaart 53,7% van de oorspronkelijke variatie, de tweede

getrouwheidswaarde



Figuur 4: Strooidiagram resulterend uit een PCA die werd uitgevoerd op 2x3 parallelle samples (367 regels) uit de *Rijmbijbel*-handschriften B en D. De relatieve frequenties van de 100 meest frequente bigrammen werden gebruikt voor de voorstelling van deze samples. De positie van de samples wordt weergegeven voor de eerste twee PCs, die tezamen 89,7% van de oorspronkelijke informatie vrijwaren – een uitzonderlijk hoge score. De spreiding van de samples in dit diagram vertoont een overduidelijke correlatie met de index van de samples, zodat de spreiding sterk inhoudsgerelateerd lijkt. Een verband met de tekstdrager waaruit een sample werd geëxtraheerd is volledig afwezig.

dimensie 36%. Omdat beide dimensies gedecorreleerd zijn, mag men deze waardes gewoon optellen om een maat te bekomen voor de globale 'getrouwheid' van het afgebeelde diagram. In dit geval is het strooidiagram nog voor 89,7% (53,7%+36%) getrouw aan de oorspronkelijke structuur, hetgeen een uitzonderlijk hoge score is. De samenvatting in dit diagram is dus een veel meer compacte maar toch nog erg getrouwe benadering van onze oorspronkelijke tabel.[15]

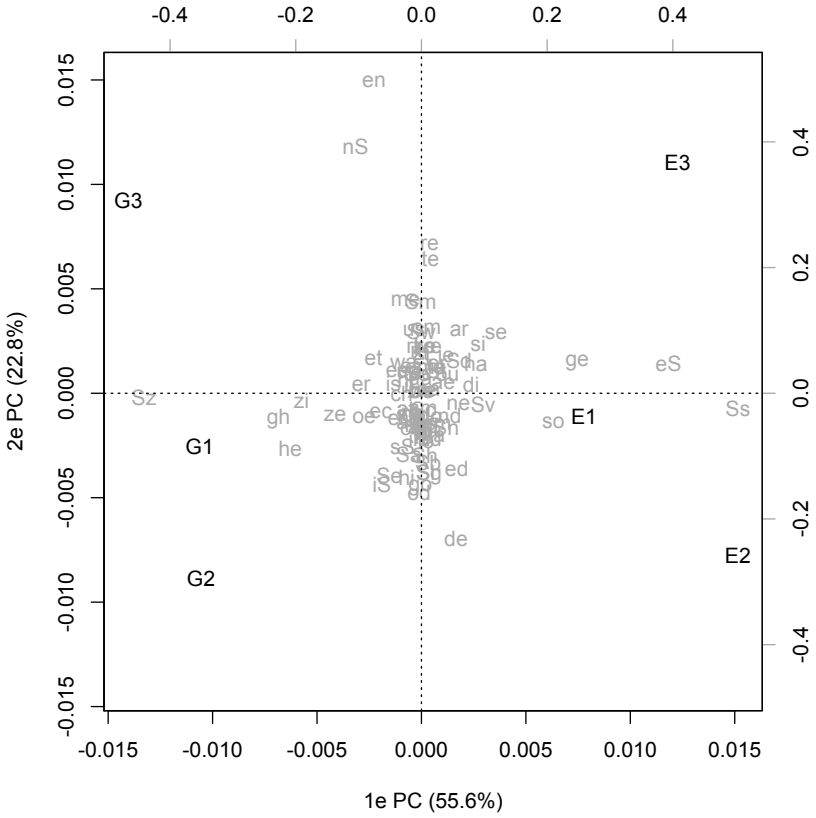
15 In principe mag men alle componenten die meer dan 5% van de oorspronkelijke variatie vrijwaren in een interpretatie van de PCA betrekken (Baayen 2008, §5.1.1). Een andere vuistregel is soms dat men alle componenten betrekt tot wanneer de getrouwheidswaarden plots een markante daling vertonen. Ik zal mij in deze studie

Een dergelijk strooidiagram kunnen we nu gebruiken voor de interpretatie van het resultaat van de PCA. Vaak gebruikt men bij de interpretatie van een strooidiagram impliciet een *nearest neighbour*-redenering, bijvoorbeeld in auteursherkenning (Argamon 2008, 144). Als de samples van een anonieme tekst opvallend dicht bij die van een andere, gekende auteur liggen, levert dit steun voor de hypothese dat het anonieme sample door dezelfde auteur werd geschreven. Als een groep samples opvallend ver van de samples van een gekende auteur ligt, kan een tegengestelde redenering worden opgezet, waarbij het auteurschap van de bewuste auteur wordt verworpen. In onze grafiek kunnen we vanuit een *nearest neighbour*-redenering een duidelijk clustering vaststellen wat betreft de index van de samples: de samples met eenzelfde index (bv. B_2 en D_2 rechts in het diagram) liggen duidelijk veel dichter bij elkaar dan samples met een verschillende index. Zo ligt het rechtse clustertje met B_2 en D_2 duidelijk ver verwijderd van het linkse clustertje met B_3 en D_3.

Hoewel de index van de samples een duidelijke correlatie vertoont met de sample-spreiding in dit diagram, is een dergelijk verband met het handschrift waaruit de samples komen volledig afwezig. Men zou dus opgelucht kunnen ademhalen: de factor 'handschrift' of 'kopiist' schijnt de PCA niet te verstoren. De samples schijnen veeleer te clusteren op basis van hun parallelle inhoud zodat er op het eerste zicht geen 'functorverstoringen' schijnen te zijn opgetreden ten aanzien van de oorspronkelijk auteurstekst. De willekeurige combinatie van de handschriften B en D is echter misleidend. Hieronder worden twee voorbeelden weergegeven waaruit een heel ander resultaat naar voor komt, hoewel een volledig analoge analyse werd uitgevoerd. De eerste figuur (Fig. 5, p. 62) contrasteert 6 parallelle samples (van 353 regels) uit de handschriften G en E, opnieuw op basis van de relatieve frequenties van de 100 vaakst voorkomende bigrammen in deze samples. De tweede figuur (Fig. 6, p. 63) biedt hetzelfde experiment maar voor de reeds besproken handschrift-combinatie A en D uit de eerdere sectie (6 samples van 368 regels).

In deze figuren (Fig. 5, p. 62 en Fig. 6, p. 63) is eenzelfde effect merkbaar. De index-structuur is nog steeds duidelijk in de sample-spreiding maar is teruggedrongen naar de tweede en dus minder belangrijke component. Men ziet hoe samples met dezelfde index in de verticale dimensie consequent op een vergelijkbare hoogte worden geplaatst. In de eerste en belangrijkste component heeft de inhoudsgerelateerde spreiding plaats moeten maken voor een heel ander effect. In de horizontale dimensie vallen de samples uiteen aan weerszijden van de verticale stippellijn door het nulpunt van de eerste hoofdcom-

consequent beperken tot een bespreking van de eerste twee dimensies in het geval dat deze beide tenminste 5% variatie verklaren. Dat is niet alleen gebruikelijk in de stylometrie, maar zo wil ik ook vermijden dat men op geforceerde wijze een (minder belangrijke) dimensie uitzoekt waarin de spreiding van samples het best aan een vooropgestelde hypothese beantwoordt.

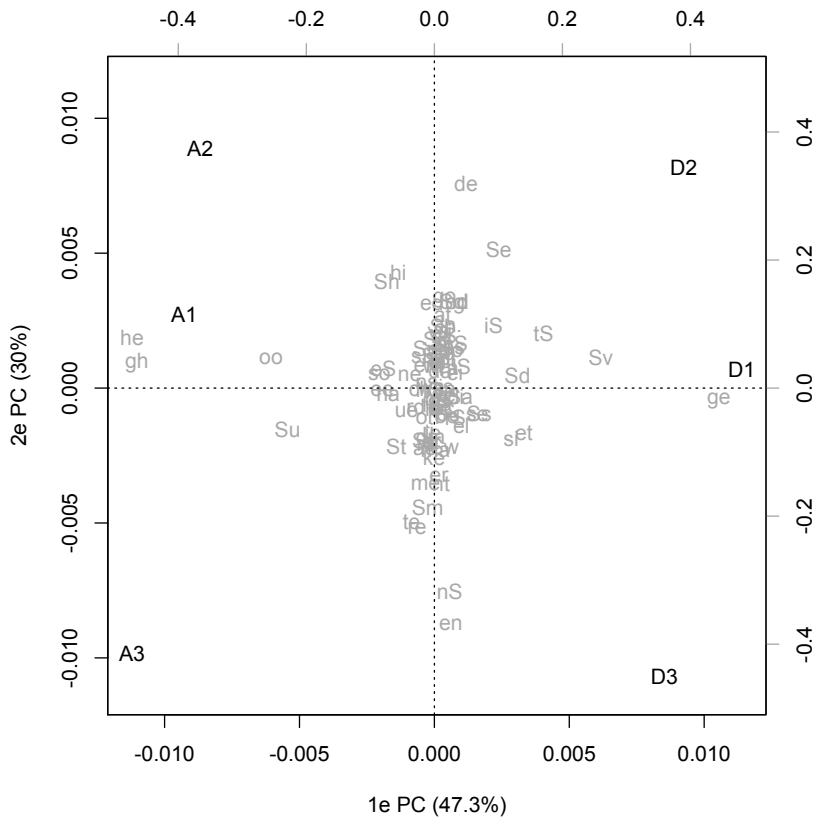


Figuur 5: Strooidiagram resulterend uit een PCA die werd uitgevoerd op 2x3 parallelle samples (van 353 regels) uit de *Rijmbijbel*-handschriften G en E. De parameters voor deze analyse zijn exact dezelfde als die van het vorige experiment. De spreiding van de samples in dit diagram vertoont eveneens in de eerste plaats een correlatie met de factor 'tekstdrager'. De inhoudsgerelateerde spreiding is nog aanwezig maar is naar de tweede, verticale component teruggedrongen.

ponent. De eerste component in deze analyses biedt zo een uitstekend model van de factor 'handschrift' of 'kopiist' in de bewuste samplecollecties. In deze experimenten vertonen de parallelle handschriften kennelijk wel een aanzienlijk verschil in het gebruik van functoren, zoals gereflecteerd in de gebruikte bigram-frequenties.

Op dit moment is het interessant om ook de lichtgrijze lettergroepen in deze diagrammen te beschouwen. Deze corresponderen met de oorspronkelijke bigrammen en hun positie reflecteert de zogenaamde *component loadings* van deze items. Deze 'ladingen' drukken uit hoe zwaar een bepaald bigram heeft doorgewogen bij de constructie van een bepaalde component. Concreet betekent dit dat we uit de figuur kunnen aflezen hoe specifieke bigrammen de spreiding van bepaalde samples beïnvloed hebben. In de laatste twee diagrammen vinden we de meeste bigrammen centraal in de figuur, als een lichtgrijze letter-

component loadings



Figuur 6: Strooidiagram resulterend uit een PCA die werd uitgevoerd op 2x3 parallelle samples (van 368 regels) uit de *Rijmbijbel*-handschriften A en D. De parameters voor deze analyse zijn exact dezelfde als die van het vorige experiment. De spreiding van de samples in dit diagram vertoont in de eerste plaats een correlatie met de factor 'tekstdrager'. De inhoudsgerelateerde spreiding is nog aanwezig maar is naar de tweede, verticale component teruggedrongen.

wolk waarin de individuele bigrammen amper nog te onderscheiden zijn. Daaruit kunnen we afleiden dat deze bigrammen geen bijzondere rol hebben gespeeld in de weergegeven sample-spreiding. Individuele bigrammen maken zich echter los uit deze centrale letterwolk en bezetten een meer uitgesproken positie in de figuur.

In het geval van de figuur met A en D (Fig. 6, p. 63) vinden we uiterst links in de figuur de bigrammen *gh* en *he*. Daaruit blijkt dat de PCA het frequente gebruik van deze bigrammen typisch acht voor de samples die in een vergelijkbare regio in de plot voorkomen en dus eveneens uiterst links zijn afgebeeld. De bigrammen *gh* en *he* worden daarmee typisch geacht voor de A-samples. Opmerkelijk genoeg treffen we aan de overzijde in dit diagram het bigram *ge* aan. Waar de PCA de bigrammen *gh* en *he* met de D-samples associeert, wordt het gebruik van *ge* typisch bevonden voor de D-samples. De analyse schijnt zo een duidelijke orthografische oppositie tussen A en D-samples bloot

te leggen. De *gh*-spelling (bv. *tghebod*, vgl. Tab. 3, p. 44) wordt immers vrij consequent geprefereerd in A, daar waar D de voorkeur geeft aan de alternatieve *g*-spelling (bv. *tgebod*). De PCA heeft de correlatie tussen de frequenties van deze bigrammen opgepikt en hieraan een grote 'lading' toegekend in de constructie van de belangrijkste component. Uit de confrontatie van E en G blijken gelijkaardige spellingsopposities. Naast de oppositie *gh* versus *g* die ook hier een rol speelt, is vooral het contrast tussen *Sz* en *Ss* erg uitgesproken. In de *Anlaut* van woorden verkiest G blijkbaar consequent het grafeem *z-* terwijl E de voorkeur geeft aan het corresponderende bigram *Ss* in deze positie. Zo leest G in de regel *zonder*, *zine* en *zi* waar het handschrift G zich van de spellingen *sonder*, *sine* en *si* bedient.

In de laatste experimenten schijnt het gebruik van functoren – of toch zeker hun oppervlakkige verschijningsvorm – dus grote verschillen te vertonen tussen handschriften onderling. De vraag is nu of we in de laatste experimenten voor geïsoleerde gevallen staan. Misschien zijn dit net twee extreme voorbeelden waarin de combinatie van twee handschriften grote orthografische verschillen vertoont. Hieronder is een tabel afgebeeld die jammer genoeg het tegengestelde doet vermoeden. In deze tabel (Tab. 7, p. 66) wordt voor elke combinatie van twee handschriften het resultaat van een statistische test weergegeven: de *Kruskal-Wallis*-toets. Met deze niet-parametrische statistische test ben ik nagegaan of de confrontatie van twee handschriften in de eerste dimensie een sample-spreiding opleverde die correleerde met de factor 'tekstdrager', zoals dat in de laatste twee experimenten het geval was. Een dergelijke test zal een score teruggeven die een maat is voor de mate waarin de horizontale sample-spreiding correleert met de factor 'tekstdrager'. Voor de laatste drie experimenten is die maat respectievelijk 0,4286, 3,8571 en 3,8571. Deze maat is echter minder belangrijk dan de *p*-waarde die volgens de test geassocieerd is met deze score. De *p*-waarde geeft aan hoe groot de kans is dat toeval deze uitkomst heeft veroorzaakt. Hoe lager deze waarde, hoe kleiner die kans en hoe significanter het resultaat. In de humane wetenschappen gaat men meestal uit van de (vrij willekeurige) drempelwaarde van 0,05. Als de *p*-waarde voor de uitslag van een test onder deze grens ligt, wordt aangenomen dat de uitslag 'significant' is. In het eerste experiment is deze spreiding zoals te verwachten net niet significant, in de laatste twee net wél.

Kruskal-Wallis-toets

p-waarde

In onderstaande tabel (Tab. 7, p. 66) is deze test steeds uitgevoerd voor een handschriftcombinatie waarbij 3 zo groot mogelijke samples werden geanalyseerd op basis van een representatie aan de hand van de relatieve frequenties van de 100 meest frequente bigrammen.[16] Dit overzicht leert dat de gevallen waarin een stylometrisch onderscheid

16 Merk op dat de waarden in deze tabel moeilijk te vergelijken zijn. Voor ieder handschrift-paar wordt immers gewerkt met verschillende sample-groottes, omdat niet alle handschriften evenveel parallelle verzen bezitten. Bovendien is het toepassen van een statistische test pas echt zinvol op een grotere verzameling meetpunten.

valt te maken tussen twee handschriften zeker niet in de minderheid zijn. Het mag overigens benadrukt worden dat PCA een 'ongesuperviseerde' aanpak is (*unsupervised approach*). De analyse weet op geen enkel moment uit welk handschrift een sample afkomstig is en bijgevolg wordt geen bewuste poging ondernomen om de samples van elkaar te scheiden. Dat de handschriftelijke herkomst van zoveel samples toch makkelijk is te achterhalen, benadrukt hoe grondig handschriften soms verschillen.

unsupervised approach

De stylometrische techniek die hierboven werd toegepast blijkt in veel gevallen dus erg gevoelig voor kopiisteninterferentie met een tekst. Doordat kopiisten zichzelf een tekst op verschillende niveaus (tenminste orthografie en dialect) 'toe-eigenden', zijn zij niet rechtstreeks vergelijkbaar wat betreft de stijl. Dat benadrukt hoe gevaarlijk het is om Brinkmans Eerste Gebod te schenden. Teksten van dezelfde auteur die overgeleverd zijn in afschriften van verschillende kopiisten zullen mogelijk een artificieel grote afstand vertonen, indien onaangepaste stylometrische methodes worden toegepast. Volgens diezelfde onaangepaste methodes kan omgekeerd blijken dat teksten van verschillende auteurs, overgeleverd in een afschrift van dezelfde kopiist, artificieel dicht bij elkaar zullen liggen. Men loopt met andere woorden het gevaar om de analyse van auteursstijl te *overfitten* op de kopiist van een afschrift.

3.5 KOPIISTENHERKENNING?

In dit hoofdstuk heb ik toegelicht hoe teksten numeriek kunnen worden voorgesteld om ze geschikt te maken voor een stylometrische analyse (§ 3.2). Die representatie werd geïllustreerd aan de hand van een verzameling parallelle handschriften van de *Rijmbijbel*. Nadien werd een relatief veelgebruikte techniek voor stylometrische tekstvergelijking geïntroduceerd die ook hieronder nog vaak zal gebruikt worden: *Principal components analysis* (§ 3.3). Deze techniek werd toegepast op parallelle tekstgetuigen van de *Rijmbijbel* (§ 3.4). Uit een reeks experimenten werd duidelijk dat de techniek in veel gevallen gevoelig was voor de variatie die tussen handschriften bestaat. Kopiisten eigenden zich een exemplaar op verschillende niveaus toe. Dat lijkt zeker het geval voor het dialect en de spelling van een tekst.

Natuurlijk mag men aan deze casus geen grotere conclusies verbinden dan strikt mogelijk. Functorverstoringen traden lang niet in alle handschriftcombinaties (in dezelfde mate) op. Bovendien kan men deze bevindingen niet zomaar extrapoleren naar de overlevering van andere teksten. Waar het in dit hoofdstuk vooral om ging, was het illustreren van het feit dat de *kans* bestaat dat een middeleeuwse kopiist zijn legger grondig heeft aangepast. Van veel Middelnederlandse

Het gaat mij hier slechts om het illustreren van een effect, zonder dat ik voor ieder experiment een strooidiagram moet weergeven.

	A	B	D	E	F	G	J	K	L	M	N
A		-	+	-	-	+	+	+	-	-	+
B			-	-	-	+	+	-	-	-	-
D				-	+	+	+	-	+	-	-
E					+	+	+	-	+	-	-
F						+	+	+	-	-	+
G							+	+	+	+	+
J								+	+	+	+
K									+	-	-
L										-	+
M											-
N											

Tabel 7: Overzicht van een reeks experimenten waarbij telkens 3 zo groot mogelijke samples uit twee *Rijmbijbel*-handschriften werden vergeleken in een PCA. Analoog aan de voorbeelden hierboven werden steeds 100 bigram-frequenties gebruikt. In de eerste dimensie werd voor ieder paar met behulp van een *Kruskal-Wallis*-toets getest of de sample-spreiding correleerde met het handschrift waaruit de samples afkomstig waren. De plus- en min-symbolen geven aan of de uitslag respectievelijk wel of niet significant was (voor een p-niveau van 0,05 en df=1).

teksten resten er te weinig parallelle tekstgetuigen om de oorspronkelijke auteurstekst te reconstrueren, zodat de status van de individuele handschriften onzeker blijft. Daarmee blijft het momenteel onduidelijk hoe men een stylometrisch onderzoek kan opzetten naar de Middelnederlandse letterkunde, zonder Brinkmans Eerste Gebod op grote schaal te schenden.

De stemmatologische aspecten van de *Rijmbijbel*-traditie zijn hierboven opzettelijk genegeerd. In dat licht moet natuurlijk wel opgemerkt worden dat het logisch is dat teksten die dichter bij elkaar in het stemma liggen, ook moeilijker onderscheidbaar kunnen zijn. De spelling van een legger kan in bepaalde gevallen nog doorschemeren in de spelling van het afschrift, zodat de afstand tussen exemplaar en kopie niet noodzakelijk heel groot is. Een ander probleem waaraan ik in dit verkennend experiment geen aandacht heb besteed, is dat een specifiek handschrift misschien door één enkele kopiist is afgeschreven, maar op basis van een legger die wél door meerdere kopiisten is vervaardigd. Zeker in dit geval is het goed mogelijk dat de latere kopie nog verschillen kan verraden tussen de verschillende kopiistenhanden in de oudere legger. Deze problematiek moet in toekomstig onderzoek hoe dan ook in rekening worden gebracht. Bovendien zal men een onderscheid moeten maken tussen het gedrag van individuele kopiisten,

want ongetwijfeld heeft niet iedere afschrijver altijd even doortastend ingegrepen in zijn of haar legger.

Toch kan men zich afvragen of het in de toekomst mogelijk zal zijn om middeleeuwse kopiisten te herkennen aan de hand van stylometrische technieken, net zoals dat vaak mogelijk blijkt voor de auteurs van modernere teksten. De resultaten in dit hoofdstuk suggereren namelijk sterk dat technieken voor auteursherkenning in sommige gevallen zonder de minste aanpassing geschikt zijn om de kopiisten van parallelle teksten te onderscheiden.

Deze vraagstelling is vooral vernieuwend omdat de stylometrie op basis van louter tekstuele, linguïstische metingen opereert. Traditioneel beperkt men zich in de paleografie echter tot een analyse van de uiterlijke kenmerken van een handschrift, bijvoorbeeld voor de gelijkstelling van kopiistenhanden in verschillende codices.[17]. Het identificeren van schrijfhanden is echter vaak een betrekkelijk controversiële aangelegenheid, omdat de gehanteerde werkwijzes vaak nogal (inter-)subjectief zijn en gebaseerd worden op de expertise van individuele specialisten die hun methodologie soms onvoldoende expliciteren. Het staven van kopiistenattributies is daarom vaak een hachelijke praktijk.[18]

Dat is zeker het geval wanneer men handen tracht te herkennen over de grenzen van een specifieke schriftsoort heen: als een kopiist zowel teksten in een *littera textualis* als in een *littera cursiva* heeft geschreven, is het vaak moeilijk om dit vast te stellen op basis van uiterlijke schriftkenmerken. Als men zou beschikken over een computationele methodologie die handen kan herkennen op basis van louter intern-linguïstische kenmerken, zou dat een grote methodologische voordelen bieden, want deze benadering is:

OBJECTIEF – de methode is minder afhankelijk van variabelen die een subjectieve interpretatie vergen die soms moeilijk kan worden geëxpliciteerd;

GEAUTOMATISEERD – aangezien de methode geen manuele analyse door een expert vergt, kunnen makkelijk grote databases worden doorzocht op mogelijk identificaties voor een nieuw afschrift – hetgeen veel moeilijker en meer tijdrovend is voor een mens (vgl. Bulacu & Schomaker 2007, 701);

SCHRIFT-ONAFHANKELIJK – de methode opereert niet op basis van visuele schriftkenmerken: in theorie maakt dat kopiistenherkenning mogelijk over de grenzen van verschillende schriftsoorten.

17 De problematiek komt bijvoorbeeld aan bod bij Kwakkel (2002, 136) of bij Klein (1995, 2). Invloedrijke recente casussen (met verwijzingen naar relevant voorgaand werk) zijn Kwakkel & Mulder (2001, in het bijzonder p. 151) of Brinkman & Mulder (2003, 27-28)

18 Vandaar dat binnen de Artificiële Intelligentie ook onderzoek wordt uitgevoerd naar kopiistenherkenning via beeldverwerking (bv. Bulacu & Schomaker 2007).

Bovendien kan men de technieken toepassen op edities van handschriften die mogelijk zijn verloren gegaan na het verschijnen van de editie;[19]

EXPERT-ONAFHANKELIJK – de methode vereist in beginsel geen tussenkomst van een expert en kan dus worden uitgevoerd door een paleografische 'leek' die geen uitgebreide kennis en ervaring heeft inzake middeleeuwse schriftsoorten.

Vanzelfsprekend is er nog veel meer onderzoek nodig naar stylometrische kopiistenherkenning. Zo is het bijvoorbeeld nog maar de vraag of men eenzelfde kopiistenhand in een afschrift van verschillende teksten zou kunnen herkennen. Omdat in dat geval niet meer met parallelle teksten wordt gewerkt, kan het voor de analyse moeilijker zijn om te focussen op de eigenheden van kopiisten. In deze studie wordt vanaf het volgende hoofdstuk niettemin enkel nog aandacht besteed aan auteursherkenning in plaats van kopiistenherkenning. Daarbij zal het in de eerste plaats zaak zijn om de 'stilistische' invloed van een kopiist te neutraliseren.

19 Ik dank deze suggestie aan E. de Bruijn (Antwerpen), met wie ik een artikel voorbereid over kopiistenherkenning in de parallelle handschriften van de Middelnederduitse *Flos unde Blankeflos*. Binnen die casus zijn exact dezelfde resultaten te verkrijgen als voor de traditie van de *Rijmbijbel*. Een niet-gerapporteerd onderzoek naar de parallelle incunabelen van *Karel ende Elegast* suggereerde zelfs dat deze analysemethodes bruikbaar zijn voor het herkennen van de zetters van jongere drukken.

Deel II

HET NOODRIJM

In het voorgaande deel van deze studie heb ik geïllustreerd dat de toepassing van stylometrische methodes op Middelnederlandse teksten niet vanzelfsprekend is. Voor een betrouwbare analyse van de schrijfstijl in teksten, moet men de variatie die kopiisten in afschriften hebben aangebracht, tot op zekere hoogte buiten spel kunnen zetten. In dit deel wordt beargumenteerd dat hoogfrequente rijmwoorden een geschikte uitvalsbasis bieden voor auteursherkenning in berijmde narratieve teksten. In een eerste hoofdstuk wordt een experiment besproken met twee auteurs van de *Spiegel historiael*, Maerlant en Utenbroeke. In het volgende hoofdstuk wordt ook de bijdrage van Lodewijk van Velthem aan de *Spiegel* besproken. Het laatste hoofdstuk van dit deel biedt een stylometrische analyse van het bijna volledige oeuvre van Jacob van Maerlant. Het doel van dit deel is het ijken van bepaalde stylometrische technieken die in het volgende deel zullen worden toegepast op teksten waarvan het auteurschap onzeker is.

4

DE SPIEGEL

4.1 HET RIJM ALS REDDINGSBOEI

Hierboven heb ik besproken waarom stylometrici voor auteursherkenning zoveel belang hechten aan functoren (§ 2.1).[1] In tegenstelling tot lager frequente items, zijn functoren goed gespreid over teksten en minder gebonden aan een specifiek onderwerp of genre. Op middeleeuwse teksten is dit inzicht niettemin moeilijk toepasbaar, omdat van deze teksten meestal geen autografen zijn overgeleverd, maar slechts jongere afschriften (van afschriften) daarvan. Hierboven is geïllustreerd dat middeleeuwse handschriften van dezelfde auteurstekst een grote variatie kunnen vertonen. Analyses van de parallelle handschriften van de *Rijmbijbel* toonden hoe kopiisten zich een tekst konden toe-eigenen (§ 3.4). Het duidelijkst is dat voor de relatief oppervlakkige variatie die men aantreft op het niveau van orthografie en dialectkenmerken. Bovendien leert stemmatologisch onderzoek dat afschriften ook op het lexicale niveau onderling kunnen afwijken als gevolg van (meer of minder) bewuste ingrepen van opeenvolgende kopiisten. Bij gebrek aan een uitgebreide handschrifttraditie is het voor veel Middelnederlandse teksten onbegonnen werk om de oorspronkelijke auteurstekst te reconstrueren.

Toch is één categorie woorden in het verleden aangeduid als betrouwbaar voor het herkennen van Middelnederlandse auteurs: *rijmwoorden*.[2] De Middelnederlandse narratieve literatuur of epiek is namelijk voor het leeuwendeel berijmd in coupletten (aabbccdd...).[3] Dit rijmwoord zou een betrekkelijk taai en stabiel element zijn in de overlevering van teksten en bestendig of 'robuust' tegenover de variatie die latere kopiisten in een tekst konden introduceren.[4] Bekend is het beeld dat wie kijkt naar de rijmen van een epische tekst, de oorspronkelijke dichter recht in het aangezicht staart (vgl. Besamusca 1991, 165). De schier eindeloze ketting van coupletten ligt structureel aan de basis van berijmde Middelnederlandse teksten (Van Driel 2007, 37). Een kopiist kon dan wel makkelijk morrelen aan de woorden binnenin een vers maar wat het rijmwoord betreft, zat een afschrijver redelijk vast

1 Een eerdere versie van dit hoofdstuk werd reeds gepubliceerd als artikel (Kestemont 2011d).

2 Dit gegeven wordt vaak vermeld in de secundaire literatuur, bijvoorbeeld bij Besamusca (1991, 165-166), Biemans (1997, 207-228), Westgeest (2001, 15-16), Van Driel (2007, 164-167) en recent nog Caers (2011, 224).

3 Zie Lie (1994) of recenter Besamusca (2007).

4 Representatief voor deze zienswijze zijn de vele publicaties van E. van den Berg, waaronder Van den Berg (1992b, 200ff), Van den Berg (1983) of Van den Berg (1986, 305-306).

aan de grondtekst. Immers, als een afschrijver een rijm wou aanpassen, zou hij op betrekkelijk omslachtige wijze een deel van de grondtekst moeten herwerken. Hoewel de spelling van rijmwoorden soms aangepast werd, werd het onderliggende rijmwoord veelal intact gelaten (Van den Berg 1986, 305-306). De parallelle passages uit de handschriften in het vorige hoofdstuk illustreren treffend hoe handschriften die aan de oppervlakte uitgesproken verschillen vertonen, wat betreft hun rijmwoordenschat merkwaardig sterke overeenkomsten bezitten (vgl. Tab. 2, p. 42 & Tab. 3, p. 44).

Rijmwoorden lijken op die manier eilandjes van stabiliteit te vormen in Middelnederlandse teksten. Als er van de oorspronkelijke auteurstekst nog iets is overgebleven in latere kopieën, zal dat in de eerste plaats in het rijm zijn terug te vinden. In deze studie zal ik daarom de vraag centraal plaatsen of het mogelijk is om auteursherkenning uit te voeren in de Middelnederlandse epiek op basis van rijmwoordenschat. Als het rijmprofiel in een jonger afschrift inderdaad nog representatief is voor dat van de oorspronkelijke auteur, leveren rijmwoorden een uitgelezen kans om Brinkmans Eerste Gebod vakkundig te omzeilen bij het vergelijken van teksten. Hoewel het uitgangspunt van deze studie goed is te verantwoorden op basis van de secundaire literatuur, blijft voorzichtigheid vooralsnog geboden. De aanname dat iedere Middelnederlandse auteur te herkennen zou zijn aan een individuele rijmstijl is immers nog erg hypothetisch en hetzelfde geldt voor de sterk veralgemenende stelling dat latere kopiisten het rijmgebruik in een tekst altijd ongemoeid lieten. In dit hoofdstuk zal ik daarom via een relevante casus aantonen hoe het rijmwoord inzetbaar is voor auteursherkenning in de Middelnederlandse letterkunde.

Eerste Gebod: Eenheid van kopiist

4.1.1 *Een atelier te Damme*

De casus waarmee ik in dit hoofdstuk zal werken, is de *Spiegel historiael*. Deze tekst vormt een kolossale bewerking in Middelnederlandse verzen van de Latijnse wereldgeschiedenis *Speculum historiale* van Vincentius van Beauvais.[5] Het initiatief voor dit grote project moet ca. 1280 zijn genomen door Jacob van Maerlant. In navolging van de tekstgeleding van zijn brontekst voorzag Maerlant aanvankelijk vier grote tekstblokken of Partieën. Deze Partieën werden onderverdeeld in boeken die op hun beurt weer bestonden uit kleine hoofdstukjes, kapittels genaamd.

Maerlant was wel de architect van de *Spiegel* maar heeft het project niet alleen uitgevoerd. Hij schreef de Eerste Partie (die de geschiedenis verhaalde van de Schepping tot Nero) maar sloeg de tweede over. De Tweede Partie werd geschreven door Filip Utenbroeke, waarin de geschiedenis werd verhaald tot het jaar 381. Aangezien Maerlant

5 Zie voor deze algemene gegevens de belangrijke studies van Biemans (1997, 19ff) en Van Oostrom (1996, 307ff).

wel de Derde Partie schreef (de geschiedenis tot net voor Karel de Grote), moet hij dus geweten hebben dat Utenbroeke de tweede voor zijn rekening zou nemen, hoewel dat in zijn deel van de *Spiegel* niet expliciet kenbaar wordt gemaakt (Biemans 1997, 20-23).

Maerlant begon daarna ook aan de Vierde Partie maar moest zijn werkzaamheden voortijdig staken, wellicht om gezondheidsredenen. De Vierde Partie werd later afgemaakt door de Brabantse priesterdichter Lodewijk van Velthem, die op eigen initiatief ook een Vijfde Partie aan het geheel toevoegde (Besamusca et al. 2009a, 10ff). Het is overigens uit Velthems werk dat we Utenbroekes naam kennen, want die wordt bij Velthems voorgangers niet vermeld (Biemans 1997, 23). In het voorlaatste kapittel van het slot dat Velthem aan de Vierde Partie breide, lezen we hoe Maerlant en Utenbroeke reeds overleden waren op het moment dat Velthem de tekst afrondde (in augustus 1315):

> Des danke ic [i.e. Velthem] Gode ende sinte Marie,
> Dat ic gheënt hebbe dese pertie,
> Die Jacob van Merlant began,
> Die deerste entie derde vort an
> Maecte, ende oec in die vierde
> Hi syns levens faelierde;
> Entie ander, die hi achterliet,
> Die maecte een, die men hiet
> Philip ane den Dam Uten Broke,
> Ende oec starf na den boeke.
> In die vierde heeft gemaect nu
> Haer Lodewijc van Veltem segic u;
> Van daert Jacob liet ter stont,
> So heeft hijt vort gemaect cont
> (De Vries & Verwijs 1982, dl. 3, 432, vs. 41-54)

De relatie tussen Maerlant en Utenbroeke is intrigerend: Maerlant en Utenbroeke moeten elkaar van nabij gekend hebben.[6] Beide West-Vlamingen zouden in het laatste kwart van de dertiende eeuw professioneel actief zijn geweest in of rond Damme. Maerlant was er mogelijk schepenklerk en werd misschien in die functie opgevolgd door Filip, die stamde uit een in de streek belangrijke familie (Biemans 1997, 23-24). Algemeen wordt aangenomen dat Maerlant, die toen reeds naam en faam moet hebben gehad, de Tweede Partie als het ware uitbesteedde aan de jongere Filip. Het heeft er alles van dat Utenbroeke als een soort stagiair-assistent in de leer ging in het atelier van meester Maerlant (Biemans 1997, 23-24). Een dergelijk samenwerkingsverband kennen we uit de schilderkunst en ook van Vincentius

6 Voor de relatie tussen beiden, zie vooral Van Oostrom (1992, 203-204), in het bijzonder noot 63 voor de figuur van Utenbroeke, die weinig is bestudeerd. Het daar vermelde onderzoek door Els Sneep is tot op heden (jammer genoeg) niet gepubliceerd.

weten we dat die hulp kreeg van assistenten (Mantingh 2009, 207, n. 2). De exacte relatie tussen beide dichters is nog onduidelijk maar de verschillende verwijzingen in Maerlants Derde Partie naar Utenbroekes aandeel doen alleszins vermoeden dat het duo nauw heeft samengewerkt.

De combinatie Maerlant-Utenbroeke is in veel opzichten een uitstekende casus voor auteursonderscheiding. In het onderzoek wordt vaak benadrukt dat het in tekstvergelijking voor auteursherkenning belangrijk is alle factoren, buiten het auteurschap, zo constant mogelijk te houden. De vergeleken teksten verschillen beter zo weinig mogelijk in genre, inhoud en dialect en de auteurs delen bij voorkeur ook hun opleidingsniveau en sociale achtergrond (vgl. Besamusca 1991, 140). Op die manier wordt verzekerd dat eventuele verschillen tussen teksten slechts teruggaan op de factor 'auteur' en niet op andere variabelen. De casus Maerlant-Utenbroeke benadert in veel opzichten dit ideaal want van weinig Middelnederlandse dichters kan aangetoond worden dat zij zo dicht bij elkaar stonden. De dichters waren afkomstig uit dezelfde streek, deelden hun dialect, moeten een gelijkaardige opleiding genoten hebben en hadden heel waarschijnlijk dezelfde professionele achtergrond.

Ook hun teksten lenen zich uitstekend tot deze verkennende casus: ze vertrokken vanuit dezelfde brontekst en hadden duidelijk de bedoeling om in samenwerking tot een coherent geheel te komen. Hier zullen de Tweede en Derde Partie worden vergeleken: de kans is groot dat zij gelijktijdig, op dezelfde plek en met dezelfde instelling aan deze teksten werkten. Ook zijn de teksten op dezelfde wijze gestructureerd in kapittels en boeken, zodat deze als een organische eenheid voor tekstvergelijking kunnen dienen. Bovendien is beweerd dat van Maerlants rijmgedrag een dwingende invloed is uitgegaan op zijn navolgers (Van Loey 1946, 42-43). Als het mogelijk blijkt om op basis van de rijmwoorden de Tweede en Derde Partie te onderscheiden, gaat dit onderscheid voor het merendeel terug op het auteursverschil aangezien andere variabelen hieronder op erg natuurlijke wijze zo goed als constant worden gehouden.

Utenbroekes aandeel is slechts min of meer volledig overgeleverd in Wenen, Ö.N.B, Cod. 13.708, dat zijn toenaam – Tweede Partie-handschrift – aan de aanwezigheid van deze tekst dankt.[7] Dit handschrift kwam aan het einde van de veertiende eeuw tot stand in de kartuis te Herne. Hoewel het handschrift niet de volledige Tweede Partie bevat, vormt deze tekstgetuige voor Utenbroeke het geijkte startpunt. Van de Hernse kartuizers is geweten dat zij werkten met het grootste respect voor de brontekst en het bewuste afschrift kan als erg degelijk worden beschouwd.[8] Het handschrift bevat 366 van de ca.

7 Handschrift 64 bij Biemans (1997, via de index, maar ook p. 116ff). Recent over het handschrift: Kwakkel (2002, 128ff) en Kestemont (2009).

8 Deze claim werd getoetst aan de hand van de casus van Velthems kapittels in het handschrift (Kestemont 2009).

460 oorspronkelijke kapittels verspreid over zeven boeken. Maerlants Derde Partie is volledig overgeleverd in handschrift DEN HAAG, KA XX, een prachtig geïllumineerd handschrift dat Maerlants aandeel in de *Spiegel historiael* volledig dekt.[9] Uit deze Derde Partie worden hieronder 366 kapittels – evenveel als Utenbroeke – gebruikt, met name alle kapittels tot en met de Derde Partie, Boek 7, kapittel 60. Digitale, kritische edities van de tekst van beide auteurs zijn integraal te vinden op de *Cd-rom Middelnederlands*; deze bestanden worden hieronder gebruikt.

4.1.2 *Lemmatisering*

Hierboven werd al opgemerkt dat kopiisten eerder zelden de rijmwoorden van teksten zouden hebben aangepast, zodat de lexemen in rijmpositie veelal behouden werden in latere afschriften. Het is opvallend dat kopiisten vaak zelfs rijmparen uit een grondtekst overnamen als die eigenlijk niet goed 'klonken' in hun eigen dialect. Een goed voorbeeld daarvan zijn de beruchte *bem*-rijmen uit het middeleeuwse graafschap Vlaanderen. De eerste persoon enkelvoud van de tegenwoordige tijd van het werkwoord ZIJN kende in de Vlaamse variant van het Middelnederlands de vorm *ic bem*. In andere dialectgebieden, zoals bijvoorbeeld in het hertogdom Brabant, stond daar de vorm *ic ben* tegenover, zoals we die nu ook in de Nederlandse standaardtaal terugvinden. Dat betekent dat de combinatie van *(ic) bem* met *hem* (van het pronomen HIJ) voor Vlamingen een perfect acceptabel rijmpaar opleverde, maar buiten het graafschap – als assonantie – minder goed moet hebben geklonken.

De Vlaamse *bem*-rijmen vormen dan ook een dankbaar lokaliseringsargument: als een *bem*-rijm in een tekst optreedt, is het bijvoorbeeld goed mogelijk dat de tekst werd geschreven door een auteur die in het graafschap Vlaanderen is opgegroeid.[10] Vlaamse auteurs waren zich goed bewust van de gemarkeerdheid van hun *bem*-rijmen. In de proloog van *Sinte Franciscus leven* richt de Vlaming Jacob van Maerlant zich tot een publiek van Utrechtse minderbroeders aan wie hij dit heiligenleven opdroeg. Maerlant verontschuldigt zich voor het feit dat zijn Vlaamse tongval nog in de tekst zal merkbaar zijn. Spelenderwijs doet hij precies voor die formulering een beroep op een *bem*-rijm:

9 Handschrift 1 in Biemans (1997, via de index, maar ook p. 197ff).

10 Dat is echter niet noodzakelijk het geval. Een bv. Brabantse auteur kan immers zijn dialect hebben aangepast met het oog op een Vlaams doelpubliek (Besamusca & Postma 1997, 14). Bovendien kunnen we eveneens te maken hebben met een later geïmmigreerde auteur. Een laatste probleem is dat het Vlaamse dialect zich in het Middelnederlands al snel tot een supra-regionale variant zou ontwikkelen, zodat Vlaamse taalkenmerken geleidelijk aan ook elders in het taalgebied optreden. Deze problematiek wordt aan de orde gesteld in de recente studies van Willaert (2010) en Caers (2011).

Ende om datic Vlaminc bem
Met goeder herte biddic hem
Die dit Dietsche sullen lesen
Dat si mijns genadich wesen (vs. 125-128)

Wanneer Vlaamse teksten in Brabant werden afgeschreven, werden dergelijke *bem*-rijmen vaak oppervlakkig aangepast. In de *Lancelot-compilatie*, bijvoorbeeld, werden veel in oorsprong Vlaamse teksten in een Brabants afschrift opgenomen (vgl. § 8.2). Daarin treedt ettelijke keren het rijmpaar *ben* : *hem* op. Hoewel dergelijke rijmparen door de bewuste scribenten dus oppervlakkig zijn aangepast, is het oorspronkelijk uitzicht van deze rijmen nog makkelijk te reconstrueren (Besamusca & Postma 1997, 14). Het abstracte lexeem is in dergelijke afschriften dus ongewijzigd, maar de oppervlakkige vorm is kennelijk wel aan variatie onderhevig. Ook de parallelle handschriften van de *Rijmbijbel* tonen dat kopiisten graag in de spelling van rijmwoorden ingrepen. De parallelle eerste verzen uit de *Eva*-passage (zie Tab. 2, p. 42) bevatten in rijmpositie wel allemaal hetzelfde lexeem UUR, maar de schrijfwijze van dit woord wisselt sterk tussen de handschriften (*vren*, *uren*, *veren*, *vre*).

Voor een betrouwbare tekstvergelijking is het dus nodig om de rijmwoordenschat van teksten te normaliseren (Knowles & Mohd Don 2004, 70). In deze studie is dat gebeurd via lemmatisering, een proces waarbij men een token koppelt aan een lemma, of de genormaliseerde hoofdvorm van een woord zoals we die aantreffen in een woordenboek (Francis & Kucera 1982, 1). Aan tokens die onderling slechts verschillen inzake spelling, vervoeging of verbuiging, wordt bijgevolg hetzelfde lemma toegekend (Lyras et al. 2008, 1043ff). De bedoeling is natuurlijk dat we abstractie kunnen maken van de verschillen tussen tokens die niet wezenlijk of zelfs hinderlijk zijn voor het uitvoeren van een specifieke tekstanalyse (Crystal 1997). We willen in het geval van rijmwoorden immers op zinnige wijze kunnen generaliseren over de verschillende oppervlakte-vormen van dezelfde lexemen.

lemma

De lemmatisering van de rijmwoorden voor deze studie is uitgevoerd via speciaal hiertoe ontwikkelde software. Deze techniek is beschreven door Kestemont et al. (2010) en wordt in appendix achterin nader toegelicht. Ik beperk mij in de hoofdtekst tot een bespreking van de voornaamste aspecten van deze lemmatisering. De gebruikte *lemmatiser* is getraind op het literaire deel van het digitale *Corpus-Gysseling* dat wordt ontwikkeld door het Instituut voor Nederlandse Lexicologie (INL).[11] In dit corpus zijn de Middelnederlandse woorden

lemmatiser

11 Zie http://www.inl.nl. Momenteel worden voorbereidingen getroffen om het digitale *Corpus-Gysseling* online beschikbaar te maken, zoals reeds is gebeurd met verschillende woordenboeken. In het digitale Middelnederlandse woordenboek op de site van het INL kan men voorbeelden vinden van de lemma's (of moderne pendanten) die in deze studie zijn gebruikt. In het geval van twijfel bij het lemmatiseren heb ik steeds de recentste versie van dit online woordenboek geconsulteerd.

verbonden met een modern lemma, of toch een lemma in een moderne spelling in het geval dat er geen moderne pendant meer voorhanden is (Van der Voort van der Kleij 2005). In deze studie zal uitsluitend worden gewerkt met tellingen op basis van dergelijke lemma-labels.

Volgende basisprincipes zijn gehanteerd bij het lemmatiseren van teksten. Homonieme lemmata werden niet expliciet onderscheiden: dat wil bijvoorbeeld zeggen dat HEER in de betekenis van 'meester' niet formeel wordt onderscheiden van HEER in de betekenis van 'leger'. Een te verwaarlozen aantal weesrijmen is uit de teksten verwijderd, tenminste als er duidelijk sprake was van tekstverlies. Alle rijmwoorden en slechts de rijmwoorden zijn gelemmatiseerd: slechts die woorden die door de dichter gebruikt worden om tussen twee opeenvolgende verzen een rijm tot stand te brengen, kregen een lemma toegewezen.

In de praktijk gaat het meestal gewoon om het laatste woord van een regel, maar niet altijd. In het geval van proclisis bijvoorbeeld wordt het enclitische token meestal niet in de lemmatisering betrokken. Het versfinale token *tsweert* (in een rijm met *eert*) bijvoorbeeld, krijgt slechts het lemma ZWAARD en niet het lemma DAT+ZWAARD omdat het clitische *t-* geen rol speelt in de totstandkoming van het rijm en kan teruggaan op een kopiist. In principe kan een auteur oorspronkelijk immers *dat sweert* gebruikt hebben. In enkele gevallen kan toch sprake zijn van een combinatie van lemmata: in een stilistisch verfijnd rijmpaar als *vicaris* : *waer is* is ook het voorlaatste woord *waer* in het tweede vers cruciaal voor het rijm en wordt in dit geval het samengestelde lemma-label WAAR+ZIJN gebruikt.

(pro)clisis

In technische termen kregen dus alle woorden een lemma die de eerste beklemtoonde vocaal van een rijm bevatten, alsook alle woorden die daarop volgen. In drie gevallen is van deze annotatie afgeweken: om de inhoud van deze teksten van meet af aan zoveel mogelijk buiten spel te zetten, zijn drie soorten woorden afwijkend gecodeerd. Zij kregen geen echt inhoudelijk lemma maar eerder een vage aanduiding van hun woordsoort:

woordsoort

EIGENNAMEN – alle eigennamen (persoonsnamen, geografische aanduidingen, boektitels, ...) kregen het lemma PRNAME. Een uitzondering vormen enkele hoogfrequente (eigen)namen als GOD en CHRISTUS, die zo algemeen zijn in Middelnederlandse teksten dat ze amper met de inhoud van teksten samenhangen;

TELWOORDEN – alle hoofd- en rangtelwoorden (buiten 1, 2, 3 en afgeleiden) kregen het label NUMB. Deze betekenissen werden verbleekt omdat zij vaak inhoudsgevoelig zijn (bijvoorbeeld jaartallen) en omdat de numerieke inhoud die zij uitdrukken daarom in de regel minder van belang lijkt voor het stijleigen van auteurs;

BASTAARDWOORDEN – alle expliciet anderstalige woorden, zoals enkele Latijnse substantieven (bv. *deus*) of Franstalige woordgroe-

pen, kregen het lemma FOREIGN. Dat gold niet voor erg gangbare leenwoorden.

Een laatste uitzondering op de voorgaande regels betreft woordscheiding: in het Middelnederlands moeten sommige woorden gecombineerd worden om er een betekenisvol lemma aan te kunnen toekennen: de versfinale tokens *daer ave* worden daarom samengenomen onder de letterreeks *daer+ave* en die krijgt het lemma DAARAF. Daarbij werden zo goed mogelijk de richtlijnen gevolgd die op het INL werden gebruikt voor de annotatie van het *Corpus-Gysseling*.

Het corpus werd dus via software automatisch geannoteerd en werd nadien oppervlakkig gecorrigeerd om de meeste evidente fouten te verwijderen. Dit is slechts gebeurd door één persoon en aangezien het om een aanzienlijke hoeveelheid data gaat, kan niet worden gegarandeerd dat de annotatie geen fouten meer bevat. Over het algemeen is de annotatie-kwaliteit echter hoog genoeg om betrouwbaar te zijn (> 95% accuraatheid). Problematisch is natuurlijk dat verschillende lemmatiseringsfenomenen voor interpretatie vatbaar zijn. In het algemeen wordt er echter een zekere consistentie gegarandeerd door het feit dat de *lemmatiser* teruggaat op het *Corpus-Gysseling*, dat in grote mate consistent geannoteerd is. Als er fouten zijn gemaakt, moeten deze dus relatief consistent zijn gemaakt. Ook het feit dat de berekeningen hieronder teruggaan op hoogfrequente, en daarom meestal makkelijk te interpreteren fenomenen leidt ertoe dat er slechts in een enkel geval aanleiding tot discussie zal zijn. Onderstaande tabellen illustreren op welke wijze de teksten voor deze studie zijn verrijkt en bieden een weergave van de eerste tien verzen (van het eerste kapittel van het eerste boek) van respectievelijk Utenbroekes Tweede Partie (Tab. 8, p. 79) en Maerlants Derde Partie (Tab. 9, p. 80).

annotatie

4.1.3 *De Wetten van Zipf*

In de vorige hoofdstukken is toegelicht waarom stylometrici in het bijzonder geïnteresseerd zijn in de functoren van teksten. Deze grammaticale morfemen zouden onbewust worden gebruikt door auteurs en omdat zij zo vaak in verschillende teksten optreden, blijken zij vaak indicatief voor het auteurschap van teksten. In deze studie kan echter niet worden gewerkt met de functoren in Middelnederlandse teksten omdat deze sterk onderhevig zijn aan kopiistenvariatie in jongere afschriften van teksten. Afgaande op de vele suggesties in de secundaire literatuur is er daarom voor gekozen om in de plaats van met gewone woorden, met (gelemmatiseerde) rijmwoorden te werken. Deze woorden zijn immers veel stabieler in de overlevering van teksten en daarom principieel meer betrouwbaar voor auteursherkenning in Middelnederlandse teksten. De vraag rijst nu natuurlijk of de inzichten uit de internationale stylometrie wel zonder meer toepasbaar zijn op Middelnederlandse rijmwoorden.

Van Nerons goeden beghinne
(Tweede Partie, 1e boek, 1e kapittel, vs. 1-10, Filip Utenbroeke)

Oorspronkelijk vers	Lemma rijmwoord
in der ierster partien inden	EIND
mogedi van claudiusse vinden	VINDEN
hoe hi was keyser ende here groot	GROOT
dese heeft gedaen so vor sine doot	DOOD
dat nero keyser na hem blive	BLIJVEN
bi den rade van sinen wive	WIJF
ende der heren daer hi in desen	DEZE
te seer af scheen bedwongen wesen	WEZEN
octavien siere dochter man	MAN
was nero ende also dan	DAN

Tabel 8: Deze tabel illustreert op welke wijze de teksten voor deze studie zijn verrijkt en biedt een weergave van de gelemmatiseerde eerste tien verzen van het eerste kapittel van het eerste boek van Utenbroekes Tweede Partie.

In deze context is het cruciaal dat veel onderzoekers zich laatdunkend hebben uitgelaten over de rijmstijl van Middelnederlandse dichters (Van Driel 2003). Het is bijzonder vaak opgemerkt dat Middelnederlandse dichters maar weinig variatie hebben aangebracht in hun rijmwoordenschat: uitentreuren worden immers dezelfde rijmwoorden gebezigd. Dergelijke rijmwoorden komen vaak voor als 'noodrijmen' in stilistische constructies die in de vakliteratuur 'stoplappen' worden genoemd. Deze rijmwoorden zijn zelden semantisch geladen (Van Driel 2007, 19ff & 37ff) en hun nut schijnt meestal beperkt tot het makkelijk tegemoet komen aan de vormelijke eis van het gepaarde eindrijm in Middelnederlandse epiek (vgl. Besamusca 1991, 165). Stoplappen krijgen in de vakliteratuur dus meestal een 'functionele' interpretatie: zij dragen niet bij aan de inhoudelijke voortgang van een verhaal maar worden gebruikt als een louter vormelijke formule (bv. *sijt seker des*, *sijt seker*, *dats waer*, *dat seggic u*). Deze rijmconstructies dienen zo als 'kolom-vulsel' of *column fillers*: het gaat om weinigzeggende formulaire elementen waarmee het einde van de verskolom wordt opgevuld (Windelberg & Miller 1980).

noodrijmen

stoplappen

column fillers

Natuurlijk treffen wij mogelijk eerder zelden echte functoren (bv. lidwoorden) aan in rijmpositie, want de rijmwoordenschat in Middelnederlandse teksten wordt gedomineerd door inhoudswoorden zoals substantieven en werkwoorden. Toch kan men de vraag stellen of hoogfrequente rijmwoorden zich tot op zekere hoogte gedragen als functoren in moderne teksten. Daarbij kan men in de eerste plaats denken aan de typische noodrijmen in stoplapconstructies, maar ook andere rijmwoorden komen mogelijk in aanmerking. Ook hoogfrequente rijmwoorden als HEER, ZIJN of GROOT zijn misschien geen echte

Hoe Rome eerst dalen began

(Derde Partie, 1e boek, 1e kapittel, vs. 1-10, Jacob van Maerlant)

Oorspronkelijk vers	Lemma rijmwoord
nu helpet moeder ende maghet vrie	VRIJ
ic beginne die derde paertie	PARTIJ
vanden spieghele Ystoriale	PRNAME
ende hevet in bi ghetale	GETAL
viii bouke ende toghet al clare	KLAAR
van cccc ende xx jare	JAAR
wat overginc den roemscen rike	RIJK
dat teersten so mogendelike	MOGENDLIJK
al meest aldie werelt dwanc	DWINGEN
nu ginct dalen an sinen ganc	GANG

Tabel 9: Deze tabel illustreert op welke wijze het corpus is verrijkt en biedt een weergave van de gelemmatiseerde eerste tien verzen van het eerste kapittel van het eerste boek van Maerlants Derde Partie.

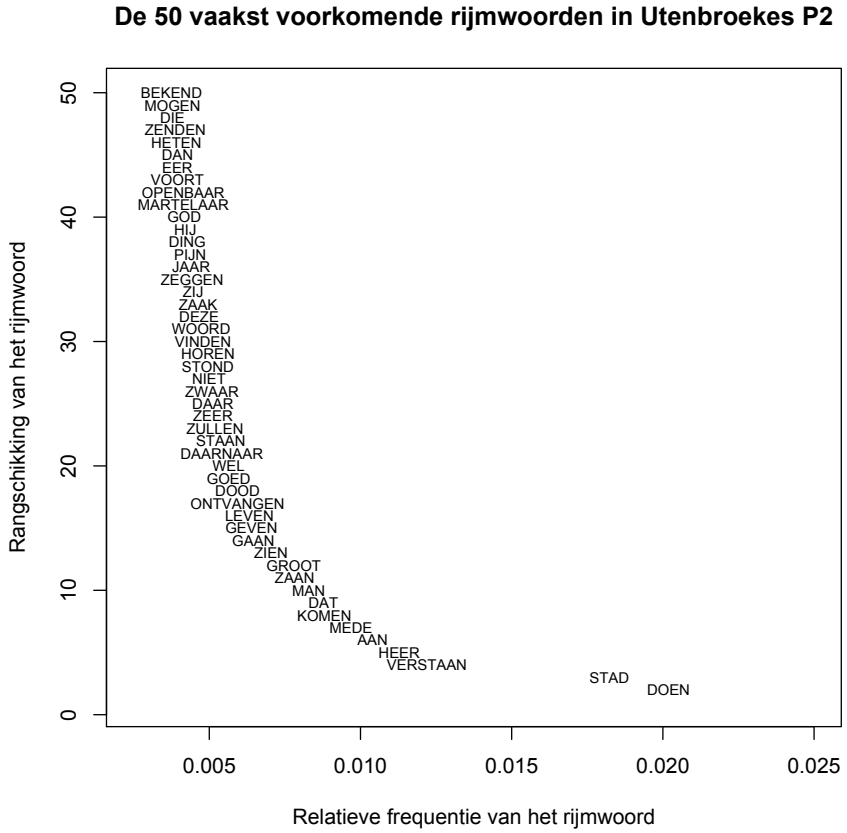
functoren vanuit linguïstisch opzicht, maar zij lijken toch niet rechtstreeks verbonden aan een specifieke tekstinhoud. Vanzelfsprekend hebben rijmwoorden als DOEN, KOMEN of GAAN ook in isolatie nog enige betekenis: zij drukken immers steevast een specifiek type handeling uit. Het is echter belangrijk dat deze handelingen zo algemeen zijn dat zij niet verraden in welke context zij geschiedden. Zowel in een passage over een heilige monnik als een rondetafelridder kunnen woorden met een dergelijk algemene betekenis immers optreden. Zo bekeken zijn deze hoogfrequente rijmwoorden niet noodzakelijk inhouds*loos*, maar wél betrekkelijk inhouds*onafhankelijk*.

Op het eerste zicht schijnen hoogfrequente rijmwoorden dus dezelfde voordelen als functoren te bieden: ze zijn functioneel en amper gebonden aan het taalgebruik van individuele dichters. Janssens (1988, 97) stelde bijvoorbeeld dat veel dichters putten uit eenzelfde poel van algemeen bekende rijmwoorden. Dat beperkte rijmregister noemde hij een 'pre-tekstuele potentie'. Bovendien zijn bepaalde rijmwoorden erg frequent in teksten – anders had men zich trouwens nooit eraan geërgerd – en mogelijk ook tamelijk inhoudsonafhankelijk. Als wij deze poëticale aspecten koppelen aan de stabiliteit van het rijmwoord in de overlevering, dan wordt deze woordcategorie erg aantrekkelijk voor auteursonderzoek. Via deze rijmwoorden kunnen we immers trachten tegemoet te komen aan Brinkmans Tweede Gebod, dat van de inhoudsonafhankelijkheid van de bestudeerde woorden.

Tweede Gebod: inhoudsonafhankelijkheid

Laat ons verkennen of het mogelijk is om stylometrische technieken ontwikkeld voor moderne teksten toe te passen op de hoogfrequente rijmwoorden in Middelnederlandse teksten zoals de *Spiegel historiael*. Onderstaande figuren geven weer welke 50 rijmwoorden cumulatief gesproken het meest frequent zijn in de Tweede Partie (Fig. 7, p. 81)

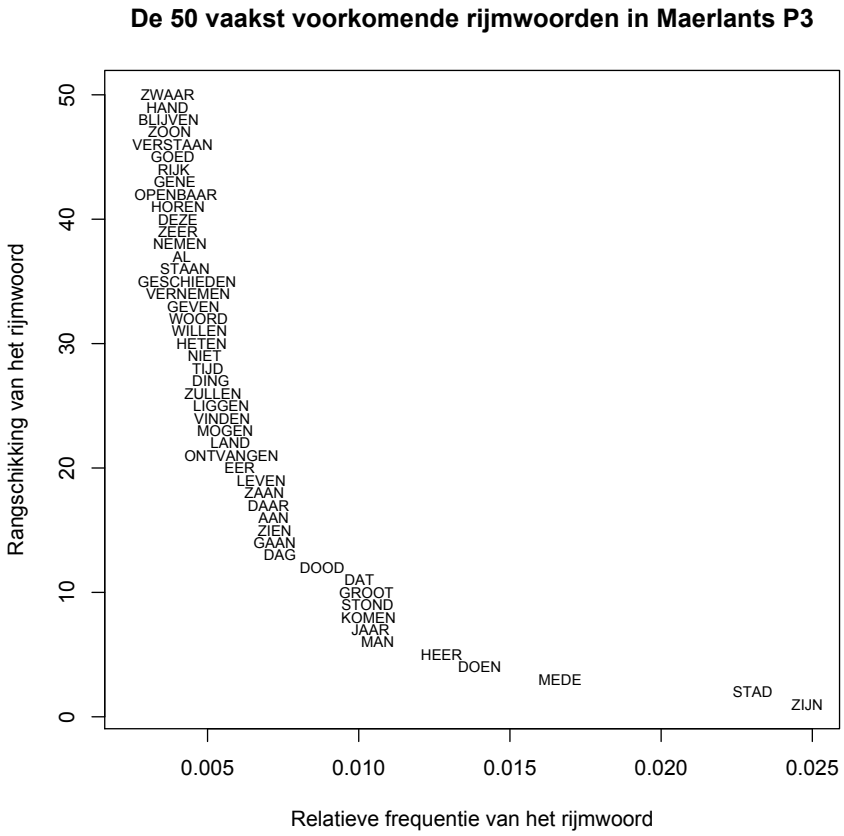
en Derde Partie (Fig. 8, p. 82), respectievelijk het aandeel van Utenbroeke en van Maerlant. Hoewel slechts een deel van de volledige rijmwoordenschat in deze teksten is afgebeeld, is nu reeds duidelijk dat beide figuren een gelijkaardige kromme vertonen. Slechts een klein aantal rijmwoorden komt heel frequent voor in deze teksten en deze hoogfrequente rijmwoorden kennen bijgevolg een erg beperkte semantische lading. Een veel groter aantal rijmwoorden (met een relatieve frequentie ver onder 0,005) komt bijgevolg veel minder vaak voor en valt buiten deze afbeeldingen.



Figuur 7: Een grafische voorstelling van de vijftig meest frequente rijmwoorden in Utenbroekes Tweede Partie. (Het meest frequente rijmwoord (PRNAME) is niet meegenomen in deze analyse.) De rijmlexemen zijn verticaal (y-as) gerangschikt op basis van hun relatieve frequentie (x-as). Slechts het frequentie-interval 0,0025–0,025 is op de x-as weergegeven zodat het bij Utenbroeke erg frequente rijmlexeem ZIJN buiten de grafiek valt.

In linguïstische termen lijkt zo sprake van een 'Zipf-achtige kromme'. Zipf (1949) was een twintigste-eeuwse taalkundige die fundamenteel onderzoek heeft uitgevoerd naar de regelmatigheden inzake de frequenties van woorden in corpora (zie ook § 6.2). Hij heeft een aantal wetten geformuleerd die uitdrukken hoe slechts een relatief klein aantal items in een taal (bv. functoren) een relatief groot aandeel in

de wetten van Zipf



Figuur 8: Een grafische voorstelling van de vijftig meest frequente rijmwoorden in Maerlants Derde Partie. (Het meest frequente rijmwoord (PrName) is niet meegenomen in deze analyse.) De rijmlexemen zijn verticaal (y-as) gerangschikt op basis van hun relatieve frequentie (x-as).

de totale frequentie-massa van teksten vertegenwoordigen. Hoewel daar hieronder nog dieper wordt op ingegaan (zie § 6.2), zij nu reeds opgemerkt dat de distributie van deze rijmwoorden eveneens aan Zipf-achtige regelmatigheden schijnt te gehoorzamen.

Deze voorstellingen lijken opnieuw de hypothese te rechtvaardigen dat stylometrische technieken kunnen worden toegepast op Middelnederlandse hoogfrequente rijmwoorden, mogelijk op een wijze die analoog is aan de manier waarop deze technieken worden toegepast op reguliere functoren in tekst. Als we bijvoorbeeld alle rijmwoorden uit de Tweede en Derde Partie extraheren en deze vervolgens rangschikken op basis van hun cumulatieve frequentie blijken volgende 50 rijmwoorden het vaakst voor te komen:

1 – PrName, 2 – zijn, 3 – stad, 4 – doen, 5 – mede, 6 – heer, 7 – komen, 8 – man, 9 – dat, 10 – groot, 11 – aan, 12 – verstaan, 13 – stond, 14 – zaan, 15 – dood, 16 – jaar, 17 – zien, 18 – gaan, 19 – leven, 20 – daar, 21 –

ONTVANGEN, 22 – GEVEN, 23 – ZULLEN, 24 – VINDEN, 25 – DAG, 26 – EER, 27 – NIET, 28 – STAAN, 29 – GOED, 30 – WOORD, 31 – DING, 32 – MOGEN, 33 – ZEER, 34 – WEL, 35 – HOREN, 36 – ZWAAR, 37 – DEZE, 38 – HETEN, 39 – DAARNAAR, 40 – LAND, 41 – WILLEN, 42 – OPENBAAR, 43 – TIJD, 44 – VERNEMEN, 45 – ZAAK, 46 – GESCHIEDEN, 47 – NEMEN, 48 – PIJN, 49 – AL, 50 – WEZEN

De meest frequente rijmwoordenschat van Utenbroeke vertoont alvast geen frappante verschillen met die van Maerlant. Dat bevestigt op het eerste zicht de hypothese dat beide dichters zich inderdaad van een gelijkaardige poel aan rijmwoorden bedienen. Bovendien verklaart dit waarom de samenstelling van bovenstaande lijst van 50 rijmwoorden ook geen noemenswaardige afwijkingen vertoont in vergelijking met de hoogfrequente rijmwoordenschat van de individuele auteurs. Merk op dat verschillende van de rijmwoorden in bovenstaande lijsten en grafieken geassocieerd kunnen worden met typische stoplapconstructies. De rijmwoorden VERSTAAN of VERNEMEN zullen bijvoorbeeld vaak voorkomen in formulaire constructies waarin een auteur zijn bronnen toelicht, eventueel om het waarheidsgehalte van zijn tekst te benadrukken – bv. *also ict versta* of *also ict vernam*. Ook ZAAN of MEDE zijn rijmwoorden die traditioneel vaak als noodrijmen beschouwd worden, onder meer omdat zij in onze ogen moeiteloos uit een tekst verwijderd kunnen worden zonder de betekenis van een passage fundamenteel te wijzigen. De in deze studie onderzochte rijmwoorden zullen daarom in verschillende gevallen geassocieerd zijn met constructies die men in het verleden noodrijmen of stoplappen heeft genoemd.

Het merendeel van de hoogfrequente rijmwoorden die hier zijn afgebeeld, zijn echter geen items die men in de vakliteratuur met de stilistische categorieën van stoplappen of noodrijmen zou verbinden. Een goed voorbeeld zijn de werkwoorden DOEN, KOMEN of GAAN. Binnen een louter interpretatieve benadering die geen rekening houdt met de frequentie van woorden is er immers geen reden om deze inhoudswoorden met formulair taalgebruik te verbinden. Pas wanneer men de daadwerkelijke frequentie van deze woorden bekijkt, valt op hoe vaak zij eigenlijk voorkomen. Deze hoge frequentie rechtvaardigt de hypothese dat ook deze rijmwoorden niet aan een specifieke inhoud zijn verbonden.

Deze studie zal daarom op twee belangrijke punten afwijken van het voorgaand stilistisch onderzoek naar formulair taalgebruik in de medioneerlandistiek. Ten eerste zal hieronder *enkel* met rijmwoorden worden gerekend. In veel gevallen zullen de geselecteerde hoogfrequente rijmwoorden samenhangen met het optreden van stilistische constructies, zoals stoplappen die meer woorden in het vers kunnen dekken dan enkel het rijmwoord. In deze studie zal ik echter enkel de rijmwoorden uit dergelijke constructies beschouwen, vanwege het hier-

boven toegelichte 'dogma' dat het rijmwoord en *slechts* het rijmwoord in Middelnederlandse teksten beschermd was tegen kopiistenvariatie.

Ten tweede is het selectiecriterium voor rijmwoorden dat ik hieronder zal aanwenden zuiver gebaseerd op de frequentie van rijmwoorden en niet op de menselijke interpretatie van hun betekenis. Dat betekent dat ook veel rijmwoorden in het onderzoek zullen worden betrokken die traditioneel niet als noodrijmen of stoplappen worden beschouwd. Een dergelijke aanpak kan men data-gedreven (*data-driven*) noemen, omdat deze vertrekt van een objectief frequentiecriterium dat gestoeld is op de kwantitatieve eigenschappen van de teksten zelf en niet van de subjectieve menselijke interpretatie die men aan rijmwoorden oplegt. Natuurlijk is daarmee nog niet aangetoond dat deze benadering ook daadwerkelijk effectief is voor auteursherkenning in Middelnederlandse rijmteksten. Deze aanpak zal daarom in de rest van dit hoofdstuk worden getoetst.

data-driven

4.2 CIJFEREN VOOR GEVORDERDEN

In het vorige hoofdstuk is reeds een techniek voor stylometrische auteursherkenning geïntroduceerd, namelijk *Principal components analysis* (vgl. § 3.3). Dit is echter zeker niet de enige techniek die in de hedendaagse stylometrie wordt aangewend. In dit hoofdstuk introduceer ik bondig het paradigma van de tekstclassificatie. Vervolgens wordt beschreven hoe een eenvoudige classificatietechniek kan helpen bij het toetsen van de hypothese dat hoogfrequente Middelnederlandse rijmwoorden een geschikte uitvalsbasis vormen voor stijlgebaseerde auteursherkenning.

4.2.1 *Tekstclassificatie*

In de computationele filologie wordt de studie van auteurschap vaak betrekkelijk nuchter opgevat als een vorm van tekstclassificatie: een bepaalde hoeveelheid tekst moet geclassificeerd worden of een label krijgen dat uitdrukt wie de auteur ervan is (Stamatatos et al. 2000, 472). Tekstclassificatie is zeker niet de enige methode in de hedendaagse stylometrie, maar wel een dominante. Tekstclassificatie kent momenteel veel toepassingen, ook buiten de stijlstudie (Sebastiani 2002). Het bekendst is wellicht *spam filtering*, de geautomatiseerde taak waarbij een computer aan een e-mailbericht een label toekent (*spam* of *geen spam*) en op basis daarvan het bericht verder verwerkt (verwijderen, doorsturen, de gebruiker alarmeren, ...). In welke klasse een nieuwe tekst moet worden ondergebracht, wordt beslist door een *classifier*, een software-toepassing die erop getraind is om labels aan dergelijke teksten toe te kennen.[12]

tekstclassificatie

spam filtering

12 Een introductie tot de algemene gegevens in deze alinea wordt bijvoorbeeld geboden door Alpaydin (2004), Daelemans & Van den Bosch (2005) of Sebastiani (2002).

De notie van het trainen is ontleend aan het wetenschapsveld *Machine Learning*, een subdomein van de Artificiële Intelligentie. Hier veronderstelt men dat een van de belangrijkste kenmerken van natuurlijk intelligente wezens de mogelijkheid is om te leren: dat wil zeggen, om op basis van vroegere ervaringen kennis op te doen die het wezen in staat stelt zijn toekomstig gedrag te optimaliseren. Een kind dat zich eenmaal aan een lucifer heeft verbrand, zal de volgende keer tweemaal nadenken als het met een lucifer speelt door de gelijkenis van de tweede met de eerste lucifer. Voorbeeld- of trainingsmateriaal speelt in Machine Learning dan ook een grote rol: de software ('het kind') wordt met een reeks voorbeelden geconfronteerd ('brandende lucifer', 'ijsje', 'niet aangestoken lucifer', ...) die van een label zijn voorzien ('gevaarlijk' of 'niet gevaarlijk'). De bedoeling is de software aan het verstand te brengen welke reactie in een nieuwe situatie het meest gepast zou zijn ('gevaarlijk' > niet aanraken, 'niet gevaarlijk' > aanraken).

Machine Learning

Artificiële Intelligentie

Een dergelijke aanpak kan ook worden toegepast op tekst: in het voorbeeld van de *spam filtering* kan een classifier de kennis opdoen dat een buitensporige hoeveelheid schuttingtaal of het voorkomen van het woord *lottery* meer dan wel minder dwingend vraagt om het label *spam* (Sebastiani 2002, 7). De representatie van de voorbeeld-mails (de trainingsinstanties) is de sleutel tot succes. De bedoeling is dat om het even welke e-mail kan worden voorgesteld aan de hand van een beperkt aantal kenmerken die relevant kunnen zijn voor de classificatie. In tekstclassificatie – bijvoorbeeld het automatisch onderbrengen van nieuwsitems in de categorie *sport* of *economie* – wordt vaak gewerkt met een tekstrepresentatie onder een *term-frequency vector space model*. Een voorbeeld van een dergelijke voorstelling wordt hieronder geboden (Tab. 10, p. 86). De bedoeling is natuurlijk dat een classifier op basis van een software-gestuurd leerproces kennis kan opbouwen over de manier waarop het best een nieuwe, ongeziene instantie wordt gelabeld. Een voorbeeld van een dergelijke leermethode wordt hieronder in detail besproken.

trainingsinstanties

Ook in het onderzoek naar auteurs wordt vaak tekstclassificatie toegepast. Men gaat bijvoorbeeld na of op basis van trainingsmateriaal in de vorm van het oeuvre van een gekende auteur voorspeld kan worden of andere anonieme teksten ook aan deze auteur kunnen worden toegeschreven. In dergelijk onderzoek wordt zelden gewerkt met teksten van een betwiste signatuur. Men wil namelijk eerst zeker zijn dat een methode werkt, vooraleer men uitwijkt naar teksten waarvan het auteurschap niet zeker is. Om de bruikbaarheid van een classifier te testen, is in Machine Learning een interessante evaluatiemethode bedacht (Alpaydin 2004, 327ff). Het probleem met teksten van een onbekende signatuur is dat een classifier getraind op voorbeelden van het werk van gekende auteurs wel een uitspraak zal doen over het auteurschap van de onbekende tekst, maar dat men de correctheid

trainingsmateriaal

evaluatiemethode

	loterij	X?x!	tijdschrift	**Klasse**
Voorbeeld 1	Ja	Nee	Nee	SPAM
Voorbeeld 2	Nee	Nee	Ja	GEEN SPAM
Voorbeeld 3	Nee	Nee	Nee	GEEN SPAM
Voorbeeld 4	Ja	Ja	Nee	SPAM

Tabel 10: Deze tabel biedt een fictief voorbeeld van hoe e-mailberichten kunnen worden voorgesteld in een classificatiesysteem voor *spam filtering*. Iedere rij bevat een e-mail (1 tot 4), voorgesteld aan de hand van de aan- of afwezigheid van drie specifieke woorden (de kolommen 2 tot 4). Uiterst rechts is de klasse van elk bericht weergegeven (SPAM of GEEN SPAM). Op basis van een dergelijke voorstelling kan een classifier leren wat typerend is voor het onderscheid tussen SPAM- en GEEN SPAM-berichten. Als het woord 'loterij' in een e-mailbericht voorkomt, zal de classifier het nieuwe bericht waarschijnlijk als SPAM beschouwen.

van die uitspraak niet kan evalueren. Om toch een zicht te krijgen op de bruikbaarheid van de classifier wordt als volgt gewerkt.

Stel dat men wil onderscheiden tussen twee auteurs en dat van elk van beide auteurs tien romans beschikbaar zijn. Om het nut van de classifier te testen werkt men met evaluatiemethodes als *leave one out validation* (Daelemans & Van den Bosch 2005, 47-48). In deze methode wordt telkens één van de romans uit het beschikbare materiaal gelicht als test case. Vervolgens wordt de classifier getraind op de negentien overgebleven romans, waarna de classifier wordt gevraagd aan wie van beide auteurs hij het nog 'anonieme' testexemplaar zou toeschrijven. Dit proces van 'kruisvalidering' wordt twintig keer herhaald, voor elk van de beschikbare items, waarbij steeds wordt gesimuleerd dat één van de romans van onbekende signatuur is. Na twintig rondjes heeft men zo een goed zicht op de accuraatheid van de classifier, zodat men ook een inschatting kan maken van de betrouwbaarheid van de classifier als die een label moet toekennen aan een echt werk van betwiste signatuur.

leave one out validation

Een dergelijke methode wordt momenteel druk toegepast op auteursonderscheiding. Belangrijk aan dit soort onderzoek is dat men met dergelijke simulaties inzichten kan opdoen over het specifieke van een schrijfstijl. Zo kan blijken dat in het geval van de twee auteurs hierboven een bepaald kenmerk A de *leave one out test* erg goed doorstaat maar een ander kenmerk B helemaal niet. Het ligt dan voor de hand dat kenmerk A typerender is voor het onderscheid tussen de stylomen van beide auteurs dan kenmerk B. Zeker in dergelijke experimenten is bijvoorbeeld gebleken dat functoren het erg goed doen in het onderscheiden van de schrijfstijl van verschillende auteurs (vgl. § 2.1).

simulaties

4.2.2 *Geheugen-gebaseerd leren*

Laat ons nu de Tweede en Derde Partie benaderen vanuit de tekstclassificatie met een *leave one out evaluation*. In dit scenario wordt dus vertrokken vanuit een aantal voorbeeldtekstjes of samples per auteur. Telkens wordt één sample uit het geheel gelicht en aan de kant gelegd. Vervolgens traint men een classifier op de overgebleven samples en wordt er daarna vastgesteld of het tot dan ongeziene sample aan de correcte auteur wordt toegeschreven. De classifier kent de auteur van het ongeziene sample vanzelfsprekend niet op voorhand maar wij wel. De mate waarin een bepaald experiment succesvol of 'accuraat' is (Daelemans & Van den Bosch 2005, 48-51), wordt bepaald door het aantal correcte toeschrijvingen. De accuraatheid (*accuracy*) van een experiment is gelijk aan het aantal correct toegeschreven samples gedeeld door het totale aantal samples.

accuracy, accuraatheid

Het rijmgebruik van een dichter in een sample zullen we hieronder voorstellen onder een *term-frequency vector space model*, waarbij we onze representatie beperken tot de relatieve frequenties van een relatief klein aantal hoogfrequente rijmwoorden. Ook hier is natuurlijk ruimte voor experimenteren: men kan zich beperken tot bijvoorbeeld de 10 meest frequente rijmwoorden, maar ook tot de 125 meest frequente. Een te klein aantal rijmwoorden kan leiden tot een al te beperkte vertegenwoordiging van de samples in kwestie. Te veel lager frequente rijmwoorden betrekken kan eveneens nefaste gevolgen hebben voor de classificatie. Als rijmwoorden minder frequent zijn kan dat immers betekenen dat zij gebonden zijn aan een specifieke tekstinhoud, zodat de classifier het gevaar loopt te *overfitten* op een specifiek onderwerp.

Cruciaal voor deze experimenten is natuurlijk de classifier. Omwille van de duidelijkheid wordt hier gewerkt met een erg intuïtieve classificatiemethode. In *memory-based learning* wordt de nadruk gelegd op het 'geheugen' van het lerende systeem.[13] Een classifier zal de voorbeelden die worden aangereikt in de trainingsfase gewoon opslaan in een grote tabel in het geheugen, zonder daarbij in principe een onderscheid te maken tussen meer of minder relevante voorbeeld-instanties. De kracht van het algoritme zit in de classificatiefase. Als de classifier een nieuw ongezien sample krijgt aangeboden, inspecteert hij de tabelwaarden van het ongeziene sample en gaat vervolgens op zoek naar de voorbeeldinstantie in zijn geheugen die het meest lijkt op het nieuwe sample (Daelemans & Van den Bosch 2005, 29ff). De classifier kent aan het nieuwe sample de klasse toe van het meest gelijkaardige sample in zijn geheugen. Dit is de *nearest neighbour*-methode: de klasse van

memory-based learning

nearest neighbour

13 De software die voor het hier beschreven onderzoek werd gebruikt, is TIMBL (*Tilburg Memory-Based Learner*, een gratis en goed gedocumenteerd software-pakket, dat vrijelijk is te downloaden op `http://ilk.uvt.nl/`. De software gaat vergezeld van een erg inzichtelijke handleiding bij de programmatuur en inleiding tot het geheugengebaseerd leren (Daelemans et al. 2010b). Wie meer wil lezen, kan terecht bij het boek *Memory-Based Language Processing* (Daelemans & Van den Bosch 2005).

een ongezien sample wordt bepaald op basis van de *nearest neighbour* in het geheugen van het systeem.

Belangrijk is hoe wordt bepaald hoe ongelijkend samples zijn: dat gebeurt via de berekening van de afstand (*distance*) tussen beide (Daelemans & Van den Bosch 2005, 28). Een dergelijke afstandsmaat kan in deze experimenten heel makkelijk worden geïmplementeerd, aangezien met numerieke, continue waarden wordt gewerkt: de relatieve frequentie van hoogfrequente rijmwoorden (een eerder klein decimaal getal tussen 0 en 1, vgl. Tab. 6, p. 55). Als twee samples worden vergeleken, wordt het rijtje rijmwoorden afgegaan en berekent de classifier het absolute verschil voor de frequentie van ieder rijmwoord in beide samples.[14] Vervolgens worden deze verschillen gewoon bij elkaar geteld. De resulterende waarde, de zogenaamde *Manhattan distance*, zal hoog zijn in gevallen van weinig gelijkende samples maar klein in het geval van sterk op elkaar gelijkende samples. Het sample in het geheugen op een minimale afstand van het ongeziene sample zal als *nearest neighbour* worden aangewezen. De classifier zal dus voorspellen dat aan het ongeziene sample dezelfde klasse moet worden toegewezen als die van de *nearest neighbour*.

distance

Manhattan distance

In de onderstaande tabel wordt weergegeven hoe de trainingssamples in het geheugen worden voorgesteld aan de hand van de relatieve frequentie van rijmwoorden, hier de vijf meest frequente (Tab. 11, p. 89). Aan deze tabel is toegevoegd welk sample voor elk ander sample de *nearest neighbour* zou zijn tijdens *leave one out*-validatie en wat dan de afstand tussen beide samples is. Meteen wordt duidelijk hoe krachtig deze ogenschijnlijk eenvoudige classificatiewijze is: zelfs als men slechts de vijf meest frequente rijmwoorden in beschouwing neemt, wordt slechts één boek (P2B7) aan een verkeerde auteur toegewezen (Maerlant in plaats van Utenbroeke).[15] De auteur van de 13 andere boeken wordt wel correct herkend, wat voor dit experiment een gemiddelde accuraatheid van ca. 92% oplevert (13 van de 14 toeschrijvingen correct).

4.2.3 *Spiegel-experimenten*

Hieronder kunnen we experimenteren met de wijze van sampling. Hoe men het materiaal aan de classifier voert, is immers van groot belang. Voor de *Spiegel historiael*-casus zal ik aanvankelijk werken met twee

14 Hieronder wordt een voorbeeld gegeven van deze berekeningswijze met concrete getallen (Daelemans & Van den Bosch 2005, 28-29). Let wel: TIMBL zal de verschillen per feature-kolom 'schalen' tussen de hoogste en laagste waarde binnen één kolom. Wie de distances in deze tabel zelf wil narekenen, moet zich natuurlijk rekenschap geven van deze manier van schalen. Deze berekeningswijze wordt verder uitgediept in § 7.2.

15 Merk op dat in deze setting geen sprake hoeft te zijn van 'transitiviteit': als sample *b* de *nearest neighbour* is van sample *a*, hoeft sample *a* niet noodzakelijk de *nearest neighbour* van sample *b* te zijn, doordat gewerkt wordt met *leave one out*-validatie.

PrName	zijn	stad	doen	mede	**Boek**	**NN**	**afstand NN**
0,0498	0,0298	0,0139	0,0202	0,0058	P2B1	P2B6	0,9626
0,0632	0,0269	0,0178	0,0211	0,0076	P2B2	P2B3	0,7658
0,0651	0,0321	0,0193	0,0183	0,0071	P2B3	P2B6	0,6373
0,0678	0,0304	0,0212	0,0230	0,0125	P2B4	P2B2	0,1779
0,0458	0,0340	0,0200	0,0174	0,0079	P2B5	P2B3	0,6373
0,0531	0,0344	0,0158	0,0198	0,0097	P2B6	P2B5	0,8112
0,0194	0,0259	0,0132	0,0142	0,0111	**P2B7**	**P3B3**	0,5991
0,0781	0,0220	0,0146	0,0119	0,0152	P3B1	P3B6	0,1949
0,0359	0,0248	0,0204	0,0110	0,0169	P3B2	P3B3	0,4389
0,0362	0,0257	0,0174	0,0091	0,0170	P3B3	P3B2	0,4610
0,0454	0,0209	0,0260	0,0149	0,0160	P3B4	B3B7	0,4322
0,0812	0,0273	0,0220	0,0141	0,0137	P3B5	P3B6	0,9095
0,0842	0,0206	0,0252	0,0142	0,0151	P3B6	P3B4	0,8482
0,0430	0,0220	0,0233	0,0158	0,0158	P3B7	P3B4	0,4167

Tabel 11: Illustratie van het *nearest neighbour*-algoritme, toegepast op de boeken van de Tweede Partie en Derde Partie. Elke rij stelt een volledig boek voor aan de hand van de relatieve frequentie van de vijf meest frequente rijmwoorden (kolommen 1 tot 5). Onder 'Boek' wordt gespecifieerd om welk boek het in de rij gaat. 'P3B6' slaat bijvoorbeeld op het zesde boek ('B6') van de Derde Partie ('P3'). Onder 'NN' wordt voor elk boek aangeduid welk ander boek uit het trainingsmateriaal er het meest op lijkt tijdens *leave one out*-validatie. Het test-sample wordt toegeschreven aan de auteur van de *nearest neighbour*. De attributie is slechts één keer fout (P2B7). De mate van gelijkenis wordt bepaald door de *distance* (kolom uiterst rechts). (Getallen weergegeven tot op vier cijfers na de komma.)

wijzen van sampling: op boek-niveau en op kapittel-niveau. Op boek-niveau bestaan de aangeboden samples uit hele boeken per auteur: 7 van Utenbroeke en 7 van Maerlant of 14 in totaal. Op kapittel-niveau is het kapittel de eenheid van vergelijking en wordt gewerkt met samples die een (variabel) aantal kapittels bevatten. Zo kan flexibel gewerkt worden met ofwel een groot aantal samples die een klein aantal kapittels bevatten, ofwel een klein aantal samples die een groot aantal kapittels bevatten. Men wil immers niet enkel achterhalen of beide auteurs te onderscheiden zijn, maar ook hoeveel tekst daarvoor nodig is. Het valt te verwachten dat uit een klein stuk tekst als één kapittel geen significant rijmprofiel valt te distilleren maar misschien wel uit een groep van 8 kapittels.

Let wel: in dit hoofdstuk zal dus worden gewerkt met samengestelde samples die niet exact evenveel rijmwoorden tellen zijn, aangezien het aantal verzen per kapittel kan variëren. (Een kapittel in het *Spiegel*-corpus telt gemiddeld 86 verzen.) Het voordeel van deze aanpak is dat

we kunnen werken met intuïtieve, literaire eenheden ('het kapittel', 'het boek') waarvan we precies kunnen zeggen waar zij voorkomen in de tekst. Zoals nog verderop in deze studie zal blijken, wordt dat laatste veel moeilijker als men teksten ophakt in abstracte reeksen van rijmwoorden, die niet meer corresponderen met de klassieke, literair-historische bestanddelen van teksten. In de secundaire literatuur wordt trouwens nog fel bediscussieerd of het methodologisch wel is aangewezen om teksten met een ongelijke omvang voor stylometrische doeleinden te vergelijken. Sommige studies stellen dat men slechts samples van exact dezelfde grootte mag vergelijken (cf. Brinkmans 4e Gebod), terwijl een andere reeks studies – de zogenaamde *profile-based approaches* – eigenlijk het tegenovergestelde claimt (Stamatatos 2009, 546-548). Daarom is het hoe dan ook van bijzonder groot belang dat we onze samples hieronder vergelijken op basis van de *relatieve* frequenties van rijmwoorden.

Zoals gezegd kunnen we ook het aantal rijmwoorden in onze *Spiegel*-experimenten variëren. Onderstaande tabel geeft experimenten weer op boekniveau waarbij stelselmatig het aantal hoogfrequente woorden in het experiment wordt opgevoerd (Tab. 12, p. 91). Belangrijk is het nulexperiment: de *baseline* van een experiment wordt gevormd door een soort domme classifier (Daelemans & Van den Bosch 2005, 51). Als men in een experiment steeds Utenbroeke als label zou kiezen, zou de accuraatheid van het experiment 50% zijn. Aangezien de helft van de samples in het experiment van Utenbroeke is, zou de domme toeschrijving in de helft van de gevallen toch correct zijn. Wil de aanpak een meerwaarde bieden, moet de accuraatheid natuurlijk gevoelig hoger liggen dan de baseline (50%). Dat is op boekniveau meteen het geval: al vanaf één enkel woord (het relatieve voorkomen van eigennamen (PrName) in een boek) heeft het algoritme genoeg om beter dan kans te presteren (78%). Vanaf 8 rijmwoorden blijkt de classificatie foutloos (100%) en wordt de auteur feilloos herkend.

baseline

Hier werken we op boekniveau met een weliswaar klein aantal samples maar wel samples die een grote hap tekst vertegenwoordigen en waaruit dus relatief makkelijk algemene tendensen kunnen worden opgemaakt. Hoewel het met dergelijke grote tekstblokken dus relatief makkelijk blijkt twee auteurs te onderscheiden, is het verstandig meteen deze aanpak te problematiseren. Voor veel fragmentarisch overgeleverde teksten beschikt men immers niet over dergelijke grote hoeveelheden materiaal. Een interessante vraag is daarom hoe groot een sample moet zijn vooraleer het kan toegewezen worden aan de juiste auteur.[16] Onderstaande tabel beschrijft een experiment waarbij de sample-grootte stelselmatig artificieel wordt opgevoerd (Tab. 13, p. 96). Het vertrekpunt is 732 samples (366 per auteur) die elk één ka-

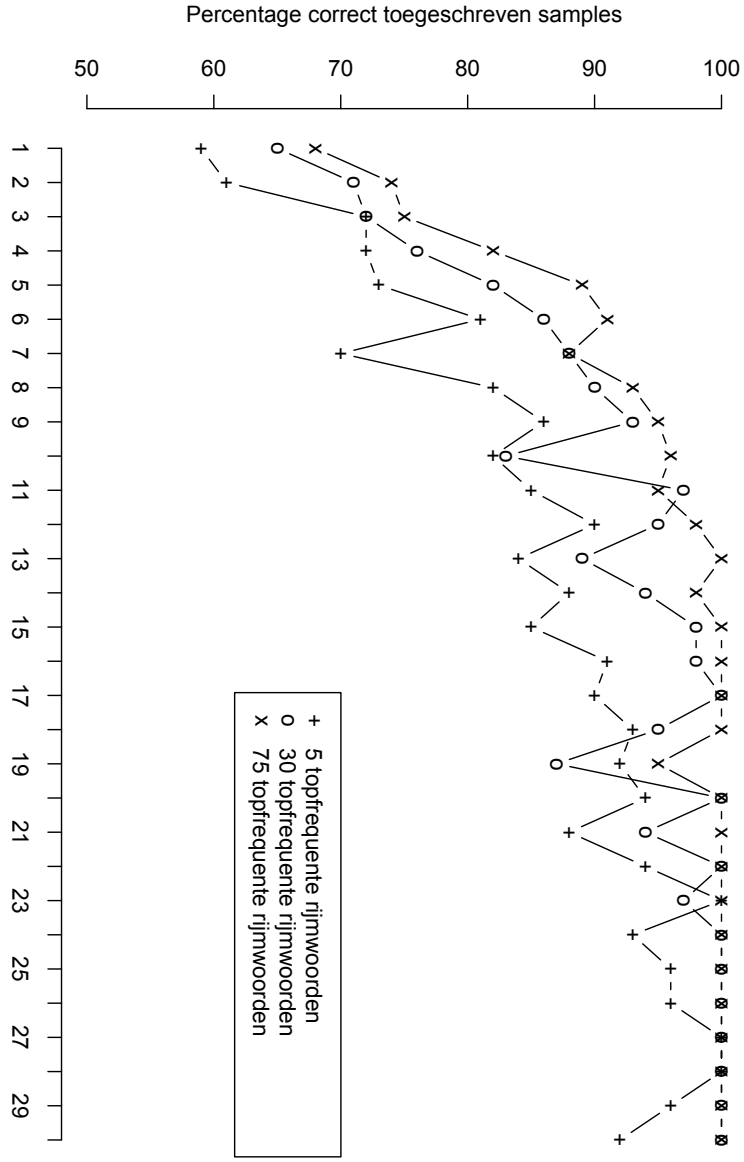
16 Vergelijkbaar zijn de experimenten met data-grootte bij bv. Luyckx & Daelemans (2011). Let wel: hier wordt geëxperimenteerd met de grootte van de samples en niet met het aantal beschikbare samples (hoeveel trainingsmateriaal er voor een bepaald auteur voorhanden is).

Aantal rijmwoorden	Toegevoegd rijmwoord	Accuraatheid (%)
0	-	50% (= baseline)
1	+PRNAME	78
2	+ZIJN	78
3	+STAD	78
4	+DOEN	92
5	+MEDE	92
6	+HEER	92
7	+KOMEN	92
8	+MAN	100
9	+DAT	100
10	+GROOT	100
11	+AAN	100
12	+VERSTAAN	100
13	+STOND	100
14	+ZAAN	100
15	+DOOD	100

Tabel 12: Resultaten van een *leave one out*-experiment op boekniveau. De werkwijze is exact dezelfde als in de vorige tabel, alleen wordt hier gewerkt met een variabel aantal hoogfrequente rijmwoorden. Stelselmatig wordt het aantal betrokken rijmwoorden in de analyse opgevoerd, beginnend bij geen rijmwoorden (de *baseline*) tot de vijftien meest frequente rijmwoorden. Meer rijmwoorden betrekken, leidt tot een grotere accuraatheid (het aantal correct toegeschreven boeken, weergegeven als een percentage zonder decimalen in de kolom uiterst rechts). Reeds vanaf acht rijmwoorden blijkt de attributie op boekniveau in dit experiment foutloos.

pittel bevatten. Vervolgens 366 samples met elk twee kapittels, daarna 244 samples van drie kapittels, enzovoorts. De accuraatheid (aantal correct toegeschreven samples) wordt weergegeven in functie van het aantal hoogfrequente rijmwoorden dat wordt meegenomen in ieder experiment. Onderstaande figuur biedt een grafische voorstelling van de curves voor de experimenten met respectievelijk 5, 30 en 75 hoogfrequente rijmwoorden (Fig. 9, p. 92).

Deze experimenten tonen dat de classifier aan een sample van slechts enkele kapittels weinig heeft: zeker tot een sample-grootte van tien kapittels (ca. 860 rijmwoorden) is de gemiddelde accuraatheid van de meeste toeschrijvingen eerder laag (< 90%). Toch presteert de classifier van meet af aan boven het kansniveau (> 50%), want geen enkele van de resultaten gaat onder de baseline. Hoewel de scores initieel soms slechts marginaal boven de baseline zitten, toont dit aan dat er toch reeds enige stilistische regelmaat wordt opgepikt, ook bij een relatief kleine sample-grootte. De accuraatheid stijgt vervolgens



Figuur 9: Deze grafiek bevat een grafische voorstelling van de waarden uit Tab. 13 (p. 96) voor de experimenten met respectievelijk 5, 30 en 75 hoogfrequente rijmwoorden (zie legende). Het percentage correcte toeschrijvingen wordt op de y-as weergegeven in functie van de samplegrootte (in aantal kapittels) op de x-as.

gevoelig als ook het aantal kapittels dat een sample voorstelt, wordt opgedreven: zoals bijvoorbeeld af te lezen uit Fig. 9 (p. 92) vertonen de leercurves een stijgende tendens, naarmate ook de sample-grootte toeneemt. Afgezien van enkele uitschieters blijkt de classificatie pas vanaf twintig kapittels (gemiddeld ca. 1740 verzen) in de meeste experimenten nauwkeurig (95%-100%), al komen foute toeschrijvingen ook nog bij grotere samples voor.

leercurve

Een andere belangrijke parameter is natuurlijk het aantal rijmwoorden dat in de analyse wordt betrokken. Uit deze experimenten blijkt dat de accuraatheid van de auteursherkenning in de experimenten hoger is, indien ook meer hoogfrequente rijmwoorden worden beschouwd: de top 75 doet het consequent beter dan de top 50, die het op zijn beurt consequent beter doet dan de top 30 enzovoorts. Niettemin is het interessant dat de classifier ook met een relatief klein aantal rijmwoorden bij voldoende grote samples eigenlijk geen slechte resultaten rapporteert: bijvoorbeeld voor de top 10 rijmwoorden ligt de accuraatheid meestal boven de 90%, indien samples van vijftien kapittels en groter worden gebruikt (ca. 1300 verzen). De gemiddeld meest robuuste toeschrijvingen vinden we bij het gebruik van de 75 meest frequente rijmwoorden: indien al deze rijmwoorden in het experiment worden betrokken, levert dat in deze setting voor alle sample-groottes de beste resultaten op.

parameter

Voor samples van minder dan twintig kapittels (ca. 1.740 verzen) blijven de resultaten acceptabel maar verre van foutloos, zeker als men minder dan het maximum aantal beschikbare rijmwoorden gebruikt. Natuurlijk is de verleiding op dat moment groot om ook het aantal gebruikte rijmwoorden op te voeren (meer dan de 75 hier gebruikte). Waarschijnlijk zit daar nog heel wat auteur-gerelateerde informatie in die de classifier sneller hoge scores kan laten halen.

Toch is dat methodologisch niet altijd even makkelijk te verantwoorden. Uit de rangschikking van de hier gebruikte rijmwoorden blijkt dat men dan al snel rijmwoorden zou meenemen die iets van inhoud schijnen door te laten. Een uitstekend voorbeeld is het rijmwoord MARTELAAR op positie 87 in de ranglijst: als we dit rijmwoord zouden meenemen in de classificatie zou de accuraatheid sterk stijgen want Utenbroeke gebruikt het veel vaker dan Maerlant. Methodologisch is dit niettemin minder verantwoord, aangezien de inhoud van de Tweede Partie (met veel passages over martelaren) mogelijk veel vaker aanleiding heeft gegeven tot het gebruik van het woord en het geconstateerde verschil niet (enkel) aan de auteur kan verbonden worden.

4.3 CONCLUSIE

In bovenstaand hoofdstuk werd de basis voor de rest van deze studie gelegd. In hoofdstuk 3 is geïllustreerd dat middeleeuwse handschrif-

ten van dezelfde auteurstekst een grote variatie kunnen vertonen. Zoals aangehaald in Brinkmans Eerste Gebod, De eenheid van kopiist, is het daarom onduidelijk hoe men afschriften van verschillende kopiisten kan betrekken in een stylometrisch onderzoek naar middeleeuwse auteursteksten. Niettemin is door veel medioneerlandici opgemerkt dat de rijmwoordenschat van teksten wél relatief stabiel bleef in de overlevering van een tekst (§ 4.1). Mogelijk kan het rijmwoord in Middelnederlandse teksten daarom dienen als een 'reddingsboei' voor de stylometrische analyse van het auteurschap in middeleeuwse literatuur. Daarvoor is het wel noodzakelijk om rijmwoorden te ontdoen van de (relatief oppervlakkige) spellingvariatie die kopiisten hebben geïntroduceerd. Door het lemmatiseren van teksten kan men het uitzicht van rijmwoorden normaliseren en analyses beperken tot het optreden van 'lexemen' of 'lemmas' in rijmpositie (§ 4.1.2).

In de stylometrie werkt men graag met functoren omdat deze items erg frequent en goed gespreid zijn over teksten zodat zij een stilistische analyse toelaten die relatief onafhankelijk is van de inhoud of het genre van teksten. Hierboven werd beargumenteerd dat dergelijke methodologische voordelen van functoren ook gelden voor de meest frequente rijmwoorden in de Middelnederlandse literatuur (§ 4.1.3). Ik heb betoogd dat verschillende hoogfrequente rijmwoorden zo vaak voorkomen in diverse Middelnederlandse teksten dat hun optreden amper nog kan worden verbonden met de specifieke inhoud van teksten. Daarom lijken Middelnederlandse hoogfrequente rijmwoorden principieel geschikt voor auteursherkenning. Daarbij werd ook verwezen naar het vroegere onderzoek naar stoplappen en noodrijmen, formulaire constructies die Middelnederlandse dichters vaak als weinigzeggend 'kolom-vulsel' gebruikten. In veel van deze inhoudsarme constructies maakt men gebruik van hoogfrequente rijmwoorden.

In dit hoofdstuk werd dan ook getest of hoogfrequente rijmwoorden geschikt zijn voor auteursherkenning in Middelnederlandse rijmteksten. Daarvoor deed ik beroep op het aandeel van Maerlant en Utenbroeke in de *Spiegel historiael* (§ 4.1.1). Omdat deze dichters zo dicht bij elkaar moeten hebben gestaan, was dit een ideale eerste testcasus. Voor de analyse van deze tekst maakte ik gebruik van een methodologie uit de tekstclassificatie, een paradigma waarin men algoritmes 'traint' of aanleert om automatisch bepaalde labels aan teksten toe te kennen (§ 4.2). Wij gebruikten hierboven een eenvoudige vorm van *Memory-based learning* waarin het concept van de *nearest neighbour* een belangrijke rol speelt (§ 4.2.2). Aan dit paradigma kunnen bovendien aantrekkelijke testscenario's worden ontleend, zoals *leave one out validation*.

In experimenten met de Tweede en Derde Partie van de *Spiegel historiael* bleek dat een eenvoudig classificatie-algoritme eigenlijk maar weinig moeite had om de rijmstijlen van beide auteurs te onderscheiden (§ 4.2.3). Daarvoor moest het algoritme wel beschikken over

samples die groot genoeg waren. Bovendien moesten voldoende hoogfrequente rijmwoorden worden gebruikt. In dit hoofdstuk presteerde de toeschrijvingstechniek hoe dan ook ruim boven het kansniveau van de *baseline*. De vraag is natuurlijk of dergelijke technieken ook goed presteren als meer dan twee auteurs in de analyses betrokken worden. Bovendien moet de claim inzake de inhoudsonafhankelijkheid van hoogfrequente rijmwoorden nog beter onderbouwd worden. Dergelijke kwesties komen aan bod in de twee volgende hoofdstukken van dit deel.

Samples	Top 5	Top 10	Top 20	Top 30	Top 50	Top 75
732s x 1k	59	62	66	64	66	67
366s x 2k	61	63	64	71	71	74
244s x 3k	72	72	72	71	67	75
182s x 4k	71	72	73	75	77	81
146s x 5k	73	76	80	81	82	89
122s x 6k	81	75	88	86	88	90
104s x 7k	70	79	85	87	83	88
90s x 8k	82	76	81	90	88	93
80s x 9k	86	86	91	92	95	95
72s x 10k	81	80	90	83	93	95
66s x 11k	84	84	89	96	92	95
60s x 12k	90	83	93	95	96	98
56s x 13k	83	83	83	89	96	**100**
52s x 14k	88	86	88	94	98	98
48s x 15k	85	93	97	97	97	**100**
44s x 16k	90	90	**100**	97	97	**100**
42s x 17k	90	85	97	**100**	**100**	**100**
40s x 18k	92	92	95	95	95	**100**
38s x 19k	92	94	94	86	92	94
36s x 20k	94	97	86	**100**	**100**	**100**
34s x 21k	88	85	**100**	94	**100**	**100**
32s x 22k	93	96	96	**100**	96	**100**
30s x 23k	**100**	**100**	**100**	96	**100**	**100**
30s x 24k	93	93	**100**	**100**	**100**	**100**
28s x 25k	96	**100**	96	**100**	**100**	**100**
28s x 26k	96	**100**	**100**	**100**	**100**	**100**
26s x 27k	**100**	**100**	**100**	**100**	**100**	**100**
26s x 28k	**100**	**100**	**100**	**100**	**100**	**100**
24s x 29k	95	**100**	95	**100**	**100**	**100**
24s x 30k	91	**100**	**100**	**100**	**100**	**100**

Tabel 13: *Leave one out*-experiment waarbij een classifier stukken tekst uit de Tweede en Derde Partie op basis van de frequenties van rijmwoorden aan de correcte auteur moet toeschrijven (respectievelijk Utenbroeke en Maerlant). Hier wordt op twee variabelen gefocust: ten eerste, het effect van de wijze waarop het beschikbare materiaal in samples wordt verdeeld (s(amples) x k(apittels)) en, ten tweede, het effect van het aantal hoogfrequente rijmwoorden uit deze samples dat in de toeschrijving wordt betrokken (van 5 tot 75 rijmwoorden). Het opvoeren van zowel de sample-grootte als het aantal betrokken rijmwoorden heeft een positief effect op het aantal correcte attributies (percentage zonder decimalen), zodat de hoogste scores zich onderaan rechts in de tabel bevinden. Zie ook Fig. 9 (p. 92).

5 VELTHEM ET AL.

5.1 DE VOORTZETTING

In het vorige hoofdstuk heb ik beargumenteerd dat het rijmwoord een geschikte uitvalsbasis vormt voor auteursherkenning in Middelnederlandse narratieve teksten. Deze tekstelementen bleven immers betrekkelijk stabiel in de overlevering (§ 4.1). De piloot-experimenten met de *Spiegel historiael* bevestigden dit: met een simpele classificatietechniek bleek het mogelijk om twee auteurs van deze omvangrijke rijmtekst te onderscheiden op basis van hun meest frequente rijmwoorden (§ 4.2.3). Utenbroeke (Tweede Partie) bleek relatief makkelijk te onderscheiden van Maerlant (Derde Partie), zelfs al moeten deze dichters dicht op elkaar geleefd en geschreven hebben. Zoals reeds aangehaald heeft Lodewijk van Velthem de *Spiegel* later voortgezet. Dit hoofdstuk betrekt ook Velthems aandeel in de stylometrische analyse van de *Spiegel historiael*.[1]

5.1.1 *Velthem*

Lodewijk van Velthem was een Brabantse priester-dichter die zijn herkomstnaam ontleende aan het kleine dorpje nabij Leuven waar hij vanaf 1312 als priester was aangesteld.[2] De weinige biografische informatie die we over Velthem bezitten, kennen we uit zijn eigen werken. Zo weten we dat Velthem zich al vroeg in de hoogste adellijke kringen begaf, want in de winter van 1293-1294 verbleef hij te Parijs in het gevolg van hertog Jan I van Brabant. Voor zijn aanstelling in Veltem-Beisem, was Lodewijk werkzaam in Zichem, nabij Diest. In 1327 voltooide Velthem zijn laatst bekende werk en men gaat ervan uit dat hij niet lang daarna gestorven is.

Zoals gezegd heeft Jacob van Maerlant zelf niet meer de voltooiing van de *Spiegel historiael* mogen meemaken. In het voorlaatste kapittel van de Vierde Partie lezen we immers dat Lodewijk van Velthem de *Spiegel* heeft voortgezet vanaf het punt waar Maerlant zijn werk – waarschijnlijk om gezondheidsredenen – had moeten staken (*van daert Jacob liet*). In 1315 was Velthem klaar met de Vierde Partie en droeg hij deze op aan Maria van Berlaar, de edelvrouw die als mecenas voor het werk is opgetreden. Opvallend genoeg heeft Velthem geheel op

1 Dit hoofdstuk steunt in belangrijke mate op een eerdere publicatie (Kestemont 2010a).

2 Tegenwoordig is het plaatsje Velthem een deel van de gemeente Veltem-Beisem. Alle biografische informatie die ik in deze sectie vermeld, is terug te vinden in de recente bundel studies *De boeken van Velthem* (Besamusca et al. 2009a). Nadere literatuurverwijzingen vindt men daar in de inleiding (Besamusca et al. 2009b).

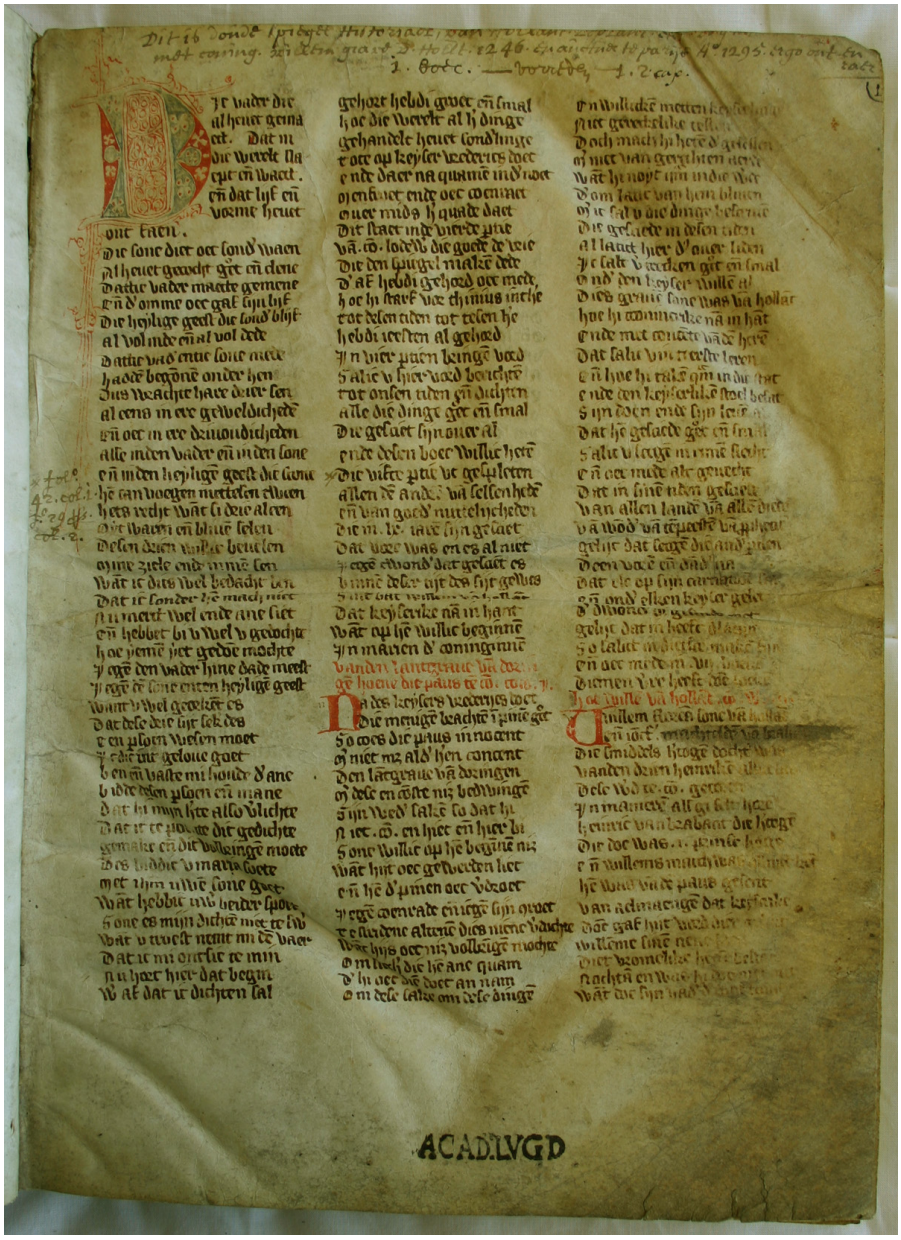
eigen initiatief een Vijfde Partie aan de *Spiegel* toegevoegd, waarin hij de periode 1247-1316 behandelde. Deze slotpartie moest het geschiedverhaal doortrekken tot aan zijn eigen tijd, want het *Speculum* van Vincentius liep zo ver niet door. Velthem kon zich voor de Vijfde Partie dus niet meer bedienen van de Latijnse brontekst en moest te rade gaan bij andere, veelal volkstalige bronnen en informanten. Omstreeks 1316-1317 moet Velthem de Middelnederlandse *Spiegel historiael* hebben afgerond.

Velthem was dan wel een Brabander die iets later schreef dan zijn West-Vlaamse voorgangers, maar toch vormt zijn Voortzetting uitstekend vergelijkingsmateriaal voor de eerder besproken Tweede en Derde Partie. Velthem had immers expliciet de bedoeling om de *Spiegel* in dezelfde lijn door te zetten en beriep zich voor de Vierde Partie zelfs op dezelfde brontekst. Bovendien was Velthem een groot bewonderaar van Maerlant, zelfs in die mate dat men hem heeft aangeduid met het epitheton 'executeur van Maerlants literaire testament' (Van Oostrom 1982, 35). Het hoeft dan ook niet te verbazen dat men wel beweerd heeft dat Maerlant een grote stilistische invloed heeft uitgeoefend op zijn epigoon (Van Loey 1946, 42-43). Stylometrisch gesproken zou het op het eerste gezicht dus niet heel makkelijk mogen zijn om Velthem van zijn voorgangers te onderscheiden.

Velthems aandeel in de Vierde Partie is jammer genoeg niet volledig overgeleverd: de brokstukken die door de jaren heen zijn opgedoken, zijn veelal afkomstig uit erg disparate handschriften en dekken vaak niet meer dan een beperkte reeks kapittels (Biemans 1997; Warnar 2009). Velthems Vijfde Partie is daarentegen veel beter overgeleverd en zelfs nagenoeg volledig in het handschrift LEIDEN, BPL E 14. De vervaardiging van dit zogenaamde Vijfde Partie-handschrift (zie Fig. 10, p. 99) wordt vaak in de onmiddellijke omgeving van de dichter gesitueerd, als het al niet in opdracht van de dichter zelf werd aangelegd (Biemans 2009). De overlevering van Velthems Vijfde Partie maakt dit dichtwerk zo tot het ideale startpunt om Velthems stijl met die van zijn voorgangers te vergelijken, zeker als we rekening houden met Brinkmans Achtste Gebod dat stelt dat een geanalyseerde tekstgetuige best zo dicht mogelijk staat bij de autograaf. Velthems Vijfde Partie is bovendien eveneens gestructureerd in boeken en kapittels, zodat we deze op een organische wijze kunnen vergelijken met die van de Tweede en Derde Partie.

Achtste Gebod: Minimale afstand tot de autograaf

Toch is het niet zinvol op dit moment om de volledige Vijfde Partie in het onderzoek te betrekken. De Vijfde Partie telt in totaal namelijk acht boeken, maar slechts de eerste zes hiervan zijn zuiver geschiedkundig. In de laatste twee boeken waagt Velthem zich aan een bespreking van toekomstvisioenen en biedt hij een vooruitblik op het einde der tijden (vgl. Sleiderink 2009). Omdat de toon en stijl in deze laatste twee boeken sterk zouden kunnen afwijken van die in de voorgaande



Figuur 10: Het openingsfolium van het zo goed als complete Vijfde Partie-handschrift (Leiden, Universiteitsbibliotheek, bpl e 14). Als hij al niet in opdracht van Velthem zelf werd aangelegd, kan de codex in tijd en ruimte (Brabant, ca. 1325) dicht bij de auteur zijn vervaardigd (Biemans 2009).

zes, leggen we het zevende en achtste boek best nog even aan de kant, zodat de geanalyseerde teksten maximaal vergelijkbaar zijn.

Een ander punt is Velthems verslag van de Slag bij Woeringen in 1288. Op verschillende plaatsen in de eerste drie boeken van zijn Vijfde Partie besteedt Velthem ruim aandacht aan deze belangrijke veldslag, maar ook aan de aanloop en nasleep van dit militaire treffen tussen de hertog van Brabant en de bisschop van Keulen. Deze passages blijken echter niet op Velthem zelf terug te gaan: de auteur moet ze ontleend hebben aan een andere rijmtekst, *De rijmkroniek van Woeringen*, en hij heeft ze na een beperkte herwerking in zijn eigen geschiedverhaal geïntegreerd (Mantingh 2009). Broeder Jan van Heelu schreef de *Rijmkroniek* omstreeks 1290 in opdracht van hertog Jan I (Sleiderink 2003, 87-97). Omdat Heelu's kroniek ook in een ander handschrift is overgeleverd, kunnen we de overgenomen passages makkelijk isoleren (Besamusca 2009, 83). De interpolaties betreffen de volgende passages:

- Boek 1, vs. 2.824-3.543 = kapittels 40 t.e.m. 46
- Boek 2, vs. 3.220-4.032 = kapittels 44 t.e.m. 55
- Boek 3, vs. 71-1.388 = kapittels 2 t.e.m. 21

Velthem geeft in de lopende tekst nooit aan wanneer zijn eigen werk in dat van Heelu overvloeit of omgekeerd. Sterker nog: op geen enkel moment vermeldt hij expliciet zijn schatplichtigheid aan Heelu's werk. Op het einde van zijn kroniek geeft hij wel aan dat hij diverse verslagen van ooggetuigen heeft geraadpleegd. Met die opmerking verwees hij mogelijk ook naar de *Rijmkroniek*, aangezien Heelu op verschillende plaatsen in zijn tekst kenbaar maakte dat hij aanwezig was op het slagveld in 1288.

Toch moeten we oppassen met het afschilderen van Velthems hergebruik van teksten als kwaadwillig plagiaat. Wijzelf kennen de Heelu-toeschrijving van de *Rijmkroniek* pas uit een jongere proloog die in de vijftiende eeuw aan de tekst werd toegevoegd (Sonnemans 1992, 334). Daarom is het goed mogelijk dat de *Rijmkroniek* zoals Velthem die ca. 1316 voor zich had liggen nog anoniem was. In dat geval kón Velthem de tekst simpelweg niet aan Heelu hebben toegeschreven. Anderzijds is het moeilijk voorstelbaar dat onze doorgaans goed ingelichte geschiedschrijver slechts enkele jaren later geen weet meer had van deze auteursnaam. Sterker nog: als Velthem ca. 1294 contacten had met het hertogelijke entourage is het bijna onmogelijk dat hij de Brabantse hofdichter niet heeft gekend. Of kwaadwillige motieven (persoonlijke afgunst?) een rol hebben gespeeld in het verzwijgen van Heelu's naam blijft zo moeilijk uit te maken. Aangezien Velthem de Heelu-interpolaties slechts in beperkte mate heeft bewerkt, is het niettemin aangewezen om deze stilistisch 'vervuilde passages' uit de Vijfde Partie te verwijderen, alvorens die aan een stylometrisch onderzoek te onderwerpen.

5.1.2 *Een verwarrende matrix*

Wanneer we de kapittels uit Velthems Vijfde Partie van de standaard-editie op de *Cd-rom Middelnederlands* oogsten, levert dat 258 kapittels op – na aftrek van de laatste twee boeken en de Heelu-kapittels. Dat zijn er iets minder dan voor Utenbroeke en Maerlant maar nog ruim voldoende voor een betrouwbaar experiment. We kunnen de rijmwoorden in deze kapittels verrijken met lemma's, volledig analoog aan onze werkwijze voor de Tweede en Derde Partie. Op dit moment kunnen we *leave-one-out*-experimenten uitvoeren met drie in plaats van twee *Spiegel*-auteurs. Uit de internationale literatuur is bekend dat attributie-technieken in de regel aan accuraatheid inboeten, als het aantal auteurs in experimenten wordt opgevoerd (bv. Luyckx & Daelemans 2011). De kans op verwarring is vanzelfsprekend groter als we meerdere auteurs beschouwen, zodat de finale accuraatheid meestal lager ligt. Onderstaande tabel geeft de resultaten weer voor enkele van deze experimenten (Tab. 14, p. 101).[3]

Samples	**Top 25**	**Top 50**	**Top 75**
75s x 10k	86	92	94
69s x 11k	88	92	88
63s x 12k	88	92	90
57s x 13k	84	87	92
54s x 14k	90	98	90
51s x 15k	98	98	98
48s x 16k	95	100	97
45s x 17k	97	97	97
42s x 18k	92	97	100
39s x 19k	82	87	89
36s x 20k	94	100	100

Tabel 14: Accuraatheid (weergegeven als een percentage zonder cijfers na de komma) van een serie *leave one out*-experimenten waarbij een classifier stukken tekst uit de Tweede, Derde en Vijfde Partie op basis van de frequenties van rijmwoorden aan de correcte auteur moet toeschrijven (respectievelijk Utenbroeke, Maerlant en Velthem).

Deze tabel leert dat de accuraatheid over het algemeen niet slecht is en zeker boven het kansniveau ligt, want in het geval van een toeschrijvingsexperiment met drie auteurs (met elk evenveel samples) veronderstellen we een baseline van 33%. Toch schijnt de accuraatheid

3 Merk op dat het aantal samples per auteur wordt gelijk gehouden. Van Velthem hebben we slechts 258 kapittels wat in het geval van samples van 10 kapittels 25 samples oplevert (=(258-8)/10). In dat geval worden analoog slechts de eerste 25 samples van de twee controle-auteurs gebruikt – in totaal dus 75 (=3x25) samples van 10 kapittels – om geen van onze drie auteurs kwantitatief te bevoordelen.

van de attributies iets lager te liggen dan in de experimenten met twee auteurs, zelfs met een grotere sample-omvang. Opvallend is bijvoorbeeld het rijtje waarden voor de 45 samples van 17 kapittels: met zowel de 25, 50 als 75 meest frequente rijmwoorden wordt telkens één sample verkeerd toegeschreven.

error rate

Natuurlijk kunnen we naast de accuraatheid ook aandacht besteden aan de *error rate* in een experiment, i.e. het percentage verkeerd toegeschreven werken. Een analyse van de verkeerd toegeschreven samples levert mogelijk nieuwe inzichten op over het experiment. Een zogenaamde 'verwarringstabel' of *confusion matrix* is een waardevol hulpmiddel bij de foutenanalyse van een experiment. Hieronder zijn enkele voorbeelden van dergelijke tabellen afgebeeld (Tab. 15, p. 102).

confusion matrix

Top 25			
	Utenbroeke	Maerlant	Velthem
Utenbroeke	**15**	0	0
Maerlant	0	**15**	0
Velthem	1	0	**14**
Top 50			
	Utenbroeke	Maerlant	Velthem
Utenbroeke	**15**	0	0
Maerlant	0	**15**	0
Velthem	0	1	**14**
Top 75			
	Utenbroeke	Maerlant	Velthem
Utenbroeke	**15**	0	0
Maerlant	0	**15**	0
Velthem	0	1	**14**

Tabel 15: *Confusion matrices* voor *leave one out*-experimenten waarbij een classifier telkens 45 samples van 17 kapittels uit de Tweede, Derde en Vijfde Partie aan de correcte auteur moet toeschrijven (respectievelijk Utenbroeke, Maerlant en Velthem). De matrix wordt getoond voor het experiment op basis van de frequenties van de respectievelijk 25, 50 en 75 meest frequente rijmwoorden.

De kolommen in deze matrices stellen de effectief toegekende auteurslabels toe, terwijl de rijen de correcte labels voorstellen (de *gold standard labels*). Deze tabellen maken duidelijk dat de 'lekken' in onze toeschrijvingen enkel Velthems samples treffen. Het zijn immers uitsluitend samples van de Brabander die foutief worden toegeschreven aan een van zijn voorgangers. Anders dan men misschien zou verwachten op basis van hun geografische of chronologische nabijheid zijn het niet de Vlamingen Utenbroeke en Maerlant die stilistisch verward worden. Een nadere inspectie leert dat de foutief toegekende

Velthem-samples niet in ieder experiment dezelfde zijn maar wel één opvallend aspect gemeen hebben: zij bestaan voor het merendeel uit samples van het vierde boek van de Vijfde Partie. Dat een naïeve classifier als de onze fouten kan maken, hoeft niet te verrassen. Wel is het opvallend dat zich – ook in andere, hier niet-gerapporteerde proeven – een patroon voordoet waarbij de foutief toegekende samples steeds voor het grootste deel bestaan uit samples van het vierde boek van de Vijfde Partie. Het is dan ook een interessante vraag of dit patroon zich verder zet bij analyses op het niveau van samples die volledige boeken dekken.

Top 25			
	Utenbroeke	Maerlant	Velthem
Utenbroeke	**7**	0	0
Maerlant	0	**7**	0
Velthem	1	0	**5**
Top 50			
	Utenbroeke	Maerlant	Velthem
Utenbroeke	**7**	0	0
Maerlant	0	**7**	0
Velthem	0	2	**4**
Top 75			
	Utenbroeke	Maerlant	Velthem
Utenbroeke	**7**	0	0
Maerlant	0	**7**	0
Velthem	0	1	**5**

Tabel 16: *Confusion matrix* voor een *leave one out*-experiment waarbij een classifier telkens samples bestaande uit hele boeken van de Tweede, Derde en Vijfde Partie aan de correcte auteur moet toeschrijven (respectievelijk Utenbroeke, Maerlant en Velthem). De matrix wordt getoond voor het experiment op basis van de frequenties van de respectievelijk 25, 50 en 75 meest frequente rijmwoorden.

In bovenstaande tabel zijn de verwarringsmatrices weergegeven voor enkele experimenten op boek-niveau (Tab. 16, p. 103). In alle drie de gevallen zijn het steeds boeken van Velthem die foutief worden toegeschreven. Erg opvallend is dat boek 4 telkens foutief wordt toegeschreven, één keer aan Utenbroeke (top 25) en twee keer aan Maerlant (top 50 en top 75). In het geval van de top 50 wordt ook Velthems derde boek foutief aan Maerlant toegeschreven. Nu hoeft het niet te verbazen dat onze naïeve classifier fouten maakt: zeker door het betrekken van drie in plaats van twee auteurs viel immers te verwachten dat het algoritme makkelijker zou verward zijn tussen de stijlen van deze auteurs. Boek 3 bijvoorbeeld wordt in de meeste an-

dere experimenten wel gewoon aan de correcte auteur toegeschreven, zodat men zich niet al te veel zorgen hoeft te maken over dergelijke, geïsoleerde uitschieters.

De kapittels uit Boek 4 zijn echter meer zorgwekkend: deze worden maar zelden aan Velthem toegeschreven in de experimenten. Het algoritme lijkt daarbij geen expliciete voorkeur voor Utenbroeke of Maerlant te hebben, zodat de toeschrijving van het algoritme eerder luidt 'elke auteur behalve Velthem'. De fout zet zich bovendien onveranderd door als men met een grotere sample-omvang werkt, hoewel de stylometrische literatuur net benadrukt dat een grotere sample-omvang meestal de accuraatheid van proeven verbetert (Luyckx & Daelemans 2011). Een dergelijke, recurrente fout is dus mogelijk minder toevallig en verdient een nadere analyse.

5.2 HET VIERDE BOEK

5.2.1 *De Guldensporenslag*

Laat ons de lading van het vierde boek van de Vijfde Partie eens nader inspecteren. De gebeurtenissen die in dit boek worden besproken, handelen voor het leeuwendeel over de bekende Guldensporenslag. Op 11 juli 1302 vond een belangrijk militair treffen plaats tussen de zwaarbewapende ruitertroepen van de Franse koning en een onervaren guerilla-leger van landbouwers en (voornamelijk) stedelingen uit het middeleeuwse graafschap Vlaanderen.[4] De slag werd uitgevochten aan de rand van de welvarende stad Kortrijk. Hoewel de slag in se slechts een van de vele conflicten was in een langdurig conflict tussen de Franse kroon en het opstandige leen Vlaanderen, wordt hij niet zelden beschouwd als een mijlpaal in de Vlaamse, en ook Europese socio-politieke geschiedenis. Voor de allereerste keer bleek een leger van lichtbewapend voetvolk immers in staat om de onverslaanbaar geachte, professionele ruitertroepen van de Franse koning te verslaan.

Het nieuws van de verpletterende Franse nederlaag verspreidde zich als een lopend vuurtje door Europa. Het is bijvoorbeeld tekenend dat paus Bonifatius VIII zeven dagen later in het midden van de nacht werd gewekt, toen een boodschapper in Rome aankwam met het verrassende nieuws over de vernederende Franse nederlaag. Uit de Britse eilanden is zelfs een lied overgeleverd dat kort na de slag moet zijn gecomponeerd, waarin de Vlaamse heldendaden werden bezongen (Van Bossele 2002). De slag had zulke grote socio-economische gevolgen (bv. inspraak van ambachten in het beleid van steden) dat deze wordt beschouwd als een keerpunt in de geschiedenis van het graafschap

4 In het Frans heet de slag *Bataille des éperons d'or*, in het Engels *Battle of the Golden Spurs*. Een recente Engelstalige studie – hoewel met enkele feitelijke onjuistheden – biedt Fegley (2002). Een andere recente reeks studies naar aanleiding van het eeuwfeest van de slag vindt men bij Trio et al. (2002). Iets ouder maar nog bijzonder relevant is het werk van Verbruggen (1952).

Vlaanderen. De Guldensporenslag wordt vandaag de dag nog steeds herdacht op de jaarlijkse feestdag van de Vlaamse gemeenschap.[5]

De Guldensporenslag is naar middeleeuwse maatstaven erg goed gedocumenteerd in contemporaine bronnen. Toch zien we dat één bron er met kop en schouders bovenuit steekt: het vierde boek van Velthems Vijfde Partie biedt een uitzonderlijk uitgebreid en gedetailleerd verslag van de veldslag. Zoals gezegd, wordt het vierde boek van de Vijfde Partie toegeschreven aan de Brabantse priester-dichter Lodewijk van Velthem. Dit boek vormt misschien wel de meest bestudeerde passage uit zijn hele oeuvre. De passage is meerdere keren afzonderlijk uitgegeven en vertaald, hetgeen de aanhoudende relevantie van Velthems relaas benadrukt (Waterschoot 1979; Jongen & Piters 2002).

Toch wekt het verbazing dat precies deze belangrijke passage uit de Vijfde Partie via de stylometrie moeilijk aan Velthem toe te schrijven blijkt. Kennelijk wijkt de stijl in dit boek af van de rest van de Vijfde Partie en vertoont het gebruik van rijmwoorden in dit boek meer gelijkenissen met het werk van de Vlamingen Utenbroeke of Maerlant, dan met dat van Velthem. Deze situatie wekt argwaan. Natuurlijk is het goed mogelijk dat ons attributie-algoritme gewoon niet goed werkt en ons misleidt. De gebruikte techniek is immers relatief naïef en is nog maar amper getest op Middelnederlandse teksten. Toch kon dit algoritme op betrekkelijk eenvoudige wijze het verschil vatten tussen de stijlen van Utenbroeke en Maerlant, twee auteurs die voor zover wij weten toch relatief dichter bij elkaar stonden dan bijvoorbeeld Utenbroeke en Velthem. De vraag werpt zich daarom op of de veronderstelde *ground truth* in ons *Spiegel historiael*-corpus wel zo een gouden standaard biedt als wij veronderstellen. Hoewel ik dit deel van mijn studie oorspronkelijk wilde beperken tot een analyse van teksten waarvan het auteurschap niet betwijfeld wordt, zie ik mij op dit punt toch genoodzaakt volgende vraag te stellen: is het vierde boek uit de Vijfde Partie eigenlijk wel door Velthem geschreven?

Deze vraag wordt vooral gerechtvaardigd door de observatie dat Velthem niet vies was van het betere knip- en plakwerk. Zijn weinig expliciete interpolatie van de Heelu-passages in de Vijfde Partie, bijvoorbeeld, werd voorlopig nog genegeerd in de proeven, maar toont wel aan dat we Velthems auteurschap voor de Vijfde Partie niet altijd zomaar mogen aannemen. Vanzelfsprekend bieden de Heelu-passages waardevol vergelijkingsmateriaal: wij weten dat deze passages stilistisch zouden kunnen afwijken van Velthems normale styloom, omdat hij deze ontleende van een andere auteur.[6] Het loont daarom de moeite om deze passages opnieuw in te voegen in ons *Spiegel*-corpus.

5 Jaarlijks vindt de viering plaats op 11 juli. Zie `http://www.vlaanderen.be` voor meer details over de feestelijkheden.

6 Een terzijde: ik gebruik het woord styloom in deze studie soms ter aanduiding van de individuele rijmstijl van een auteur maar heb niet de intentie om daarmee onrechtstreeks de beruchte Styloom-hypothese (Van Halteren et al. 2005) te onderschrijven dat dergelijke stylomen in alle gevallen perfect van elkaar te onderscheiden zijn.

Bovendien is het nu ook aangewezen om ook de laatste twee boeken uit de Vijfde Partie bij het onderzoek te betrekken: de toon en inhoud van dit boek wijken wel af van de rest van de Partie maar zover wij weten is er geen reden om Velthems auteurschap voor deze boeken in twijfel te trekken.

Op analoge wijze heb ik ook de resten van Velthems Vierde Partie bij elkaar geharkt en verrijkt met rijm-lemmata. De overlevering van deze partie is misschien fragmentarisch, maar biedt even goed belangrijk referentie-materiaal voor Velthems rijmstijl in het historiografische genre. In totaal zijn nog 103 kapittels uit Velthems aandeel in de Vierde Partie te verzamelen (zie bv. Fig. 18, p. 118). De meeste daarvan zijn in een digitale versie op de *Cd-rom Middelnederlands* te vinden en de gebroeders Geirnaert waren zo vriendelijk om mij een digitale editie te bezorgen van recenter opgedoken fragmenten (Geirnaert & Geirnaert 2006).

5.2.2 *Opnieuw de dichtste buur*

Laat ons even terugkeren naar de interne werking van de *Memory-Based Learner*. Een dergelijke classifier wordt eenvoudigweg getraind door een aantal voorbeeld-instanties op te slaan in het geheugen. Onze voorbeeld-instanties zijn een verzameling kapittels door dezelfde auteur die worden voorgesteld aan de hand van de relatieve frequenties van een aantal hoogfrequente rijmwoorden. Een dergelijke leermethode wordt ook 'lui' genoemd omdat de classifier voor een dergelijke *lazy learning* in principe geen verregaande ingrepen uitvoert in het trainingsmateriaal, bijvoorbeeld door minder betrouwbare trainingsinstanties weg te gooien. Alle instanties worden gewoon bijgehouden: de achterliggende gedachte is dat het schadelijk kan zijn om geïsoleerde (groepjes) samples te verwijderen, omdat deze soms belangrijke uitzonderingen bevatten (Daelemans et al. 1999).

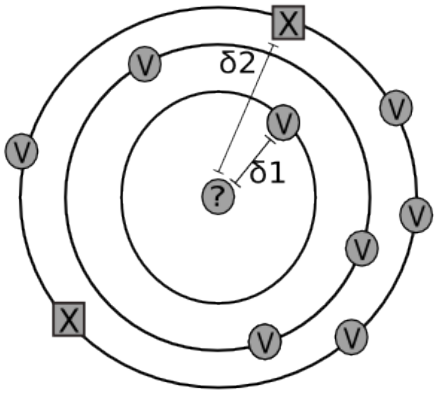
lazy learning

Als de classifier een nieuwe, nog ongelabelde instantie krijgt aangeboden, wordt deze een test item genoemd. Het gaat dan bijvoorbeeld om een sample van opeenvolgende kapittels uit Velthems Vijfde Partie. De classifier gaat in zijn geheugen op zoek naar het trainingssample dat er het meest op lijkt. De mate van gelijkenis – of eigenlijk ongelijkenis – tussen twee samples wordt bepaald door een afstand tussen beide samples te berekenen. In het eenvoudigste geval – het zogenaamde IB1-algoritme (Aha et al. 1991) – zal de classifier de afstand berekenen tussen het test item en elk trainingsitem. Vervolgens zal één trainingsitem als *nearest neighbour* worden aangeduid, namelijk het trainingsitem dat op de kortste afstand ligt van het bewuste test item. De eigenlijke classificatie (bv. het toekennen van een auteurslabel) bestaat er dan uit dat de classifier aan het test item hetzelfde klasse-label toekent als de *nearest neighbour*. Deze laatste stap noemt men de extrapolatie van het klasse-label naar het test item.

test item

IB1-algoritme

Onderstaande afbeelding biedt een visuele, naïeve voorstelling van het classificatie-proces binnen een 'dichtste buur'-redenering (Fig. 11, p. 107). Het ongeziene test item (?) is centraal afgebeeld met daarrond op steeds verder gelegen schillen de trainingsitems in het geheugen van de classifier.[7] Deze behoren tot twee klasses, namelijk (V) (bv. Velthem) en (X). De afstand tot het eerste V-sample op de binnenste schil is δ1 en is duidelijk kleiner dan de afstand tot het dichtsbijzijnde sample van de andere klasse X (δ2) op de buitenste schil. Bijgevolg zal het test item gelabeld worden als een V-sample na extrapolatie van het klasse-label van de dichte buur.



Figuur 11: Naïeve tweedimensionele voorstelling van classificatie volgens het *nearest neighbour*-model. Het test item (?) is centraal afgebeeld met daarrond op steeds verder gelegen schillen de trainingsitems in het geheugen van de classifier (V en X). De afstand δ1 tot het eerste V-sample op de binnenste schil is duidelijk kleiner dan de afstand δ2 tot het dichtsbijzijnde sample van de andere klasse X op de buitenste schil. Bijgevolg zal het test item gelabeld worden als een V-sample na extrapolatie van het klasse-label van de dichte buur.

Hieronder bespreek ik een reeks experimenten waarbij gebruik is gemaakt van de relatieve frequenties van de 30 vaakst voorkomende rijmwoorden in Velthems Vierde Partie. Deze relatief context-onafhankelijke rijmwoorden – vaak duidelijk stoplappen – zijn hieronder, gerangschikt naar cumulatieve frequentie, weergegeven.

1 – PrName, 2 – zijn, 3 – mede, 4 – stad, 5 – doen, 6 – zaan, 7 – gereed, 8 – dat, 9 – daar, 10 – verstaan, 11 – daarnaar, 12 – godweet, 13 – gaan, 14 – groot, 15 – ding, 16 – voorwaar, 17 – komen, 18 – waan, 19 – man, 20 – nu, 21 – niet, 22 – heer, 23 – met, 24 – gij, 25 – gene, 26 – stond, 27 – al, 28 – staan, 29 – deze, 30 – dood

7 Dit nieuwe test item wordt aangeduid met een vraagteken omdat de classifier de klasse (c.q. de auteur) ervan niet kent.

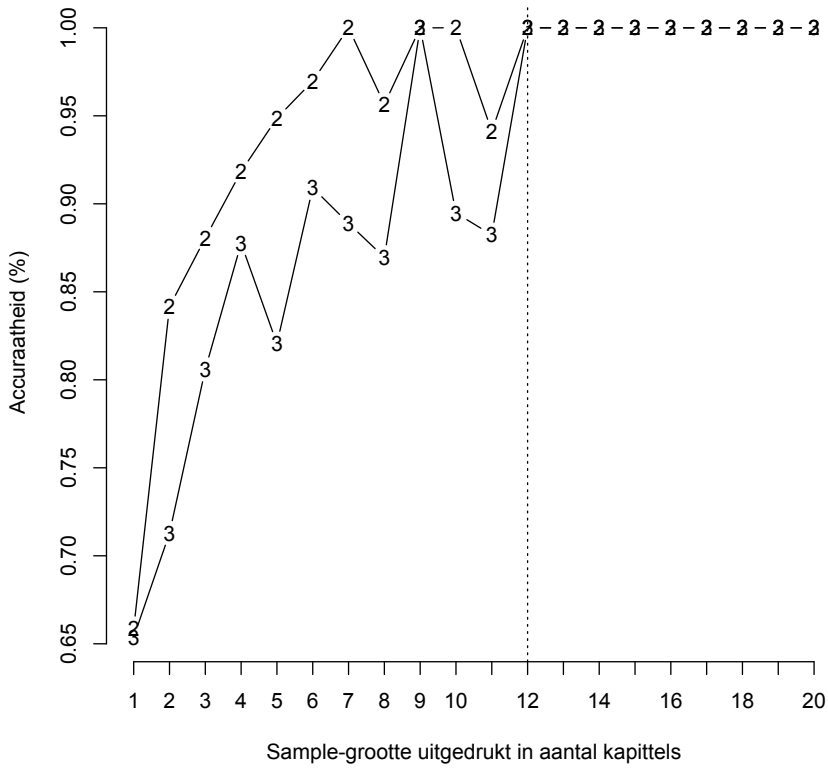
Voor deze techniek is gebruikt gemaakt van een iets ingewikkeldere classificatie-methode uit de TIMBL-software, dan de simpele optelling die tot nu toe is gehanteerd.[8] Het belangrijkste verschil is dat niet elke rijmwoord-frequentie even zwaar doorweegt in het bepalen van de afstand tussen twee samples. Op basis van de trainingssamples kan het classificatie-algoritme immers kennis opdoen over het belang van specifieke woorden in het stilistische onderscheid. Het rijmwoord GODWEET is bijvoorbeeld erg kenmerkend voor de stijl van Velthem. Via *feature weighting* kan men ervoor zorgen dat een dergelijk, goed onderscheidend kenmerk zwaarder doorweegt in het bepalen van een afstand tussen twee samples. Ik ga hier nog even voorbij aan een meer gedetailleerde bespreking van *feature weighting*, omdat hieronder nog een uitgebreidere analyse volgt van een eenvoudige manier van *feature weighting* (vgl. § 7.2).

Fig. 12 (p. 109) beeldt een serie experimenten uit waarin Velthems 103 kapittels uit de Vierde Partie werden geconfronteerd met de Tweede en Derde Partie van Utenbroeke en Maerlant. Daarbij heb ik de sample-grootte gevarieerd (x-as). De verschillen in globale accuraatheid (y-as) die dit opleverde zijn uitgezet als een functie hiervan. Deze leercurve laat duidelijk zien dat het gebruik van grotere samples een positief effect heeft op het resultaat van de classificatie. Zowel in het geval van Utenbroeke als Maerlant, blijkt dat Velthems stijl zo goed als feilloos wordt afgezet van de controle-auteur wanneer men een sample-grootte van 12 (of meer) kapittels hanteert (vgl. horizontale stippellijn).

5.2.3 *Goede buren en valse vrienden*

Wanneer we een classifier toepassen op een verzameling voorbeeld-instanties, bijvoorbeeld via *leave-one-out*-validering, is onze eerste bekommernis natuurlijk de uiteindelijke accuraatheid van het experiment. Toch levert dit cijfer niet altijd het beste inzicht op in de structuur van onze data: we krijgen immers gewoon een getal terug dat uitdrukt in hoeveel van de gevallen de dichtste buur van een test item het gewenste klasse-label had. Dat de dichtste buur die werd aangewezen van de juiste klasse is, mag dan wel interessant zijn, maar minstens even belangrijk is de vraag *hoe dicht* deze buur dan precies bij het

8 Technische noot: voor de afstandsberekening werd gebruikt van de *cosine distance* i.p.v. van de standaard *Manhattan distance* omdat deze beter omgaat met taken in tekstclassificatie waar documenten in lengte kunnen verschillen. Als *feature weighting* werd de standaard *Information Gain* gekozen. Deze parameters werden manueel bepaald en bleken optimaal na empirische proeven met een kruisvalidering van het trainingsmateriaal. Vanzelfsprekend werd de *k*-parameter voor de berekening van de *nearest distances* hier nog op 1 vastgezet. De proeven waarin de $\delta\delta$-waarden (zie onder) werden geëxtraheerd via de opties voor *neighbourhood inspection* werden in een apart experiment bepaald met $k <= 206$ (i.e. het maximum aantal trainingssamples in deze experimenten). Zie voor deze opties de handleiding door Daelemans et al. (2010b).



Figuur 12: Grafische weergave van een reeks experimenten waarbij Velthems *P4* in een serie experimenten is afgezet tegen Utenbroekes *P2* en Maerlants *P3*. De accuraatheid na *leave-one-out*-validering is als een percentage uitgezet als een functie van de gebruikte sample-groottes. Het gebruik van grotere samples heeft een positief effect op de classificatie en deze blijkt feilloos vanaf een sample-grootte van 12 kapittels (verticale stippellijn).

test item lag én – misschien nog belangrijker – hoe ver de eerste 'valse vriend' was verwijderd van het test item, dat wil zeggen: het dichtsbijzijnde trainingsitem met een ongewenst klasse-label.[9]

Onderstaande figuren maken deze redenering inzichtelijker (Fig. 13, p. 111). Een test item (van de gewenste klasse V) wordt in de vier subpanelen telkens op een iets andere wijze behandeld. In subpaneel (a) wordt het met grote zekerheid en correct aan V toegeschreven; in (b) wordt het ook nog correct toegeschreven, maar met een kleinere marge. In subpaneel (c) wordt het test item met beperkte zekerheid en foutief toegeschreven aan X en in (d) wordt dezelfde foute toeschrijving gemaakt maar met een grotere zekerheid. In deze afbeeldingen is $\delta 1$ consequent gelijkgesteld met de afstand van het test item tot de eerste

9 Deze zienswijze vindt men ook terug bij Hoover (2004a, 471), die stelde: 'large changes in Delta [...] from the likeliest to the second likeliest author are characteristic of reliability'.

voorbeeld-instantie van het (gewenste) type V, terwijl $\delta 2$ steeds de afstand aanduidt tot het eerste kapittel van de (ongewenste) klasse X. De subtiele verschillen tussen deze vier situaties kunnen betrouwbaar beschreven worden met één enkele variabele die we hier de 'Dubbele Delta' ($\delta\delta$) zullen noemen. Deze waarde definiëren we simpelweg als: $(\delta 1/\delta 2) - 1$.[10]

De lezer moet even de tijd nemen om zichzelf van het volgende te overtuigen. Als $\delta\delta$ negatief is betekent dit dat een test item aan de correcte klasse is toegewezen, want $\delta 1$ was in dat geval kleiner dan $\delta 2$. Hoe lager deze waarde, hoe zekerder die voorspelling. Indien $\delta\delta$ echter positief is, wordt het test item aan de verkeerde of toch ongewenste klasse toegewezen, want $\delta 1$ was in dat geval groter dan $\delta 2$. Hoe hoger de waarde in dit geval, hoe zekerder de classifier is van die voorspelling.

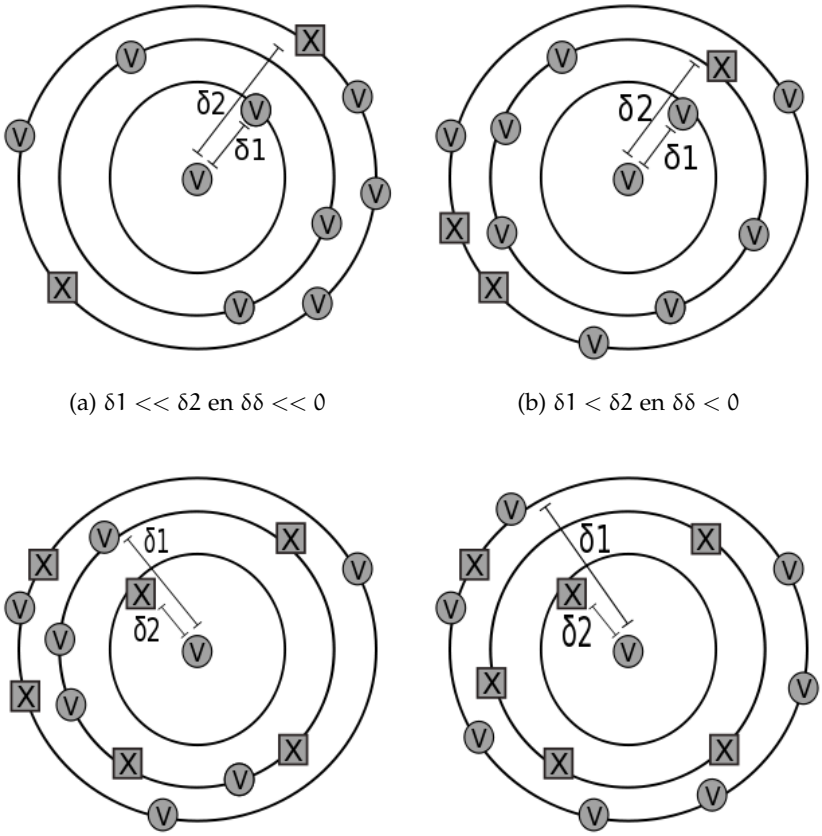
Aan de hand van de $\delta\delta$-waarde kunnen we nu proberen het onderscheid tussen de rijmstijlen van Utenbroeke, Maerlant en Velthem te bestuderen. Aangezien we specifiek geïnteresseerd zijn in Velthems styloom, zullen we hieronder werken met een vergelijkingstechniek die voor de Velthem-samples een zo negatief mogelijke $\delta\delta$ oplevert ten aanzien van samples van Utenbroeke en Maerlant. Dat betekent immers dat de vergelijkingstechniek een duidelijk verschil in nabijheid blootlegt tussen Velthems stijl en die van zijn twee voorgangers.

Onderstaande figuren bieden een grafische voorstelling van twee experimenten waarin Velthems Vierde Partie (afgekort als *P4*) wordt afgezet tegen de Tweede Partie van Utenbroeke (*P2*) en de Derde Partie van Maerlant (afgekort *P3*; Fig. 14, p. 112 en 15, p. 113). In plaats van de globale accuraatheid is nu echter de gemiddelde $\delta\delta$-waarde voor Velthems samples weergegeven en dit voor verschillende sample-groottes. Naast deze gemiddelde waarde (μ) is telkens ook het gemiddelde min en plus de standaarddeviatie (σ) voor de bewuste $\delta\delta$-waardes weergeven, hetgeen een blik werpt op de stabiliteit van deze waardes in elk experiment.[11] Beide figuren bevestigen opnieuw dat de gemiddelde $\delta\delta$-waardes voor Velthems samples een acceptabel

standaarddeviatie

10 Een technische noot: deze berekeningswijze gaat terug op een suggestie van een anonieme peer reviewer van een eerste versie van een artikel voor het tijdschrift *Queeste* (vgl. Kestemont 2010a, 29). Aanvankelijk had ik $\delta\delta$ gewoon gedefinieerd als $\delta 2 - \delta 1$ maar de formule voorgesteld door de reviewer is veel genuanceerder, omdat deze niet alleen het verschil tussen beide δ-waardes beschouwt, maar ook de *eigenlijke* δ-waardes. Immers: voor het geval (a) waarin $\delta 1 = 100$ en $\delta 2 = 101$ kregen we anders dezelfde $\delta\delta = 1$ dan voor geval (b) met $\delta 1 = 1$ en $\delta 2 = 2$. De nieuwe berekeningswijze levert voor (a) $\delta\delta = -0,0099$, hetgeen een toeschrijving met beperkte zekerheid impliceert en voor (b) $\delta\delta = -0,5$ hetgeen relatief hogere zekerheid impliceert. De nieuwe berekeningswijze benadert de classificatie-werkelijkheid meer getrouw omdat *beide* samples in (a) relatief verder liggen van het test item dan in (b).

11 De 'standaarddeviatie' is een wijd verbreide statistische maat die uitdrukt wat de gemiddelde afwijking is van een reeks waardes ten opzichte van het gemiddelde van die waardes; vgl. Van Dalen-Oskam & Van Zundert (2007). De maat geeft bijgevolg een beter inzicht in de algemene variatie tussen de waardes. Zie ook § 7.2 en § 6.2 (in het bijzonder Tab. 19, p. 140).



(a) $\delta 1 << \delta 2$ en $\delta\delta << 0$

(b) $\delta 1 < \delta 2$ en $\delta\delta < 0$

(c) $\delta 1 > \delta 2$ en $\delta\delta > 0$

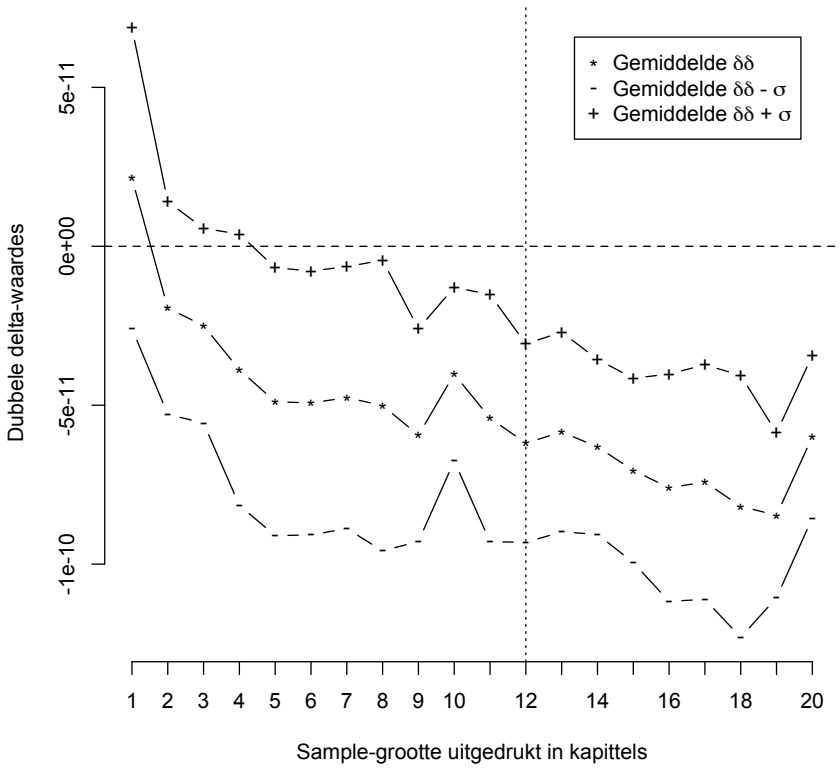
(d) $\delta 1 >> \delta 2$ en $\delta\delta >> 0$

Figuur 13: Naïeve visuele voorstelling van enkele fictieve situaties binnen een *nearest neighbours*-redenering. Het test item van de klasse V wordt in subpaneel (a) met grote zekerheid correct aan V toegeschreven, in (b) met een kleinere zekerheid. In subpaneel (c) wordt het test item met beperkte zekerheid foutief toegeschreven aan X en in (d) wordt dezelfde fout gemaakt maar met een grotere zekerheid.

dieptepunt bereiken vanaf een sample-grootte van 12 kapittels – vgl. de verticale stippellijn in Fig. 14 (p. 112) en 15 (p. 113).

5.2.4 *De Vijfde Partie*

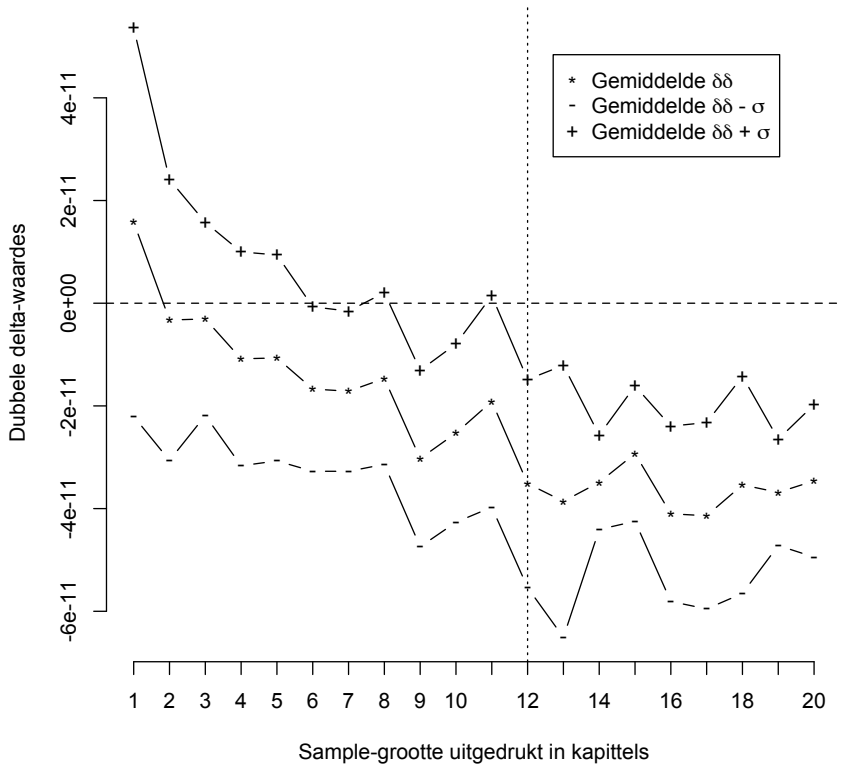
Om een lang verhaal kort te maken: experimenten met de Tweede, Derde en Vierde Partie tonen aan dat ons classificatie-algoritme vanaf een sample-grootte van 12 kapittels een betrouwbaar onderscheid weet te maken tussen het styloom van Velthem en zijn twee voorgangers. De $\delta\delta$-metriek die werd voorgesteld blijkt bovendien in staat om te meten hoe typisch 'Velthemiaans' een sample is: hoe lager deze waarde voor een sample, hoe meer het lijkt op Velthems stijl. Hoe hoger deze waarde voor een sample, hoe verder het is verwijderd van Velthems typische stijl in de Vierde Partie. Gewapend met deze techniek kunnen



Figuur 14: Grafiek voor een reeks *leave-one-out*-experimenten waarin Velthems Vierde Partie wordt afgezet tegens Utenbroekes Tweede Partie. Verschillende delta-waardes zijn uitgezet als een functie van het aantal kapittels dat per sample werd gebruikt. Naast de gemiddelde δδ-waardes is telkens ook dit gemiddelde min en plus de standaarddeviatie (σ) in de δδ-waardes afgebeeld. Vanaf 12 kapittels (verticale stippellijn) bevindt het gemiddelde min σ zich voldoende onder nul.

we nu trachten te achterhalen of er zich merkwaardige stijlovergangen in Velthems Vijfde Partie voordoen.

Voor het samplen van de Vijfde Partie kunnen we gebruik maken van een werkwijze die eerder werd gebruikt door Van Dalen-Oskam & Van Zundert (2007). Daarbij worden de kapittels van de Vijfde Partie in opeenvolgende, niet-overlappende samples of *windows* verdeeld van een gelijke grootte. Hierboven is aangetoond dat we goede resultaten verkregen vanaf een sample-grootte van 12 kapittels dus verdelen we ook de beschikbare samples uit de Vijfde Partie in opeenvolgende groepjes van 12 kapittels. Deze verdeling houdt het midden tussen, enerzijds, een zo betrouwbaar mogelijke auteursherkenning en, anderzijds, de granulariteit van de analyse, want té grote samples leveren metingen op die niet fijn genoeg zijn. Vervolgens gaan we door de verkregen samples en berekenen steeds de δδ-waarde van een sample,



Figuur 15: Grafiek voor een reeks *leave-one-out*-experimenten waarin Velthems Vierde Partie wordt afgezet tegens Maerlants Derde Partie. Verschillende delta-waardes zijn uitgezet als een functie van het aantal kapittels dat per sample werd gebruikt. Naast de gemiddelde δδ-waardes is telkens ook dit gemiddelde min en plus de standaarddeviatie (σ) in de δδ-waardes afgebeeld. Vanaf 12 kapittels (verticale stippellijn) bevindt het gemiddelde min σ zich voldoende onder nul.

gegeven het trainingsmateriaal waarin Velthem telkens met een van zijn voorgangers wordt geconfronteerd.

Het resultaat van deze werkwijze wordt in onderstaande grafieken weergegeven (Fig. 16, p. 115 en Fig. 17, p. 116). In deze grafieken stelt de horizontale as de Vijfde Partie voor, zoals onderverdeeld in opeenvolgende *windows* van 12 kapittels. De afgebeelde curve geeft de δδ-waarde weer voor elk van deze samples ten aanzien van het trainingsmateriaal (*P2* vs. *P4* en *P3* vs. *P4*). De horizontale nul-lijn weerspiegelt de afwezigheid van enige zekerheid in de meting. De andere horizontale lijnen bieden een leeshulp en stellen de gemiddelde δδ-waarde voor in dit experiment, alsook dit gemiddelde min en plus de standaarddeviatie. De grijze zones in de grafiek markeren zones in de tekst die ons bijzonder interesseren. Achtereenvolgens gaat het om

de eerste Heelu-interpolatie, de tweede en derde Heelu-interpolatie en het hele vierde boek.[12]

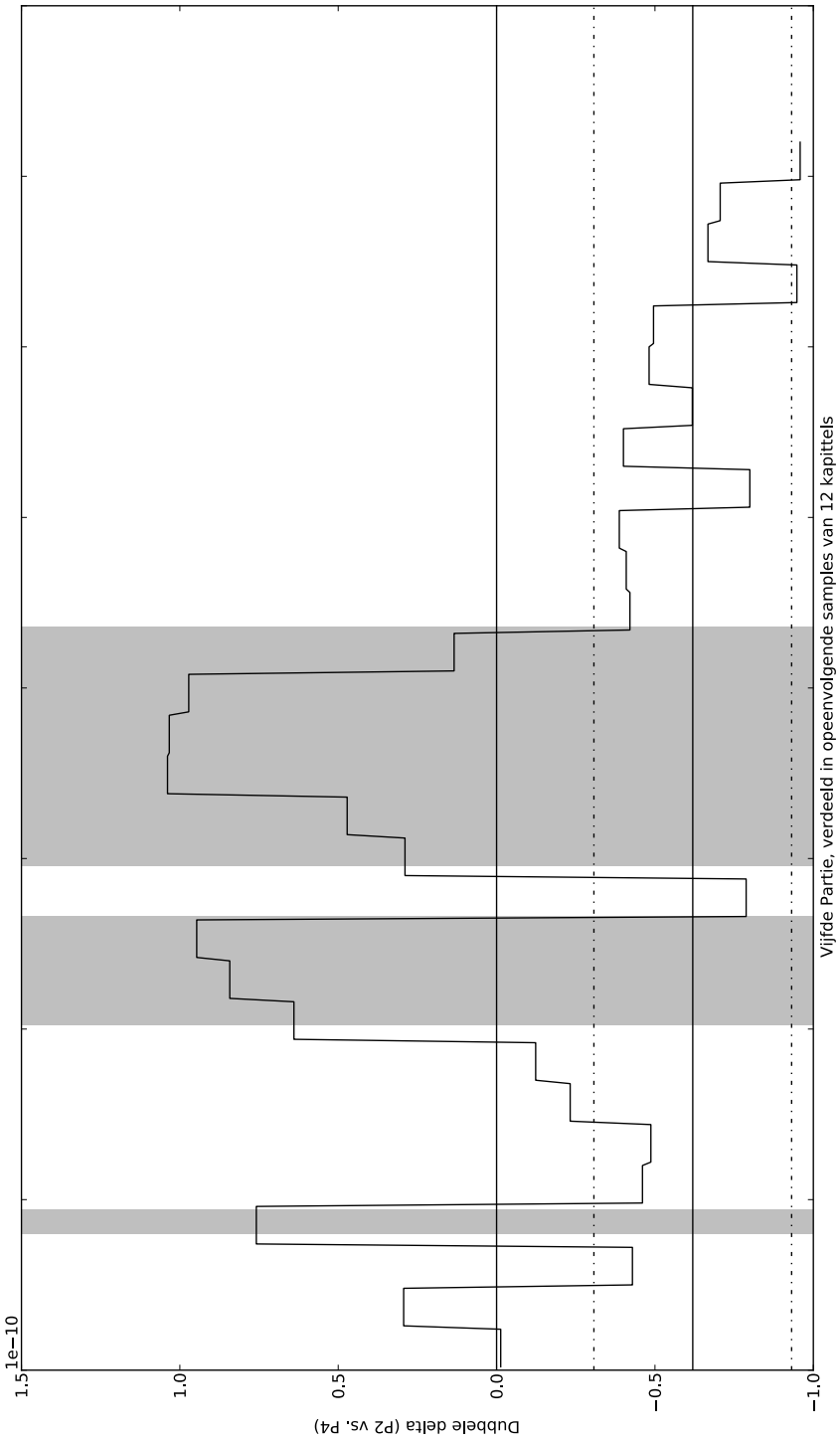
Het is opvallend dat beide figuren subtiele afwijkingen vertonen, maar dat uit zowel het experiment met Utenbroeke als dat met Maerlant dezelfde trends zijn af te lezen (Fig. 16, p. 115 en Fig. 17, p. 116). Beschouwen we de eerste twee grijze zones links in de grafiek: van deze zones in de tekst staat vast dat Velthem ze heeft ontleend aan een andere auteur. De $\delta\delta$-waardes voor de samples in deze zones zijn onveranderlijk bijzonder hoog. In beide experimenten is de classifier erg zeker van het verwerpen van Velthems auteurschap, hetgeen aan onze verwachtingen beantwoordt. Als we anderzijds de laatste boeken van de Vijfde Partie bekijken, zien we dat de meeste $\delta\delta$-waardes erg negatief zijn, waaruit we mogen concluderen dat de classifier erg overtuigd is van Velthems auteurschap voor deze samples. Dat is opvallend, want in deze laatste boeken schrijft Velthem een inhoudelijk afwijkend soort kapittels met onder meer toekomstvisioenen over het einde der tijden. Toch wijzen de extreem lage $\delta\delta$-waardes uit dat deze passages stilistisch nauw aansluiten bij de rijmstijl van de Vierde Partie.

Over het algemeen beantwoordt deze analyse dus aan onze verwachtingen want de methode detecteert relatief netjes de auteursovergangen die bijvoorbeeld met de Heelu-passages zijn gemoeid. De zaken staan er echter geheel anders voor in het geval van de derde grijze zone, die correspondeert met het vierde boek van de Vijfde Partie. Hier zijn de $\delta\delta$-waardes opvallend en onverwacht hoog, soms zelfs hoger dan voor de voorgaande Heelu-passages, waarvan wij zeker weten dat Velthem ze voor het grootste deel niet zelf heeft geschreven. Beide classifiers verwerpen zo met grote zekerheid en unaniem Velthems auteurschap voor dit vierde boek, in het bijzonder het middengedeelte van dit boek dat samenvalt met eigenlijke relaas van de slag. Dit resultaat sterkt ons in de twijfel die door eerdere experimenten al werd gevoed: een methode die Velthems stijl in de Vierde Partie feilloos kan onderscheiden van het werk van twee representatieve voorgangers verwerpt zijn auteurschap voor het vierde boek van de Vijfde Partie. Hoe mogen wij een dergelijk uitgesproken resultaat duiden?

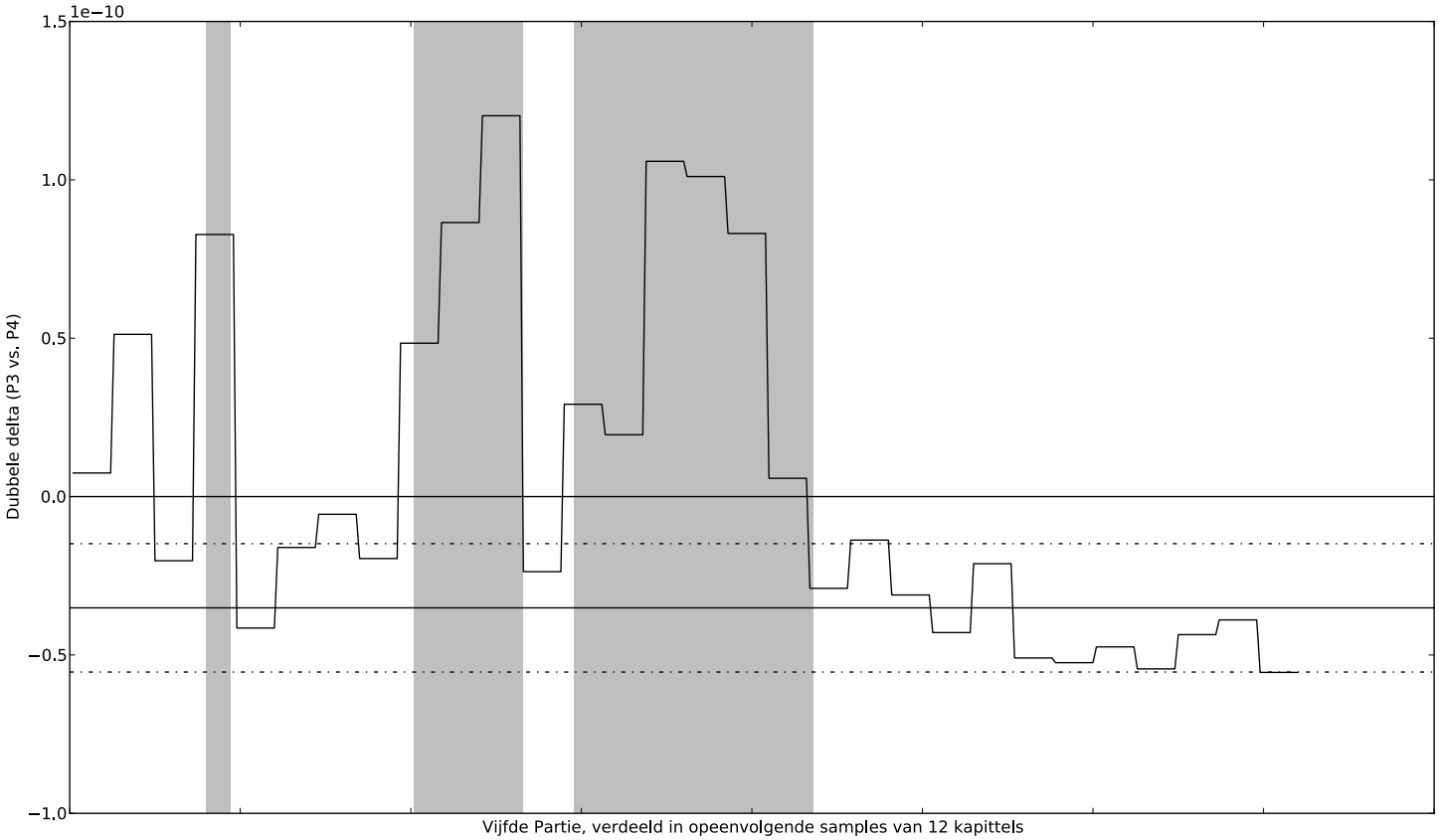
5.2.5 *Een palimpsest?*

Het is gevaarlijk om Velthems auteurschap voor het vierde boek van de Vijfde Partie al te haastig te verwerpen. Een cruciale verwijzing in dezen is een artikel door Van den Berg (1992a), waarin deze Ne-

12 De eerste grijze zone bevat dus de Heelu-interpolatie in Boek 1, vs. 2824-3543 (kapittels 40-46). De tweede grijze zone omvat de Heelu-interpolaties in boek 2, vs. 3220-4032 (kapittels 44-55) en boek 3, vs. 71-1388 (kapittels 2-21). Zie voor de interpolaties de studies van Besamusca (2009, 83) en Mantingh (2009). Het eerste kapittel van het derde boek dat traditioneel wél aan Velthem wordt toegeschreven is hier omwille van de duidelijkheid niet onderscheiden. Het bevindt zich ongeveer halfweg in de tweede grijze zone.



Figuur 16: In deze grafiek stelt de horizontale as P_5 voor, zoals onderverdeeld in opeenvolgende samples van 12 kapittels. De afgebeelde curve geeft de $\delta\delta$-waarde weer voor elk van deze samples ten aanzien van het trainingsmateriaal (P_2 vs. P_4). De horizontale nul-lijn weerspiegelt de afwezigheid van enige zekerheid in de meting. De andere horizontale lijnen bieden een leeshulp en stellen de gemiddelde $\delta\delta$-waarde voor in dit experiment, alsook dit gemiddelde min en plus de standaarddeviatie. De grijze zones stellen volgende gebieden voor: de eerste Heelu-interpolatie, de tweede en derde Heelu-interpolatie en het vierde boek.



Figuur 17: In deze grafiek stelt de horizontale as $P5$ voor, zoals onderverdeeld in opeenvolgende samples van 12 kapittels. De afgebeelde curve geeft de $\delta\delta$-waarde weer voor elk van deze samples ten aanzien van het trainingsmateriaal ($P3$ vs. $P4$). De horizontale nul-lijn weerspiegelt de afwezigheid van enige zekerheid in de meting. De andere horizontale lijnen bieden een leeshulp en stellen de gemiddelde $\delta\delta$-waarde voor in dit experiment, alsook dit gemiddelde min en plus de standaarddeviatie. De grijze zones stellen volgende gebieden voor: de eerste Heelu-interpolatie, de tweede en derde Heelu-interpolatie en het vierde boek.

derlandse filoloog de resultaten besprak van een stilistische analyse van een groot corpus Middelnederlandse epische teksten. Van den Berg merkte op dat in het vierde boek van de Vijfde Partie opvallend minder gebruik werd gemaakt van noodrijmen en epitheta dan in de rest van de Vijfde Partie. Van den Bergs duiding van deze spitsvondige observatie was kort maar boeiend. Hij dacht niet meteen aan een auteursverschil maar suggereerde een andere interpretatie:

> Is hij bij de weergave van een gebeurtenis die op hem als tijdgenoot diepe indruk gemaakt moet hebben, beter op dreef geweest dan bij die van verderweg gelegen stof? En moet de eenvormigheid van zijn epitheta [...] in hetzelfde licht gezien worden? (Van den Berg 1992a, 213)

Van den Berg vroeg zich dus af of Velthem zijn stijl – meer of minder bewust – had aangepast aan het onderwerp van de slag. Het is alleszins duidelijk dat Velthem in de Guldensporenslag een belangrijke historische gebeurtenis herkende, zodat men de mogelijkheid niet kan uitsluiten dat Velthem – die elders bekend staat als een slordige rijmelaar – zijn stijl voor de gelegenheid extra verfijnd heeft.[13]

Van den Bergs vroege detectie van de afwijkende stijl in het vierde boek is lovenswaardig, maar de interpretatie van dit gegeven blijft vooralsnog moeilijk. We hebben namelijk maar een beperkt zicht op de stilistische schommelingen die zich konden voordoen binnen het oeuvre van een middeleeuwse auteur. Het is erg onzeker, zoniet twijfelachtig, of een auteur zijn rijmstijl wel zo flexibel kon aanpassen aan een tekstonderwerp als Van den Berg aanneemt. Beschouwen wij bijvoorbeeld Velthems erudiete behandeling en anthologie van Bernardus van Clairvaux in de Vierde Partie (zie Fig. 18, p. 118). Velthem blijkt een groot liefhebber van Bernardus' werk te zijn geweest en trekt een ongewoon groot aantal verzen voor de bespreking van deze figuur uit. Het is daarenboven duidelijk dat Velthem in verschillende van de Bernardus-passages een meer hoogstaande literaire stijl nastreeft (Warnar 2009, 128-129). Toch heeft deze poging geen noemenswaardig effect gehad op het groot aantal stoplappen in deze passages. Dat maakt het twijfelachtig of de stilistische afwijkingen in boek 4 wel zo makkelijk kunnen worden afgedaan als een poging van Velthem om de stilistische kwaliteit van tekst te verbeteren, vanwege het belang dat hij aan het onderwerp hechtte.

Toch blijft voorzichtigheid geboden. In mijn duiding van de eerder besproken grafieken stelde ik dat de opeenvolgende $\delta\delta$-scores goed de auteursovergang wisten te vatten tussen de oorspronkelijke Velthem-passages en de passages die Velthem aan Heelu ontleende. Meestal zien we immers hoe de classifier een positieve $\delta\delta$-waarde aangaf voor de samples in de grijze zones, maar een negatieve voor

13 Ik dank deze observatie aan Remco Sleiderink, met wie ik boeiende conversaties heb mogen voeren over dit onderwerp.

Figuur 18: Fol. 71v-72r in het Heber-Serrure-handschrift dat een van de belangrijkste bronnen vormt voor Velthems aandeel in de Vierde Partie (Gent, Universiteitsbibliotheek, hs. 1476). De afgebeelde folia komen uit Velthems behandeling van het werk van Bernardus van Clairvaux. Hoewel het handschrift – kartuis van Herne, ca. 1395 Kwakkel (2002, op index) – enkel berijmde teksten bevat, zijn de verzen doorlopend neergeschreven als proza, hetgeen eind veertiende eeuw ongebruikelijk was.

de samples buiten de grijze zones. Dat is echter niet het geval voor de eerste twee samples uit de Vijfde Partie, aangezien deze of niet aan Velthem worden toegeschreven ($\delta\delta > 0$) of maar nipt wel ($\delta\delta \approx 0$). Dit effect schijnt prominenter in het Maerlant- dan in het Utenbroeke-experiment (vgl. Fig. 16, p. 115 en Fig. 17, p. 116). De bewuste samples vallen grotendeels samen met een relatief trouw vertaalde passage uit het *Speculum historiale*, met name de kapittels 4 tot 19 uit het eerste boek.

Deze passage valt op omdat zij een gedetaileerde geografische beschrijving biedt in het kader van enkele kapittels over de geschiedenis van het Midden-Oosten.[14] Dat leidt ertoe dat de rijmwoordenschat in deze passages danig afwijkt van Velthems gangbare rijmstijl en bijvoorbeeld ongewoon veel geografische eigennamen (PrName) in rijmpositie bevat. Vanzelfsprekend moet men daarom niet meteen geneigd zijn om Velthems auteurschap voor deze kapittels op de helling te zetten. Veeleer schijnt de onzekerheid van de classifier te worden veroorzaakt door het feit dat deze passages erg gelijken op erg beschrijvende samples van Utenbroeke of Maerlant die mogelijk ook veel eigennamen bevatten. De onderscheidende rijmwoorden worden in de globale frequentie-massa van de bewuste kapittels immers onderdrukt door de vele eigennamen in rijmpositie.

Dit roept de vraag op of onze techniek wel enkel het styloom van een auteur meet en ongevoelig is voor andere factoren in een tekst zoals het onderwerp (bv. geografische beschrijving). Ook deze observatie noopt tot voorzichtigheid in het verwerpen van Velthems auteurschap van het vierde boek. Toch loont het de moeite om ook via traditioneel onderzoek na te gaan hoe waarschijnlijk het is dat Velthem het vierde boek zelf heeft geschreven. De bijzonder hoge $\delta\delta$-waardes in boek vier suggereren immers bijzonder sterk dat wij in het vierde boek mogelijk voor een 'palimpsest' staan en dat Velthem een reeds bestaande tekst van een andere auteur zonder grote aanpassingen in zijn Vijfde Partie heeft ingebouwd.

In de medioneerlandistiek staat Lodewijk van Velthem hoe dan ook reeds bekend als een fervent compilator.[15] De casus van Heelu is reeds uitvoerig aan bod gekomen maar ook uit Velthems Arturroman, de *Merlijn-continuatie* weten we dat de Brabander niet terugdeinsde voor geavanceerd *copy-pasten* (vgl. § 8.1).[16] Bovendien wordt Velthem meestal gelijkgesteld met de samensteller van de *Lancelotcompilatie*. Een anonieme compilator – waarschijnlijk Velthem – heeft in deze omvangrijke codex een collectie reeds bestaande Arturromans samengevoegd tot één grote bloemlezing (vgl. § 8.2). Hoewel de compilator de reeds bestaande romans niet zelf kan hebben geschreven, worden (de namen van) de oorspronkelijke auteurs door de compilator ner-

14 Zie o.m. De Vries & Verwijs (1982, dl. 1, LXXXIII) en Besamusca (2009, 78).
15 Zie Besamusca (2009, 84-85) en Mantingh (2009).
16 Hij integreerde een tekst over Maskeroen in een tekst van Jacob van Maerlant; zie Besamusca et al. (2009b, 12-13) of Sleiderink (2009, 128, n. 29).

gens vermeld. In dat opzicht zou het alvast niet verbazen dat Velthem, klaarblijkelijk een 'literaire meesterdief', de mosterd voor zijn vierde boek elders is gaan halen.

Een ander gewichtig argument tegen Velthems auteurschap betreft de chronologie. We weten dat Velthem de ca. 23.000 verzen van zijn Vijfde Partie rond kreeg in een periode die maar een goed jaar besloeg.[17] Voor dit werk had Velthem geen uniforme brontekst meer ter beschikking. Hij moest zelf al zijn informatie zien te verzamelen uit disparate bronnen en deze zien te combineren tot een coherent geheel. Men heeft zich al vaker afgevraagd hoe mensen als Velthem zulke grootschalige initiatieven binnen een dergelijk beperkte tijdsspanne wisten af te ronden.[18] Opnieuw zou het niet verbazen dat Velthem hierin tenminste deels is geslaagd, doordat hij grote lappen tekst aan andere werken kon ontlenen.

In datzelfde opzicht is het merkwaardig dat het vierde boek uitzonderlijk lang is: met zijn 5.329 verzen is het opvallend uitgebreider dan het tweede langste boek, namelijk het eerste, dat amper 4.307 verzen telt. Het zou dus een zinvolle veronderstelling zijn dat Velthem op een gegeven moment een vrij uitgebreide brontekst over de Guldensporenslag in handen kreeg en die gulzig opnam in zijn Vijfde Partie. Mantingh (2009, 221, ook n. 43) heeft bovendien betoogd dat de uitzonderlijke lengte van het vierde boek een soort 'domino-effect' teweeg heeft gebracht in de compositie van de Vijfde Partie: Velthem had kennelijk niet verwacht dat dit boek zo lang zou uitvallen, waardoor hij moest afwijken van de structuur die hij oorspronkelijk had voorzien en moest gaan schuiven met andere tekstdelen.

Een ander argument is reeds besproken door historici: Velthem is voor een Brabander die meer dan een decennium na de eigenlijke feiten schrijft verdacht goed ingelicht (Trio 2002, 100-105). Hij kan zelfs in detail de weersomstandigheden op het slagveld beschrijven. Historici hebben zich afgevraagd hoe Velthem aan dergelijk specifieke informatie kan zijn geraakt. Heeft hij, bijvoorbeeld, Vlaamse oudstrijders kunnen interviewen? Opnieuw doet dit gegeven vermoeden dat er een brontekst in het spel is, misschien zelfs een brontekst die relatief kort na de slag kan zijn geschreven. Voor de auteur van een dergelijke (Vlaamse?) brontekst zou het veel logischer zijn dat deze een ooggetuige was geweest van de veldslag en zo beschikking had over betrouwbare informatie uit de eerste hand – net als Heelu (vgl. Sleiderink 2003, 87-97). We kunnen immers veronderstellen dat Velthem Heelu's *Rijmkroniek* tot de ooggetuigenverslagen rekende die hij naar eigen zeggen als bronnen heeft geconsulteerd voor zijn Vijfde Partie (Besamusca 2009, 75-76). Als Velthem ook voor zijn weergave van de Guldensporenslag teruggaat op de (toch bijzonder gedetailleerde)

17 Hij rondde de Vierde Partie af op 3 augustus 1315 en de eerste zes boeken van de Vijfde Partie waren af op 14 augustus 1316. Zie Sleiderink (2005) en Besamusca et al. (2009b, 11-12).

18 Zie Trio (2002, 90-92), Mantingh (2009, 207-211) en Van Oostrom (1984).

brontekst van een andere auteur, kan hij deze bron ook impliciet onder de noemer van de 'ooggetuigenverslagen' hebben ondergebracht.

5.2.6 *De Vlaamse brontekst*

Als we ervan uitgaan dat Velthem voor zijn vierde boek een beroep heeft gedaan op een reeds bestaande brontekst, is het interessant om dit boek eens als een zelfstandige literair-historische eenheid uit de Vijfde Partie te lichten. Het relaas van de Guldensporenslag is bijvoorbeeld doorspekt met referenties aan de Karelepiek, het legendarische middeleeuwse verhalencomplex rond Karel de Grote.[19] Zo lezen wij bijvoorbeeld:

> En bleven niet in Alisant
> te genen wige daer Rolant vacht
> te Roncevale met groter cracht,
> alsi vor Cortrike daden
> (Jongen & Piters 2002, 154, kap. 30, vs. 18-21)

Het paard van de graaf van Artois blijkt bovendien de naam 'Morel' te dragen, hetgeen een naamgeving is die werd ontleend aan de Karelepiek.[20] Elders wordt het bloed van de Fransen 'Saraceens' bloed genoemd, wat een ander element is dat nogal Karolingisch aandoet.[21] Even verderop lezen we zelfs:

> Men seit vele van Roncevale,
> maer dits die jammerlixste tale,
> daermen yegeren af mag tellen.
> (Jongen & Piters 2002, 156, kap. 31, vs. 7-9)

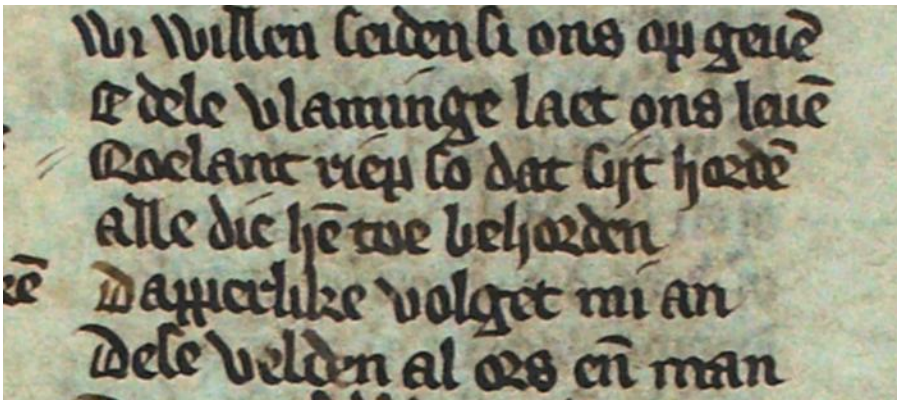
Het meest verrassende element volgt kort op de plaats in de tekst waar deze expliciete link met de Karelepiek wordt gelegd. Het gaat om een klein maar frappant detail in de Leidse Vijfde Partie-codex dat illustreert hoe sterk de impact van dergelijke intertekstuele referenties op het contemporaine publiek kan zijn geweest. Bovenaan de derde kolom van fol. 49r handschrift LEIDEN, UNIVERSITEITSBIBLIOTHEEK, BPL E 14 lezen we (in een diplomatische weergave):

> Roelant riep so dat sijt horde*n*
> Alle die he*m* toe behorden
> Dapperlike volget mi an

19 Hieronder verwijs ik consequent naar de jongste kritische editie van het vierde boek door Jongen & Piters (2002).

20 Zie kapittel 24 bij Jongen & Piters (2002, 130).

21 Zie vers 46 in kapittel 27 bij Jongen & Piters (2002, 140). Trio (2002, 101) ontwaart bovendien nog veel meer referenties aan Karelstof, maar het is moeilijk om hem daarin consequent bij te treden. Erg twijfelachtig lijkt mij bijvoorbeeld de vermeende referentie naar *Van den bere Wisselau*.



Figuur 19: Uitvergrote afbeelding van het bovenste deel van de derde kolom op fol. 49r in het enige complete handschrift van de Vijfde Partie (Leiden, Universiteitsbibliotheek, bpl e 14), dat in de onmiddellijke omgeving van Velthem zou zijn ontstaan. De kopiist schreef *Roelant* waar men *Rinesse* zou verwachten. Dit lijkt een verschrijving van een (dagdromende?) kopiist onder invloed van een Karolingische referentie enkele verzen eerder.

Het eerste vers in dit citaat (zie ook Fig. 19, p. 122) vermeldt als spreker Roeland, de legendarische neef van Karel de Grote die vooral bekend is uit het *Chanson de Roland*.[22] Deze vermelding van *Roelant* is op dit moment in de tekst klinkklare nonsens, want tot dan is gewoon nog geen sprake geweest van dit personage. Het ziet ernaar uit dat hier eigenlijk een verwijzing moest komen naar Jan van Renesse, een van de aanvoerders in het Vlaamse leger. De jongste edities (Waterschoot 1979; Jongen & Piters 2002) hebben het verrassende *Roelant* dan ook begrijpelijk geëmendeerd tot *Rinesse*. Omdat deze vermelding slechts enkele regels na de expliciete vermelding van Karelepische stof wordt aangetroffen, is de kans groot dat we hier te maken hebben met een verschrijving van de (dagdromende?) kopiist die de figuur van Renesse associatief met *Roelant* zal hebben verward.

Wat nog het meeste opvalt aan de Karolingische referenties is dat zij vrij ongebruikelijk zijn in de context van Velthems oeuvre en zeker geen karakteristiek van zijn poëtica lijken. We weten wél dat Velthem een groot adept van de Arturepiek was (Besamusca 2009, 84-85). We zien zelfs hoe Velthem zijn geschiedverhaal in de *Spiegel historiael* niet zelden met 'Arturiaanse tinten' bijkleurt (Summerfield 2009). Het is daarom opvallend dat precies het vierde boek doorweven is met knipoogjes naar de Karelepiek, terwijl Velthem – voor zover wij weten – geen bijzondere band met die verhaalstof had. We weten natuurlijk niet zeker of Velthem de Arturepiek boven de Karelepiek verkoos, maar opnieuw is het toch veelzeggend dat deze afwijkende

22 Het vers wordt in de marge voorafgegaan door een klein merkteken (zie Fig. 19, p. 122) dat waarschijnlijk een paragraafteken moest worden, maar uiteindelijk nooit is uitgevoerd (Sleiderink 2009, 181).

intertekstuele referenties samenvallen met de aanzienlijke deviaties in stijl in het vierde boek.[23]

Het wordt stilaan duidelijk dat er met het vierde boek in de Vijfde Partie iets vreemd aan de hand is. Het lijkt steeds waarschijnlijker dat Velthem een reeds bestaande brontekst ca. 1316 in zijn eigen kroniek heeft ingebouwd. Het blijft natuurlijk mogelijk dat Velthem deze brontekst vroeger in zijn loopbaan zelf heeft geschreven, maar ook dat is gegeven de bovenstaande argumenten niet heel waarschijnlijk. Alleszins wijkt het vierde boek stilistisch zo sterk af van de rest van de Vijfde Partie die ca. 1315-1316 tot stand kwam, dat we de mogelijkheid moeten openhouden dat Velthem voor dit boek gebruik kon maken van een reeds bestaande blauwdruk, waarschijnlijk van een andere, anonieme auteur.

De hypothese omtrent Velthems brontekst (en het uitzicht daarvan) kan overigens verder worden uitgewerkt door een blik te werpen op contemporaine historiografische geschriften aangaande de Guldensporenslag aan Franse zijde.[24] Onmiddellijk na de nederlaag bij Kortrijk, vatte de Franse koning, Philippe le Bel, het plan op om strafexpedities op de been te brengen en de opstandige Vlaamse bevolking terug in het gareel te dwingen. Dat moest zo snel mogelijk gebeuren, aangezien het nieuws over de nederlaag zich snel verspreidde. De nederlaag plaatste Philippe in een uitermate slecht daglicht en kon aanvallen uitlokken van vijanden die hun kans schoon zagen om het verzwakte koninkrijk aan te vallen.

Voor die strafexpedities had de Franse kroon fondsen nodig, maar de publieke opinie keerde zich al snel tegen Philippe. De vernederende nederlaag lag nog vers in het geheugen van de teleurgestelde belastingbetalers. Zij stonden daarom weigerachtig tegenover de zware belastingen die nodig waren om de strafexpedities te financieren. Om het vertrouwen in zijn beleid te herstellen, schakelde Philippe zijn PR-machine in. Nog voor de herfst van 1302 vaardigde het hof een officiële verklaring uit voor de ongelukkige afloop in Kortrijk. De Vlamingen zouden volgens die 'officiële berichtgeving' op een onridderlijke, oneervolle manier gestreden hebben. Op het slagveld zouden zij heimelijk verraderlijke grachten hebben gegraven waarin de Franse ruiters bij het chargeren ten val waren gekomen en verdronken. Het was door deze sluwe en oneerlijke list dat de Vlamingen de slag hadden kunnen winnen. In de bronnen voorafgaand aan deze 'pamfletten' is echter nooit sprake van dergelijke verraderlijke grachten.

23 Een element dat ik zijdelings vermeld, is het voorkomen van volledige Franstalige verzen in het vierde boek; zie bijvoorbeeld (Jongen & Piters 2002, 150-152, kap. 29, vs. 71-75). Dergelijke frases – overigens indicatief voor een tweetalige dichter en/of publiek – heb ik nog niet elders in Velthems oeuvre aangetroffen.

24 Voor onderstaande paragrafen heb ik vooral gebruik gemaakt van Verbruggen (1952) voor mijn informatie over de Franse historiografische traditie (p. 60-63), de *Chronique Artésienne* (p. 63-68), Guiart (p. 73-84) en Trio (2002), in het bijzonder voor de grachten.

De eerste historiografische bron over de Guldensporenslag aan Franse zijde was de *Chronique Artésienne* (ca. 1304-1305). Deze stond duidelijk onder de invloed van de officiële berichtgeving en bevestigde veelal de Franse versie van de feiten. Enkele jaren later schreef een burger uit Orléans, Guillaume Guiart, een kroniek die hij opdroeg aan de Franse koning: *La branche des royaus lignages*. Deze auteur ging zich daarvoor uitgebreid documenteren in de abdij van Saint-Denis. Opvallend genoeg stelt de auteur in zijn kroniek dat hij een bevooroordeelde *Vlaamse* versie van de feiten wil weerleggen. Volgens Guiart was dit Vlaamse relaas al té suggestief en helemaal niet waarheidsgetrouw, omdat deze versie voorbijging aan al de nederlagen die de Vlamingen vóór 1302 hadden moeten lijden.[25] Het is daarentegen wel merkwaardig dat de kroniek van Guiart expliciet beschrijft hoe er op het slagveld veel kleine beekjes aanwezig waren. Niettemin wordt nooit zelfs maar de indruk gewekt dat deze verraderlijk waren of bewust waren aangelegd door de Vlamingen.[26]

Als we kijken naar het relaas dat is overgeleverd via de Vijfde Partie, moeten we zoals gezegd rekening houden met de mogelijkheid dat deze versie een 'palimpsest' vormt en teruggaat op een Vlaamse blauwdruk. Deze Vlaamse grondtekst kan al vrij snel na de slag zijn ontstaan, want de tekst wekt soms de indruk door een ooggetuige te zijn opgesteld. Interessant is de volgende passage die suggereert dat de oorspronkelijke auteur in zijn relaas een standpunt wou innemen met betrekking tot de bedrieglijke leugens die in Franse teksten werden verspreid:

> Nu doet u selc logen verstaen,
> ende secgen van dere gracht saen
> datse die Fransoyse en wisten niet,
> ende datsi daerombe hadden tverdriet.
> Dat es sceren ende groet spel.
> Si wisten alle dese gracht wel.
> (Jongen & Piters 2002, 158, kap. 31, vs. 17-22)

Deze auteur verwerpt duidelijk de voorstelling van de feiten die was geïnitieerd in de pamflettaire geschriften van het Franse hof in de herfst van 1302. Mogelijk kunnen we dit (waarschijnlijk toch Vlaamse) relaas zo nog beter in de tijd situeren. Als Guiart in zijn kroniek ca. 1307 refereert aan het Vlaamse relaas dat ons via het palimpsest in de Vijfde Partie is overgeleverd, mogen we dit Vlaamse verslag immers nog voor 1307 dateren. Het lijkt zelfs waarschijnlijk dat Velthems Vlaamse grondtekst vrij snel na de slag kan zijn tot stand gekomen (maar dus ten vroegste in de herfst van 1302).

Het blijft natuurlijk moeilijk om uit te maken of we het bovenstaande citaat wel aan de oorspronkelijke Vlaamse auteur mogen toeschrijven

25 Vgl. met Le Muisit (Trio 2002, 85-87).

26 Wel stelt de kroniek dat de Vlamingen hadden kunnen winnen door een reeks ándere onridderlijke praktijken (Trio 2002, 84-85).

en niet aan Velthem, aangezien die makkelijk wijzigingen in de tekst kan hebben doorgevoerd.[27] Wat wel merkwaardig blijft, is dat historici het verslag van de strijd zelf in de Vijfde Partie steevast om zijn nauwkeurigheid loven maar dat de verslaggeving aangaande de aanloop en naweeën van 11 juli, bulkt van de feitelijke onjuistheden (Trio 2002, 92). Dat kan mogelijk een argument zijn voor de hypothese dat Velthem ca. 1316 slechts beschikte over een grondtekst die de Guldensporenslag, en énkel de Guldensporenslag beschreef.

Opnieuw zou de aanname van een onderliggende brontekst voor het vierde boek een aantrekkelijke verklaring voor deze situatie bieden. Als de hypotheses hierboven waar zouden blijken, staan we met deze Vlaamse brontekst voor een literair werk dat in veel opzichten een Vlaamse pendant vormt van Heelu's Brabantse *Rijmkroniek van Woeringen*. Ook hier staan we immers voor een sterk bevooroordeeld berijmd verslag van een recente overwinning, waarschijnlijk gedicht door een ooggetuige. Hoewel het vierde boek reeds bekend staat om zijn hoge ouderdom en gedetailleerdheid, is hierboven betoogd dat dit verslag nog veel ouder en authentieker zou kunnen zijn dan nu wordt verondersteld.

5.3 CONCLUSIE

In dit hoofdstuk is de classificatie-techniek die in het vorige hoofdstuk werd toegepast op de hoogfrequente rijmwoorden van de Tweede en Derde Partie, uitgebreid naar de Voortzetting van de *Spiegel historiael* door Lodewijk van Velthem. Al in de eerste experimenten bleek uit de zogenaamde *confusion matrixes* dat het vierde boek van de Vijfde Partie stilistisch afweek van de rest van het dichtwerk, zodat de stylometrische resultaten vragen opwierpen over Velthems auteurschap voor dit deel van het werk (§ 5.1.2). Dit vierde boek bevat een van de vaakst bestudeerde passages uit Velthems oeuvre, namelijk de erg gedetailleerde beschrijving van de Guldensporenslag (1302, § 5.2.1). Een nadere analyse van de Vijfde Partie werd gebaseerd op de metriek van de 'Dubbele Delta' (§ 5.2.3) die aanleiding gaf tot een verdieping van onze behandeling van het geheugen-gebaseerd leren (*Memory-based learning*, § 5.2.2). Deze analyse verstevigde de observatie dat de stijl

27 Een problematiek waar ik hier niet verder op inga maar die zeker moet uitgeklaard worden is de precieze relatie tussen de Middelnederlandse tekst en de Franse historiografie. Historici hebben benadrukt dat er hier en daar fascinerende overeenkomsten bestaan tussen de Middelnederlandse tekst en Franse parallellen, bijvoorbeeld inzake de aantallen en namen van de gesneuvelden aan Franse zijde (Trio 2002, 91). Verder verwijst de Middelnederlandse tekst tot driemaal toe naar een geschreven bron: Waterschoot (1979, 19) vermeldt de regels 466, 489 en 1230 in zijn editie. Mij komt het voor dat voor elk van deze verwijzingen drie interpretaties mogelijk zijn: (a) Velthem verwijst naar een (Franse?) brontekst; (b) Velthems brontekst verwijst naar een (Franse?) brontekst; (c) Velthem verwijst naar zijn eigen (Middelnederlandse) brontekst.

van het vierde boek abnormaal sterk afwijkt van het styloom dat we elders in de Vierde en Vijfde Partie voor Velthem aantreffen (§ 5.2.4).

Daarom werd verkend of Velthem voor dit boek een beroep heeft kunnen doen op een reeds bestaande volkstalige rijmtekst die kort na de slag in Vlaanderen zou zijn ontstaan (§ 5.2.6). Er bestaan namelijk goede argumenten voor deze aanname, waaronder de ongewone lengte van het boek, de verrassende intertekstualiteit met de Karelepiek en de onverwachte gedetailleerdheid van het relaas (§ 5.2.5). Het heeft er daarom alles van dat Velthem in zijn dichtwerk zonder expliciete bronvermelding een reeds bestaande anonieme tekst heeft geïntegreerd na een oppervlakkige bewerking. Mogelijk biedt Velthems erg gedetailleerde grondtekst zelfs het verslag van een Vlaamse ooggetuige in de slag bij Kortrijk. Velthems brontekst lijkt zo een Vlaamse parallel te vormen voor de Brabantse Woeringen-kroniek van Heelu, die Velthem eveneens zonder bronvermelding in zijn geschiedverhaal verwerkte. Het verslag in de Vijfde Partie dat reeds heel betrouwbaar werd geacht, lijkt zo nog ouder en authentieker dan tot nu toe werd verondersteld.

6 MAERLANT IN MEERVOUD

6.1 EEN OEUVRE VAN FORMAAT

In het vorige hoofdstuk heb ik enige reserve uitgedrukt met betrekking tot de stabiliteit van de stijl van middeleeuwse auteurs (§ 5.2.5). Het is immers helemaal niet bewezen dat middeleeuwse auteurs gedurende hun carrière en voor teksten in verschillende genres consequent dezelfde stijl aanhielden. Daarom zou het interessant zijn om het oeuvre van één auteur eens vanuit een stylometrische invalshoek te benaderen.[1] Oeuvres of een verzameling teksten die allemaal met zekerheid aan dezelfde auteur kunnen worden toegeschreven, zijn echter schaars in de Middelnederlandse berijmde literatuur. Als er al het vermoeden bestaat dat één auteursnaam mag verbonden worden aan verschillende teksten, zijn deze toeschrijvingen niet zelden onzeker en meestal gaat het om een klein aantal teksten.

Het oeuvre van de dertiende-eeuwse Vlaming Jacob van Maerlant steekt niettemin met kop en schouders boven de rest uit. Zowel qua omvang als diversiteit is de verzameling rijmteksten die aan deze dichter wordt toegeschreven uniek. Zelfs in Europees opzicht is de omvang van Maerlants oeuvre relatief ongewoon. Hoewel Maerlants teksten zelden expliciet gedateerd zijn, staat de chronologie ervan nog maar zelden ter discussie.[2] Zijn vroegst bekende werk is *Alexanders Geesten*, een oosterse ridderroman over het leven van Alexander de Grote, die Maerlant in hoofdzaak bewerkte naar de Latijnse *Alexandreis* van Gauthier de Chatillon. Aan het slot van het vijfde boek, verwijst Maerlant naar een tolgeschil nabij Rupelmonde dat in 1257 ontstond. Omdat de Brabantse hertog die in dit geschil als scheidsrechter optrad in 1261 is gestorven, moeten wij *Alexanders geesten* dus in de periode 1257-1261 situeren (Van Oostrom 1996, 113-114).

Maerlant moet kort daarop een tweede ridderroman hebben geschreven, namelijk de Arturiaanse dubbelroman *Historie van den Grale – Boek van Merline* over de vroegste regeringsjaren van de legendarische koning Artur en zijn bekende raadgever/tovenaar Merlijn. Maerlant vertaalde deze dubbelroman naar de Franse proza-adaptaties van de *Estoire dou Graal* en *Merlin en prose* door Robert de Boron. Omdat Maerlant in de proloog van deze bewerking expliciet terugwijst

1 De inzichten in dit hoofdstuk ontleen ik in grote mate aan een inspirerend onderzoek dat ik mocht uitvoeren met mijn tweede promotor, Walter Daelemans, en mijn goede collega Dominiek Sandra (Kestemont et al. 2012). Ik ben hen zeer erkentelijk voor de inspirerende samenwerking.

2 Onderstaande bespreking gaat deels terug op die door Caers & Kestemont (2011). Ik laat Maerlants kortere (strofische) gedichten verder onbesproken.

Figuur 20: Een auteursportret uit het prachtige Haagse handschrift van Maerlants aandeel in de *Spiegel historiael* (Den Haag, Koninklijke bibliotheek, ka xx, fol. 3v). Waarschijnlijk wordt Maerlant hier zelf afgebeeld, hoewel niet valt uit te sluiten dat het om Vincentius van Beauvais gaat, de auteur van de brontekst van de *Spiegel*.

naar *Alexanders geesten* staat de volgorde van deze twee werken vast (Van Oostrom 1996, 117). Aangezien de *Merlijn*tekst in de proloog is opgedragen aan Albrecht, heer van Voorne, weten we dat deze tekst na 1259 is geschreven, omdat Albrecht pas in dat jaar zijn vader is opgevolgd als heer van Voorne.

Maerlants derde tekst was een ridderroman over de Rondetafelridder *Torec*, die hij bewerkte naar een nu verloren gegane Franse Arturroman. De proloog van Maerlants daaropvolgende tekst, de *Historie van Troyen*, vermeldt immers dat de dichter op dat moment reeds *Allexander*, *Merlyn* en *Toerecke* had geschreven. De omvangrijke *Historie van Troyen* bewerkte Maerlant naar de Franse hoofse ridderroman *Roman de Troie*, een populair verslag door Benoît de Saint-Maure van de Trojaanse oorlog. Een zekere Segher Diengotgaf had volgens Maerlants tekst reeds een deeltje van de *Roman de Troie* vertaald (Kestemont 2010b). Maerlant vermeldt in zijn proloog dat hij deze vertaling zal overnemen en in de hoofdtekst zien we inderdaad hoe hij precies aangeeft waar hij Seghers verzen interpoleert.

Men gaat ervan uit dat de *Historie van Troyen* is geschreven vóór de *Heimelijkheid der Heimelijkheden*. Men redeneert dat de *Heimelijkheid* anders wel zou vermeld zijn in de proloog van Maerlants Trojeroman. De *Heimelijkheid* is een vorstenspiegel die Maerlant bewerkte naar het Latijnse *Secretum secretorum*. De tekst biedt de gefictionaliseerde weergave van een dialoog tussen Alexander de Grote en diens bekende leermeester Aristoteles. Hoewel Maerlant hiermee deels teruggrijpt naar stof die hij eerder reeds heeft behandeld, is de *Heimelijkheid* geen ridderroman. Men kan de tekst veeleer situeren binnen het didactische genre. Men gaat ervan uit dat Maerlant deze tekst in 1266 zou hebben geschreven, mogelijk naar aanleiding van het begin van de regering van de jonge Floris v, graaf van Holland. Hoewel die datering een geloofwaardige optie is, mag men niet vergeten dat hiervoor geen sluitend bewijs bestaat.[3]

Maerlants volgende tekst is *Der naturen bloeme*, een bewerking op rijm van *De natura rerum* door Thomas van Cantimpré. Deze omvangrijke natuurencyclopedie is opnieuw eerder in het didactische genre te situeren. Deze tekst zou Maerlant ca. 1270 hebben geschreven, kort voordat hij in 1271 aan zijn *Rijmbijbel* begon. De *Rijmbijbel* behandelt Bijbelse geschiedenis en is zo sterk historiografisch van inslag. Een dergelijk geschiedkundige karakter is ook eigen aan Maerlants laatste twee narratieve teksten, namelijk *Sinte Franciscus Leven* en de *Spiegel historiael*. Traditioneel projecteert men een driedeling op Maerlants oeuvre, want zijn werk vertoont een duidelijke ontwikkeling: hij heeft zijn carrière aangevat met een aantal verhalende ridderromans. Halfweg zijn oeuvre schakelde hij over op een aantal belerende teksten die eerder in het didactische genre thuishoren. Op het einde van zijn loopbaan schreef Maerlant vooral teksten met een historiografische inslag.[4]

In dit hoofdstuk zal ik deze driedeling voor waar aannemen, al moet ik daarbij opmerken dat deze niet zonder problemen is. Het is bijvoorbeeld juist dat Maerlants eerste drie romans alle behoren tot wat wij vandaag de dag gemakshalve 'ridderepiek' noemen. Niettemin bestaan er aanzienlijke verschillen inzake stofsoort tussen zijn eersteling, een oosterse Alexanderroman, en de twee Britse Arturromans die erop volgen. Dat *Der naturen bloeme* en *Heimelijkheid* beide belerende trekjes hebben, neemt bovendien niet weg dat er tussen deze teksten zowel naar aard, inhoud en lengte duidelijke verschillen bestaan. Het etiket 'didactiek' voor deze teksten is daarom wel zinvol,

3 Te Winkel (1922, 476-477) opperde als eerste dat de *lieve neve* aan wie de dichter het werk opdraagt, Floris v was. De 1266-these (vgl. Gerritsen 1979, 82) vormt in veel opzichten de ruggengraat van Van Oostroms invloedrijkste publicaties, zoals zijn beeldvormend artikel 'Hoe snel dichtten middeleeuwse dichters?' (Van Oostrom 1984).

4 Maerlant schreef gespreid over zijn carrière ook een aantal kortere, strofische gedichten waaronder bijvoorbeeld de *Martijn*-gedichten of *Van den lande van oversee* (Van Oostrom 1996, *passim*). Ik laat deze verder onbesproken omdat zij wegens hun beperkte lengte niet meteen bruikbaar zijn voor mijn onderzoek.

maar maskeert ook veel van de eigenheid van deze teksten. Ook de teksten uit Maerlants derde fase vertonen onderling meer verschillen dat het naïeve label 'historiografie' suggereert. Het type geschiedschrijving dat we aantreffen in de *Spiegel historiael* durft bijvoorbeeld sterk af te wijken van dat in de *Rijmbijbel* en *Sinte Franciscus Leven*, twee teksten die misschien makkelijker in de categorie 'hagiografie' zijn onder te brengen. De aanname dat Maerlants werken in drie duidelijk te onderscheiden genres uiteenvallen, mag daarom met een korrel zout worden genomen, al was het maar omdat het begrip 'genre' als bijzonder problematisch en zelfs anachronistisch wordt ervaren in de mediëvistiek. In de woorden van Dewhurst (2003, 43-44):

> [R]esearch into medieval criteria for classifying literary works, as well as into the circumstances of medieval textual production and transmission, has demonstrated that such isolating and normative conceptions of genre cannot be applied to the texts of the Middle Ages without jarring anachronism. Medieval rhetorical treatises show a remarkable lack of concern with generic classification, concentrating instead on the poetic styles and repertoires that stood at the disposal of all authors. Few generic terms circulate, and those that do can be employed with bewildering looseness, cutting across the boundaries laid by modern criticism.

Ik benadruk daarom dat ik de term 'genre' in deze studie wil gebruiken in de algemene en hopelijk niet al te geladen betekenis van 'tekstsoort'. Ik wil daarbij geen zwaarwichtige claims maken inzake de middeleeuwse tekstbeleving of normatieve waarde-oordelen. Afgezien van deze reserves met betrekking tot het begrip 'genre' in deze studie, moet men ook goed beseffen dat genre en chronologie in Maerlants oeuvre onlosmakelijk verbonden zijn, want Maerlant heeft de drie hierboven besproken tekstsoorten slechts in aparte, opeenvolgende fases van zijn carrière beoefend. Als ik hieronder dus uitspraken doe over de impact van de factor 'genre' op Maerlants rijmwoordenschat spreekt het daarom voor zich dat deze uitspraken noodzakelijk ook een chronologische dimensie zullen hebben.

6.2 SPREIDING

Wanneer wij de standaardedities van Maerlants teksten verzamelen van op de *Cd-rom Middelnederlands* en verrijken met lemmata, levert dat een omvangrijke data set op die in onderstaande tabellen is beschreven (Tab. 17, p. 131 en Tab. 18, p. 132). Ik laat Maerlants *Torec* in dit hoofdstuk nog even buiten beschouwing, omdat de tekst slechts is overgeleverd in de *Lancelotcompilatie*. Lodewijk van Velthem zou de *Torec* in dit handschrift zodanig bewerkt hebben dat deze versie

Maerlants oorspronkelijke stijl nog maar amper weerspiegelt (Claassens 2009). Ook Maerlants *Merlijn*teksten zijn overgeleverd in een versie die tot op zekere hoogte door Lodewijk van Velthem kan zijn bewerkt (zie hoofdstuk 8). Omdat Velthems bemoeienis met deze tekst doorgaans echter minder ingrijpend wordt geacht, zal ik deze tekst wel in de analyses betrekken. Uit deze dubbelroman heb ik wel de zogenaamde Maskeroen-passage verwijderd, omdat duidelijk is dat deze niet oorspronkelijk kan zijn (Warnar 2009). Verder zijn ook de passages verwijderd uit de *Historie van Troyen* die Maerlant naar eigen zeggen aan Segher Diengotgaf heeft ontleend. Voor de *Spiegel historiael* beperk ik mij tot de reeds aan bod gekomen Derde Partie. Ook voor *Alexanders Geesten* beperk ik mij tot een representatieve steekproef.

Titel	**Afkorting**	**Datering**	**Genre**
Alexanders geesten	*AlexG*	1257-1261	Ridderepiek
Historie van den Grale etc.	*HvdGrale*	1259-1266	Ridderepiek
Historie van Troyen	*HvTroy*	ca. 1261-1266	Ridderepiek
Heimelijkheid der Heimelijkheden	*HdHeim*	1266	Didactiek
De naturen bloeme	*dNbloem*	ca. 1270	Didactiek
Rijmbijbel	*RijmB*	ca. 1271	Historiografie
Sinte Franciscus leven	*SFrancL*	1270-1280	Historiografie
Spiegel historiael (Derde Partie)	*SpH3*	ca. 1284	Historiografie

Tabel 17: Een overzicht van de teksten (ca. 1259 - ca. 1284) die in dit hoofdstuk worden gebruikt voor de analyse van het oeuvre van Jacob van Maerlant.

Onderstaande tabel biedt enkele eenvoudige tellingen op basis van de lemmata die in deze teksten optreden in rijmpositie (Tab. 18, p. 132). Wat meteen opvalt is de grote rijmarmoede in Maerlants oeuvre: hoewel de bestudeerde teksten in totaal 156.339 rijmwoorden tellen, worden slechts 5.262 verschillende lexemen gebruikt om deze rijmslots te vullen. 1.937 van de unieke lemma's zijn hapaxen, oftewel rijmlexemen die slechts één keer in het hele corpus worden aangetroffen. Dat sterkt op het eerste zicht de eerder vermelde hypothese dat veel rijmwoorden zich gedragen als functoren: het karakter van verschillende rijmwoorden is immers zo formulair dat zij herhaaldelijk kunnen worden gebruikt.

Onderstaande figuur geeft weer welke 50 rijmlexemen cumulatief gesproken het meest frequent zijn in Maerlants teksten (Fig. 21, p. 133). Een dergelijke figuur hebben we hierboven reeds weergegeven voor de bijdrages van Utenbroeke en Maerlant aan de *Spiegel historiael* (Fig. 7, p. 81 en Fig. 8, p. 82). Ook in onderstaande figuur wordt duidelijk hoe het vooral weinig inhoudsgevoelige rijmwoorden zijn die komen bovendrijven, als men rijmwoorden rangschikt op basis van hun cumulatieve frequentie en slechts de meest frequente bekijkt.

Tekst	Rijmwoorden	Unieke lemmas	Hapaxen
AlexG	14.237	1.792	698
HvdGrale	8.601	1.224	533
HvTroy	38.391	2.652	999
HdHeim	2.156	793	439
dNBloem	16.672	2.298	1.025
RijmB	34.708	2.454	957
SFrancL	10.494	1.461	584
SpH3	31.080	2.454	957
Totaal	**156.339**	**5.262**	**1.937**

Tabel 18: Een overzicht van enkele algemene tellingen gebaseerd op de gelemmatiseerde rijmwoorden in de geanalyseerde teksten van Jacob van Maerlant. In de hapax-kolom wordt per tekst weergegeven hoeveel rijmwoorden in een bewuste tekst slechts één keer voorkomen, terwijl in de onderste rij is weergegeven hoeveel rijmwoorden in het hele oeuvre slechts één keer voorkomen.

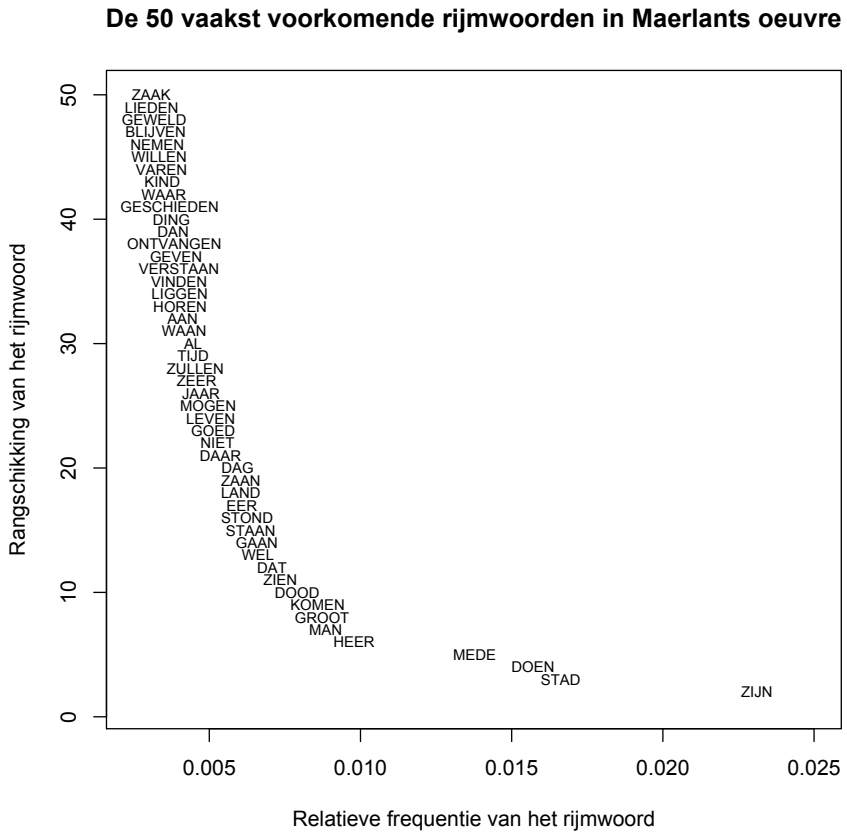
In deze weergave is voor de leesbaarheid van de rijmwoorden de rangorde van de rijmwoorden op de verticale as geprojecteerd, terwijl de eigenlijke frequenties zijn weergegeven op de horizontale as. In de kwantitatieve linguïstiek is het echter veel gebruikelijker om deze voorstelling om te draaien en de eigenlijke frequentie op de verticale as weer te geven.

Onderstaande figuur stelt Maerlants rijmwoorden voor volgens deze alternatieve weergave (Fig. 22, p. 134). De afgebeelde curve bestaat uit kleine cirkeltjes die elk een rijmwoord voorstellen dat in het vlak van de figuur is geprojecteerd op basis van zijn cumulatieve relatieve frequentie (verticale as) en de rangorde van het rijmwoord volgens deze frequentie (horizontale as). Merk op dat slechts een kleine verzameling rijmwoorden zich bovenin de rangschikking door een hoge frequentie onderscheidt. Het merendeel van de andere rijmwoorden is veel minder frequent en daarom moeilijk te onderscheiden in de snel dalende kromme.

Omdat de kromme in bovenstaande grafiek nogal snel daalt en daarom moeilijk leesbaar is, zal men in de kwantitatieve linguïstiek niet vaak de eigenlijke waardes gebruiken voor de voorstelling van woordfrequenties. Men verkiest om het logaritme van de frequentiewaardes en rangordes te nemen. Men heeft namelijk opgemerkt dat een dergelijke logaritmische omzetting een veel mooiere curve oplevert die bovendien een opvallende eigenschap heeft. Onderstaande figuur is een zogenaamde 'Zipf-curve' (Fig. 23 (p. 135), zie ook § 4.1.3), vernoemd naar de belangrijke taalkundige G. Zipf, die de vorige eeuw enkele belangrijke statistische wetten heeft geformuleerd over het gedrag van woordfrequenties in tekst (Zipf 1949). In deze figuur is elk

logaritme

Zipf-curve



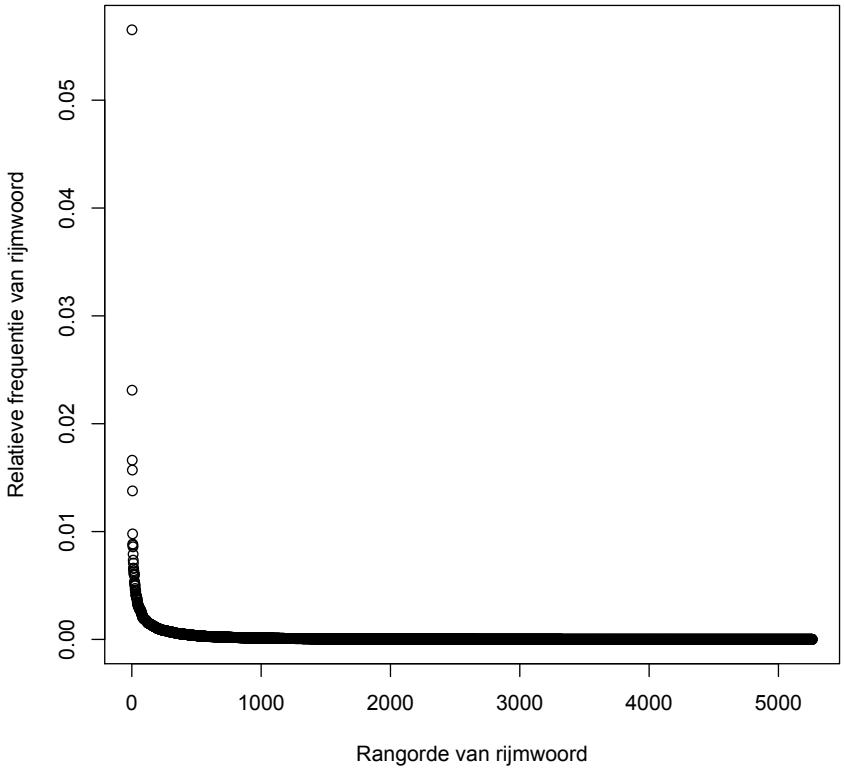
Figuur 21: De 50 rijmwoorden die cumulatief gesproken het meest frequent zijn in de geanalyseerde teksten van Jacob van Maerlant. De relatieve frequentie van de rijmwoorden is op de horizontale as uitgezet in functie van hun rangorde volgens deze frequentie op de verticale as. Slechts het frequentie-interval 0,0025–0,025 is weergegeven zodat het hoogfrequente label PRNAME buiten de weergave valt.

rijmwoord als een cirkeltje weergegeven dat is uitgezet volgens het logaritme van zijn cumulatieve relatieve frequentie en het logaritme van zijn rangorde volgens deze frequentie.

Anders dan bij de vorige figuur (Fig. 22, p. 134), zorgt de logaritmische omzetting van de waardes op beide assen ervoor dat we een kromme bekomen die minder bruusk daalt. Bij benadering vormt die stippenreeks een rechte met een hellingsgraad van ca. -1. De schuine stippellijn die in de figuur is weergegeven komt overeen met de regressielijn die we bekomen als we een lineair model van de reeks waarden maken. Die analyse komt er eigenlijk op neer dat we een rechte proberen te definiëren die de bewuste cirkelcurve zo goed mogelijk volgt. De hellingsgraad van de regressielijn is in dit geval $\approx -1,6$.[5] Die

regressielijn
lineair model

5 De gebruikte logaritmes zijn berekend op basis van het grondtal 10. De exacte formule van de regressielijn is *log frequentie* = 70,74 – 1,6 * *log rangorde*. De significantie die dit

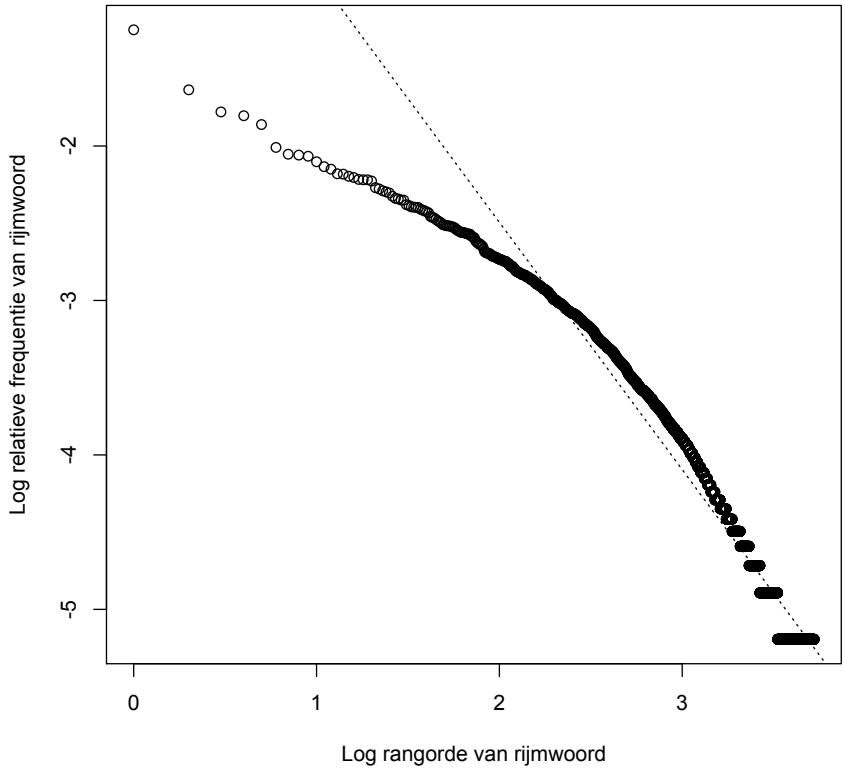


Figuur 22: Grafiek die alle rijmwoorden uit Maerlants teksten als cirkeltjes in een assenstelsel projecteert op basis van hun cumulatieve relatieve frequentie en hun rangorde volgens deze frequentie.

hellingsgraad is iets lager dan in de klassieke formulering van Zipf waar zij op -1 is gesteld. Dat duidt erop dat in Maerlants vocabulaire relatief weinig rijmwoorden laagfrequent zijn (vgl. hapaxen), althans in vergelijking met de distributies van de woordfrequenties die we in andere corpora aantreffen.

Dat de Maerlants rijmwoordenschat een Zipf-achtige curve vertoont is interessant, maar eigenlijk weinigzeggend. De hoge frequentie van een rijmwoord in dit oeuvre zegt immers niets over de spreiding van dat rijmwoord in het oeuvre. Stel bijvoorbeeld dat een specifiek inhoudsgerelateerd rijmwoord bijzonder frequent is in één lange tekst, maar in andere teksten niet of nauwelijks voorkomt. In dat geval zou het rijmwoord toch een hoge cumulatieve frequentie hebben, hoewel het erg slecht is gespreid over de verschillende teksten. Als we de frequentie van een dergelijk rijmwoord zouden meenemen in de stylometrische analyse van de stijl van een auteur, lopen we het

lineair model rapporteert, is groot ($F(1, 5260)$, $p < 0{,}001$, *adjusted* $R^2 = 0{,}95$) maar dit is tot op zekere hoogte een *non-issue*, aangezien rangorde hier wordt bepaald op basis van frequentie, zodat beide variabelen gecorreleerd zijn.



Figuur 23: Grafiek die alle rijmwoorden uit Maerlants teksten als cirkeltjes in een assenstelsel projecteert in functie van het logaritme van hun cumulatieve relatieve frequentie en het logaritme van hun rangorde volgens deze frequentie. De schuine stippellijn stelt de regressielijn voor die het best aan de resulterende curve beantwoordt.

gevaar ons stilistisch model te *overfitten* op dit inhoudsgerelateerde rijmwoord.

Daarom is het van groot theoretisch belang om te onderzoeken of de hoge cumulatieve frequentie van een rijmwoord in een oeuvre zoals dat van Maerlant, correleert met een goede spreiding van dat rijmwoord over de delen van dit oeuvre. Een dergelijke test wordt zelden uitgevoerd in de stylometrie, omdat daar gewerkt wordt met gewone functoren waarvan men op voorhand mag veronderstellen dat deze goed gespreid zijn over teksten. Dat is voor rijmwoorden in deze studie echter niet noodzakelijk het geval.

Hieronder stel ik daarom twee testmethodes voor om de spreiding van woorden te onderzoeken. Deze testen zijn beide sterk gerelateerd aan het concept polytextualiteit (*polytextuality* of *polytexty*) uit de kwantitatieve linguïstiek, meerbepaald het paradigma van de Synergetische Linguïstiek (Köhler 1986, 2005). Binnen dit paradigma wordt gesteld

polytextuality, polytexty

dat de frequentie van een (bv. lexicaal) item in een tekst een functie is van zijn communicatieve relevantie. Woorden of functoren kunnen dus maar hoogfrequent zijn, als zij ook een grote communicatieve relevantie hebben. Daaruit volgt dat de hoge frequentie van een woord noodzakelijk voortkomt uit de goede spreiding van een woord. De wetten uit de Synergetische Linguïstiek beklemtonen echter dat het tegengestelde niet noodzakelijk waar is: de hoge frequentie van een woord komt immers niet noodzakelijk voort uit de goede spreiding van een woord.

6.2.1 *Inverse Document Frequency*

Inverse Document Frequency

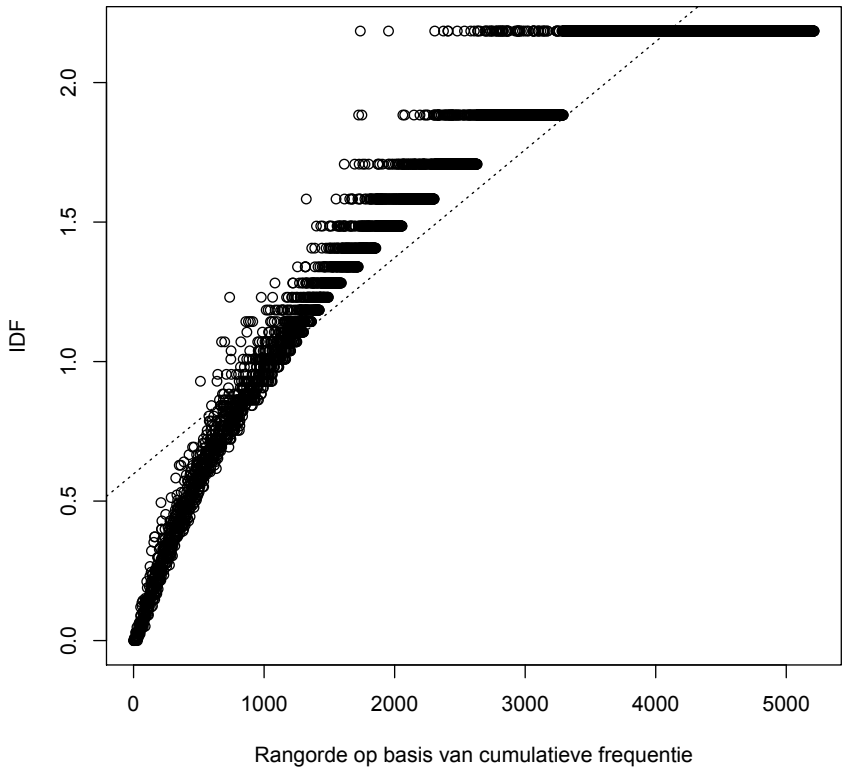
De eerste test die men kan gebruiken om de spreiding van een woord te testen, is gebaseerd op de *Inverse Document Frequency* (IDF), een notie uit de *Information Retrieval* (Manning et al. 2008, 117-118). Het is een veel gebruikte wegingsmethode om na te gaan hoe specifiek het optreden van een woord is in een bepaald document, temidden een grotere verzameling documenten. De traditionele berekeningswijze voor de IDF van een term t is erg eenvoudig. Men gaat eerst na hoeveel documenten (N) in de totale collectie voorkomen en in hoeveel documenten t voorkomt ($\mathtt{df}_t$). Men stelt de IDF van een term t ($\mathtt{idf}_t$) traditioneel gelijk aan het logaritme van de deling van N door $\mathtt{df}_t$.[6] Woorden die erg document-specifiek zijn hebben een hoge IDF, terwijl woorden die veel beter gespreid zijn over teksten een erg lage IDF zullen hebben, aangezien de $\mathtt{df}_t$ in dit geval veel hoger is.

Om deze testmethode toe te passen op Maerlants oeuvre verdelen we elke tekst in niet-overlappende samples van 1.000 rijmwoorden. In totaal levert dat 153 samples op. Vervolgens bepalen we de cumulatieve frequentie van elk rijmwoord dat in deze samples optreedt en rangschikken we deze rijmwoorden op basis van deze frequenties. In een tweede stap bepalen we de IDF van elk rijmwoord met de bovenvermelde berekeningswijze, waarbij ieder sample als een onafhankelijk document wordt beschouwd. De hypothese die we willen testen is of de rangorde van een rijmwoord significant correleert met de spreiding van het rijmwoord over de individuele samples.

In onderstaande figuur is het resultaat van deze analyse grafisch weergegeven (Fig. 24, p. 137). De cirkeltjes in deze grafiek geven elk een rijmwoord weer dat is uitgezet in functie van zijn rangorde in de cumulatieve frequentielijst (horizontale as) en zijn individuele IDF (verticale as). Onze hypothese wordt door deze analyse bevestigd: de hoge frequentie van een rijmwoord in het corpus komt duidelijk overeen met een goede spreiding (i.e. een lage IDF). Zeker wat de hoogst gerangschikte rijmwoorden betreft, kunnen we vaststellen dat

6 Traditioneel berekent men het logaritme op basis van het grondtal 10. Deze logaritmische omzetting (alsook het gebruikte grondtal) zijn echter niet-essentieel omdat zij de onderlinge rangschikking van de resulterende waardes niet veranderen.

deze onveranderlijk goed gespreid zijn over de individuele samples uit Maerlants oeuvre. De corresponderende regressielijn die als een schuine stippellijn is weergegeven, benadrukt de sterke band tussen frequentie en spreiding.[7]

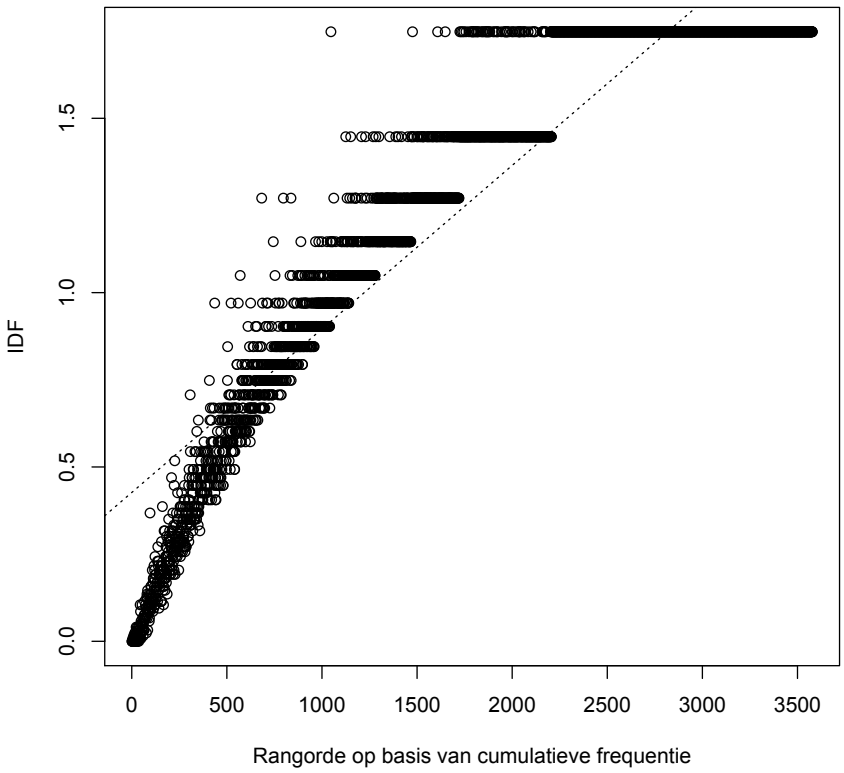


Figuur 24: Grafische voorstelling van een analyse van 153 samples van 1.000 rijmwoorden uit Maerlants oeuvre. De IDF van de rijmwoorden (verticale as) wordt weergegeven als een functie van de rangschikking van de rijmwoorden op basis van hun cumulatieve frequentie (horizontale as). De schuine stippellijn is de corresponderende regressielijn.

Mogelijk is deze analyse niettemin bedrieglijk: de individuele teksten lopen immers sterk uiteen qua lengte. Op die manier kan de enorme *HvTroy* (38 samples van 1.000 rijmwoorden) veel meer gewicht in de schaal leggen dan de korte *HdHeim* waaruit we slechts 2 samples van 1.000 woorden kunnen extraheren. Daarom is een bijkomende analyse noodzakelijk. Daarvoor kunnen we de korte *HdHeim* negeren en ons voor de overige teksten beperken tot de 8.000 eerste rijmwoorden uit elke tekst – de lengte van de tweede kortste tekst, nl. *HvdGrale*. Op die manier kan elk van de overgebleven teksten slechts

7 Technische noot: het resultaat van deze test is bijzonder significant ($F(1, 5.211)$, $p < 0{,}001$, *adjusted* $R^2 = 0{,}87$).

8 samples inbrengen en zijn de langere teksten niet langer kwantitatief bevoordeeld. Onderstaande figuur maakt duidelijk dat er in deze analyse mogelijk iets meer speling zit op de spreiding van de lager gerangschikte rijmwoorden (Fig. 25, p. 138). Die speling doet zich echter amper voor in de kruin van de kromme uiterst links die de meest frequente rijmwoorden voorstelt. Bijgevolg rapporteert de regressie-analyse ook hier een meer dan significante correlatie ($p <$ 0,001). Ook in deze analyse hebben de meeste rijmwoorden een IDF die nul benadert, hetgeen impliceert dat deze rijmwoorden in nagenoeg elk sample van 1.000 rijmwoorden voorkomt.



Figuur 25: Grafische voorstelling van een analyse van 56 samples van 1.000 rijmwoorden uit Maerlants oeuvre. De analyse negeert de korte tekst *HdHeim* en beschouwt slechts de eerste 8.000 rijmwoorden uit de andere teksten, zodat de langere teksten geen kwantitatief overwicht meer hebben. De IDF van de rijmwoorden (verticale as) wordt weergegeven als een functie van de rangschikking van de rijmwoorden op basis van hun cumulatieve frequentie (horizontale as). De schuine stippellijn is de corresponderende regressielijn.

6.2.2 *De variatie-coëfficient*

De *Inverse Document Frequency* is een handige metriek maar heeft één uitgesproken nadeel: er wordt slechts gewerkt met een binaire waarde die uitdrukt of een woord al dan niet aanwezig is in een document. De methode is bijgevolg ongevoelig voor de eigenlijke frequentie van een woord, zodat het niet uitmaakt of een woord 1, 200 of 20.000 keer in een document voorkomt. Een andere, meer genuanceerde techniek om de spreiding van woorden te meten is de variatie-coëfficient (*coefficient of variation*, CV). Deze metriek (Lewontin 1966) kunnen we toelichten aan de hand van onderstaande tabel (Tab. 19, p. 140). De CV meet de relatieve variabiliteit in een reeks waarden, bijvoorbeeld de relatieve frequentie van een woord in de verschillende samples van een corpus. Voor de frequenties van een bepaalde term *t* wordt de CV gedefinieerd als de deling van de standaarddeviatie σ_t in deze frequenties door het gemiddelde van deze waarden μ_t. Formeel is dat: $cv_t = \sigma_t/\mu_t$. Onderstaande tabel laat duidelijk zien dat de CV een aantrekkelijke manier is om de spreiding van een woord te meten (Tab. 19, p. 140). Woorden met een hogere frequentie en betere spreiding (bijvoorbeeld Woord A) hebben consequent een lagere CV dan woorden die minder frequent en minder goed gespreid voorkomen in dit fictieve corpusje.

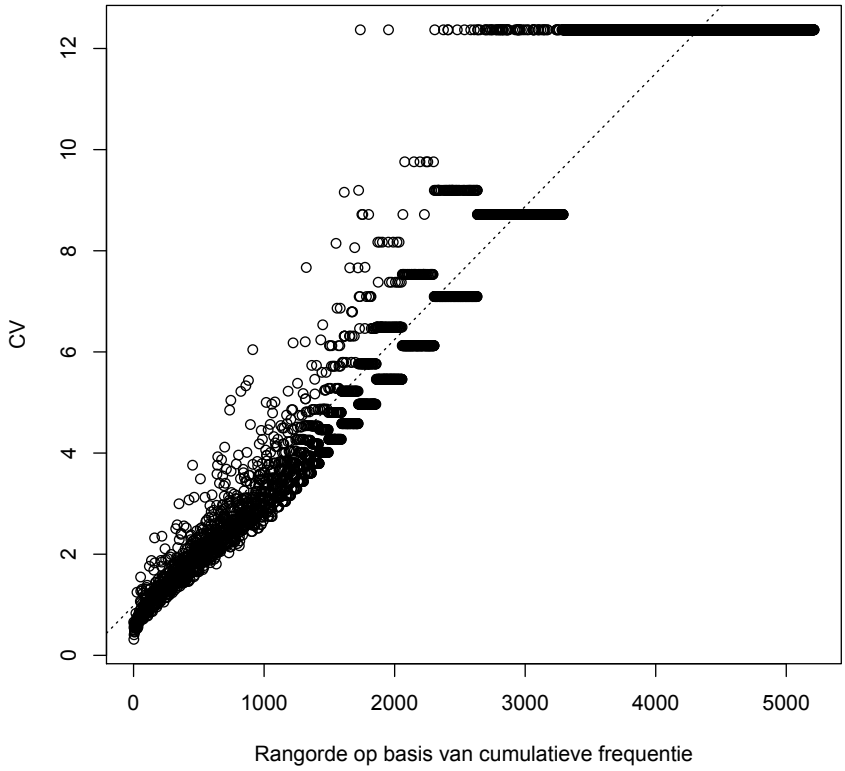
coefficient of variation

We kunnen de experimenten uit de vorige sectie herhalen met de variatie-coëfficient in plaats van de *Inverse Document Frequency*. Ons doel blijft daarbij ongewijzigd: we willen nog steeds de hypothese testen dat rijmwoorden die cumulatief gesproken erg frequent in het corpus voorkomen, een goede spreiding kennen over kleinere delen van dat corpus. Rijmwoorden met een cumulatief hoge frequentie, zouden een relatief lage variatie-coëfficient moeten hebben en omgekeerd.

Onderstaande figuur geeft een dergelijke analyse weer voor de 153 samples die we in totaal uit Maerlants oeuvre kunnen extraheren (Tab. 26, p. 141). De resulterende curve bevestigt opnieuw onze verwachting: rijmwoorden met een hogere cumulatieve frequentie in het corpus hebben in de regel ook een lagere CV, hetgeen een betere spreiding over de individuele samples van het corpus betekent. Als we deze analyse herhalen voor de 56 samples van 1.000 woorden (zonder *HdHeim*), zien we dat deze trend grosso modo behouden blijft (Fig. 27, p. 142). De regressielijnen die in beide grafieken zijn weergegeven, melden dan ook weinig verbazend een erg significantie correlatie tussen beide variabelen. De meer genuanceerde CV-metingen bevestigen met andere woorden het resultaat van de analyses op basis van de *Inverse Document Frequency*.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	**Totaal**	**Standaarddeviatie**	**Gemiddelde**	**CV**
Woord A	70	60	65	75	69	339	5,630275	67,8	0,08304
Woord B	20	10	15	25	20	90	5,700877	18	0,31671
Woord C	10	20	15	0	10	55	7,416198	11	0,67419
Woord D	0	10	5	0	0	15	4,472136	3	1,49071
Woord E	0	0	0	0	1	1	0,447213	0,2	2,23606
Totaal in sample	**100**	**100**	**100**	**100**	**100**	**500**			

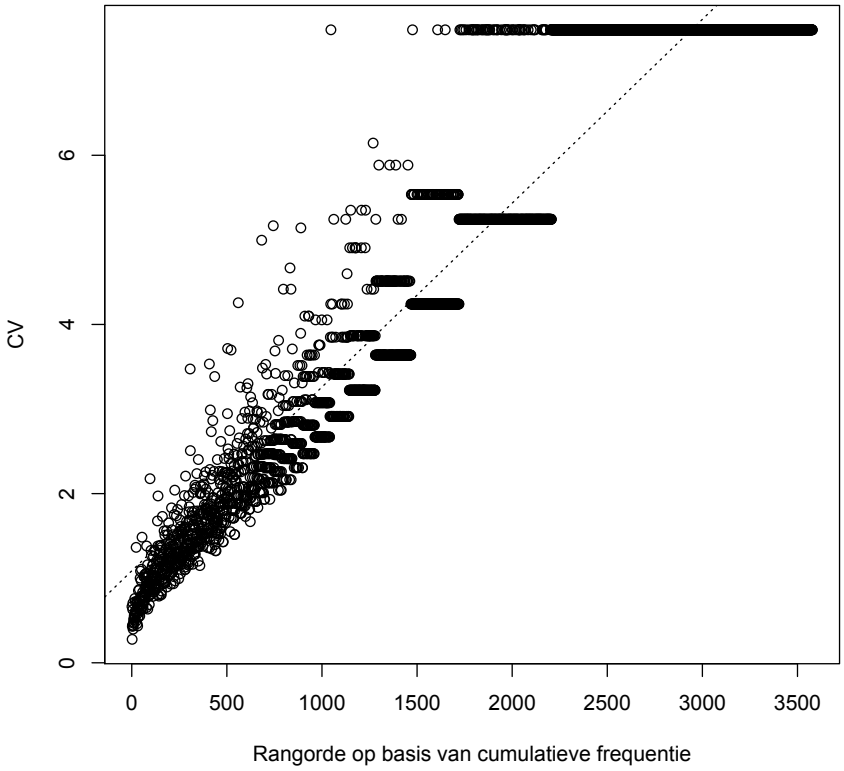
Tabel 19: Illustratie van de variatie-coëfficient. Deze tabel stelt een fictief corpus van 5 samples (1-5) die elk honderd instanties van 5 verschillende woorden (A-D) tellen. Voor elk woord is de gemiddelde frequentie weergegeven, de standaarddeviatie en de variatie-coëfficient (CV). Woorden met een betere spreiding, worden gekenmerkt door een relatief lagere variatie-coëfficient.



Figuur 26: Grafische voorstelling van een analyse van 153 samples van 1.000 rijmwoorden uit Maerlants oeuvre. De CV van de rijmwoorden (verticale as) wordt weergegeven als een functie van de rangschikking van de rijmwoorden op basis van hun cumulatieve frequentie (horizontale as). De schuine stippellijn is de corresponderende regressielijn.

6.3 MAERLANTS GENRE-DRIEHOEK

Maerlants oeuvre vertoont zoals gezegd een belangrijke driedeling, zowel qua genre als chronologie. De dichter vatte zijn carrière aan met een aantal (kort na elkaar geschreven) ridderromans. Halfweg zijn carrière vertonen zijn teksten een sterkere didactische inslag, terwijl hij zijn loopbaan beëindigde met een aantal teksten die onder een historiografische noemer kunnen worden samengenomen. De metingen in de vorige sectie wijzen uit dat Maerlant zich van een betrekkelijk stabiel rijmvocabularium heeft bediend. Als we Maerlants rijmen rangschikken op basis van hun cumulatieve frequentie in zijn oeuvre, observeren we het bestaan van een kleine kruin van hoogfrequente rijmwoorden die keer op keer in zijn teksten voorkomen. Het optreden van dergelijke functoren lijkt daarom op het eerste zicht relatief



Figuur 27: Grafische voorstelling van een analyse van 56 samples van 1.000 rijmwoorden uit Maerlants oeuvre. De analyse negeert de korte tekst *HdHeim* en beschouwt slechts de eerste 8.000 rijmwoorden uit de andere teksten, zodat de langere teksten geen kwantitatief overwicht meer hebben. De CV van de rijmwoorden (verticale as) wordt weergegeven als een functie van de rangschikking van de rijmwoorden op basis van hun cumulatieve frequentie (horizontale as). De schuine stippellijn is de corresponderende regressielijn.

onafhankelijk van de inhoud, het genre of de ontstaansperiode van een tekst.

Hierboven werd reeds *Principal components analysis* toegelicht, een techniek voor het statistisch reduceren van het aantal 'dimensies' in een frequentietabel (§ 3.3). In veel voorgaand stylometrisch onderzoek is gebleken dat de strooidiagrammen die uit dergelijke analyses resulteren, het auteurschap in een sample-collectie voorstellen. Wanneer men dimensie-reductie toepast op op de functor-frequenties in stukken tekst van verschillende auteurs, zullen de samples van één auteur vaak clusteren, terwijl stukken tekst met een niet-identiek auteurschap verder uit elkaar zullen liggen.

Het is een interessante vraag wat er zou gebeuren als men dimensie-reductie toepast op een aantal teksten van dezelfde auteur. De in dit hoofdstuk geanalyseerde teksten worden allemaal toegeschreven aan

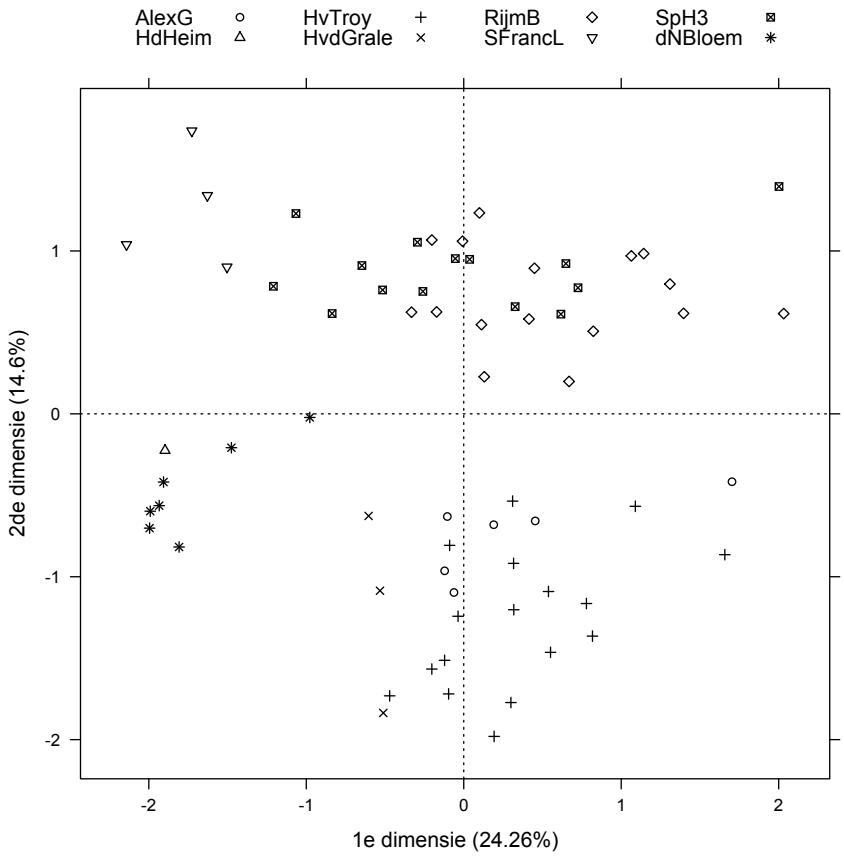
Jacob van Maerlant. Gegeven de ogenschijnlijke stabiliteit van functoren zou men eigenlijk verwachten dat de spreiding van de teksten na dimensie-reductie geen significante verschillen tussen teksten aan het licht zal brengen. We verkeren immers in de veronderstelling dat functor-frequenties slechts de stijl van een auteur in kaart in brengen en relatief ongevoelig zijn voor andere factoren.

Laat ons de teksten uit Maerlants oeuvre opnieuw opdelen in 68 samples. We kiezen daarbij voor een sample-grootte van precies 2.156 rijmen zodat we *HdHeim*, de kleinste tekst in het corpus, nog net als één sample kunnen voorstellen. Vervolgens bepalen we welke 50 rijmwoorden het vaakst voorkomen in deze teksten en we stellen deze samples voor als vectoren die de relatieve frequenties van deze 50 rijmwoorden in de samples bevatten. Hierboven is *Principal components analysis* reeds besproken als een techniek voor dimensie-reductie. Dit is echter lang niet de enige techniek die men hiervoor kan gebruiken. Een andere veelgebruikte methode in stylometrisch auteursonderzoek is correspondentie-analyse (*correspondence analysis*, CA). Buiten een iets andere berekeningswijze verschilt deze methode aan de oppervlakte maar weinig van PCA (Nenadić & Greenacre 2007). Ook deze techniek levert een samenvatting van een frequentietabel in de vorm van een (meestal tweedimensioneel) strooidiagram (Baayen 2008, § 5.1.3). Onderstaande figuur biedt het resultaat van een correspondentie-analyse van de 68 samples uit Maerlants oeuvre, voorgesteld aan de hand van de 50 meest frequente rijmwoorden erin (Fig. 28, p. 144).

correspondentie-analyse

Een eerste banale observatie is dat samples die uit dezelfde tekst afkomstig zijn elkaar opzoeken in dit strooidiagram. Hoewel de CA op zich geen weet heeft van de herkomst van een sample, worden samples uit dezelfde tekst wel relatief dicht bij elkaar geplaatst in deze twee-dimensionele samenvatting. Dit gegeven relativeert op zich reeds de hypothese dat functor-frequenties *enkel* de onbewuste stijl van een auteur zouden weergeven, want in dit diagram kan de analyse vrij goed achterhalen uit welke tekst een sample afkomstig is.

Nog veel opvallender is dat dit strooidiagram een driedeling schijnt te vertonen. De ridderromans *AlexG*, en *HvdGrale* en *HvTroy* zouden volgens deze analyse uitgesproken overeenkomsten in rijmgebruik vertonen, want zij worden erg dicht op elkaar afgebeeld. Hetzelfde kan gezegd worden van de historiografische cluster, bestaande uit de samples van *RijmB*, *SFrancL* en *SpH3*. De sample-wolk uit het encyclopedische dNBloem wordt relatief goed onderscheiden van de voorgaande clusters; het enige sample uit *HdHeim* zoekt het gezelschap op van *dNBloem*. Als we de samples voorstellen met een symbooltje dat hun genre voorstelt (zie Fig. 29, p. 145), is dit effect nog makkelijker waarneembaar. Het moet benadrukt worden dat een correspondentie-analyse net als PCA geen gesuperviseerde methode is: de analyse heeft a priori geen weet van het genre waaruit een tekst afkomstig is. Bijgevolg wordt er ook geen doelgerichte poging ondernomen om

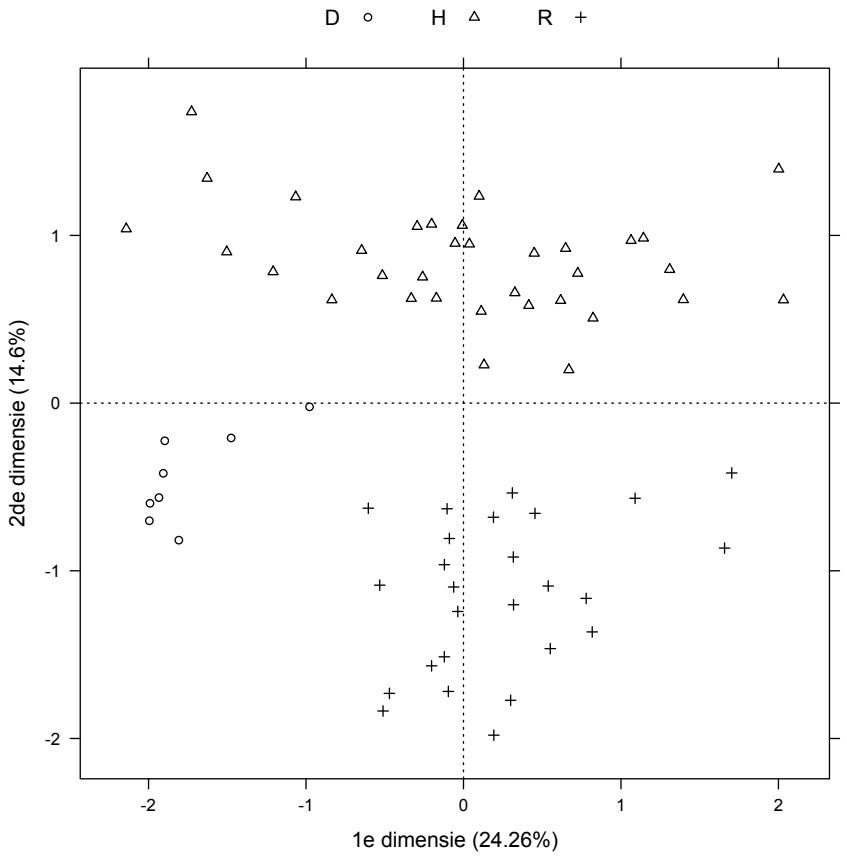


Figuur 28: Strooidiagram (eerste twee dimensies) voor een correspondentie-analyse van 68 samples van 2.156 rijmwoorden uit Maerlants oeuvre. De samples zijn voorgesteld aan de hand van de relatieve frequentie van de 50 vaakst voorkomende rijmwoorden erin.

de teksten uit te splitsen op basis van een dergelijke, op voorhand gegeven classificatie. De driedeling komt als het ware uit de teksten zelf naar voren.

Vanzelfsprekend kan de interpretatie van een strooidiagram een subjectieve aangelegenheid zijn. Om na te gaan of een dergelijke clustering wel significant is, worden er in de stylometrie soms testen op de spreiding van samples uitgevoerd. In sommige gevallen is die test overbodig: als een clustering visueel duidelijk aanwezig is in een diagram zal de statistische test die observatie meestal gewoon bevestigen (en omgekeerd).

Ter illustratie kunnen we hieronder een dergelijke test op de genre-driedeling uitvoeren. We hebben in dit geval een statistische test nodig die de posities van de samples uit elk van onze tekstsoorten onderling vergelijkt en nagaat of deze verschillen niet aan het toeval kunnen te wijten zijn. Dat moet gebeuren voor de coördinaten van onze samples in elk van de twee dimensies waarin onze samples zijn geprojecteerd. Een significant onderscheid tussen twee sample-groepen in de ene



Figuur 29: Strooidiagram (eerste twee dimensies) voor een correspondentie-analyse van 68 samples van 2.156 rijmwoorden uit Maerlants oeuvre. De samples zijn voorgesteld aan de hand van de relatieve frequentie van de 50 vaakst voorkomende rijmwoorden erin. De samples worden weergegeven met een symbooltje dat hun genre uitdrukt: R(idderroman), D(idactiek) en H(istoriografie).

dimensie, impliceert immers niet noodzakelijk dat beide types ook in de andere dimensie significant zullen verschillen.

Tukey's Honestly Significant Difference

Een geschikte test hiervoor is *Tukey's Honestly Significant Difference* (Baayen 2008, §5.1.3). Deze test (afgekort HSD) is een veeleisende proef omdat deze erg conservatieve *p*-waardes rapporteert: een effect moet bijgevolg erg uitgesproken zijn, wil het door de test niet aan het toeval worden toegeschreven. De uitslag van deze statistische proef kan visueel aantrekkelijk worden weergegeven. In onderstaande figuur hebben we de HSD berekend voor elk paar van twee genres (Fig. 30, p. 147). De coördinaten van de samples uit twee van de drie genres zijn steeds onderling vergeleken, wat betreft de eerste dimensie (linkse subpaneel) en de tweede dimensie (rechtse subpaneel). Het resultaat van elk van deze vergelijkingen wordt weergegeven als een betrouwbaarheidsinterval in de vorm van een verticaal lijnstuk. Als

dit lijnstuk niet door nul gaat (vgl. verticale stippellijn), mag men het verschil als significant beschouwen.

Het resultaat van de HSD-berekening is erg uitgesproken: in de meeste gevallen zijn de verschillen in positie tussen twee genres – (R)idderepiek, (D)idactiek en (H)istoriografie – significant in beide dimensies. Dat biedt dus een statistische bevestiging van onze eerdere interpretatie van het strooidiagram. Slechts in de eerste dimensie blijken de posities van de samples uit (R)idderepiek en (H)istoriografie niet significant van elkaar te verschillen.[8] Het uitblijven van een significant verschil tussen beide genres is hier misschien zelfs betekenisvol: in deze genres ligt de nadruk immers veel sterker op een echt 'verhaal', dan in de didactische teksten.

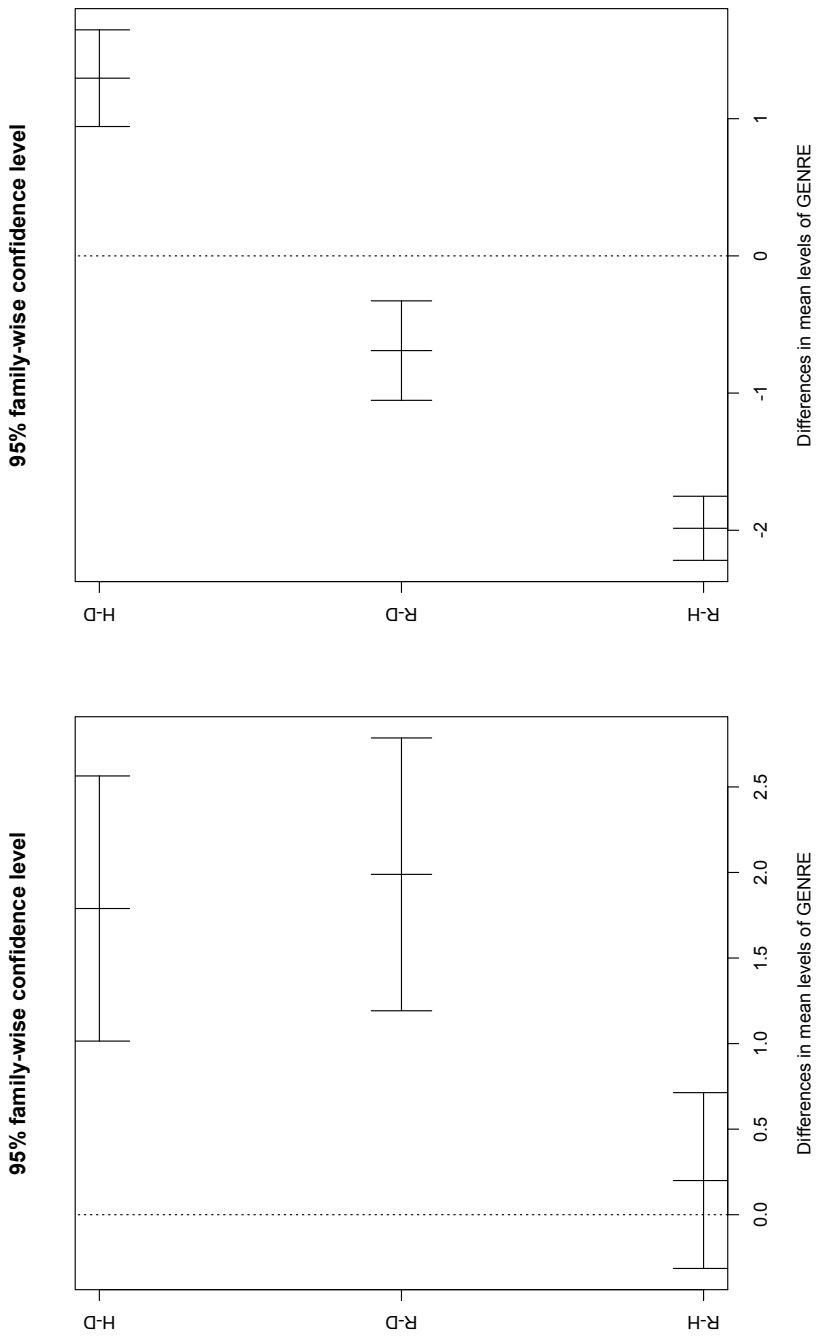
Een correspondentie-analyse biedt net als een PCA de mogelijkheid om uit te beelden hoe individuele rijmwoorden hebben bijgedragen aan de spreiding van teksten in een diagram. In het geval van Maerlant is dat interessant, want misschien hebben specifieke rijmwoorden een te interpreteren effect gehad op de sample-spreiding. Onderstaande afbeelding maakt een betrekkelijk rommelige indruk en is moeilijk leesbaar (Fig. 31, p. 148). Toch nemen enkele rijmwoorden (driehoekjes) een dergelijk extreme positie in, dat hun effect op de spreiding van de samples (bolletjes) valt te interpreteren.

Erg geprononceerd is de positie uiterst links van PLEGEN; dit rijmwoord schijnt relatief vaker voor te komen in teksten als *dNBloem* en *HdHeim*, maar kennelijk ook in het hagiografische *SFrancL*. Dat we hier te maken hebben met teksten met een meer didactische (soms zelfs ethische) inslag lijkt hier relevant, aangezien PLEGEN in de eerste plaats wordt gebruikt om een bepaalde gewoonte of gangbare zede te beschrijven. De positie van de rijmwoorden GOED en VERSTAAN (centraal links) kan men mogelijk ook in verband brengen met het didactische karakter van de werken links van horizontale stippellijn. Rechts boven in de grafiek vinden we het rijmwoord JAAR dat weinig verbazend eerder typisch blijkt voor Maerlants historiografische werk (bv. *SpH3*). De rol van dergelijke rijmwoorden valt dus te interpreteren vanuit hun specifieke semantiek. Anderzijds zijn er ook veel rijmwoorden die een iets meer uitgesproken positie innemen (vgl. MEDE, AAN of DAN) maar op het eerste zicht moeilijker vanuit de inhoud van de teksten vallen te duiden.

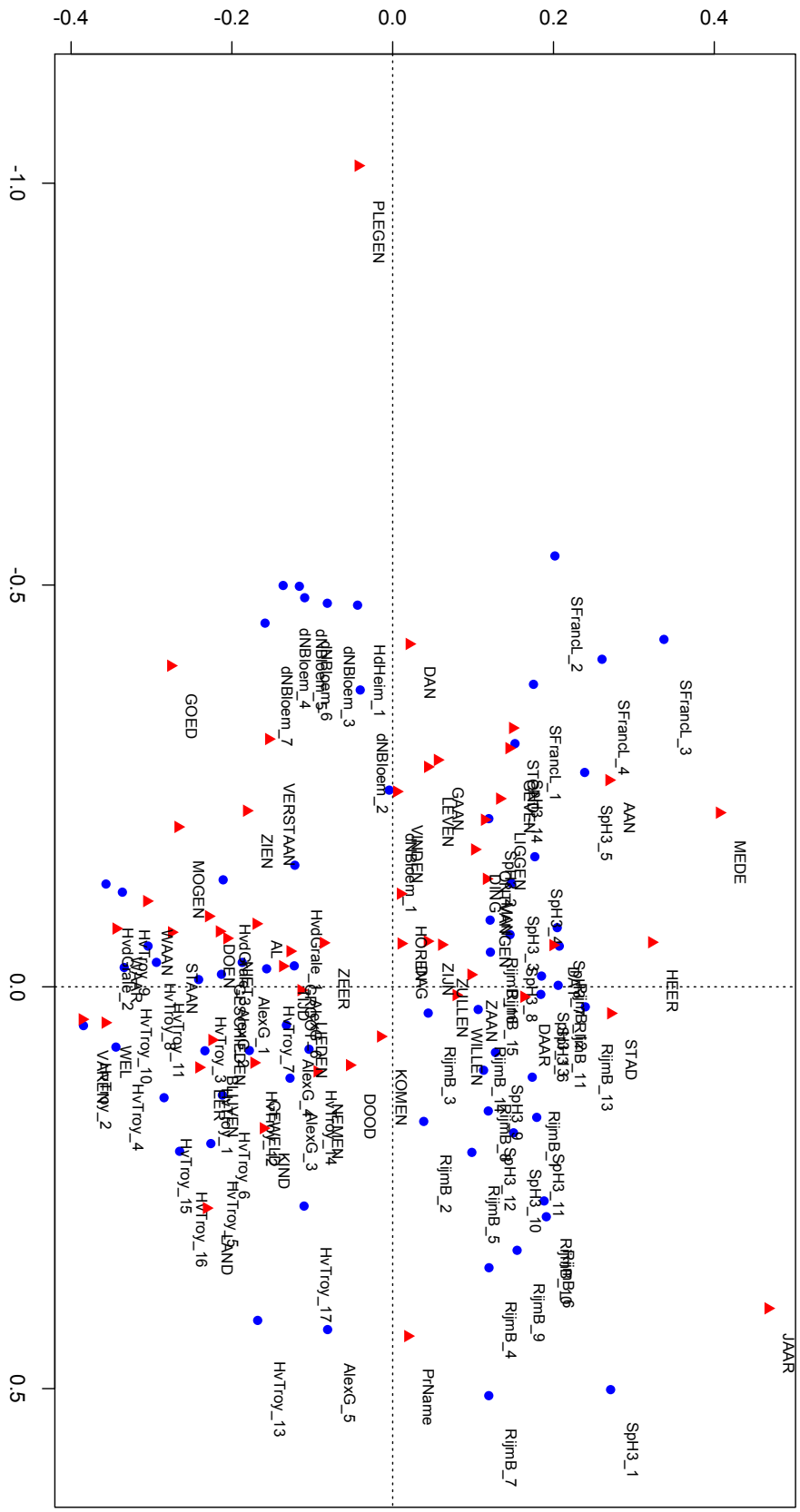
6.4 HET GENRE-GEVAAR

In dit hoofdstuk werd een groot deel van het oeuvre van Jacob van Maerlant (§ 6.1) stylometrisch geanalyseerd. Eerst werd aangetoond dat de meest frequente rijmwoorden uit Maerlants oeuvre een relatief

8 Een technische noot: voor de zekerheid is deze test ook herhaald met een niet-parametrische Kruskal-Wallis-toets, maar deze leverde geen minder significant resultaat op.



Figuur 30: Visuele weergave van een HSD waarin de positie van samples uit de drie genres in Maerlants oeuvre – (R)idderepiek, (D)idactiek en (H)istoriografie – worden vergeleken op basis van hun coördinaten in de hogerop afgebeelde correspondentie-analyse (vgl. Fig. 29, p. 145). Lijnstukken die niet door nul gaan (verticale stippellijn) duiden op een significant verschil tussen twee genres.


0.4
0.2
0.0
-0.2
-0.4
-1.0
-0.5
0.0
0.5
JAAR
MEDE
SFrancL_3
HEER
SpH3_1
STAD
RijmB_13
SFrancL_4
AAN
SpH3_5
SFrancL_2
SFrancL_1
RijmB_7
RijmB_4
RijmB_9
RijmB_5
RijmB_2
RijmB_3
PrName
KOMEN
DOOD
HvTroy_17
AlexG_5
HvTroy_13
ZEER
HvTroy_4
HvTroy_2
HvTroy_15
HvTroy_16
HvTroy_6
HvTroy_5
HvTroy_1
HvTroy_3
HvTroy_7
HvTroy_10
HvTroy_4
HvTroy_8
HvTroy_11
AlexG_1
AlexG_3
AlexG_4
KIND
WEL
WAAN
STAAN
MOGEN
GOED
VERSTAAN
ZIEN
AL
DOEN
HvdGrale_2
HvdGrale_3
HvdGrale_4
dNBloem_3
dNBloem_4
dNBloem_7
HdHeim_1
dNBloem_2
PLEGEN
DAN
GAAN
LEVEN
ZAAN
WILLEN
ZIJN
ZULLEN
HORDEN
DAAR
SpH3_8
SpH3_10
SpH3_12
SpH3_11
SpH3_13
SpH3_14
SpH3_15
RijmB_11
RijmB_15
RijmB_16
RijmB_10
RijmB_12
RijmB_14
RijmB_8
RijmB_6
RijmB_1
RijmB_2
RijmB_3
RijmB_4
RijmB_5
RijmB_7
RijmB_9
RijmB_10
RijmB_11
RijmB_12
RijmB_13
RijmB_14
RijmB_15
RijmB_16
SpH3_1
SpH3_2
SpH3_3
SpH3_4
SpH3_5
SpH3_6
SpH3_7
SpH3_8
SpH3_9
SpH3_10
SpH3_11
SpH3_12
SpH3_13
SpH3_14
SpH3_15
SpH3_16
LIGGEN
STOELEN
DING
VINDEN
dNBloem_1
dNBloem_2
dNBloem_3
dNBloem_4
dNBloem_5
dNBloem_6
dNBloem_7
HvdGrale_1
HvdGrale_2
HvdGrale_3
HvdGrale_4
HEDEN
HEEFT
HEBBEN
HEEFT
HEEFT
HEEFT

goede spreiding kennen en daarom interessant zijn voor een stylometrische analyse van de stijl van de auteur (§ 6.2). Daarvoor werd een beroep gedaan op twee testen, op basis van de *Inverse Document Frequency* (§ 6.2.1) en de *coefficient of variation* (§ 6.2.2). Het resultaat toonde aan dat een selectietechniek van rijmwoorden op basis van een frequentiedrempel een relatief betrouwbaar methode is om de hoogfrequente rijmwoorden in een corpus te selecteren die het sterkst onafhankelijk van inhoud zijn.

Vervolgens werd een correspondentie-analyse – een techniek voor dimensie-reductie – uitgevoerd van een groot aantal samples uit Maerlants oeuvre (§ 6.3). Deze analyses werden uitgevoerd op de frequenties van slechts 50 erg frequente rijmwoorden en werden beperkt tot het werk van één auteur. De analyses legde niettemin de (sterk genre-gerelateerde) driedeling bloot die ook in het traditionele onderzoek in Maerlants oeuvre wordt gezien. Een statistische proef (*Tukey's Honestly Significant Difference*) bevestigde de relevantie van deze genre-clusters. Deze resultaten nopen tot een belangrijke nuancering: hoewel het gebruik van rijmwoorden tot op zekere hoogte indicatief kan zijn voor het auteurschap van een tekst, mogen wij niet de naïeve illusie koesteren dat deze rijmwoorden volledig inhoudsonafhankelijk zouden zijn. Hoewel rijmwoorden tot op zekere hoogte geschikt zijn voor de analyse van de individuele stijl van een auteur, mogen wij op voorhand niet uitsluiten dat genre-kenmerken zullen interfereren met auteurskenmerken. In de rest van deze studie zullen wij de stilistische spanning tussen auteur en genre daarom voortdurend in het achterhoofd moeten houden.

Deel III

ONGEBAANDE WEGEN

In het vorige deel zijn een aantal stylometrische technieken toegepast op Middelnederlandse teksten waarvan het auteurschap zelden wordt betwijfeld. In dit deel worden deze methodes verder uitgediept en toegepast op enkele bekende Middelnederlandse teksten waarvan het auteurschap controversieel is. Bij die analyses begeef ik mij dus op ongebaande wegen en daarbij probeer ik zoveel mogelijk rekening te houden met de beperkingen van technieken die in het vorige deel proefondervindelijk aan het licht zijn gekomen. Het eerste hoofdstuk behandelt de teksten uit de zogenaamde 'Antwerpse School', waarvan sommige werken worden toegeschreven aan de schepenklerk Jan van Boendale. In het volgende hoofdstuk analyseer ik de Middelnederlandse Arturroman met veel aandacht voor de Haagse *Lancelotcompilatie*. In het laatste hoofdstuk van dit deel wordt ook de Middelnederlandse Karelepiek in het onderzoek betrokken.

7 BOENDALES SCHOOL

7.1 JAN VAN BOENDALE

Zoals uit de vorige hoofdstukken al ruimschoots is gebleken, werken stylometrici het liefst met langere teksten omdat deze de duidelijkste resultaten opleveren in kwantitatieve experimenten inzake auteursherkenning. Om die praktische reden heb ik tot nu vooral historiografische werken behandeld, omdat die tot de langste teksten in de Middelnederlandse letterkunde behoren. Wie op zoek gaat naar wereldlijke oeuvres in de medioneerlandistiek die qua omvang en diversiteit min of meer vergelijkbaar zijn met bijvoorbeeld dat van Maerlant, komt automatisch uit bij Jan van Boendale. Zeker voor de fascinerende figuur van Boendale is het verleidelijk, maar ook gevaarlijk, om te vervallen in een zeker biografisme: hij is een van de vroegste auteurs uit de Nederlandse literatuurgeschiedenis voor wie wij over uitgebreide biografische informatie beschikken – men weet bijvoorbeeld zelfs in welk huis hij woonde. Hoewel dit toelaat om (het oeuvre van) deze dichter op een verrassende manier te contextualiseren, zij meteen opgemerkt dat de vele biografische informatie niet wegneemt dat de afbakening van Boendales oeuvre controversieel is. Deze problematiek staat in dit hoofdstuk centraal.[1]

7.1.1 *Jan, gheheten Clerk*

Jan van Boendale was een Middelnederlandse dichter, geboren in de streek rond Tervuren – Boondaal is een nabijgelegen gehucht in het huidige Elsene (Claassens 2001). Uit administratieve bronnen is bekend dat de Boendales een groep woningen (een *bloc*) in Tervuren bezaten en tot de sociale middenklasse behoorden. Maurice Wynants, een Tervuurs heemkundige, rapporteerde enkele jaren geleden de vondst van een *Jan* en *Margriete Boendaels* in lokale bronnen, die men mogelijk mag gelijkstellen met Boendales ouders (Wynants 1994). Over Boendales vroegste levensjaren is amper iets bekend. De hypothese van Geert Warnar dat hij net als Jan van Ruusbroec zijn opleiding heeft genoten aan de Brusselse kapittelschool is wegens de geografische nabijheid van deze instelling niet onaannemelijk, maar kan bij een gebrek aan relevante bronnen voor geen van beide auteurs worden

1 Dit hoofdstuk steunt in belangrijke mate op twee reeds verschenen artikels (Kestemont 2011a, 2012a). Graag dank ik Bram Caers, Renée Gabriël, Daniëlle Prochowski, Dirk Schoenaers en Robert Stein voor hun zorgvuldige opmerkingen bij eerdere versies van deze bijdrage. Ook Remco Sleiderink wil ik danken voor de vele aangename conversaties over dit onderwerp.

geverifieerd.[2] Opvallend is wel dat een gelijkaardige achtergrond en scholing zijn gesuggereerd voor Jacob van Maerlant aan het kapittel te Brugge (Van Oostrom 1996, 29-80). Een dergelijk parcours is voor Boendale natuurlijk voorstelbaar maar blijft bij de huidige stand van het onderzoek erg speculatief. Interessant is wel dat Boendale in een van zijn teksten, de *Lekenspiegel*, mogelijk refereert aan een Brusselse dichter (een vroegere schoolleraar?), Hendrik van Aken van Brussel, van wie hij twee verzen heeft onthouden.[3]

Van Boendale is bekend dat hij zeker één literair werk schreef: *Jans Teesteye*, een moralistisch gedicht dat een discussie weergeeft tussen Boendales alter ego en zijn fictionele gesprekpartner Wouter over het moreel verval in de samenleving (Snellaert 1869). Het werk werd in de vroege jaren dertig van de veertiende eeuw opgedragen aan Rogier van Leefdaal (en diens echtgenote), toen een van de belangrijkste hertogelijke raadgevers (gestorven in 1333).[4] Of de edelman uit Leefdaal (nabij Tervuren) als latere mecenas van Boendale een rol heeft gespeeld in de opleiding of vroegste tewerkstelling van de dichter is moeilijk uit te maken. In de proloog van *Jans Teesteye* vinden we belangrijke biografische informatie:

> Alle die ghene die dit werc
> Sien, lesen ende horen
> Die gruetic, Jan, gheheten Clerc
> Vander Vueren gheboren.
> Boendale heetmen mi daer
> Ende wone te Andwerpen nu,
> Daer ic ghescreven hebbe menech jaer
> Der scepenen brieve, dat segghic u.
> (Snellaert 1869, vs. 1-8)

Jan van Boendale blijkt naar eigen zeggen in de vroege jaren dertig al een tijd werkzaam in de Antwerpse stadsadministratie als schepenklerk. De eerste volledige stadsrekening uit Antwerpen is overgeleverd uit het jaar 1324: daarin blijkt ene 'Jan de klerk' – hoogstwaarschijnlijk Boendale (vgl. *Jan, gheheten Clerc*) – inderdaad erg actief binnen de Antwerpse stadsadministratie als hoofdklerk – of toch: hoogst bezoldigde klerk. Boendale stond al vroeger in dienst van de stad (cf. *menech jaer*). Er moet bijvoorbeeld een stadsrekening uit 1313-1314

2 Zie Warnar (1997, 140); vgl. Vermassen (2011, 85-86).

3 Vgl. Reynaert (2003, 10-12). De indruk van een 'Brusselse connectie' wordt verstevigd door het feit dat deze verzen wel nog o.m. voorkomen als marginale notities in enkele Ruusbroec-handschriften (Lievens 1960, 65-66). Het citaat komt ook als spreuk voor in de Geraardsbergse codex (Govers et al. 1994, fol. 110r). Wel moet men rekening houden met de mogelijkheid dat deze latere attestaties gewoon teruggaan op de *Lekenspiegel* zelf. Zie voor Hendrik ook § 10.2.

4 De datering van *Jans Teesteye* is onzeker: men gaat er van uit dat de *Teesteye* vóór 1333 (de dood van Rogier) werd geschreven, maar ná de *Lekenspiegel* (voltooid in 1330), omwille van een allusie op dit laatste werk in de *Teesteye* (Kinable 1997, 25, n. 25 en 33, n. 45).

bestaan hebben waarvan slechts een deel overgeleverd is, maar waaruit verschillende zeventiende-eeuwse historici aantekeningen geput hebben die wel bewaard zijn. Meestal wordt voor Boendale verwezen naar de volgende rekening-ingang, opgetekend door L. van Caukercken (1629-1701): *Heer Jan Willemaers, dat hij was te Vueren met Jan de Clerck met de andere steden 8 lb. 10 s., feria sexta post Epiphaniam*.[5] De Antwerpse schepen reisde bij deze gelegenheid naar Tervuren voor een van de statenvergaderingen volgend op de dood van hertog Jan II. De *Jan de Clerck* in dit gezelschap wordt doorgaans geïdentificeerd als Jan van Boendale, meestal onder sterke impuls van de vermelding van zijn geboorteplaats Tervuren.[6] Het lijkt hier inderdaad te gaan om onze Jan van Boendale, al is voorzichtigheid geboden omdat er in die periode wel meerdere 'Jannen' als klerk voor de stad Antwerpen werken.[7] Of Jan van Boendale reeds vóór deze datum dan wel naar aanleiding van deze bijeenkomst in een Antwerps dienstverband trad, kan niet uit de vermelding worden opgemaakt.

In latere bronnen komen wij meer te weten. Boendale verschijnt in de jaren twintig en dertig als de stadsklerk in Antwerpse rekeningen, die naast het toezicht op schepen belast was met de controle op de handel in de lakenhal.[8] Op basis van administratieve bronnen kan men een kleurrijk beeld schetsen van zijn gevarieerde loopbaan: zo blijkt hij de belangen van de stad Antwerpen bij verschillende gelegenheden te hebben verdedigd, bijvoorbeeld op ettelijke diplomatieke missies in de jaren dertig. Op latere leeftijd moet Boendale in hoog aanzien hebben gestaan. Uit lokale bronnen is zoals gezegd geweten dat hij Antwerpens hoofdklerk moet zijn geweest in 1324. De *Wardrobe Accounts* van het Engelse hof onder Edward III vermelden dat het hof *Johanni le Clerc de Andewerpia* en zijn collega's had begiftigd, omdat deze Franse schepen hadden bespioneerd langs de Schelde in de jaren 1339-1340.[9] Boendale komt in één van de ingangen als volgt ter sprake: *Johanni le Clerc de Andewerpia et sociis suis vigilantibus noctanter in batellis suis ad explorandum de galeis venientibus versus Andewerpiam* (Lucas 1937, 368). In de ogen van buitenlanders schijnt de klerkenleider dus een groepje gezellen of getrouwen (*et sociis suis*) om zich te hebben verzameld.[10]

5 Naar Prims (1931, 231); op p. 232 vinden we nog een andere gekopieerde ingang die een *Jan den Clerck* vermeldt (vgl. Prims 1933, 223-224).

6 Zie Génard (1853, 170): 'Trouwens wie was geschikter om den Schepen naar *Tervueren* te vergezellen, dan de daar geborene en met de plaats van nabij bekende BOENDALE'.

7 Vgl. *Jan de Scarfhundere*, in 1302 *der stad clerc* of de 'meesters' *Jan van den Ghere* en *Jan op de porte* in 1324, tegelijk met Boendale. Zie Willems (1839, XIII-XIV), De Vreese (1900, 83-86) of Van Uytven (2002, 28).

8 Boendales administratieve en diplomatieke activiteiten worden nader beschreven bij Appelmans (2002) en Van Anrooij (1994b, 1995).

9 Zie Lucas (1937, 368). Geheel terzijde: Boendales toenaam vormt een prachtige illustratie van de flexibiliteit van de middeleeuwse persoonsnaamgeving. In Tervuren wordt hij 'van Boendale' genoemd, in Antwerpen wordt hij 'van Tervuren' genoemd en in Engeland wordt hij 'van Antwerpen' genoemd.

10 Een interessant detail – dat eerder ook door Remco Sleiderink werd opgemerkt – betreft Wouter, de ietwat naïeve gesprekspartner van Boendales alter ego in *Jans Teesteye*,

Hoewel Boendales geboorte- en sterfdatum hevig bediscussieerd zijn, blijkt de dichter in 1343-1344 zeker nog in leven. In 1343 treden *Johanne dicto Clerico et Nicolao dicto Noese, clericis* op als getuigen bij een schenking van de rijke Antwerpse koopman Hendrik Suderman.[11] Uit een akte uit 1344 blijkt dat Boendales woonst trouwens slechts door één huis gescheiden werd van die van Suderman (Prims 1933, 226-227). Uit deze gegevens is men geneigd op te maken dat Boendale en Suderman goede kennissen waren, mogelijk zelfs vrienden, aangezien uit Boendales optreden als getuige in 1343 een zekere verbondenheid met de initiatieven van zijn buurman spreekt (Génard 1853, 199). Boendale, die zich als stadsklerk tussen de Antwerpse stadselite begaf, lijkt een *selfmade* man die volop de toenmalige mogelijkheden voor sociale mobiliteit heeft benut.

Wat betreft zijn dichterlijke carrière blijft het onduidelijk of Boendale naast *Jans Teesteye* nog meer werken heeft geschreven. Voor één tekst is dat erg aannemelijk: de *Lekenspiegel* (De Vries 1844–1848). De tekst bevat een uitgebreid leerdicht van bijna 22.000 verzen, waarin uiteenlopende onderwerpen aan bod komen. De tekst is een uitstekend voorbeeld van de didactische lekenethiek en richt zich in eerste instantie tot een publiek van niet-geestelijken. De proloog van de *Teesteye* vermeldt Boendale namelijk dat zijn patroon hem bij een vroegere gelegenheid had bekritiseerd omwille van enkele vrouwonvriendelijke passages in eerder werk: *Want hi mi sulc stont heeft der af / Berespt van dien dat ik hem gaf* (vs. 100-101). Er zijn goede redenen om dit eerdere werk te identificeren als de *Lekenspiegel*. Ook deze tekst is immers opgedragen aan Rogier van Leefdaal (De Vries 1844–1848, 5, vs. 48-49). In een proloog die de auteur aan een latere versie toevoegde, draagt hij zijn tekst eveneens op aan hertog Jan III en noemt hij zichzelf *Jan, u arme clerk*.[12] Aangezien de *Lekenspiegel* inderdaad enkele sterk misogyne passages bevat, is er eigenlijk geen goede reden om de toeschrijving aan Boendale nog te betwijfelen (vgl. Brinkman 1993, 53).

---

een goedgelovige jongeling die het lastig heeft met het pareren van de opmerkingen van zijn geleerde en ervaren gespreksgenoot. Nu kennen we uit rekeningen zo'n 'Woutertje' die Boendale erg goed moet gekend hebben (Génard 1864, 179, n. 1). In de stadsrekeningen van 1324 wordt ettelijke keren gesproken van een *Woutre der scepen knape*, de klerkenhulp die wordt ingezet om de brieven van de schepenen en klerken van hot naar her te dragen. Is het niet aantrekkelijk om in Jans gesprekspartner van de *Teesteye* een vertederende verwijzing te ontdekken naar deze jonge klerkenbode? Gezien de kleine publiekskring waarop middeleeuwse werken in de regel anticipeerden, zou een dergelijke ludieke knipoog natuurlijk niet verbazen. Een contextuele benadering van 'Boendales Wereld' biedt nog heel wat mogelijkheden voor toekomstig onderzoek.

11 Zie Génard (1853, 199-200) of Van Uytven (2002, 27-28).

12 Zie Willems (1837, 341). Het standpunt van bv. Bender (1899) lijkt achterhaald.

7.1.2 *Boendale erbij*

Men kan er dus van uitgaan dat Boendales oeuvre tenminste de *Lekenspiegel* en *Jan Teesteye* bevatte. Voor dit onderzoek heb ik de standaardedities van deze teksten van de *Cd-rom Middelnederlands* gehaald en de rijmwoorden verrijkt met lemma's. In deze sectie kunnen we ook Boendales teksten betrekken in een stylometrische vergelijking met de eerder aan bod gekomen delen van de *Spiegel historiael* door Utenbroeke, Maerlant en Velthem. Vanzelfsprekend kunnen Boendales teksten relatief verder van deze *Spiegel*-teksten verwijderd zijn, dan de bijdrages van bijvoorbeeld Utenbroeke en Maerlant onderling. De *Spiegel historiael* bevat zowat uitsluitend geschiedkundig materiaal: hoewel Boendales beide teksten eveneens een sterk geschiedkundige inslag hebben, treden hier ook andere thema's (didactiek en ethiek) sterker op de voorgrond. Als er een stilistisch verschil is te bespeuren tussen de hoogfrequente rijmwoordenschat van bijvoorbeeld Utenbroeke en Boendale, kán dit dus ten dele te wijten zijn aan meer inhoudsgerelateerde factoren.

Toch moeten we de mogelijkheid tot onderscheiding van de stijl in deze teksten niet overschatten. Boendale groeide op ten oosten van Brussel tijdens de eerste helft van de veertiende eeuw: dat maakt hem een streek- en tijdgenoot van Lodewijk van Velthem. Aangezien de moedertaal van beide dichters betrekkelijk dicht bij elkaar kan hebben gelegen, is het niet a priori zeker dat men hun rijmstijlen makkelijk zal kunnen onderscheiden. Bovendien heeft Maerlant een grote invloed uitgeoefend op Boendale. De Antwerpse schepenklerk duidde Maerlant in zijn *Lekenspiegel* zelfs aan als de stamvader van de Dietse letterkunde: *Jacob van Maerlant, die vader / Is der dietscher dichtren algader* (De Vries 1844–1848, III, kap. 15, vs. 118-119). Hoogstwaarschijnlijk heeft Maerlant dus ook op deze navolger een grote stilistische invloed uitgeoefend.

Boendales *Lekenspiegel* (afgekort *LS*) telt na verrijking 21.669 rijmwoorden; de *Teesteye* 4.101 (afgekort *JT*). Deze kunnen we combineren met de reeds besproken Tweede Partie van Utenbroeke (*P2*, 32.312 rijmwoorden), de Derde Partie van Maerlant (*P3*, 31.080 rijmwoorden) en de *Spiegel*-delen van Velthem, waaruit we de Heelu-passages en het 'verdachte' vierde boek uit de Vijfde Partie hebben verwijderd (*P4P5*: 14.237 rijmwoorden). We zullen hieronder gebruik maken van de rijmwoorden die cumulatief gezien het meest frequent zijn in dit corpus. Op dit moment is het daarom interessant om te testen of deze rijmwoorden wel goed gespreid zijn in het corpus. Daarvoor doen we opnieuw een beroep op de variatie-coëfficient, de eerder besproken maat om de spreiding van een woord in een verzameling samples te meten (§ 6.2.2).

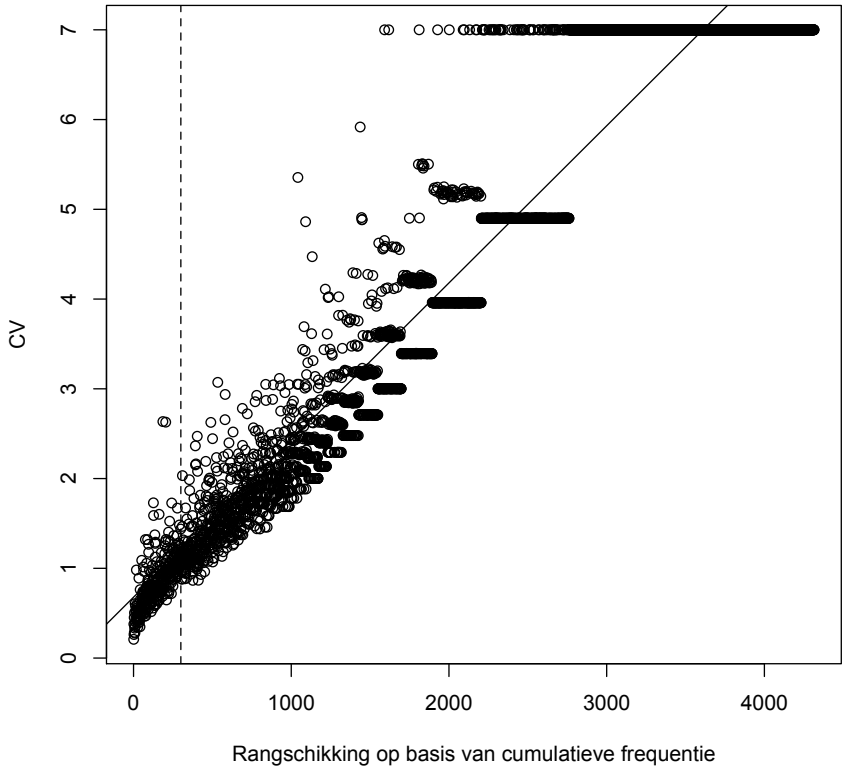
In onderstaande figuur wordt het resultaat van een dergelijke proef grafisch weergegeven (Fig. 32, p. 159). De beschikbare teksten zijn

verdeeld in samples van 2.000 rijmwoorden. Alle rijmwoorden in deze samples (cirkels) werden gerangschikt op basis van hun cumulatieve frequentie (horizontale as) en in de figuur uitgezet op basis van de variatie-coëfficient van hun frequenties in de samples (verticale as). Opnieuw kunnen we een duidelijke correlatie vaststellen tussen de rangorde van een rijmwoord en de spreiding ervan: de meest frequente rijmwoorden in de verzameling samples kennen over het algemeen ook de beste spreiding. Het lineair model in deze figuur (diagonale rechte) rapporteert dan ook een groot en significant effect. De verticale stippellijn beeldt een drempelwaarde uit van 300 rijmwoorden: als men voor deze teksten de 300 meest frequente rijmwoorden selecteert, wordt dus enkel gewerkt met de cirkeltjes links van de stippellijn die voor het leeuwendeel een relatief goede spreiding kennen.

Laat ons op basis van deze drempelwaarde van 300 rijmwoorden een *Principal components analysis* (PCA) uitvoeren van de samples in deze teksten. Zoals gezegd, kunnen dergelijke technieken voor dimensiereductie het auteurschap van teksten blootleggen als zij worden toegepast op de frequenties van hoogfrequente functoren in teksten. Hierboven is reeds getoond dat deze techniek soms in staat is om middeleeuwse kopiisten van elkaar te onderscheiden en in het vorige hoofdstuk kon de aan PCA verwante techniek correspondentie-analyse de driedeling detecteren die traditioneel op Maerlants oeuvre wordt geprojecteerd. De vraag is of PCA ook bruikbaar zal zijn om de verschillen in de rijmstijl van de betrokken vier auteurs te detecteren.

Het is echter niet aangewezen om de samples van alle vier auteurs tezamen in een PCA te steken: stylometrici hebben er op gewezen dat een PCA in de mate van het mogelijke moet beperkt blijven tot een vergelijking van drie auteurs, wil zij echt bewijskrachtig zijn (Binongo & Smith 1999, 464). Meestal wordt immers gewerkt met een visualisering van de twee hoofdcomponenten en twee variabelen kunnen maar zoveel variatie beschrijven. De achterliggende idee schijnt te zijn dat één van beide hoofdcomponenten steeds één auteur van de twee andere kan onderscheiden. Natuurlijk kán een vergelijking van meerdere auteurs in principe een auteur-gerelateerde scheiding opleveren, maar dat is niet noodzakelijk.

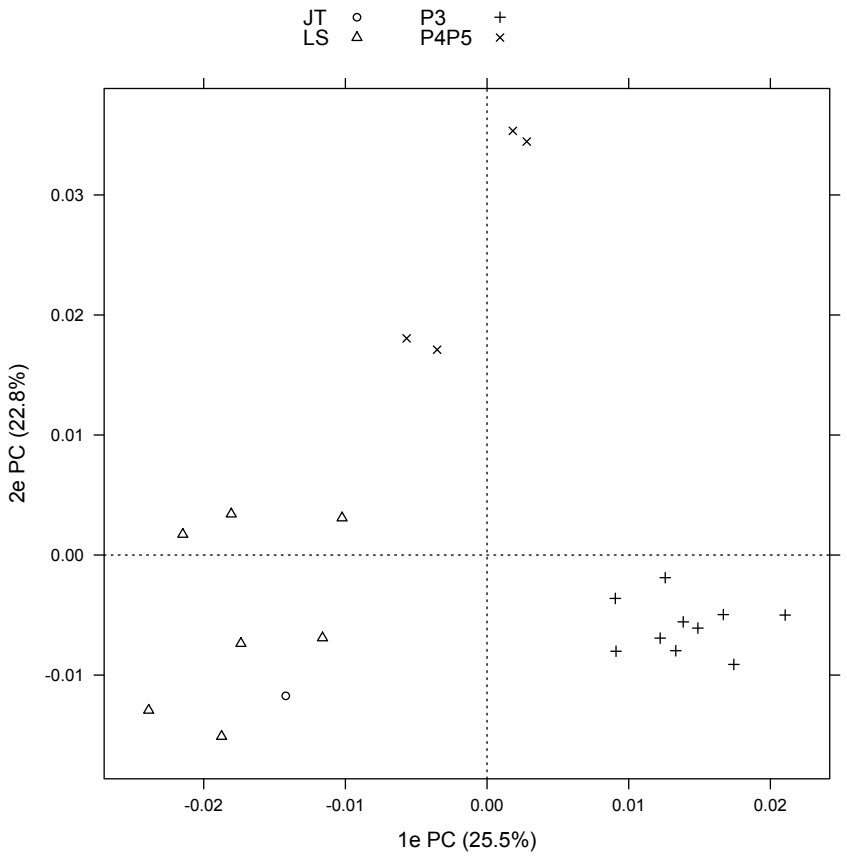
In de volgende figuren hebben we steeds de teksten van drie van de vier auteurs in een PCA vergeleken – Figuren 33 (p. 160), 34 (p. 161), 35 (p. 162) en 36 (p. 163). Voor deze piloot-experimenten gebruikten we een voldoende grote sample-grootte (3.000 rijmwoorden) en relatief veel hoogfrequente rijmwoorden (300). Met zulke 'luxueuze' parameters is de kans immers groter dat we een goede clustering van het materiaal van elke auteur zullen krijgen. In de resulterende strooidiagrammen is de positie van ieder sample in de eerste twee PCS weergegeven aan de hand van een symbooltje dat uitdrukt uit welke tekst een sample afkomstig is. Aangezien we werken met sample-



Figuur 32: Experiment van samples van 2.000 rijmwoorden uit *P2* (Utenbroeke), *P3* (Maerlant), *P4P5* (Velthem) en twee teksten van Boendale (*Lekenspiegel* en *Jans Teesteye*). De cirkeltjes in de grafiek stellen de rijmwoorden voor die zijn gerangschikt op basis van hun cumulatieve frequentie in de samples (horizontale as). De waardes op de verticale as geven de variatie-coëfficient van deze woorden weer. De verticale stippellijn (x=300) geeft een voorbeeld weer van een drempelwaarde (300 rijmwoorden). Het lineair model (diagonale rechte) rapporteert een groot en significant effect: $F(1, 4.313) = 50,050$, $p < 0{,}001$, adjusted $R^2 = 0,92$.

grootte van 3.000 rijmwoorden is er bijvoorbeeld steeds maar één *JT*-sample beschikbaar.

Op basis van deze vier figuren lijkt het redelijk te stellen dat de samples goed schijnen te clusteren in groepen met een identiek auteurschap (Fig. 33, 34, 35 en 36). Samples van verschillende teksten worden eerder zelden vermengd en vormen doorgaans goed te onderscheiden sample-wolkjes. Het mooiste voorbeeld is wellicht het strooidiagram waarin enkel samples uit de *Spiegel historiael* zijn betrokken (Fig. 36, p. 163). De samples uit Maerlants *P3* (driehoekjes) zitten beneden links in de figuur, terwijl die van Utenbroekes *P2* (cirkeltjes) bovenaan links zijn gesitueerd. Velthems samples treffen we daarentegen uiterst rechts aan (plus-tekens). In deze figuur wordt Velthems stijl als erg

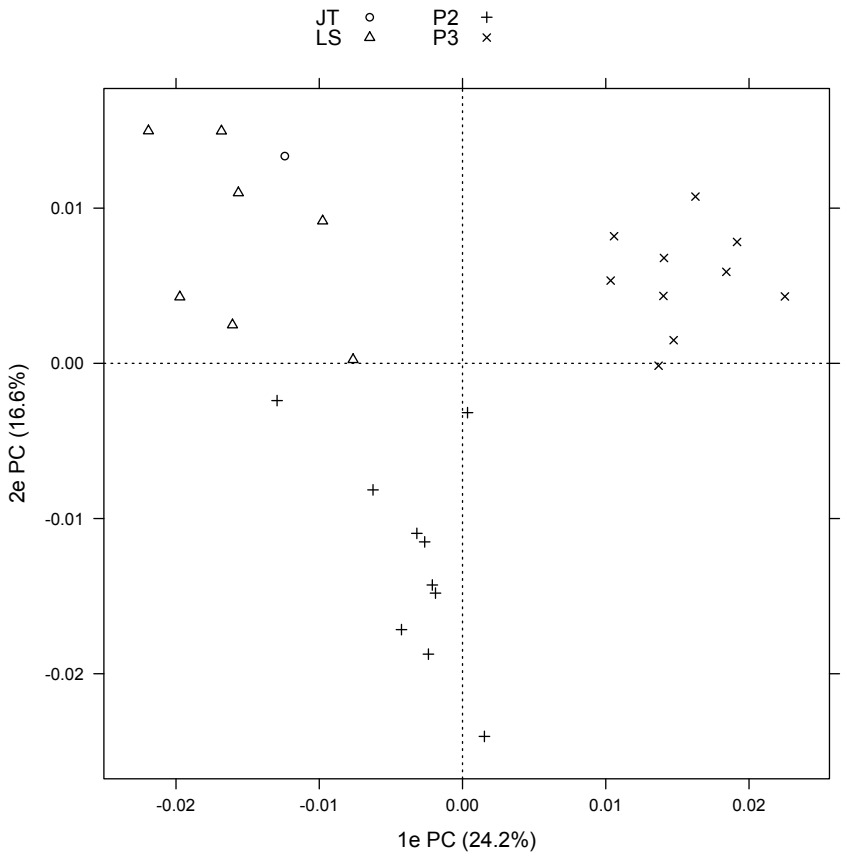


Figuur 33: Resultaat van een PCA van samples van 3.000 rijmwoorden uit *LS* en *JT* (Boendale), *P3* (Maerlant) en *P4P5* (Velthem). De samples zijn voorgesteld aan de hand van de 300 meest frequente rijmwoorden erin.

afwijkend van de andere twee auteurs gekenmerkt, want zijn samples worden – niet toevallig in de eerste PC – duidelijk onderscheiden. Het enkele *JT*-sample in deze analyses treffen we steeds temidden de *LS*-samples aan, hetgeen de stilistische affiniteit tussen beide teksten benadrukt. Mocht het auteurschap van *JT* nog niet vaststaan, dan zou men bij de interpretatie van deze strooidiagrammen een *nearest neighbour*-redenering kunnen toepassen en *JT* toeschrijven aan de auteur van *LS*, *casu quo* Boendale. Natuurlijk kan men het in dat geval oneens zijn over hoe dicht een sample bij andere samples moet liggen om daadwerkelijk tot een toeschrijving te mogen overgaan.

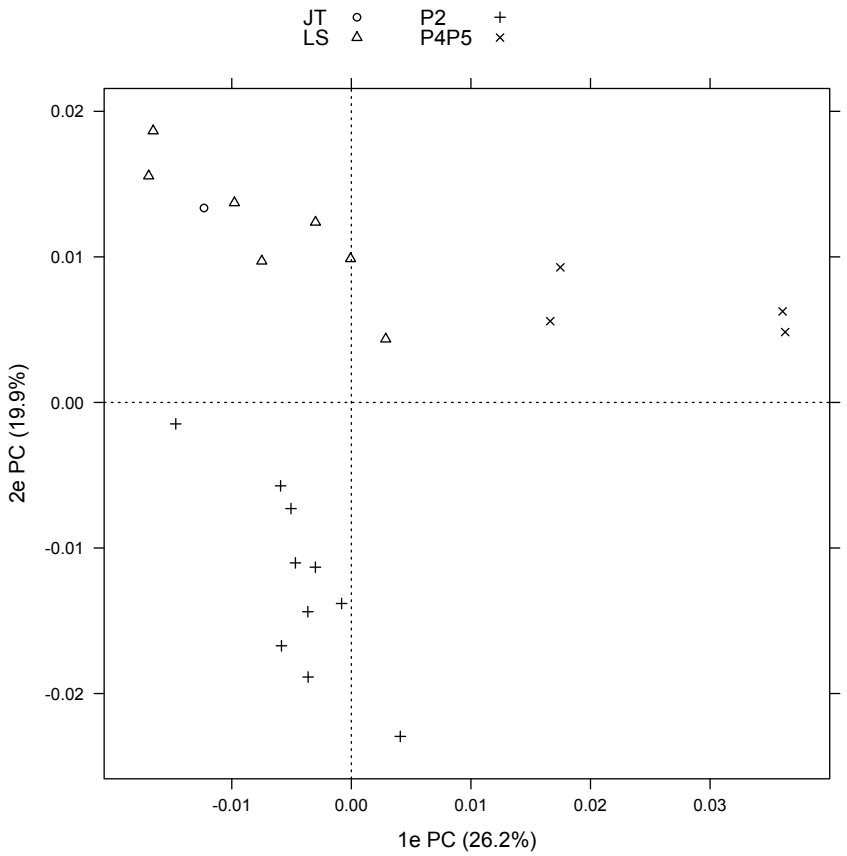
7.2 DE KLASSE VAN BOENDALE

Zoals reeds benadrukt is het grote voordeel van technieken voor dimensie-reductie dat zij uiteindelijk een grafische voorstelling van teksten opleveren. Het visueel resultaat is daarom handig en aangenaam bij het interpreteren van de stilistische relaties tussen teksten.



Figuur 34: Resultaat van een PCA van samples van 3.000 rijmwoorden uit *LS* en *JT* (Boendale), *P3* (Maerlant) en *P2* (Utenbroeke). De samples zijn voorgesteld aan de hand van de 300 meest frequente rijmwoorden erin.

Toch zijn er twee markante nadelen aan PCA. Ten eerste is de *nearest neighbour*-redenering die bij de lezing van strooidiagrammen wordt toegepast soms wat speculatief. De inschatting hoe dicht een sample bij een ander sample ligt, kan dus beïnvloed worden door de vooropgestelde hypotheses van een onderzoek. In de figuur hierboven waarin *P2*, *P3* en Boendales werken werden geanalyseerd (Fig. 34, p. 161), zijn er bijvoorbeeld samples van Utenbroeke en Boendale die vervaarlijk dicht bij elkaar komen te liggen. Een ander probleem is dat een PCA vaak slechts voor een beperkt aantal experimentele parameters kan worden getoond. Hierboven werkten we om retorische redenen bijvoorbeeld met erg grote samples en relatief veel rijmwoorden. We zijn natuurlijk ook geïnteresseerd in hoe snel ons ogenschijnlijk goede stilistische model aan kracht zou moeten inboeten als we minder genereuze parameters gebruiken. Niettemin is het vaak onbegonnen werk om voor ieder miniem verschil in de keuze van specifieke parameters een volledig strooidiagram af te beelden en te bespreken.

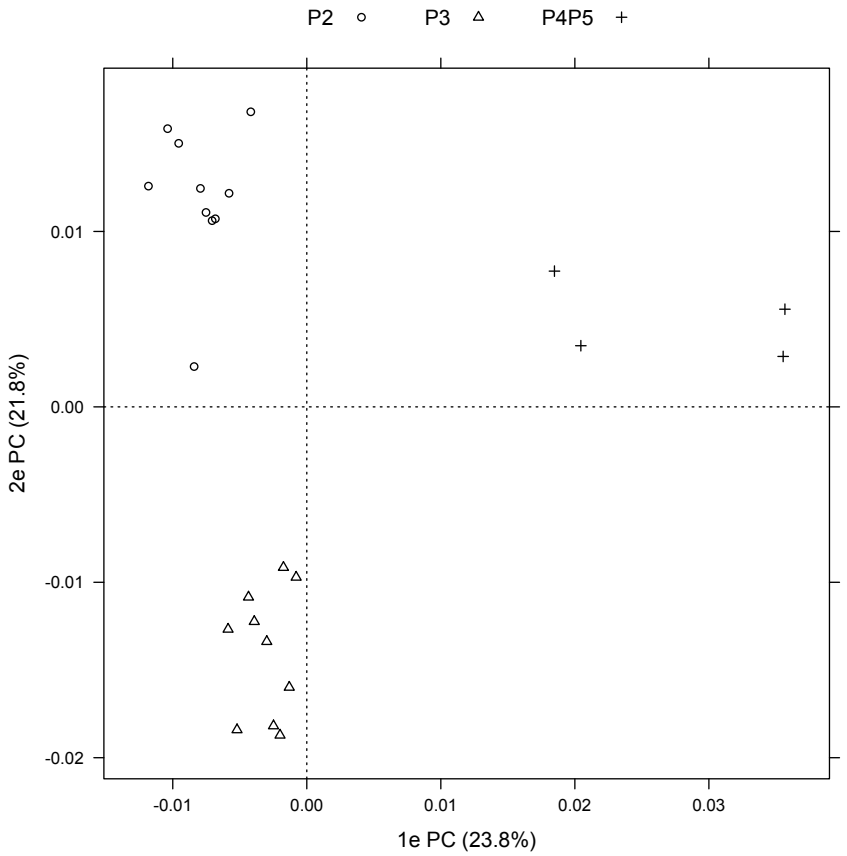


Figuur 35: Resultaat van een PCA van samples van 3.000 rijmwoorden uit *LS* en *JT* (Boendale), *P2* (Utenbroeke) en *P4P5* (Velthem). De samples zijn voorgesteld aan de hand van de 300 meest frequente rijmwoorden erin.

Deze euvels kunnen tot op zekere hoogte worden geremedieerd door het toepassen van tekstclassificatie via een *leave-one-out*-schema. Zoals gezegd heeft deze aanpak ook zijn nadelen, want deze retourneert meestal slechts een globale accuraatheid van alle toeschrijvingen in een experiment, hetgeen niet altijd nieuwe inzichten oplevert over de interne structuur van onze sample-verzameling. Dergelijk accuraatheidscijfers zijn echter wel handig om de resultaten van een grote reeks verschillen in parameters te vergelijken. Het is daarom aantrekkelijk om de tekstclassificatie-aanpak op deze vier auteurs toe te passen. Daarvoor zullen we gebruik maken van Burrows' Delta, een invloedrijke attributie-techniek die ik hieronder kort toelicht.

7.2.1 *Naïeve Delta?*

In 2001 hield de toonaangevende stylometricus J. Burrows een veelgeciteerde lezing bij het in ontvangst nemen van de *Roberto Busa Award*. Daarin presenteerde Burrows een nieuwe methode voor au-



Figuur 36: Resultaat van een PCA van samples van 3000 rijmwoorden uit *P2* (Utenbroeke), *P3* (Maerlant) en *P4P5* (Velthem). De samples zijn voorgesteld aan de hand van de 300 meest frequente rijmwoorden erin.

teursherkenning, 'Delta', die later bijzonder invloedrijk zou worden. Een herwerkte versie van zijn lezing verscheen in het tijdschrift *Literary and Linguistic Computing* (Burrows 2002). In zijn titel duidde Burrows de Delta-procedure aan als een 'Measure of Stylistic Difference and a Guide to Likely Authorship'.

In wezen gaat de methode uit van een situatie waarin men beschikt over een anonieme tekst die moet worden toegeschreven aan één auteur uit een reeks kandidaat-auteurs.[13] Van iedere van deze kandidaat-auteurs moet men beschikken over een voldoende grote hoeveelheid tekst waarvan het auteurschap niet wordt betwijfeld. Verder kan men ook gebruik maken van een (oorspronkelijk nogal vaag gedefinieerd) 'controlecorpus' met relevant tekstueel vergelijkingsmateriaal. Steeds vaker zien we echter dat men geen gebruik maakt van een extern controlecorpus, maar dit controlecorpus gelijkstelt aan de combinatie van de anonieme tekst en de teksten van de kandidaat-auteurs.

13 Mijn bespreking steunt in grote mate op de eerste bladzijden uit Argamon (2008).

Binnen deze setting kan men gebruik maken van de zogenaamde Delta-procedure, die qua opzet verrassende gelijkenissen vertoont met de *Memory-based learning* die hierboven reeds uitgebreid aan bod is gekomen (§ 4.2.2). De Delta-metriek zal de anonieme tekst vergelijken met de samples van de kandidaat-auteurs en bepalen welk kandidaat-sample het meest lijkt op het anonieme. De procedure zal het anonieme sample vervolgens toeschrijven aan de auteur van de *nearest neighbour* (vgl. klasse-extrapolatie).

Delta vergelijkt teksten op basis van de (relatieve) frequenties van een kleine verzameling van n hoogfrequente woorden (bv. n=30), zoals gebruikelijk in de stylometrie. Welke woorden worden gebruikt, bepaalde Burrows oorspronkelijk op basis van de cumulatieve woordfrequenties in het controlecorpus. Als de teksten nog geen vergelijkbare grootte hebben, zal men samples van een gelijke grootte moeten extraheren uit zowel de anonieme tekst, de teksten van de kandidaat-auteurs en het grotere controlecorpus. Vervolgens zal men voor elk van onze 30 woorden nagaan hoe vaak zij voorkomen in de samples in het controlecorpus. Ook bepaalt men de standaarddeviatie van deze frequenties. Als het controlecorpus 1.000 samples telt, zal men dus voor elk woord i 1.000 frequenties verzamelen en berekenen wat de standaarddeviatie is ten aanzien van de gemiddelde frequentie. Het gemiddelde en de standaarddeviatie duidt men voor ieder woord i respectievelijk aan als μ_i en σ_i.

Hoe de eigenlijke afstand (i.e. Delta, voor *Difference*) tussen twee samples wordt berekend, vatte Burrows als volgt samen:

> the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text (Burrows 2002, 271).

Deze ogenschijnlijk ingewikkelde berekening kan betrekkelijk eenvoudig worden geformaliseerd. Het ingewikkelde uitzicht van onderstaande formules werkt mogelijk een beetje ontmoedigend, maar bij een nadere inspectie zal blijken dat zij betrekkelijk eenvoudig zijn.

Stel dat we de Delta willen berekenen tussen de samples X en Y, waarbij X het anonieme sample voorstelt. Deze Delta-waarde drukken we uit als $\Delta(X, Y)$. Laat n het aantal hoogfrequente woorden zijn dat we willen gebruiken (bv. 30). Voor ieder woord i van deze n woorden kunnen we in ons controlecorpus de gemiddelde frequentie berekenen (μ_i) en de standaarddeviatie ten aanzien van dit gemiddelde (σ_i). Bovendien kunnen we de frequentie van woord i in sample X berekenen – deze noemen we $f_i(X)$ – en de frequentie van hetzelfde woord i in sample Y – dewelke we als $f_i(Y)$ aanduiden. De z-score waarvan Burrows gebruikt maakt, kunnen we eenvoudig weergeven.

Voor het woord met de willekeurige index i in document X is die z-score (genoteerd als $z(f_i(X))$):

$$z(f_i(X)) = \frac{f_i(X) - \mu_i}{\sigma_i} \tag{7.1}$$

De z-score van woord i in sample X is bijgevolg gelijk aan het verschil tussen de frequentie van het bewuste woord in sample X en de gemiddelde frequentie van i in het controlecorpus, gedeeld door de standaarddeviatie van het woord in het controlecorpus. De volledige Delta tussen de twee samples wordt nu berekend als de 'average absolute difference of the z-scores'. Dit gemiddelde kunnen we wiskundig formaliseren aan de hand van een een sommatieteken (Σ):

sommatieteken

$$\Delta(X, Y) = \frac{1}{n} \sum_{i=1}^{n} |z(f_i(X)) - z(f_i(Y))| \tag{7.2}$$

Via deze notatie drukken we uit dat we de bewerking rechts van de Σ uitvoeren voor elk van onze n woorden (bovenaan de Σ weergegeven), te beginnen bij het eerste van onze n woorden, namelijk het woord met de index $i = 1$ (onderaan de Σ weergegeven). De verticale streepjes ($|\ldots|$) aan weerszijden van de bewerking geeft aan dat we het absolute verschil van de bewerking willen. Als we bijvoorbeeld 5 van 3 aftrekken, bekomen we -2, maar als we 3 van 5 aftrekken krijgen we de uitkomst 2. Het *absolute* verschil tussen 3 en 5 is 2, namelijk de uitkomst zonder het min-teken in beide bewerkingen. Deze uitkomsten (n in totaal) worden vervolgens samengeteld – vandaar het sommatieteken. Deze som wordt uiteindelijk gedeeld door n, hetgeen links van de Σ als $\frac{1}{n}$ is genoteerd. Dit gemiddelde – de eigenlijke afstand of 'Delta' tussen de samples X en Y – noteren we als $\Delta(X, Y)$.

absolute verschil

In 2008 merkte Argamon (2008) op dat deze formule mathematisch kan worden vereenvoudigd. Als we de z-scores immers voluit schrijven in de bovenstaande formule geeft dat:

$$\Delta(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{f_i(X) - \mu_i}{\sigma_i} - \frac{f_i(Y) - \mu_i}{\sigma_i} \right| \tag{7.3}$$

Deze vergelijking kan vereenvoudigd worden tot:

$$\Delta(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{f_i(X) - f_i(Y)}{\sigma_i} \right| \tag{7.4}$$

En aangezien n voor alle Delta-berekeningen in één experiment hetzelfde is, kan het zelfs nog korter:

$$\Delta(X, Y) = \sum_{i=1}^{n} \frac{1}{\sigma_i} |f_i(X) - f_i(Y)| \tag{7.5}$$

Deze nieuwe formule is mathematisch gesproken veel eenvoudiger dan Burrows' oorspronkelijke formulering en kan makkelijker in mensentaal worden uitgelegd. Om de Delta tussen twee samples te berekenen, verzamelen we voor ieder van onze *n* woorden het absolute verschil in frequentie in beide samples. Dit absolute verschil delen we door de standaarddeviatie van het woord in het controlecorpus, die hier als 'wegingsfactor' wordt gebruikt. De uiteindelijke Delta is gelijk aan de som van deze *n* gewogen frequentieverschillen.

Wat vooral opvalt aan Argamons vereenvoudigde formule, is dat wij een variant ervan in deze studie reeds hebben toegepast. Daarvoor moeten even terugkeren naar het hoofdstuk waarin we een naïeve *Memory-based learner* hebben toegepast op de classificatie van samples uit de Tweede en Derde Partie van de *Spiegel historiael* (§ 4.2.2). In wezen hanteerden wij daar gewoon de volgende vergelijkingsformule:

$$\Delta(X, Y) = \sum_{i=1}^{n} |f_i(X) - f_i(Y)| \tag{7.6}$$

Afgezien van het feit dat wij geen wegingsfactor σ_i gebruikten, was onze werkwijze blijkbaar een naïeve, wegingsloze variant van het bekende Delta-algoritme.[14] In dit hoofdstuk kunnen we het gebruik van een Delta-met-weging verkennen. Als wegingsfactor voor de frequentieverschillen tussen woorden kunnen we dan de standaarddeviatie in de verschillende teksten gebruiken.[15] Het gebruik van een dergelijk wegingsfactor heeft overigens een belangrijk effect: als de standaarddeviatie voor een woord in het algemeen groot is, zullen frequentie-verschillen voor dit woord minder zwaar meetellen in de globale som. Omgekeerd zullen de frequentie-verschillen voor een woord extra gewicht in de schaal kunnen leggen, als dit woord in het algemeen een eerder beperkte standaarddeviatie in het corpus vertoont. In dat laatste geval is een frequentieverschil tussen twee samples immers extra markant.[16]

wegingsfactor

14 Merk op dat Delta over het algemeen trouwens een vrij naïef algoritme is, omdat het geen rekening houdt met de onderlinge afhankelijkheid van woordfrequenties (vgl. Argamon 2008, 140ff), waar technieken voor dimensie-reductie dat wel doen (vgl. Koppel et al. 2009, 10).

15 Een technische noot: voor het berekenen van deze wegingsfactor wordt in veel stylometrische studies geen strikt onderscheid gemaakt tussen de train- en testinstanties, hetgeen methodologisch verdacht kan lijken voor onderzoekers in Machine Learning. Toch is de werkwijze goed te rechtvaardigen, aangezien men geen gebruik maakt van de klasse-informatie van de testitems en de methode dus ook probleemloos op nieuwe, ongelabelde instanties kan worden toegepast.

16 Het gebruik van een wegingsfactor voor de Delta-procedure is onder meer verkend door Hoover. Deze merkte op dat het weglaten van de wegingsfactor in sommige gevallen zelfs een matig positief effect heeft op de attributie-resultaten (Hoover 2004b, 480). Onze (toch relatief goede) resultaten met een wegingsloze Delta voor Maerlant en Utenbroeke (§ 4.2.3) lijken te bevestigen dat men zonder wegingsfactor ook nog behoorlijke resultaten kan verkrijgen.

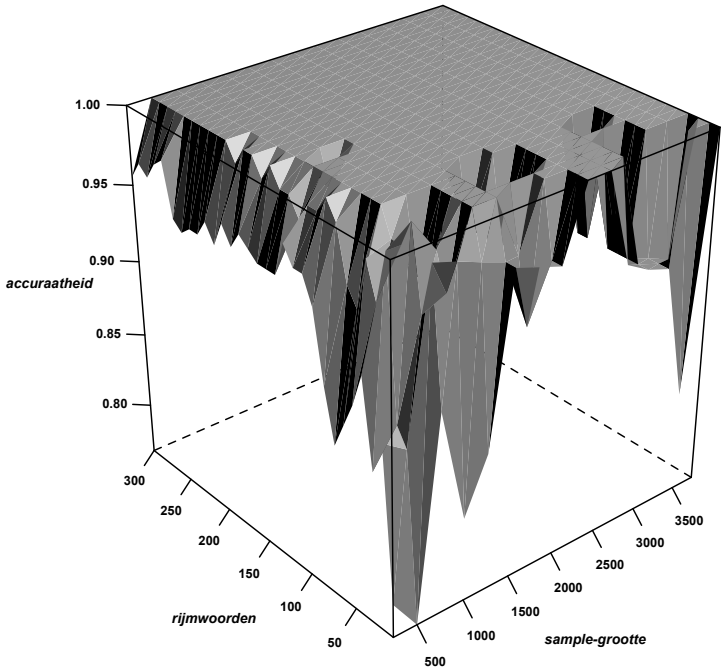
7.2.2 *Delta-experimenten*

We kunnen nu als volgt te werk gaan om het effect van verschillende parameters op onze analyses te onderzoeken. Laat ons variëren in de gebruikte sample-grootte en het effect van de volgende waardes nagaan: 250, 500, 750, ... , 3.250, 3.500, 3.750. Laat ons ook het effect nagaan van het gebruik van verschillende aantallen hoogfrequente rijmwoorden (10, 20, 30, ... , 280, 290, 300). Voor elke combinatie van deze sample-groottes en het aantal gebruikte rijmwoorden zetten we nu een apart experiment op, waarin we de teksten steeds samplen en voorstellen op basis van de specifieke parameters in dit experiment. Dat zal toelaten om na te gaan hoe de accuraatheid van een analyse varieert met deze parameters – van heel kleine samples die aan de hand van heel weinig rijmwoorden worden voorgesteld, tot erg grote samples die aan de hand van relatief veel rijmwoorden worden voorgesteld.

We gebruiken in elk experiment een *leave-one-out*-validering: telkens wordt één sample uit de beschikbare verzameling gelicht en als testinstantie aan een van de controle-auteurs toegeschreven. Als dit voor alle samples is gebeurd, gaan we na hoeveel van de samples aan de correcte auteur zijn toegeschreven, hetgeen een maat oplevert voor de globale accuraatheid van het experiment. We gebruiken Burrows' vereenvoudigde Delta als toeschrijvingsalgoritme, waarbij de testinstantie steeds wordt toegeschreven aan de auteur van de 'dichtste buur', of het sample dat de kleinste Delta vertoonde ten aanzien van het testitem.

Om een visuele voorstelling van dit resultaat te verkrijgen, gebruiken we een driedimensionele visualisering. De horizontale assen stellen daarbij de gebruikte sample-groottes voor, evenals het aantal geanalyseerde rijmwoorden. De verticale dimensie geeft voor ieder snijpunt van de horizontale assen (i.e. iedere parameter-combinatie) de accuraatheid weer, zodat een relatief hogere graad van correcte toeschrijvingen zal resulteren in een hogere, verticale waarde. Dat levert een soort berglandschap op, waarvan de veranderingen in het reliëf toelaten om de verschillen tussen specifieke parameter-combinaties in een oogopslag duidelijk te maken.

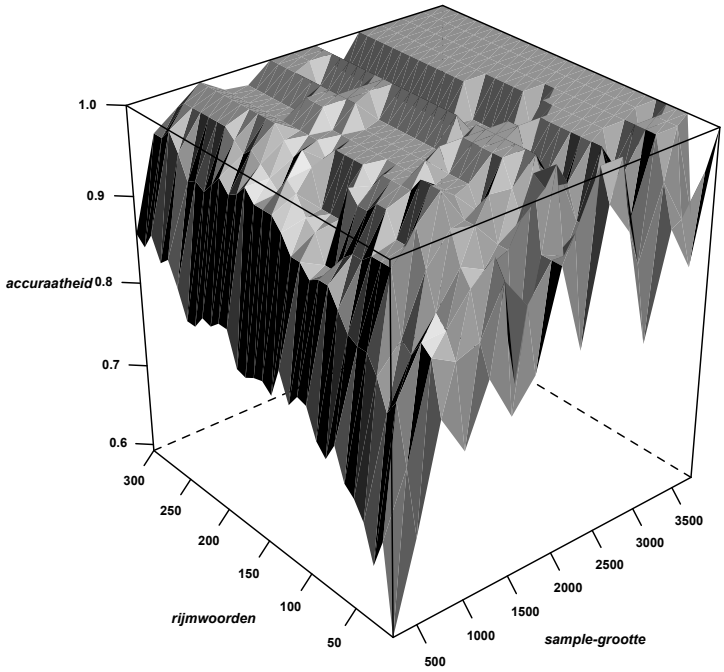
In de volgende figuren zijn de resultaten van drie verschillende experimenten weergegeven – Figuren 37 (p. 168), 38 (p. 169) en 39 (p. 170). In het eerste experiment werden enkel Maerlant en Velthem vergeleken; in het tweede en derde experiment werden ook de teksten van respectievelijk Utenbroeke en Boendale betrokken. Dat laat toe om na te gaan of het aantal geanalyseerde auteurs een belangrijk effect heeft. Alle drie de figuren maken duidelijk dat de kwaliteit van de toeschrijvingen eerder goed is en veelal significant boven het kansniveau (de *baseline*) ligt. Het effect is duidelijk waarneembaar dat hogere parameters (bv. meer rijmwoorden) een gunstig effect hebben



Figuur 37: Driedimensionele weergave van het resultaat van een *leave-one-out*-experiment op samples uit teksten van twee auteurs: *P3* (Maerlant) en *P4P5* (Velthem). De accuraatheid is weergegeven op de verticale as voor verschillende parameter-combinaties (horizontale assen).

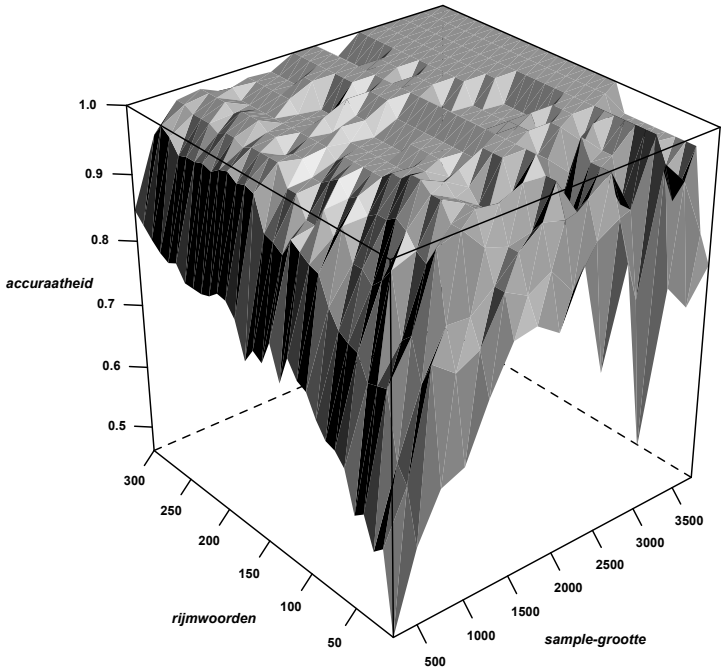
op de attributie-accuraatheid. Wel schijnt het Delta-algoritme pas foutloze resultaten op te leveren bij relatief grotere samples en een groter aantal rijmwoorden. Zeker als meer auteurs worden betrokken, worden pas goede resultaten behaald met relatief genereuze parameters. Daaruit kunnen we afleiden dat we de kracht van het Delta-algoritme zeker niet mogen overschatten: hoewel experimenten met slechts twee auteurs al snel goede resultaten opleveren, blijkt het zeker in het geval van meerdere auteurs noodzakelijk om voldoende grote stukken tekst te gebruiken. Om een echt foutloos resultaat te bereiken lijkt een sample-grootte van 2.000 woorden gewenst. Dat wil overigens niet zeggen dat met minder goede parameters onzinnige resultaten worden bereikt, want soms worden ook met bescheiden parameters nog verrassend hoge scores bereikt.

Het resultaat van de experimenten in de vorige sectie is aan één kant zeker enthousiasmerend: wanneer wij technieken voor dimensie-



Figuur 38: Driedimensionele weergave van het resultaat van een *leave-one-out*-experiment op samples uit teksten van drie auteurs: *P2* (Utenbroeke), *P3* (Maerlant) en *P4P5* (Velthem). De accuraatheid is weergegeven op de verticale as voor verschillende parametercombinaties (horizontale assen).

reductie toepassen op de hoogfrequente rijmwoordenschat van middeleeuwse dichters, blijkt het resultaat in veel gevallen betrouwbaar het auteurschap van de betrokken teksten te weerspiegelen. Aan de andere kant werd de mogelijkheid tot stylometrische auteursherkenning meteen ook gerelativeerd. Analyses blijken zelden feilloos en hangen sterk af van de keuzes voor specifieke parameters, zoals sample-grootte. Bovendien mag men niet de illusie koesteren dat de hoogfrequente rijmwoordenschat van teksten énkel met auteurschap samenhangt – dat bleek ook uit het vorige hoofdstuk. Tot op zekere hoogte zullen inhoudsgerelateerde factoren een rol hebben gespeeld in onze experimenten. Daarom mag men zeker niet claimen dat de stylometrische methode de literair-historische waarheid in pacht heeft. Wel kan de stylometrie interessante argumenten leveren die – samen met andere argumenten – kunnen helpen in het debat rond het auteurschap van



Figuur 39: Driedimensionele weergave van het resultaat van een *leave-one-out*-experiment op samples uit teksten van vier auteurs: *P2* (Utenbroeke), *P3* (Maerlant), *P4P5* (Velthem) en *LS* en *JT* (Boendale). De accuraatheid is weergegeven op de verticale as voor verschillende parameter-combinaties (horizontale assen).

teksten. Het kan immers ook benadrukt worden dat de stylometrische methode in verschillende experimenten erg zinvolle resultaten opleverde. Laat ons daarom de hierboven beschreven methodes toepassen op een auteursprobleem 'uit de echte wereld' en nagaan in hoeverre zij daarin kunnen helpen.

7.3 DE DIDACTISCHE TEKSTEN

Hoewel Boendales auteurschap moeilijk kan worden betwijfeld voor *Jans Teesteye* en de *Lekenspiegel*, is het heel goed mogelijk dat de Antwerpse schepenklerk veel meer werken heeft geschreven. Er bestaat namelijk een grote groep anonieme werken die vermelden dat zij in de eerste helft van de veertiende eeuw tot stand kwamen in Antwerpen. Het gaat om een vrij hecht corpus aan moralistische, ethische en

historiografische geschriften die in toon, inhoud en wereldbeeld erg dicht bij elkaar aanleunen.[17] Zij lijken ook vaak op elkaar te alluderen via stilistische echo's en min of meer woordelijke ontleningen. Velen denken ook voor deze teksten aan Boendale als auteur. Hard bewijs voor die toeschrijvingen ontbreekt echter.

Deze toestand heeft decennialang aanleiding gegeven tot heel wat onderzoek. Een eerste 'kamp' van onderzoekers is voorzichtig gebleven: deze 'minimalisten' stelden dat Boendales auteurschap slechts voor een relatief klein aantal werken aannemelijk kan worden gemaakt.[18] Voor het overgrote deel van deze werken was zijn auteurschap in hun ogen niet meer dan een aantrekkelijke mogelijkheid. Een tweede kamp, de zogenaamde 'maximalisten', zijn veel verder gegaan in hun toeschrijvingen. Zij stelden dat veel meer, zoniet alle Antwerpse werken door de schepenklerk moeten zijn geschreven. Het Antwerpse corpus zou volgens de maximalistische zienswijze daarom nagenoeg samenvallen met Boendales oeuvre.

De aanhoudende twijfels omtrent het auteurschap van deze teksten hebben in de vorige eeuw aanleiding gegeven tot de introductie van het begrip 'Antwerpse School'.[19] In 1989 maakte de minimalist J. Heymans voorzichtig gebruik van deze – reeds door Van Mierlo gebruikte – term om het corpus Antwerpse teksten (tezamen meer dan 60.000 versregels) aan te duiden.[20] Het is opvallend hoe makkelijk de term ingang heeft gevonden in het onderwijs en onderzoek, zeker na 1989. Het gaat natuurlijk om een methodologisch handige nomenclatuur die als relatief neutrale aanduiding de onzekerheid opvangt omtrent Boendales inbreng in het corpus Antwerpse werken. Niettemin is de term wat misleidend want hij heeft de connotatie van een echte 'school', in de zin van een goed georganiseerd samenwerkingsverband van gelijkgestemde dichters met gemeenschappelijke idealen. In die zin zou de 'Antwerpse School' het aantrekkelijke maar ietwat geromantiseerde idee kunnen oproepen van een soort voorloper van de latere rederijkerskamers. Dat is natuurlijk een hypothese waarvoor momenteel evenmin tastbare bewijzen bestaan.

In een belangrijk 'maximalistisch' artikel heeft Joris Reynaert beargumenteerd dat Boendale verschillende andere Antwerpse werken heeft geschreven (Reynaert 2002). Eén tekst die Reynaert behandelde, was de *Melibeus*, in de middeleeuwen ook bekend als *Van Melibeo ende van sinre vrouwen Prudencia of dat boec van troeste ende van rade* (uitgegeven bij Snellaert 1869, 1-133). Deze tekst vormt een verta-

17 Een karakterisering biedt Van Anrooij (1994a).

18 De opdeling van het onderzoek in een 'maximalistisch' en een 'minimalistisch' kamp ontleen ik aan Appelmans (2002). Zie dezelfde publicatie voor verdere referenties.

19 Een erg recente bespreking van de problematiek vindt men bij Riedel-Bierschwale (2009).

20 Verschillende onderzoekers stelden dat het begrip werd geïntroduceerd door Heymans (1989b) – mogelijk omdat Heymans zelf ook niet verwijst naar Van Mierlo (1940, 26ff).

ling van het Latijnse *Liber consolationis et consilii* door Albertanus van Brescia. De anonieme tekst werd halfweg april 1342 in Antwerpen afgewerkt. De tekst vertoont veel affiniteit met andere teksten uit de 'Antwerpse School' inzake tijd en plaats van ontstaan, maar verder ook in taal en stijl, zodat Boendales auteurschap voor de *Melibeus* vaak is gepostuleerd.

Hetzelfde geldt voor de *Dietsche doctrinale*, eveneens een vertaling van een tractaat van Albertanus van Brescia, *De amore et dilectione Dei et proximi et aliarum rerum et de forma vitae* (uitgegeven bij Jonckbloet 1842). De anonieme auteur vermeldt dat hij de tekst afwerkte in 1345 te Antwerpen, zodat de *Dietsche doctrinale* bijzonder snel na de *Melibeus* moet zijn ontstaan. De hypothese is dan ook aantrekkelijk dat beide gedichten van dezelfde hand zijn. De Latijnse grondteksten van deze zusterteksten komen trouwens vaak tezamen in middeleeuwse handschriften voor, zodat men zelfs heeft beargumenteerd dat de auteur van deze teksten (Boendale?) voor beide bewerkingen op dezelfde legger een beroep heeft gedaan (Reynaert 2002, 133). De aanname dat Boendale de *Dietsche doctrinale* schreef, gaat al een hele tijd mee (zie ook § 7.4.2). Reeds uit de zestiende eeuw kennen we een notitie die *den Duytschen Doctrinael* toeschrijft aan *Jan de Clerck, secretaris van Antwerpen* (Willems 1839, XIII).

De *Melibeus* en *Dietsche doctrinale* vallen beide op door hun sterk moralistische ondertoon: de teksten formuleren voortdurend morele richtlijnen voor de lezers en zijn sterk gericht op ethische onderwerpen. Datzelfde schijnt ten dele waar voor het *Boec vander wraken*, een (opnieuw anonieme) tekst die slechts zeven jaar na de *Dietsche doctrinale* tot stand kwam. De tekst licht het einde der tijden en de Dag des Oordeels toe aan de hand van diverse historiografische verhalen. Hoewel de insteek van het *Boec vander wraken* daarmee sterker geschiedkundig is, blijken ook de thema's van moraal, didactiek en ethiek nog sterk aanwezig. De geschiedenis wordt door de auteur immers aangegrepen als een aanleiding om morele richtlijnen te formuleren.

De auteur zou reeds een eerste versie kunnen hebben afgewerkt in 1346 – dus al snel na de *Doctrinale* – maar Van Anrooij (1995) heeft beargumenteerd dat de auteur de tekst ca. 1351 zou hebben gereviseerd. Ook voor het *Boec vander wraken* is Boendale een aantrekkelijke auteurskandidaat. Erg interessant is dat de auteur op het einde van de tekst vermeldt dat hij zijn werkzaamheden moest staken: hij anticipeerde dat hij binnenkort zou gaan sterven aan de gevolgen van *arbeit ende outheit*. Van Boendale wordt vaak aangenomen dat die ca. 1351 zou zijn gestorven, zodat de chronologie van het *Boec vander wraken* opvallend parallel loopt met Boendales levenslijn en mogelijk het laatst bekende werk van deze dichter vormt (Van Anrooij 1995, zie verderop ook § 7.4.2).[21]

21 Merk op dat deze teksten vaak intertekstueel verbonden zijn door (soms erg subtiele) citaatjes. Ik heb geen poging ondernomen om dergelijke kruisverwijzingen uit het

De stylometrische methodes die wij hieronder zullen toepassen, vertonen één opvallend verschil met het voorgaande stilistisch onderzoek naar deze teksten: de nadruk op hoogfrequente verschijnselen. De methodologie van bijvoorbeeld Reynaert (2002) focuste precies op het omgekeerde. Reynaert extraheerde uit deze teksten een eerder klein aantal kenmerken die hij op basis van zijn eerdere leeservaringen als 'eigenaardigheden' beschouwde, zoals ongebruikelijke substantieven of zeldzame syntactische constructies. Aan de hand van deze lijst doorzocht hij via de zoekfunctie het corpus Middelnederlandse literaire teksten op de *Cd-rom Middelnederlands*. Verschillende van de fenomenen in de *Melibeus* en *Doctrinale* die Reynaert als kenmerkend voor Boendale had getypeerd, bleken op de *Cd-rom* slechts voor te komen in andere teksten die aan Boendale worden toegeschreven. Op basis van dergelijke observaties – uitzonderlijke en laagfrequente fenomenen die slechts in de teksten uit de 'Antwerpse School' voorkomen – beargumenteerde Reynaert dat *Melibeus* en *Doctrinale* eveneens door Boendale moesten geschreven zijn.

Reynaerts conclusies zijn in de medioneerlandistiek instemmend onthaald (bv. Sleiderink 2003, 191, n. 79). Zijn methodologie is bij mijn weten amper in twijfel getrokken en heeft zelfs navolging gevonden (Hogenbirk 2009b). Vanuit de stylometrische inzichten die de basis vormen van deze studie kan de advocaat van de duivel echter retorische vraagtekens plaatsen bij Reynaerts werkwijze (vgl. § 2.1.1). Zijn werkwijze maakte bijvoorbeeld geen duidelijk onderscheid tussen teksten die 'zeker' en de teksten die 'mogelijk' aan Boendale kunnen worden toegeschreven. Op die manier wordt in de toeschrijving een gelijk gewicht toegekend aan de gelijkenissen tussen, bijvoorbeeld, de *Melibeus* en *Jans Teesteye* (zeker van Boendale) en de *Melibeus* en het *Boec vander wraken* (mogelijk van Boendale). Bovendien wordt de samenstelling van de lijst onderzochte kenmerken amper toegelicht. Afgezien van het feit dat lexicale en syntactische stijlkenmerken vrijelijk worden gecombineerd, blijft meestal onduidelijk waarom bepaalde elementen wel en andere helemaal niet worden onderzocht. Bovendien wordt helemaal niet verantwoord waarom de wel bestudeerde kenmerken typisch voor de individuele stijl van een auteur zouden zijn. Daardoor ontstaat een cirkelredenering: woorden die Reynaert typisch acht voor teksten uit de 'Antwerpse School', blijken typisch voor teksten uit de 'Antwerpse School'.

De kenmerken die Reynaert onderzocht, werden bovendien gedomineerd door uitzonderlijke, laagfrequente stijlkenmerken en dus kenmerken die eigenlijk al binnen de teksten uit de 'Antwerpse School' niet zo heel vaak voorkomen. Precies door hun lage frequentie hebben deze tekstkenmerken immers Reynaerts aandacht getrokken. Het voornaamste methodologische probleem dat hiermee is gemoeid, is

materiaal te verwijderen: zij zijn erg moeilijk te inventariseren en bovendien vaak erg subtiel (Van Anrooij 1994a, 150, n. 9).

dat deze elementen ook door middeleeuwse imitators zullen zijn opgemerkt en niet alleen door moderne onderzoekers. Als er inderdaad een 'Antwerpse School' bestond waarin dichters hun stijl op elkaar afstemden, zouden dergelijke kenmerken dus de eerste zijn geweest om onderling uit te wisselen.

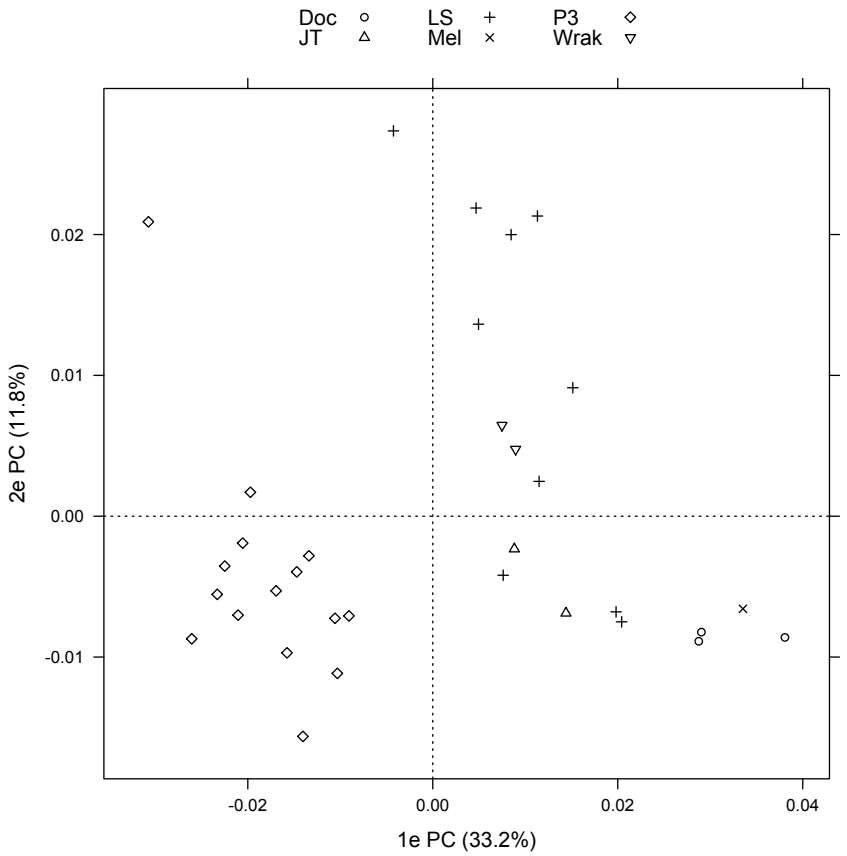
Hoewel dat niet noodzakelijk een ander beeld zal opleveren, kunnen we hieronder nagaan of een onderzoek dat zich enkel op functorachtige rijmwoorden richt, andere resultaten oplevert. Als er inderdaad meerdere dichters actief zijn geweest binnen de 'Antwerpse School', valt immers te verwachten dat het corpus Antwerpse teksten markante stilistische opposities bevat. Deze opposities zouden in dat geval immers gebaseerd zijn op tekstelementen waaraan een eventuele imitator van Boendale minder aandacht zal hebben besteed dan de opvallende stijlkenmerken die bij Reynaert aan bod zijn gekomen.

Laat ons de Antwerpse teksten in kwestie (*Mel*, *Doc* en *Wrak*) in een PCA confronteren met, ten eerste, de teksten die met zekerheid aan Boendale kunnen worden toegeschreven en, ten tweede, het controlemateriaal van de drie dichters van de *Spiegel historiael*. Op die manier proberen we recht te doen aan Brinkmans Negende Gebod, dat stelt dat teksten van twee auteurs slechts zinnig vergeleken kunnen worden als er ook een relevante derde in de vergelijking wordt betrokken. Onderstaande figuren bieden de strooidiagrammen die resulteerden uit dergelijke analyses – Figuren 40 (p. 175), 41 (p. 176) en 42 (p. 177). Ik heb hier steeds een sample-grootte van 2.000 woorden gehanteerd en de 300 meest frequente rijmwoorden uit deze teksten.[22] In de eerste drie analyses werd Boendales vermeende 'oeuvre' steeds met de tekst van één controle-auteur geconfronteerd. Dat is belangrijk aangezien het niet raadzaam is om meer dan drie auteurs in een PCA te betrekken. Als we de mogelijkheid willen openlaten dat er naast Boendale nog een tweede auteur in het Antwerpse corpus verscholen is, mogen we dus slechts één controle-auteur beschouwen.

Negende Gebod: Tertium comparationis

Deze figuren vertonen onderling sterke gelijkenissen. De samples van iedere controle-auteur vormen bijvoorbeeld een hechte samplewolk die veelal geconcentreerd is in één van de vier 'kwadranten' in het diagram. De samples uit de 'Antwerpse School'-teksten vullen daarentegen de vlakken in de andere drie (of twee) kwadranten. We stellen dus vast dat *LS* en *JT*, tezamen met *Mel*, *Doc* en Wrak steeds worden onderscheiden van de samples van de controle-auteur met een ruime (diagonale) marge. De relatieve nabijheid van de Antwerpse teksten bevestigt hoe dan ook nu reeds de traditionele aanname dat deze teksten een stilistische coherente tekstverzameling vormen. De *Lekenspiegel* blijkt bovendien de meest centrale 'Antwerpse' tekst in deze analyses: de verschillende *LS*-samples trekken meestal een soort

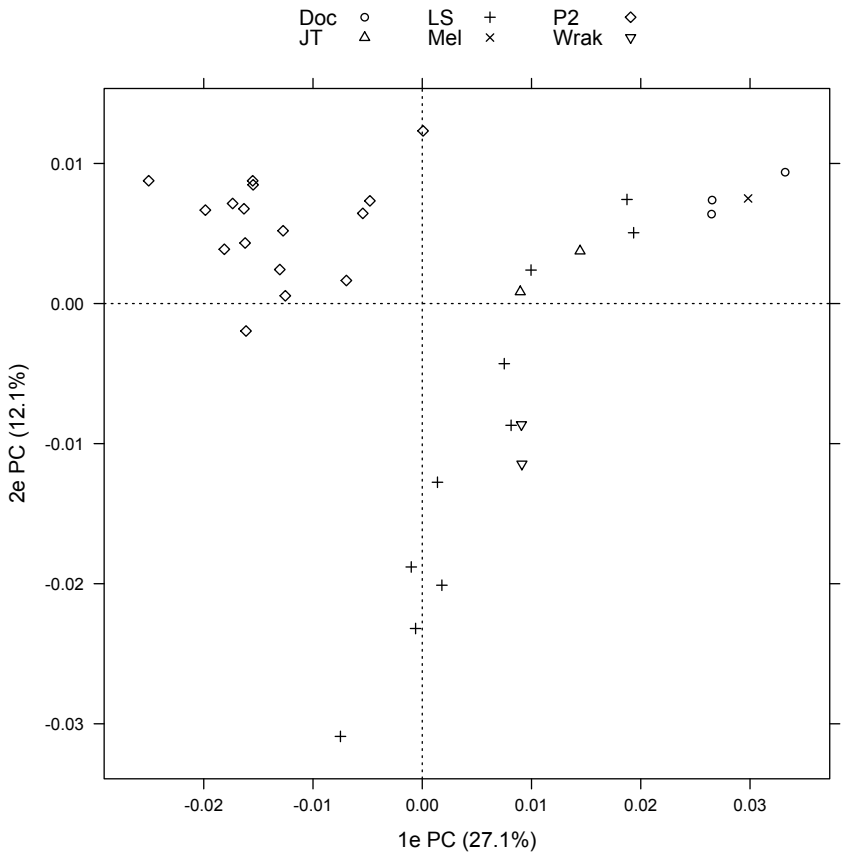
22 Andere analyses die ik hier door plaatsgebrek niet rapporteer, leverden zelden een significante afwijking op wat betreft het algemene uitzicht van deze strooidiagrammen.



Figuur 40: Strooidiagram resulterend uit een PCA van de teksten die zeker door Boendale zijn geschreven (*LS* en *JT*) en drie teksten uit de 'Antwerpse School' die mogelijk door Boendale zijn geschreven (*Mel*, *Doc* en *Wrak*). Maerlant (*P3*) is als controle-auteur gebruikt. De teksten werden onderverdeeld in samples van 2.000 rijmwoorden en voorgesteld aan de hand van de 300 meest frequente rijmwoorden erin.

diagonaal spoor door de figuren. Het is duidelijk dat de andere Antwerpse samples door dit spoor worden aangetrokken. Zeker *JT* en *Wrak* nemen overigens steeds een erg centrale positie in temidden de *LS*-samples. Als we op deze spreiding van samples een *nearest neighbour*-redenering toe te passen, zijn er dus weinig argumenten om *Wrak* niet aan Boendale toe te schrijven.

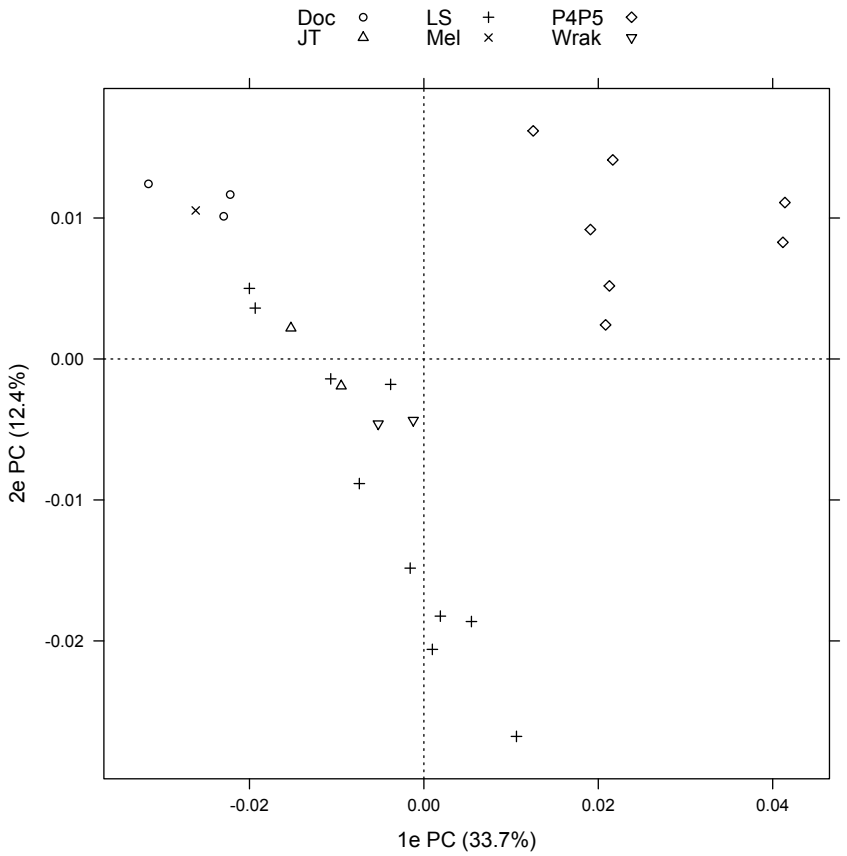
De samples uit *Melibeus* en de *Doctrinale* zijn in deze figuren iets verder verwijderd van de *LS*-samples dan bijvoorbeeld *Wrak*. Toch is deze verwijdering misschien niet zo uitgesproken als men zou verwachten in het geval dat deze teksten van een andere auteur zijn. In onderstaande figuur hebben we de analyse herhaald voor het corpus Antwerpse teksten, maar zonder controle-auteur (Fig. 43, p. 178). Daarin wordt bevestigd dat de *Melibeus* en de *Doctrinale* geen overdreven centrale plaats innemen in het corpus, want zij worden zelfs geïsoleerd in hun eigen kwadrant (linksonder). Toch lijkt de verwij-



Figuur 41: Strooidiagram resulterend uit een PCA van de teksten die zeker door Boendale zijn geschreven (*LS* en *JT*) en drie teksten uit de 'Antwerpse School' die mogelijk door Boendale zijn geschreven (*Mel*, *Doc* en *Wrak*). Utenbroeke (*P2*) is als controle-auteur gebruikt. De teksten werden onderverdeeld in samples van 2.000 rijmwoorden en voorgesteld aan de hand van de 300 meest frequente rijmwoorden erin.

dering van deze teksten in de meeste van deze experimenten niet uitgesproken genoeg om Boendales auteurschap voor deze teksten met zekerheid te verwerpen. Het probleem is natuurlijk dat we weten dat de toepassing van dimensie-reductie op een verzameling teksten van één auteur – dat is mogelijk het geval in dit laatste experiment – ook gevoelig is voor andere factoren dan auteurschap (vgl. Juola 2007). We kunnen daarom niet uitsluiten dat de (toch niet zo uitgesproken) verwijdering van *Melibeus* en de *Doctrinale* te wijten is aan andere factoren dan een niet-identiek auteurschap (vgl. § 6.3). Beide teksten zijn bijvoorbeeld kort na elkaar geschreven en zijn bovendien vertalingen uit het Latijn van grondteksten van eenzelfde auteur. Dat de teksten in de experimenten met een controle-auteur relatief dichter bij de *LS*-samples liggen, pleit in dat opzicht vóór Boendales auteurschap.

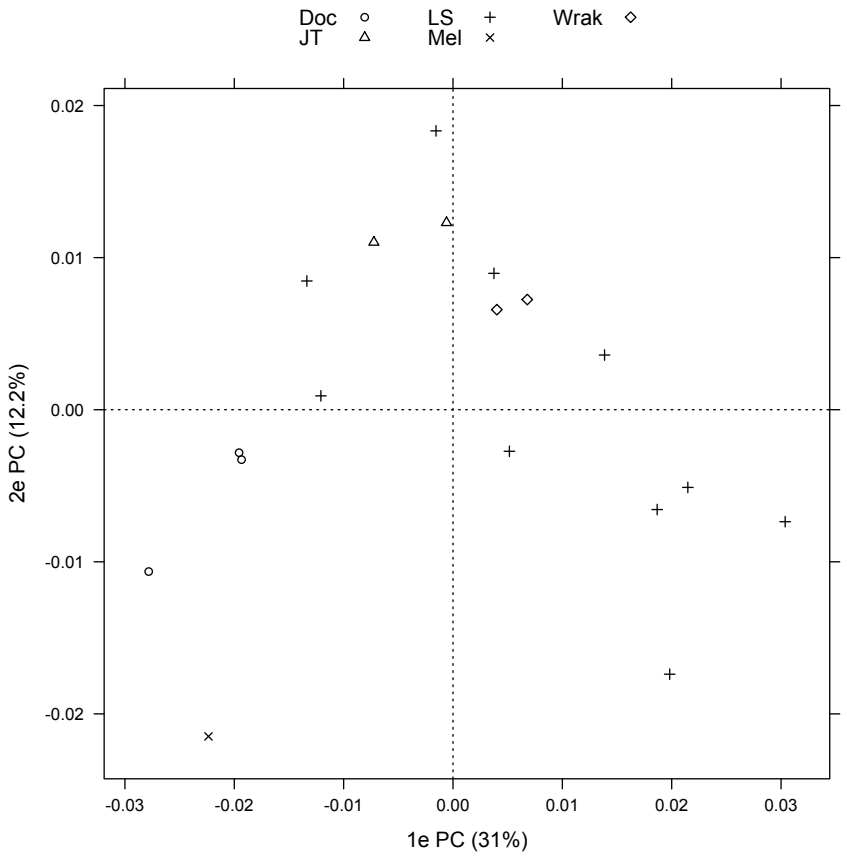
We kunnen wel nog een ultieme poging uitvoeren om Boendales auteurschap te verwerpen via een *leave-one-out*-experiment met de



Figuur 42: Strooidiagram resulterend uit een PCA van de teksten die zeker door Boendale zijn geschreven (*LS* en *JT*) en drie teksten uit de 'Antwerpse School' die mogelijk door Boendale zijn geschreven (*Mel*, *Doc* en *Wrak*). Velthem (*P4P5*) is als controle-auteur gebruikt. De teksten werden onderverdeeld in samples van 2.000 rijmwoorden en voorgesteld aan de hand van de 300 meest frequente rijmwoorden erin.

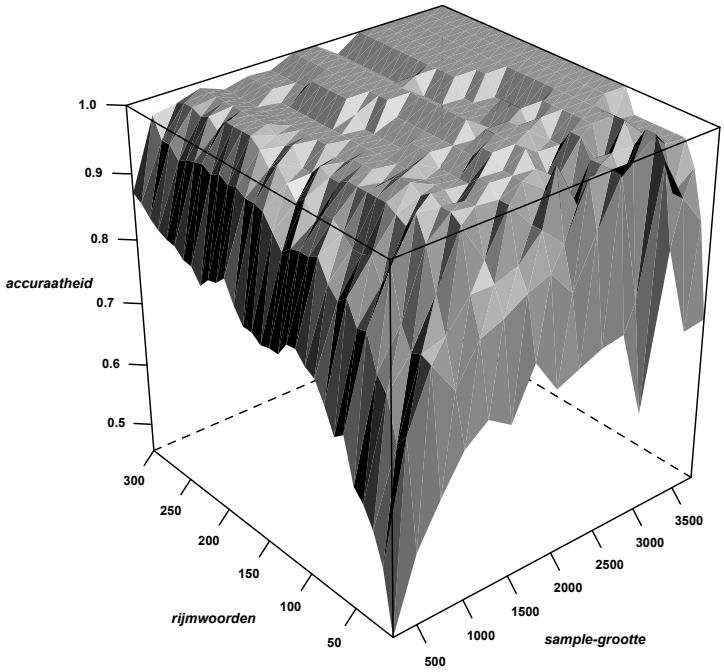
Delta-procedure. In het geval van *Doc*, *Mel* en *Wrak* kunnen we bij het bepalen van de accuraatheid veronderstellen dat een toeschrijving aan Boendale als correct moet worden beschouwd. In dat geval is een sample uit de 'Antwerpse School' immers de *nearest neighbour*. Onderstaande figuur geeft een reeks experimenten weer waarin we deze analyse hebben uitgevoerd (Fig. 44, p. 179).[23] Alle drie de controle-auteurs uit de *Spiegel historiael* werden hierbij ingezet, om de kans op verwarring bewust te vergroten. Het reliëf in deze figuur schijnt geen anomaliën te bevatten die indicatief zou kunnen zijn voor de aanname dat Boendale de Antwerpse teksten niet zelf zou hebben geschreven. De Delta-procedure schijnt over het algemeen immers geen moeite te hebben met een Boendale-toeschrijving voor *Doc*, *Mel* of *Wrak*.

23 Merk op dat de reeks waarden voor de sample-groottes beperkt is tot 3.750 rijmwoorden, zodat we steeds tenminste één sample uit de kortere *Melibeus* (3.771 rijmwoorden) konden gebruiken.



Figuur 43: Strooidiagram resulterend uit een PCA van de teksten die zeker door Boendale zijn geschreven (*LS* en *JT*) en drie teksten uit de 'Antwerpse School' die mogelijk door Boendale zijn geschreven (*Mel*, *Doc* en *Wrak*). Er werd geen controle-auteur gebruikt. De teksten werden onderverdeeld in samples van 2.000 rijmwoorden en voorgesteld aan de hand van de 300 meest frequente rijmwoorden erin.

Hieruit kunnen we concluderen dat de stylometrische methode op het eerste zicht geen tegenargumenten biedt voor Reynaerts pleidooi voor een enkelvoudig auteurschap binnen de 'Antwerpse School'-teksten. Toch mogen we die observatie wat betreft *Doc*, *Mel* en *Wrak* met een korrel zout nemen. Het enige wat deze analyse eigenlijk aannemelijk maakt, is dat van de onderzochte auteurs, Boendale de meest geschikte auteurskandidaat is voor *Doc*, *Mel* en *Wrak*. Sluitend bewijs voor Boendales auteurschap van teksten mogen we daarom zeker niet in dit resultaat zien. Onrechtstreeks kan men hierin wel een argument zien ten voordele van de Boendale-hypothese. Als er immers meerdere auteurs in het spel waren geweest, hadden we een duidelijkere tekstscheiding verwacht. Zeker voor het *Boec vander wraken* waren de rijmovereenkomsten met andere van Boendales teksten zo groot dat een meervoudig auteurschap erg onwaarschijnlijk lijkt.



Figuur 44: Driedimensionele weergave van het resultaat van een *leave-one-out*-experiment op samples uit de volgende teksten: *P2* (Utenbroeke), *P3* (Maerlant), *P4P5* (Velthem), *LS* en *JT* (zeker Boendale) en *Mel*, *Doc* en *Wrak* (mogelijk Boendale). Voor de laatste drie teksten werd slechts de toeschrijving aan Boendale als correct gerekend. De accuraatheid is weergegeven op de verticale as voor verschillende parameter-combinaties (horizontale assen).

7.4 HET COMPLEX VAN DE YEESTEN

Tot nu toe werd nog niet vermeld dat het zogenaamde corpus van de 'Antwerpse School' ook een groot aantal historiografische teksten telt. Dat had de volgende reden. In de vorige sectie werd als controlemateriaal de *Spiegel historiael* gebruikt en de toon, stijl en inhoud van deze tekst had grotere overeenkomsten kunnen vertonen met het expliciet historiografische materiaal uit de 'Antwerpse School'. Aanvankelijk had ik gevreesd dat het minder expliciet geschiedkundige materiaal uit het Antwerpse corpus zich zou verwijderen van de rest van het corpus, op basis van stijlverschillen die niet aan de factor auteurschap waren gerelateerd. Hieronder zal echter blijken dat die vrees niet ge-

grond was. De meeste historiografische teksten die hieronder worden besproken, moeten worden gesitueerd tegen het achterdoek van de *Brabantse yeesten*. Ik zal de ontstaansgeschiedenis van deze kroniek eerst grondig toelichten, aangezien het auteurschap van deze tekst een erg ingewikkelde aangelegenheid vormt. Ik hoop in de rest van dit hoofdstuk te illustreren hoe de stylometrische en traditionele onderzoeksmethode elkaar op belangrijke punten kunnen aanvullen of versterken.

7.4.1 *De ontstaansgeschiedenis van de Brabantse yeesten*

Voor ik inga op het auteurschap van de *Brabantse yeesten*, is het zinvol om de ontstaansgeschiedenis van dit werk beknopt toe te lichten. In hun totaliteit omvatten de *Brabantse yeesten* meer dan 46.000 verzen, verspreid over zeven boeken, die elk bestaan uit kleinere hoofdstukjes of kapittels. Het zogenaamde eerste deel van de kroniek (tot vers 900 van het vijfde boek) kwam getuige de proloog tot stand in opdracht van een Antwerpse patriciër en schepen, Willem Bornecolve:[24]

> Alsoe mi bat ende beval
> Van Antwerpen her Willem,
> Bornecolve noemt men hem
> (Willems 1839, I, vs. 44-46)

Verder geeft de anonieme chroniqueur te kennen dat dit deel van het werk voltooid werd *Al te Antwerpen, in die stede* (Willems 1839, I, vs. 60-62). In de proloog bespreekt de auteur van de *Yeesten* eveneens zijn bronnengebruik (vs. 21-41). De hoofdmoot van zijn informatie vond hij in de reeds uitvoerig aan bod gekomen *Spiegel historiael*. De eerste drie boeken van de *Yeesten* bevatten weinig origineel materiaal: zij vormen een vrij letterlijke, op Brabant georiënteerde bloemlezing uit dit werk – vooral uit het aandeel van Jacob van Maerlant – en behandelen de Brabantse geschiedenis vanaf de legendarische afstamming van het hertogelijk huis van de Trojanen tot het bewind van Godfried I met de baard (gestorven in 1139).[25]

In boek IV en de eerste 900 verzen van boek V gaat de kroniekschrijver meer zijn eigen weg. Reeds in de proloog had hij aangegeven dat hij de *Spiegel*-uittreksels zou aanvullen met andere *yeesten* uit een bron waar hij kennelijk relatief weinig over wist. Dit werk is op basis van de inhoud geïdentificeerd als de *Chronica de origine ducum Brabantiae*. Deze tekst werd in 1294 geschreven in het klooster van Affligem en

24 De gebruikte editie is die van Willems (1839). Voor het mecenaat van Bornecolve, zie Stein (1990, 268), Sleiderink (2003, 118-120) of Hage (1989, 147-150). Van Uytven (2002, 25) onderscheidt twee Willem Bornecolves in de bronnen: een vader (schout) en zijn zoon (schepen en volgens Van Uytven opdrachtgever van de *Yeesten*).

25 De meest uitgebreide karakterisering van de eerste drie boeken is te vinden bij Stein (1991).

behandelt de geschiedenis van de Brabantse hertogen tot aan de (toen net overleden) hertog Jan I.[26] De auteur zet zijn geschiedverhaal vanaf boek vier dus verder met een bewerking van een Latijnse brontekst.

Als de chroniqueur is gevorderd tot zijn eigen tijd (ca. 1316), wordt de tekst een eerste keer stilgelegd omstreeks regel 900 van het vijfde boek.[27] Deze cesuur wordt door de auteur als volgt verklaard (vgl. Stein 1990, 268):

> Hier met laet ict bliven
> Mijn dichten ende mijn scriven;
> Want ic nemmeer en can gheleesten
> Van der hertoghen yeesten;
> Want nu regneert dese Jan.
> Ghelevic den tijt, voert dan,
> Dat hem ghevallen eneghe saken
> Die sal ic dichten ende maken,
> Eest dat ics hebbe die stade.
> Ende waert dat ics niet en dade,
> Soe bevelic dat te maken al
> Yemenne die dan leven sal,
> Ende wachte hem wel, dats mine lere,
> Dat hi uter waerheit niet en kere.
> Nu hebbic ghedaen soe ic best can,
> Dat mi bat die goede man,
> Dien ghi mi mocht nomen horen
> In mijn beghin, hier te voren.
> (Willems 1839, V, vs. 877-894)

De auteur geeft aan dat hij zijn werk tijdelijk stopzet omdat zijn bronnenmateriaal is opgedroogd (*ic nemmeer en can gheleesten*). Dit vormt het einde van wat Heymans (1989a, 50) het eerste deel van de kroniek heeft genoemd.[28] De kroniekschrijver belooft niettemin zijn werkzaamheden later opnieuw op te nemen mocht daar aanleiding toe zijn, als hij op dat moment tenminste nog in leven en goede gezondheid is. Mocht dat niet het geval zijn, dan vraagt de auteur dat iemand

26 Appelmans (2005) schrijft de *Chronica* toe aan een monnik uit het klooster van Affligem: Hendrik van Brussel (gestorven na 1270), de auteur van enkele niet-overgeleverde traktaten over tijdsberekening.

27 Merk op dat de proloog in sommige handschriften een twijfelachtige datering (1318) bevat: *Dit boeck waert begonnen, voerwaer, / Doe men screef Jhesus Cristus jaer / Dertien hondert, achtien mede, / Al te Antwerpen, in die stede* (I, vs. 59-62). Deze komt alleszins niet overeen met de eerste cesuur (V, vers 900) rond 1316. Stein (1990) heeft beargumenteerd dat de auteur deze verzen heeft toegevoegd of aangepast bij zijn eerste continuatie in 1318. Sonnemans (1992, 331-332) stelde dan weer dat de auteur de datering in nog latere toevoegingen zelf heeft verwijderd, omdat die toch steeds aangepast moest worden. Heymans (1989a, 40) houdt deze verzen voor een kopiistentoevoeging.

28 Merk op dat de verzen betreffende de eventuele voortzetting door iemand anders (V, vs. 887-890) soms niet werden geciteerd in de onderzoeksliteratuur omdat men er van uitging dat de bewuste auteur zijn werk zelf zou hebben voortgezet (Stein 1990, 268).

anders die taak op zich zou nemen. Tot slot brengt hij nogmaals zijn opdrachtgever in herinnering. Na deze schijnbare slotopmerkingen gaan de *Brabantse yeesten* in de overgeleverde handschriften niettemin gewoon verder en blijkt het vijfde boek later met ettelijke kapittels te zijn verrijkt (tot ca. 1350). Dit vervolg wordt slechts summier ingeleid:

> Sinter dat ic desen boec liet
> Ende beloec, soe ghi hier siet,
> Soe sijn ghevallen saken,
> Die ic u cont wille maken.
> (Willems 1839, v, vs. 901-904)

Hoewel men er lang van is uitgegaan dat de tekst pas rond 1350 van een vervolg werd voorzien, heeft Stein (1990) overtuigend beargumenteerd dat de kroniek op verschillende tijdstippen tussen 1316 en 1350 moet zijn voortgezet, mogelijk al een eerste keer in 1318. Stein (en anderen met hem) lijkt nooit te hebben betwijfeld dat deze continuaties zijn toegevoegd door diegene die de kroniek was begonnen. De aanname van een enkelvoudig auteurschap voor de gehele *Yeesten* is in principe niet bewezen maar wordt wel ondersteund door de inleiding op de eerste continuatie in bovenstaand citaat. Van Anrooij (1995) heeft Stein in zijn argumentatie gevolgd en zelf nog een zesde continuatie door dezelfde auteur onderscheiden.

Problematisch is dat de handschriftelijke overlevering van de boeken IV en V een verwarrend beeld van de continuaties geeft.[29] Op een zeker moment (mogelijk al ca. 1322-1324) blijkt het vierde boek van de oorspronkelijke *Yeesten*-tekst te zijn opengebroken en aangevuld met omvangrijke excerpten uit Velthems Vijfde Partie.[30] Klaarblijkelijk heeft de oorspronkelijke dichter van de *Brabantse yeesten* in eerste instantie geen weet gehad van dit werk. Dat geldt mogelijk ook voor de rijmkroniek van Jan van Heelu over de Slag bij Woeringen, die zoals gezegd als bron heeft gediend in de Vijfde Partie en slechts langs deze weg uiteindelijk toch in de *Yeesten* is opgenomen.

Mogelijk is het de *Yeesten*-dichter zelf geweest die de Velthem-excerpten op een later moment heeft ingebouwd. Niettemin blijft dit twijfelachtig omdat de excerpten in de meeste tekstgetuigen verschillen, zowel wat betreft hun omvang als de plaatsen waar ze in de kroniek zijn ingevoegd.[31] Alleszins blijft het na ca. 1351 een tijdlang stil rond de *Brabantse yeesten*, totdat een Brusselse dichter in de eerste

29 De relatief slecht bestudeerde handschriftelijke overlevering speelt het onderzoek in het algemeen parten. Voor weinig andere Middelnederlandse teksten is een synoptische editie zo dringend gewenst als voor de *Yeesten*. Merk op dat de tekst in het handschrift-Tongerloo (C bij Willems) waarschijnlijk in een later stadium (dus inclusief enkele continuaties) wordt opgedragen aan Jan III (Willems 1839, dl. 1, 600, vs. 35-38). Er is terecht gewezen op de gelijkaardige verschuiving in het geïntendeerd publiek voor de *Lekenspiegel* (Hage 1989, 150).

30 Vgl. Stein (1990, 274-275) maar vooral ook Schoenaers (2009).

31 De hier gebruikte editie van Willems (1839) biedt enkel de zogezegde 'kale standaardtekst' zonder sporen van deze interpolaties. Het enige kapittel dat uniek is aan dit

helft van de vijftiende eeuw het werk van een update voorziet en een zesde en zevende boek aan de reeds omvangrijke kroniek toevoegt.[32] Deze auteur is lang anoniem gebleven maar is enkele jaren geleden op goede gronden geïdentificeerd als Wein van Cotthem (Sleiderink 2003, 153-155). Houthuys (2009) is recent gepromoveerd op een inspirerende studie naar de autograaf die is overgeleverd van het zesde boek van deze voortzetting en een verrassend veelzijdige blik biedt op de werkzaamheden en poëtica van deze dichter. Boeken VI en VII van de *Yeesten* zullen hieronder evenwel slechts nog zijdelings aan bod komen.

7.4.2 *De toeschrijving van de Yeesten*

De toeschrijving van de *Brabantse yeesten* – althans de boeken I-V – aan Jan van Boendale vormt een ingewikkelde aangelegenheid. De vroege receptiegeschiedenis van de *Brabantse yeesten* is nog het meest eenduidig.[33] Hoewel de tekst zelf nooit een auteursnaam prijsgeeft, hebben drie laatmiddeleeuwse tekstgetuigen marginale aantekeningen of kapittelopschriften die de tekst (impliciet) toeschrijven aan Jan van Boendale. Een eerste codex, het zogenaamde Affligemse handschrift (BRUSSEL, KONINKLIJKE BIBLIOTHEEK, IV 684-685, ca. 1470), bevat bijvoorbeeld de mededeling *Dit dichte meester Jan van Andwerpen*.[34] Toch was deze toeschrijving niet algemeen bekend in de late middeleeuwen. In deze context is de proloog van Cotthems voortzetting merkwaardig. Deze stelt dat de *Yeesten* zoals hij die in 1432 voor zich had liggen, niet door één auteur zullen zijn geschreven. Over een dergelijke enkelvoudige auteur stelt hij:

> Of si oec niet volbracht en waert
> Bi hem allene, des mach men moeden,

handschrift is hier verwijderd uit het corpus. Stein (1990, 263) acht de editie-Willems een betrouwbare weergave van het eerste deel. Wat betreft de latere continuaties is de editie-Willems dus vrij van interpolaties, wat voor dit onderzoek vanzelfsprekend handig is.

32 De boeken I tot V tellen ca. 16.000 verzen; de boeken VI en VII maar liefst ca. 30.000. De exacte aantallen variëren van handschrift tot handschrift; een overzicht bij Deschamps (1972, 108-110).

33 De receptie-informatie (en citaten) in deze paragraaf zijn ontleend aan Heymans (1989a, 36).

34 Het is onduidelijk of deze woorden van de dichter zelf zijn. Ik ben Dirk Schoenaers zeer erkentelijk voor de persoonlijke mededeling dat dit citaat (*Dit dichte meester Jan van Andwerpen*) in handschrift-Affligem (Willems sigle D) het laatste kapittelopschrift (nr. XVII) vormt van het eerste deel van de *Yeesten*. Het volgt op vers 876 van de editie-Willems (*Die soe vrome was becant*) en is als kapittelopschrift ook opgenomen in de inhoudsopgave vooraan in het handschrift. Dirk Schoenaers vroeg zich af of het opschrift oorspronkelijk een colofon was dat door een kopiist tot kapittelopschrift werd gerecycleerd. De verzen vanaf vers 877 (*Hier met laet ict blive*) volgen immers als een afzonderlijk kapittel in het Affligemse handschrift. In het handschrift-Van Damme wordt het begin van deze verzen door een lombarde gemarkeerd. Schijnt hier een 'redactie-naad' door?

Soe biddic God dat hi wil hoeden
Hare alre ziele, wi si sijn,
Of wie si waren, voer dewege pijn
Ende met den eewegen licht verlichten,
Die dese ieesten holpen dichten
(Willems 1839, VI, vs. 50-56)

Deze passage toont aan dat de *Yeesten* Wein en zijn opdrachtgevers in 1432 anoniem bereikt hadden en dat zij zich niet echt raad wisten met het auteurschap van de tekst, hoewel (de groep rond) Cotthem doorgaans goed ingelicht was. In wat volgt, blijft Wein consequent spreken over de auteurs van de vroegere *Yeesten* in het meervoud. Hier zij opgemerkt dat Wein eventueel kan gedoeld hebben op de Velthem-interpolaties die we in sommige handschriften aantreffen. Leidde men hieruit een meervoudig auteurschap af? Houthuys (2009, 36) heeft niettemin laten zien dat Wein moet gewerkt hebben met het handschrift-Kluit (of een nauw verwante redactie) dat ooit in handen was van Petrus de Thimo. Dit is echter precies een van de handschriften die de Velthem-interpolaties niet hebben, maar enkel de zogenaamde 'kale standaardtekst'. De verwarrende proloog doelt dus waarschijnlijk niet op de Velthem-interpolaties en vormt een van de (tot nu toe eerder zeldzame) argumenten tegen een enkelvoudig auteurschap voor de *Yeesten*.

Toch lijkt Cotthems onzekerheid wat betreft het auteurschap van de *Yeesten* in de laatmiddeleeuwse periode eerder uitzonderlijk. Omstreeks 1497 verzamelt bijvoorbeeld de Antwerpse drukker Roland van den Dorpe de *Croniken van Brabant*, die hij betitelt als *yerst van meester Janne de clerc van Antwerpen*. Hij bouwt kennelijk verder op de gelijkaardige toeschrijving in de proloog van *Die cornyke van Brabant int prose en int corte* (ca. 1468). Zeker deze toeschrijving heeft daarna nog verschillende kroniekschrijvers beïnvloed (Tigelaar 2006, 136-137). Ook is er een aantekening over het sterfjaar van Boendale die de Antwerpse kanunnik Gevartius tussen 1575 en 1630 aanbracht op een schutblad van het Leidse *Yeesten*-handschrift: *Anno 1351 stierf Jean de Clerck, Secretaris van Antwerpen, die den duytschen doctrinael hadde gemaeckt, In Chronico Rhytmico parvo*. Deze veelgeciteerde aantekening vermeldt wel *Jean de Clerck* als dichter van de *Dietsche doctrinael* maar schrijft anderzijds de *Brabantse yeesten* niet expliciet aan hem toe. Hoewel de vroegste attestatie van de Boendale-toeschrijving dus meer dan een eeuw jonger is dan het eigenlijke werk, lijkt op het einde van de middeleeuwen de consensus te bestaan dat ene (meester) Jan de klerk van Antwerpen de eerste vijf boeken van de *Brabantse yeesten* heeft geschreven. Het is wel merkwaardig dat precies Wein van Cothem, de vijftiende-eeuwse voortzetter van de *Yeesten*, daarvan niet op de hoogte is.

Toch komt daar in de vroege zeventiende eeuw op betrekkelijk raadselachtige wijze verandering in en worden de *Yeesten* plotseling

toegeschreven aan ene 'Niclaes de klerk'. Van blijvend grote waarde voor de historische reconstructie van deze verrassende omslag is de wetenschappelijke inleiding die Willems (1839) verzorgde bij zijn kolossale editie van de *Brabantse yeesten*. Vóór Willems' tijd was het maar liefst twee eeuwen gebruikelijk geweest om de *Yeesten* toe te schrijven aan ene Nicolaas, secretaris van de stad Antwerpen. Willems kon uiteindelijk de ene bron traceren waarop deze attributies teruggingen, namelijk Divaeus (1610, 3), die in zijn *Rerum Brabanticarum* betoogde dat de *Yeesten* werden geschreven door ene *Nicolaus Clericus, Secretarius Andouerpiensis*. In de kantlijn beriep de Leuvense geleerde zich voor die claim op een perkamenten handschrift, toen in het bezit van de Antwerpse kanunnik en humanist Jan Gevartius. Het gaat opvallend genoeg om hetzelfde handschrift met de reeds vermelde opmerking over het sterfjaar van *Jean de Clerck*.

Deze Nicolaas-toeschrijving werd even later overgenomen door Jean-Jacques Chiflet (1588-1660), de Franse humanist en lijfarts van Filips IV van Spanje. Toch blijkt deze toeschrijving van de *Yeesten* op dat moment reeds controversieel: in *Le faux Childebrand* (1659) had Chiflet een anonieme criticus van weerwoord moeten dienen in dezen. Deze *philosophe masqué* had benadrukt dat de *Yeesten* van anonieme signatuur waren (Chiflet 1659, 90-91). Chiflet kwam in zijn verdediging echter niet verder dan opnieuw een verwijzing naar Divaeus en het beruchte handschrift dat op dat moment nog bij Gevartius berustte. Dit handschrift – het handschrift-Kluit (LEIDEN, UNIVERSITEITSBIBLIOTHEEK, LTK. 1019), ca. 1425 – is overgeleverd en bevat op de titelpagina inderdaad een nota met een merkwaardige correctie:[35]

Rijm-
Chronijck Van Brabant
in Rijme gestelt bij ~~Jan~~ Niclaes
de clerck in sijnen
Leue*n*, Secretaris
der stadt
Van
Antwerpen

Iemand heeft hier de naam *Jan* doorgestreept en er *Niclaes* overheen geschreven.[36] Tot op heden blijft onduidelijk waarop deze correctie gebaseerd is: Gevartius, Divaeus noch Chiflet geven nadere bronverwijzingen. Bovendien is onduidelijk wie de oorspronkelijke aantekening

35 Diplomatisch weergegeven naar het handschrift-Kluit – met dank aan Remco Sleiderink die mij een microfilm bezorgde. Dit handschrift bevat – net als andere *Yeesten*-handschriften – nog een schat aan parateksten die hier onderbelicht blijven. De datering van het handschrift-Kluit ontleen ik aan Stein (1994, 26-27).

36 De Vreese (1894, 157-158) heeft andere aantekeningen (o.m. over het sterfjaar van Jean de Clerck) op dit folium toegeschreven aan Gevartius maar geeft geen nadere informatie over of datering van de verbetering. Ik heb in de paleografische literatuur geen expliciete toeschrijving van deze specifieke nota gevonden.

maakte en wie de correctie heeft aangebracht (tweemaal Gevartius?). Het lijkt daarom significant dat de *Niclaes*-hypothese reeds bij de lancering door een 'gemaskerde filoloog' werd gecontesteerd. Anderzijds werd de controversiële aantekening aangebracht in een geleerd, humanistisch milieu dat erg begaan was met de Brabantse historiografie.[37] Iemand moet in deze periode een overtuigende reden hebben gehad om de schimmige *Niclaes*-figuur een geschikter auteurskandidaat te vinden. Natuurlijk kan iemand bij archivalische opzoekingen toevallig op de klerkennaam Nicolaas gestoten zijn om die vervolgens met de anonieme *Yeesten* te verbinden. Niettemin: de naam Jan stond er reeds en die naam werd toch minstens even vaak in archiefstukken vermeld. Waarom was Nicolaas' kandidatuur dan plots geloofwaardiger? Willems stond voor een moeilijke keuze. Na het beschikbare materiaal zorgvuldig afgewogen te hebben, heeft hij er uiteindelijk voor gepleit om de toch erg apocrief aandoende Nicolaas-toeschrijving minder waarschijnlijk te achten dan de toch veel oudere Jan-toeschrijving. Willems' overtuigende pleidooi voor het auteurschap van Jan van Boendale is daarna enthousiast onthaald in de filologie.

Wie is toch die raadselachtige Nicolaas-figuur die in de zeventiende en achttiende eeuw te boek stond als de auteur van de *Yeesten*? In de bronnen vinden we één belangrijke kandidaat-Nicolaas, die we mogelijk al voor het eerst vermeld zien in de reeds besproken rekeninguittreksels uit 1313-1314. Daar verschijnt hij als een Antwerpse schepenknaap, *Claus der scepen cnape*, een soort brievenbode/boodschapper (Prims 1935, 231-233). Génard bedacht in 1853 reeds dat we met deze *cnape* mogelijk voor de Nicolaas staan die men een tijd lang voor de auteur van de *Yeesten* heeft gehouden.[38] Deze schepenknaap kan waarschijnlijk vereenzelvigd worden met *Nicolaes Noese*, later (vanaf 1324) een van de belangrijkere stadsklerken, die in stadsrekeningen veelvuldig wordt vernoemd in de onmiddellijke nabijheid van Boendale. In de reeds vermelde akte uit 1343 worden Boendale en Noese trouwens samen vermeld (*Johanne dicto Clerico et Nicolao dicto Noese, clericis*) als getuigen bij de schenking van Suderman.[39] Nicolaas behoorde tot het geslacht Nose of Noese, een van de oudere (schepen)families in de stad Antwerpen die oorspronkelijk carrière maakt binnen het gilde van de edelsmeden (Asaert 2005, 221-223). De familie blijkt al in de dertiende eeuw cultureel geëngageerd en verschijnt ver-

37 Stein (1994, 59) heeft er bijvoorbeeld op gewezen dat de zoon van Gevartius – Jan, die zelf stadssecretaris van Turnhout was en het handschrift-Kluit bovendien erfde – samen met Miraeus plannen had voor een editie van Brabants kroniekmateriaal.

38 Zie Génard (1853, 78, n. 1 & 169, n.1). Het is onduidelijk of het woord *cnape* in deze context wijst op de jeugdige leeftijd van de betrokkene – dat is volgens het Middelnederlands woordenboek (`http://gtb.inl.nl`) althans niet noodzakelijk het geval. Het lijkt wel aannemelijk dat de functie van schepenknaap lager was dan die van schepenklerk.

39 Zie Génard (1853, 199-200) of Van Uytven (2002, 27-28). Noese lijkt Suderman ook zelf gekend te hebben want hij schijnt er brieven voor geschreven te hebben (Diercxsens 1773, 90).

schillende keren als weldoener of stichter van religieuze instellingen.[40] Diercxsens meende dat Jan van Boendale Nicolaas had opgevolgd in het klerkenambt voor de stad Antwerpen. Op dit moment zijn er echter geen authentieke bronnen meer bekend die deze opvolging (of de Niclaes-toeschrijving) staven (Génard 1853, 170).

7.4.3 *Rondom de Yeesten*

Na Willems' editie lijkt de *Niclaes*-figuur uit het onderzoek verdwenen en werd vooral voortgebouwd op de Boendale-toeschrijving binnen de studie van de teksten uit de 'Antwerpse School'. In de regel wordt dit laatste corpus momenteel volledig toegeschreven aan Jan van Boendale. Het bestaan van verschillende, veelal woordelijke parallellen en stilistische overeenkomsten tussen deze teksten vormt het belangrijkste literair-historische argument voor een identiek auteurschap. Wim van Anrooij wees in 1995 bijvoorbeeld nog op frappante gelijkenissen tussen het *Boec van der wraken*, *Jans Teesteye* en het (laatste deel van) de *Brabantse yeesten*. De ontstaansgeschiedenis van het *Boec van der wraken* vertoont overigens opvallende overeenkomsten met die van de *Yeesten*.[41] Het stilvallen van beide werken omstreeks 1351 komt bovendien overeen met Boendales vermoedelijke overlijdensdatum.

Het zou te ver leiden om hier een uitvoerig overzicht te bieden van de parallellen tussen deze Antwerpse teksten inzake woordkeuze en een aantal specifieke zinswendingen etc. Wel zij specifiek gewezen op het bestaan van enkele kortere historiografische teksten die nauw verbonden lijken met de *Yeesten*. Ten eerste is er *Van den derden Eduwaert* (Heymans 1989a). Deze Middelnederlandse rijmtekst kwam eveneens tot stand in de Scheldestad en bespreekt de doortocht van koning Edward III van Engeland op het vasteland tijdens de eerste jaren van de Honderdjarige Oorlog (1338-1340). De *Eduwaert* vertoont nauwe intertekstuele relaties met enkele van de continuaties van de *Yeesten* en het is dan ook goed mogelijk dat de *Eduwaert* en de *Yeesten* door dezelfde auteur werden geschreven. De reeds vermelde Engelse rekeningen in de *Wardrobe Accounts* vormen trouwens een duidelijk argument voor de identificatie van deze auteur als Jan van Boendale. De nauwe band tussen Boendale en het Engelse hof onder Edward III – het hoofdonderwerp van de *Eduwaert* – maken van de schepenklerk

40 Hendrik Nose (gestorven in 1275) was deken van het Antwerpse kapittel van Onze-Lieve-Vrouw. Een voorbeeld van het cultureel engagement van leden van deze familie spreekt uit het feit dat deze deken boeken uit zijn privébezit heeft geschonken aan de Antwerpse dominicanen. Aan die schenking werd het opvallende verzoek verbonden om iedereen tot de boeken toegang te geven om te leren en te studeren (Van Anrooij 1994a, 149-150).

41 Zie Van Anrooij (1995). Elders noemt Van Anrooij (1994a, 150, n. 9) als belangrijk desideratum overigens 'het systematisch blootleggen van het netwerk van intertekstuele relaties' in de teksten van de 'Antwerpse School'.

vanzelfsprekend een extra aantrekkelijke auteurskandidaat voor deze teksten.

Daarnaast is er de *Korte kroniek van Brabant*. De auteur ervan vermeldt dat de tekst ca. 1322 in Antwerpen werd voltooid. Die datum klopt althans voor de oudste of 'korte' versie (374 verzen), want van de *Korte kroniek* moet later een tweede, 'lange' versie (1991 verzen) in omloop zijn gekomen, waarin de geschiedenis werd doorgetrokken tot 1332/1333.[42] Het is onduidelijk of de korte en lange versie van dezelfde auteur zijn, al gaat men daar vaak van uit omdat de lange versie aan de korte ontleent. Ook de tweede versie kwam in Antwerpen tot stand en net voor het einde vermeldt het kroniekje een merkwaardige bronverwijzing:[43]

> Wie dat des wil roecken,
> Dat hi dees yeesten leest voir of na,
> Ic raed hem dat hi ten boecke gha
> Dat ic maecte, ten es nyet lanc.

De auteur verwijst hier naar een eerder werk van zijn hand, gezien de inhoud en geografische nabijheid naar alle waarschijnlijkheid de *Brabantse yeesten*.[44] Stein (1990, 270, n. 23) wees erop dat de lange versie van de *Korte kroniek* afhankelijk lijkt van álle delen van de *Yeesten*. Daarom suggereert deze epiloog bijzonder sterk dat de *Yeesten* zijn geschreven door één enkele auteur, die bovendien mag gelijkgesteld worden met die van (ten minste de lange versie van) de *Korte kroniek*. Er is zelfs betoogd dat de *Korte kroniek* bij wijze van reclame een 'trailer' vormde voor de *Yeesten*.[45]

Al met al is duidelijk dat het onderzoek in de twintigste eeuw een 'maximalistische' evolutie heeft doorgemaakt, waarbij de toeschrijving aan Boendale niet alleen is gehandhaafd maar vaak ook is verstevigd door de betere inbedding van het werk in het literaire netwerk van andere teksten in de 'Antwerpse School'. Een dissidente stem in de jongste decennia was die van Heymans (1989a), die zich heeft afgevraagd of de cesuur in boek v van de *Yeesten* samenvalt met

42 Zie Van Anrooij (2002, 9-11). De datering in de tweede versie (1322) is gezien de inhoud (tot 1332) aantoonbaar fout.

43 Geciteerd naar Van Anrooij (2002, 10-11). Merk op dat ik Van Anrooij volg in zijn afwijkende interpunctie van de laatste verzen en de verzen dus als volgt begrijp: 'Wie zin heeft om deze geschiedverhalen van begin tot einde te lezen, raad ik aan om het boek te raadplegen dat ik geschreven heb – het is niet lang'. De laatste toevoeging interpreteer ik dus in navolging van Van Anrooij als 'de lengte van dit andere boek valt wel mee' in plaats van 'dit andere boek is kort'.

44 Het is wel overwogen of de korte versie niet onafhankelijk op de Latijnse bron van de *Yeesten* kan teruggaan (vgl. Appelmans 2002, 79-80), maar dat hoeft een identiek auteurschap natuurlijk niet uit te sluiten.

45 Zie Stein (1990, 11). Ik ga hier overigens voorbij aan een meer gedetailleerde bespreking van *Het Boec vander wraken*. De ontstaansgeschiedenis van dit werk is meer dan waarschijnlijk nauw vervlochten met die van de laatste bijdragen aan de *Yeesten* (Van Anrooij 1994b, 1995).

een auteursovergang. Heymans heeft zich als 'minimalist' trouwens in het algemeen bijzonder sceptisch getoond over het enkelvoudig auteurschap in de 'Antwerpse School' (vgl. ook Heymans 1989b). Volgens Heymans (1989a, 68) begint het tweede deel van de *Yeesten* in medias res, zonder de minste allusie op de belofte aan het slot van het eerste deel: 'Zonder enige tussenopmerking begint hij zijn deel van de geschiedenis'. '[W]aarom heeft hij dan op dit belangrijke moment geen enkele opmerking gemaakt van structurele aard?' Een dergelijke opmerking zou op zijn plaats zijn geweest, aangezien eerder in het werk was aangekondigd dat boek v elf kapittels zou tellen, hoewel dat er uiteindelijk zonder toelichting zestig werden.[46] Toch gaat Heymans hier deels in de fout, want hij laat na te vermelden dat de tweede continuatie wel degelijk aanvangt met een verbindende passage (v, vs. 901-904) waarin de auteur zichzelf kort maar toch duidelijk identificeert als de auteur van het eerste deel (*Sinter dat ic desen boec liet*).

Afgezien van Heymans' (soms wat vrijblijvende) argumentatie som ik tot slot nog enkele elementen op die onvoldoende zijn uitgeklaard en eventueel nog kunnen pleiten tegen Boendales auteurschap voor de (gehele) *Yeesten*. De auteur van de *Brabantse yeesten* had bijvoorbeeld aanvankelijk geen beschikking over een exemplaar van Jan van Heelu's rijmkroniek van Woeringen en kende de tekst slechts onrechtstreeks. Dat lijkt niet in overeenstemming met het profiel van de hoogopgeleide Jan van Boendale, die zijn weg kende in de plaatselijke adellijke en culturele milieus in Zuidoost-Brabant. Ook kan het vreemd lijken dat Boendale het recente Vijfde Partie-initiatief (ca. 1317) van zijn tijd- en streekgenoot Lodewijk van Velthem ten vroegste ca. 1322 leerde kennen. Het is vreemd dat beide (toch enigszins parallelle) teksten op nagenoeg hetzelfde moment werden begonnen.[47] Een ander tegenargument betreft de auteursvermeldingen. De vroegste werken uit de 'Antwerpse School' vermelden duidelijk een auteursnaam: Boendale introduceert zichzelf uitgebreid in de *Teesteye* maar ook in de *Lekenspiegel* noemt hij zichzelf (vgl. *Jan, u arme clerc*). De jongste Antwerpse geschriften ontberen echter zonder uitzondering een auteursvermelding.

De *Yeesten* (en zeker het eerste deel) is moeilijk als jeugdwerk in die ontwikkeling in te passen: indien Boendale de auteur van deze tekst was, waarom zou de recent uit Tervuren gemuteerde jongeling zich niet duidelijker in de tekst kenbaar maken? In de slotpassage van

46 Heymans (vgl. 1989a, 66 en 69), waar de onderzoeker spreekt van 'een boeiende persoonlijkheid [...], een ernstig historiograaf, een bekwaam deskundige op het gebied van de feodale verhoudingen en tegelijkertijd een bewogen en vaardig schrijver van een poëzie met bezieling, epische gloed en rijke afwisseling in presentatie en beschrijving'.

47 Deze redenering veronderstelt natuurlijk dat de *Yeesten*-auteur zelf de stukken Velthem interpoleerde, hetgeen een aantrekkelijke maar onbewezen aanname is (Schoenaers 2009, 231).

het eerste deel lijkt op het eerste zicht trouwens geen jongeling aan het woord maar een auteur die ernstig rekening houdt met het feit dat hij niet meer in leven zal zijn als de kroniek kan hervat worden (vgl. *Ghelevic den tijt*).[48] Het is opvallend hoe nadrukkelijk de auteur de eventuele continuator van de tekst toespreekt. Het lijkt verdacht dat de auteur van het eerste deel niet durft te beloven dat hij zijn tekst nog zal kunnen voortzetten, terwijl de eerste continuaties relatief kort daarna werd geschreven. Natuurlijk kunnen we wel niet uitsluiten dat hier om een typisch middeleeuwse *topos* gaat.

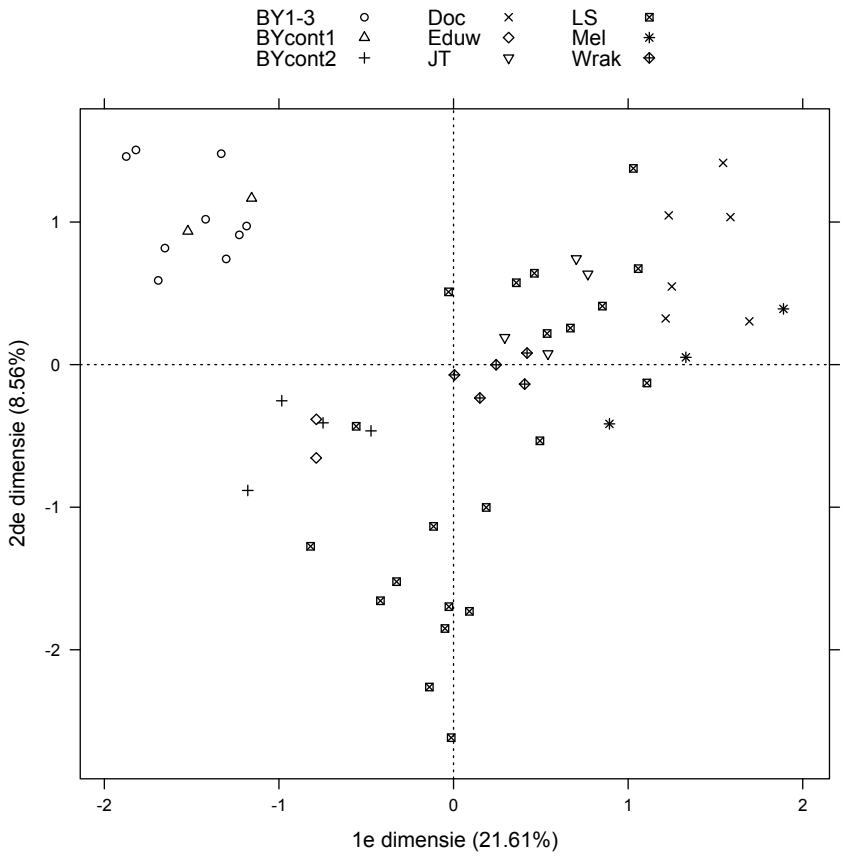
7.4.4 *Analyses*

Bij de historiografische teksten heb ik een onderscheid gemaakt tussen *Van den derden Eduwaert* (*Eduw*), de Maerlant-adaptatie in de Boeken I tot III van de *Yeesten* (*BY1-3*), het vertaalde deel tot boek 5, v. 900 (*BYcont1*) en de rest van boek V (*BYcont2*).[49] Onderstaande figuur geeft de resultaten weer van een correspondentie-analyse (Fig. 45, p. 191); deze keer van alle de voor dit onderzoek beschikbare Antwerpse rijmteksten. Als sample-grootte werd 1.000 rijmwoorden gekozen; deze samples werden voorgesteld aan de hand van de relatieve frequenties van de 150 meest frequente rijmwoorden. De volgende figuur toont het resultaat van een vergelijkbaar experiment met een sample-grootte van 2.000 rijmwoorden, zodat minder maar iets grotere (en dus mogelijk meer betrouwbare) samples worden gebruikt (Fig. 46, p. 192).

Het valt in beide figuren op dat samples uit de meeste Antwerpse teksten meestal clusteren met, ten eerste, samples uit hetzelfde werk en, ten tweede, samples uit andere Antwerpse teksten. Deze groepering wordt eens te meer gedomineerd door een diagonale sliert van samples uit de *Lekenspiegel*, die stilistisch gesproken het meest centrale van deze groep werken lijkt. Het is duidelijk dat twee historiografische teksten (*Eduw* en *BYcont2*) een coherente genre-cluster vormen – zoals te verwachten was op basis van het hierboven besproken onderzoek. In beide analyses lijkt die cluster op het eerste zicht echter niet heel ver verwijderd van bijvoorbeeld heel wat *Lekenspiegel*-samples. Die verwijdering lijkt zeker niet noemenswaardig als we de positie van

48 Ik dank deze scherpzinnige observatie aan Renée Gabriël.

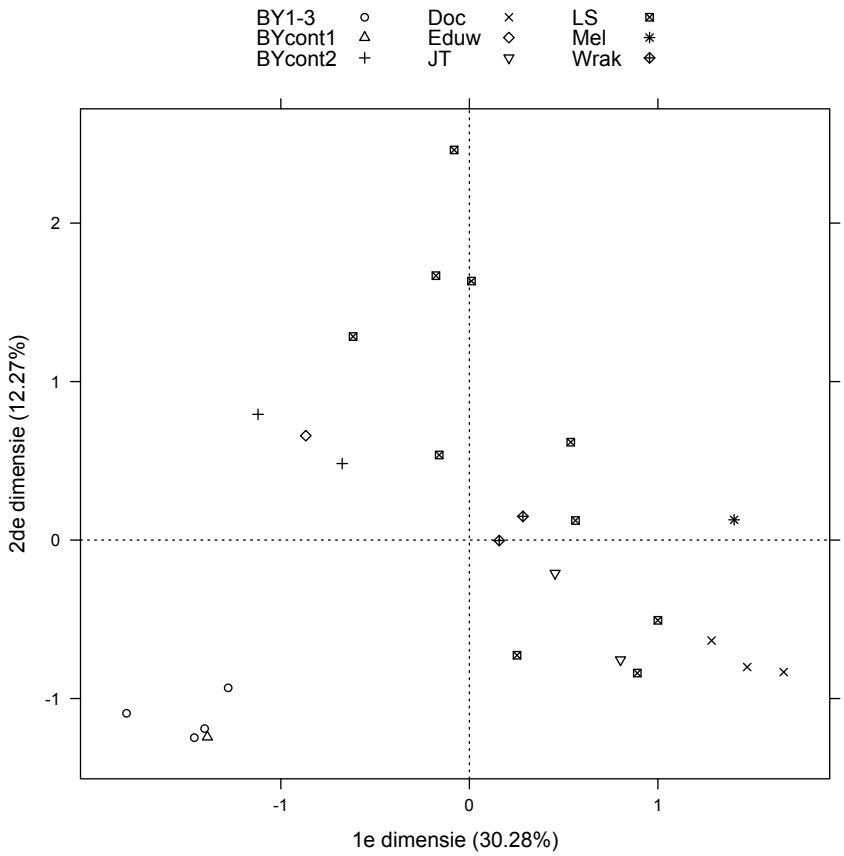
49 Het aantal gebruikte rijmparen voor de historiografische teksten is als volgt: *Eduw*: 2.018; *BY1-3*: 9.740; *BYcont1*: 2.118; *BYcont2*: 4.167. De gebruikte edities zijn de reeds vermelde zoals die te vinden zijn op de *Cd-rom Middelnederlands*. Ik heb niet onderscheiden tussen de verschillende continuaties binnen *BYcont2*. Zoals ook de analyses verderop laten zien, is er vooralsnog een reden om de vigerende opvatting te betwijfelen dat deze uitbreidingen door een en dezelfde auteur zijn aangebracht. Merk op dat de parallelle passages tussen *Eduw* en *BYcont2* niet uit de tekst zijn gelicht: dat deze teksten relatief dicht bij elkaar zullen blijken te liggen hieronder, is tot op zekere hoogte een artefact van deze situatie. Omdat geweten is dat teksten lang genoeg moeten zijn voor een stylometrische analyse, heb ik beide versies van de *Korte kroniek* jammer genoeg moeten uitsluiten in een vroeg stadium. Verder onderzoek naar deze teksten (dat zich bijvoorbeeld niet tot rijmwoorden beperkt) is zeker gewenst.



Figuur 45: Strooidiagram (eerste twee dimensies) resulterend uit een CA van alle hier bestudeerde teksten uit de 'Antwerpse School', waaronder ook de historiografische (*Eduw*, *BY1-3*, *BYcont1* en *BYcont2*). Deze teksten zijn verdeeld in samples van 1.000 rijmwoorden en voorgesteld aan de hand van de 150 meest frequente rijmwoorden erin.

Eduw en *BYcont1* vergelijken met de ietwat geïsoleerde positie van *BY1-3* en *BYcont1*. *BY1-3* en *BYcont1* vallen als duo op het eerste zicht uit de toon en worden in de resultaten tweemaal geïsoleerd van de andere samples. De genre-gebaseerde structuur waarop we geanticipeerd hadden in dit materiaal komt dus minder naar voren voor *Eduw* en *BYcont1* maar voor *BY1-3* en *BYcont2* lijkt er zich wel een effect van verwijdering voor te doen. Dit resultaat levert een opvallende bevestiging van de stilistische breuk tussen het eerste en latere delen van de *Yeesten*, zoals reeds door Heymans werd gesuggereerd. Tussentijds kunnen we wel reeds besluiten dat deze analyses Boendales auteurschap voor de latere delen van de *Yeesten* veeleer schijnen te bevestigen dan te verwerpen.

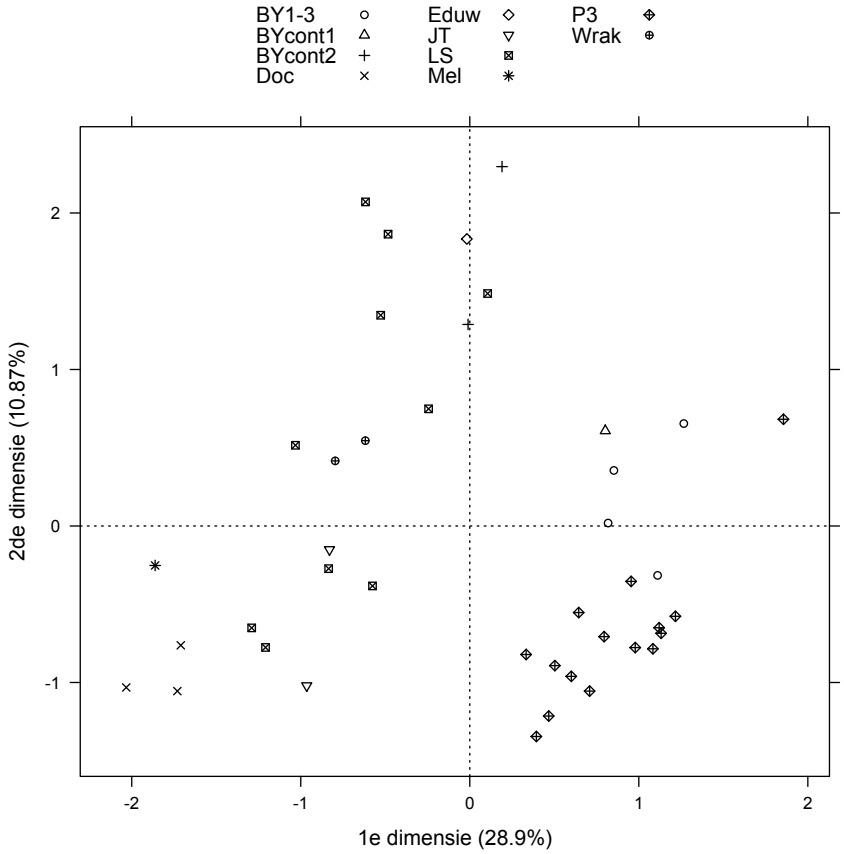
Voor *BY1-3* was deze verwijdering te verwachten: de eerste drie boeken vormen een niet meer dan oppervlakkige bewerking van Maerlants aandeel in de *Spiegel historiael*. Het spreekt dan voor zich dat



Figuur 46: Strooidiagram (eerste twee dimensies) resulterend uit een CA van alle hier bestudeerde teksten uit de 'Antwerpse School', waaronder ook de historiografische (*Eduw*, *BY1-3*, *BYcont1* en *BYcont2*). Deze teksten zijn verdeeld in samples van 2.000 rijmwoorden en voorgesteld aan de hand van de 150 meest frequente rijmwoorden erin.

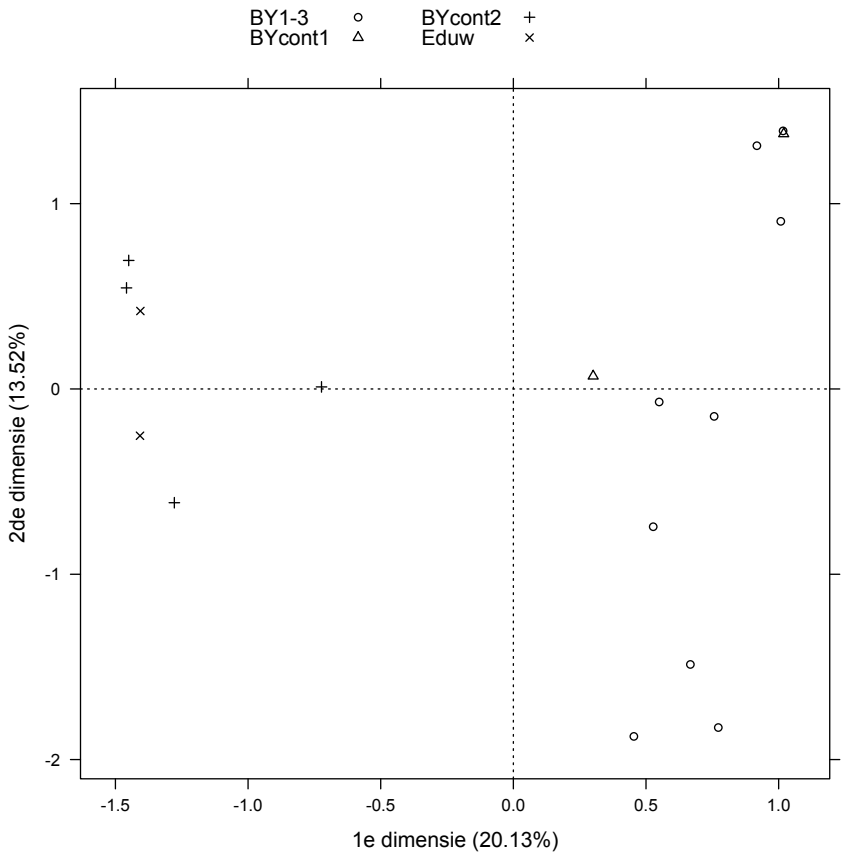
de dimensie-reductie detecteert hoe Maerlants stijl *BY1-3* domineert. *BYcont1* werd voor zover wij weten echter door de *Yeesten*-dichter zelf rechtstreeks naar het Middelnederlands omgezet. De stilistische isolatie van deze passages – ten opzichte van bijvoorbeeld *BYcont2*, de latere toevoegingen aan de *Yeesten* – is in dat opzicht opvallender. De onderstaande figuur bevestigt deze observaties (Fig. 47, p. 193). In deze analyse (met een sample-grootte van 2.000 rijmwoorden) is Maerlants *P3* uit de *Spiegel historiael* ter controle toegevoegd. Merk op dat de ontleende passages uit de *Yeesten* in de eerste drie boeken het stilistisch gezelschap opzoeken van Maerlants werk onderaan de figuur, maar dat *BYcont1* dat ook doet. Anderzijds valt op dat de andere historiografische samples (*BYcont2* en *Eduw*) wel relatief dicht aanleunen bij hun Antwerpse collega's (*Doc*, *Wrak*, *Mel*, *LS* en *JT*). In deze figuur lijkt de verticale lijn door de oorsprong van het diagram met een zekere veiligheidsmarge een model te bieden voor

het stilistische verschil tussen enerzijds samples die door Maerlant werden gedicht (of er toch door geïnspireerd werden) en anderzijds werken uit de 'Antwerpse School'. Een dergelijk model zou *BYcont1* echter niet tot de 'Antwerpse School', laat staan Boendales oeuvre rekenen.



Figuur 47: Strooidiagram (eerste twee dimensies) resulterend uit een CA van alle hier bestudeerde teksten uit de 'Antwerpse School', waaronder ook de historiografische (*Eduw*, *BY1-3*, *BYcont1* en *BYcont2*). Maerlants *P3* is toegevoegd als controlemateriaal. Deze teksten zijn verdeeld in samples van 2.000 rijmwoorden en voorgesteld aan de hand van de 150 meest frequente rijmwoorden erin.

Onderstaande figuur toont een resultaat waarbij we dieper inzoomen op het historiografisch materiaal (Fig. 48, p. 194). Door de verwijdering van niet-historiografisch materiaal in deze analyse, kunnen we genre-verschillen misschien deels buiten spel zetten en nagaan of we ook binnen het kleinere, historiografische deelcorpus een stylometrisch verschil kunnen vaststellen. Ook met een sample-grootte van slechts 1.000 verzen bekomen we hier een uitgesproken stilistische scheiding die kennelijk amper beïnvloed is door de aan- of afwezigheid van minder historiografisch getint materiaal in deze analyses.



Figuur 48: Strooidiagram (eerste twee dimensies) resulterend uit een CA van de zuiver historiografische teksten uit de 'Antwerpse School' (*Eduw*, *BY1-3*, *BYcont1* en *BYcont2*). Deze zijn verdeeld in samples van 1.000 rijmwoorden en voorgesteld aan de hand van de 150 meest frequente rijmwoorden erin.

7.4.5 *Nicolaas de clerk?*

Kunnen we het stilistisch verschil tussen deze *Yeesten*-delen in verband brengen met een auteursovergang binnen het werk? Zou de mysterieuze *Niclaes Noese* bijvoorbeeld een bijdrage aan de *Yeesten* kunnen hebben geleverd? Dat lijkt bij de huidige stand van zaken erg onwaarschijnlijk. Toch is een dergelijk meervoudig auteurschap voor de eerste vijf boeken in het verleden wel overwogen. In de inleiding tot zijn editie stelde bijvoorbeeld Willems (1837, XVII-XVIII) reeds: 'De tout cela on peut conclure, ce me semble, que JEAN DE KLERK est bien réellement l'auteur des *Brabantse Yeesten*, et non *Nicolas De Klerk*, a moins', zo voegt hij er voorzichtig aan toe, 'qu'on ne veuille supposer qu'ils y aient travaillé tous deux, ou l'un après l'autre, ce que je ne puis admettre'. Iets later stelt hij in het *Belgisch Museum* (Willems 1837, 343): 'Mogelyk dat Nicolaes de rymkronyk onmiddellijk na deze regels [bedoeld is de slotpassage van het eerste deel] heeft voortgezet.

Evenwel blyft dan nog altyd de vraeg bestaen: van waer weet men dat?' Kennelijk overwoog Willems ook constructies die het toelaten het opduiken van beide auteursnamen te verzoenen.

Een aanzet daarvoor was trouwens al in het verleden gegeven. Zoals gezegd had Diercxsens in zijn bronnen gelezen over de aantekening van Gevartius aangaande de sterfdatum van Boendale (in 1351) in het handschrift-Kluit, maar voegde daar zelf aan toe: *Quoddam chron. MS. dicit hunc Joannem successisse Nicolao Clerici in munere secretarii et obiisse anno sequenti 1352*.[50] Toch moeten we vaststellen dat het onduidelijk is op welke bronnen deze informatie teruggaat. Afgezien van de vroege vermelding als schepenknaap in 1314, lijkt *Niclaes Noese* – vaak ook *Claes*, *Claus* of *Cole(n)* genoemd – in de overgeleverde bronnen vooral gelijktijdig met Boendale of na hem op te duiken.[51]

Wanneer het eerste deel van de *Yeesten* werd geschreven (ca. 1316) was Nicolaas nog een vrij obscure schepenknaap die pas later in de stadsadministratie moet zijn doorgegroeid. In 1324 blijkt Nicolaas bijvoorbeeld veel minder te verdienen dan Boendale, wat toch de hypothese tegenspreekt dat Boendale Nicolaas zou hebben opgevolgd. Dat Nicolaas de *Yeesten* zou begonnen zijn en in zijn dichtarbeid later werd afgewisseld door Boendale, is daarom een hypothese die nauwelijks vertrouwen inboezemt. Een enkelvoudig auteurschap voor de *Yeesten* wordt bovendien ondersteund door de lange versie van de *Korte kroniek*, waarvan de auteur informatie uit de verschillende delen van de *Yeesten* put en via zijn epiloog zo impliciet het auteurschap claimt voor die gehele brontekst. De *Niclaes*-hypothese kan daarom moeilijk in verband worden gebracht met de geobserveerde stijlverschillen. Kunnen we het stylometrisch verschil tussen de *Yeesten*-delen dan op een andere manier verklaren?

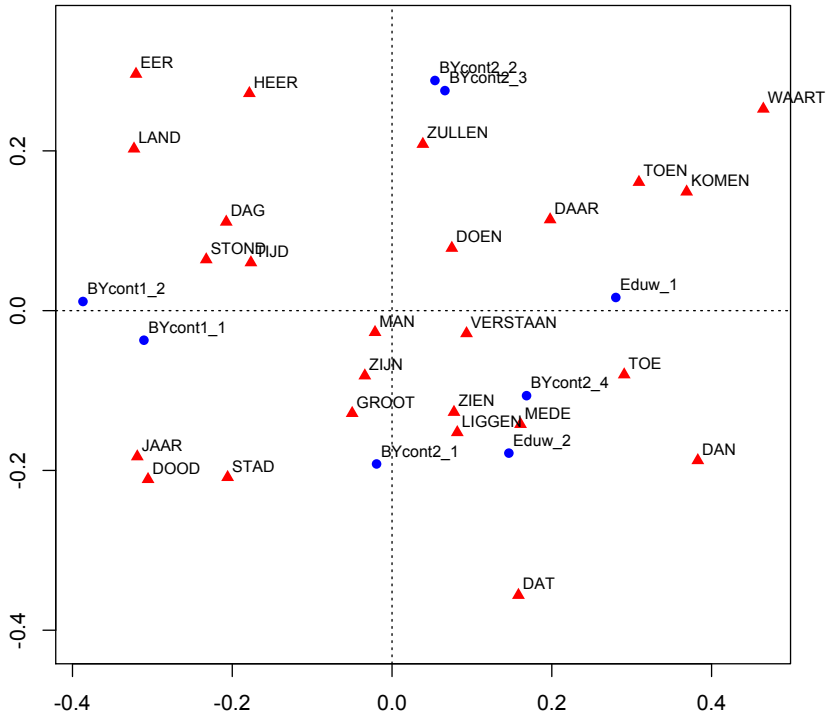
Een interessante illustratie van de stijlverschillen van *BYcont1* ten opzichte van de rest van de historiografische Antwerpse werken (*BYcont2* en *Eduw*) biedt onderstaande figuur (Fig. 49, p. 196). In dit strooidiagram wordt aangeduid welke rijmwoorden (driehoekjes) typisch zijn voor oftewel veel gebruikt worden in bepaalde samples (bolletjes).[52] De twee samples uit *BYcont1* (bolletjes) clusteren aan de linkerzijde van dit strooidiagram; hun positie correspondeert met die van specifieke rijmwoorden (driehoekjes) aan dezelfde linkerzijde (bv. JAAR, DAG, TIJD, STOND). Het eerste deel van de *Yeesten* lijkt op basis

50 Zie Diercxsens (1773, 108). Onderzoekers zoals Willems hebben er in deze periode op gewezen dat het problematisch is dat bij verschillende historici een vergelijkbare sterfdatum voor Jan en Nicolaas wordt genoemd.

51 Uit archivalische bronnen is bekend dat het Antwerpse huis 'De Mane' (Gildekamersstraat 7) in 1375-1376 bewoond werd door ene *Claus Noese der stat clerc*; zie Prims (1935, 80) en Asaert (2005, 221-223). Is dit onze Nicolaas of een familieverwant (zijn (klein)zoon?), zoals de *Claus Colenzone* die in deze periode als stadsklerk wordt vermeld (Génard 1864, 106)? Merk op dat Diercxsens (1773, 108) (zij het op basis van een onbekende bron) beweerde dat Nicolaas slechts een jaar later dan Boendale zou gestorven zijn (in 1352).

52 Het gebruik van persoonsnamen in rijmpositie werd in deze analyse genegeerd.

van dit diagram gekenmerkt door een iets armere, meer 'statische' rijmwoordenschat die erg op afgebakende tijdsmomenten lijkt gericht. De samples van de andere teksten (de bolletjes van *BYcont2* en *Eduw*) clusteren vooral aan de rechterzijde van het diagram en komen overeen met een heel ander type rijmwoorden (de driehoekjes van bv. KOMEN, ZULLEN). Deze teksten schijnen 'narratiever', met een rijkere rijmwoordenschat die ruimte laat voor uitgebreidere beschrijvingen en een dynamischer verhaalverloop.



Figuur 49: Strooidiagram (eerste twee dimensies) resulterend uit een CA van het historiografische materiaal uit de 'Antwerpse School', met uitzondering van *BY1-3*. Deze zijn verdeeld in samples van 1.000 rijmwoorden (bolletjes) en voorgesteld aan de hand van de 25 meest frequente rijmwoorden erin (de driehoekjes). De ladingen van de individuele rijmwoorden zijn eveneens weergegeven.

Gezien de sterke argumenten uit de traditionele filologie voor Boendales auteurschap voor de gehele *Yeesten* ben ik sterk geneigd om te betwijfelen of de stilistische verschillen tussen de eerste en de latere delen van de *Yeesten* met een auteursverschil samenvallen. Waarschijnlijk reflecteren deze verschillen eerder de oppervlakkige discrepanties tussen het eerste deel – gebaseerd op het kale proza van de droge *Chronica* – en de vlotter berijmde latere continuaties over de recente

geschiedenis. Het ziet er naar uit dat de auteur voor de latere delen over meer informatie beschikte en mogelijk ook meer inspiratie aan de dag kon leggen.

Bovendien behandelt het eerste deel veel verder afgelegen geschiedenis: de auteur kan meer betrokken geweest zijn bij de contemporaine geschiedenis uit de continuaties. Dit zou heel goed het algemene verschil in toon en inhoud kunnen verklaren. Het vierde boek heeft – althans naar mijn smaak – minder narratieve vaart en behandelt drogere stof, duidelijk met minder aandacht voor de *petite histoire*.[53] Stein (1991, 186) benadrukte tevens het verschil in oorsprong tussen beide delen: het eerste ontstond door een officieel mecenaat; de continuaties veelal ad hoc, naar aanleiding van incidentele gebeurtenissen en persoonlijke initiatieven van de auteur. Zoals gesuggereerd door Hage (1989, 148) kan deze inhoudelijke verschuiving ook parallel lopen met een verschuiving van het geïntendeerd publiek voor de continuaties. Bovendien zou het eerste deel van de *Yeesten* Boendales eersteling zijn geweest; mogelijk illustreert de analyse dat de onervaren auteur zich nog moeilijk los kon maken van de stijl van Maerlant. Hij had zijn grote voorbeeld immers kort daarvoor drie boeken lang zitten parafraseren: stilistische invloed op de toen nog onervaren klerk is zeker niet uit te sluiten. Merk wel op dat een zuiver diachroon verschil anderzijds minder vanzelfsprekend lijkt, want de vroege *Yeesten* liggen duidelijk verder van bijvoorbeeld de *Lekenspiegel* of *Jans Teesteye*, andere (relatief) vroege werken van Boendale. Heymans' mening dat het stilistisch onderscheid tussen de *Yeesten*-delen met een auteursverschil samenvalt, lijkt hoe dan ook een hypothese die men met de nodige scepsis mag benaderen: er zijn immers zoveel andere factoren die het geobserveerde stijlverschil kunnen verklaren. Interessant is wel dat deze analyses Boendales auteurschap voor de latere delen van de *Yeesten* niet schijnen tegen te spreken. Dat zou op zichzelf wel een aangename bevestiging van vroeger onderzoek zijn.

7.5 CONCLUSIE

Dit hoofdstuk handelde over Jan van Boendales oeuvre. Eerst werden de *Lekenspiegel* en *Jans Teesteye* behandeld, twee werken die met grote zekerheid aan Boendale kunnen worden toegeschreven (§ 7.1.1). Een *Principal components analysis* leerde dat Boendales teksten op basis van de hoogfrequente rijmwoordenschat relatief goed van controle-auteurs uit de *Spiegel historiael* kan worden onderscheiden (§ 7.1.2). Na een uitdieping van ons gebruik van *Memory-based learning*, werd ook Burrows' Delta op deze teksten toegepast (§ 7.2). Door een nadere studie van de variatie in de parameters voor experimenten, werd duidelijk dat algoritmes altijd goed presteren maar pas volledig foutloos werken indien zij over voldoende grote samples tekst beschikken die aan de hand van

53 Ik dank Dirk Schoenaers voor zijn suggesties met betrekking tot dit stijlverschil.

genoeg hoogfrequente rijmwoorden worden voorgesteld. Daarbij werd benadrukt dat hoogfrequente rijmwoordenschat slechts tot op zekere hoogte *inhoudsonafhankelijk* is en dat men steeds in het achterhoofd moet houden dat stylometrische methodes niet alleen gevoelig zijn voor auteursstijl. Daarna werden ook drie ethisch-didactische teksten uit de 'Antwerpse School' behandeld (§ 7.3) die mogelijk door Boendale werden geschreven (*Melibeus*, *Dietsche doctrinale* en *Boec vander wraken*). Dit onderzoek kon geen gewichtige tegenargumenten leveren tegen de Boendale-toeschrijving en bevestigde zo – hoewel slechts onrechtstreeks – de traditionele attributie van deze teksten aan de Antwerpse schepenklerk.

Het laatste deel van dit hoofdstuk stond in het teken van de toeschrijving van de *Brabantse yeesten* aan Jan van Boendale (§ 7.4). Een onderzoek van de hoogfrequente rijmwoordenschat van deze tekst toonde aan dat de stijl van deze tekst erg goed aansluit bij andere Antwerpse teksten die zeker of toch erg waarschijnlijk door Boendale zijn gedicht. In het bijzonder wat betreft de jongste uitbreidingen van de kroniek biedt dit stylometrische onderzoeksresultaat geen reden om de traditionele toeschrijving aan de Antwerpse schepenklerk in twijfel te trekken – integendeel. Wel bevestigt dit onderzoek het eerder geopperde bestaan van een stilistische breuk tussen de eerste en de latere delen van de kroniek. Het zogenaamde eerste deel van de *Yeesten* (voltooid ca. 1316, tot v. 900 van boek v) biedt een andere, meer statische vorm van geschiedschrijving dan de latere delen en sluit duidelijk minder goed aan bij de andere Antwerpse teksten in dit corpus. Een meervoudig auteurschap voor de *Yeesten* kan op basis van deze resultaten echter niet beargumenteerd worden. Er zijn immers te veel, meer voor de hand liggende factoren die de geobserveerde stijlverschillen kunnen verklaren (§ 7.4.5).

Het blijft wel intrigerend dat de *Yeesten* in humanistische middens in het begin van de zeventiende eeuw werden toegeschreven aan ene *Niclaes*. Waarschijnlijk doelde men op *Nicolaas Noese*, een van Boendales naaste collega's in het klerkenambt. Het meeste bronnenmateriaal spreekt de *Niclaes*-toeschrijving echter tegen, waardoor zij apocrief lijkt. Toch zijn de beweegredenen voor die verrassende toeschrijving nog onvoldoende uitgeklaard en vormen zij een interessante casus voor toekomstig onderzoek. Een groter inzicht in deze materie kan Boendales auteurschap voor de *Yeesten* ongetwijfeld nog verstevigen. Dat zal evenwel een geconcerteerde inspanning vragen van filologen en historici, die in de figuur van Boendale een gemeenschappelijke voorouder vinden.

8

EEN WOUT SONDER GENADE

In dit hoofdstuk richt ik mijn pijlen op de Middelnederlandse Arturepiek, een genre dat van oudsher sterk tot de verbeelding heeft gesproken van het grotere publiek, maar ook tot die van professionele filologen.[1] De Arturepiek is bovendien een van de Middelnederlandse tekstsoorten die momenteel de meeste aandacht trekken in het internationale onderzoek (Johnson 2005). Het corpus Middelnederlandse Arturepiek is divers en omvat zowel bewerkingen van anderstalige teksten als oorspronkelijk Middelnederlands materiaal.[2] Verschillende teksten werden recent in het Engels vertaald (Johnson & Claassens 2000a, b, 2003) en de Middelnederlandse Arturroman wordt zo steeds beter toegankelijk voor buitenlandse onderzoekers. Dat de Middelnederlandse Arturepiek momenteel relatief veel buitenlandse aandacht geniet, heeft echter ook te maken met de brugfunctie die de middeleeuwse literatuur uit de Lage Landen heeft vervuld. Veel Franse teksten schijnen immers naar het oosten van Europa te zijn gemigreerd via een Middelnederlands tussenstadium (Kooper 1994, 3-4).

De *Lancelot en prose*, bijvoorbeeld, werd tot driemaal toe in het Middelnederlands omgezet. De Duitse overlevering toont niettemin aan dat er nog twee andere Middelnederlandse bewerkingen moeten hebben bestaan die verloren zijn gegaan. Het colofon van de Blankenheim-codex wijst bijvoorbeeld op zijn afhankelijkheid van een niet-overgeleverde Vlaamse bron:

> Diss buchelin zu einer stonden
> Hain ich inn flemische geschrieben fonden,
> Von eyme ko<n>stigen meister verricht,
> Der es uss franczose darczu hait gedicht
> (naar Besamusca & Postma 1997, 5-6).

Reeds in de jaren vijftig van de vorige eeuw beargumenteerde de Finse onderzoeker P. Tilvis dat het Münchense fragment van de Middelhoogduitse proza-Lancelot invloed vertoont van een Middelnederlandse grondtekst (Klein 1994).

Niettegenstaande de populariteit van Middelnederlandse Arturepiek in de binnen- en buitenlandse filologie, worstelt het onderzoek nog met verscheidene vragen. Feitelijke informatie over de tijd van ontstaan of het auteurschap van deze werken is bijvoorbeeld amper beschikbaar of toch erg controversieel (Besamusca 2003, 5-8). Zeker

1 Dit hoofdstuk bouwt in belangrijke mate voort op een eerder geschreven artikel (Kestemont 2013).

2 Zie bv. Besamusca (2000), maar ook de themanummers in *Arthuriana* 25 (2005) en 27 (2007).

over het auteurschap van werken is druk gespeculeerd, hoewel men in de meeste gevallen geen consensus schijnt te hebben bereikt. Vanzelfsprekend staat ons gebrekkig inzicht terzake een beter begrip van dit corpus in de weg. Daarom loont het zeker de moeite om na te gaan of de hier verkende stylometrische onderzoeksmethodes een bijdrage kunnen leveren aan het debat over het auteurschap van de Middelnederlandse Arturepiek. Dat het leeuwendeel van de Middelnederlandse Arturepiek berijmd is, maakt deze tekstsoort enkel interessanter binnen mijn onderzoeksopzet.[3]

8.1 CODEX-BURGSTEINFURT

Hieronder zal al snel duidelijk worden dat een stylometrisch onderzoek naar de Middelnederlandse Arturepiek niet om de figuur van Velthem heen kan, die misschien wel de belangrijkste protagonist is geweest in de geschiedenis van het genre. Sterker nog: het zal blijken dat Velthems oeuvre haast noodgedwongen het startpunt vormt voor een dergelijk onderzoek. In het bijzonder zijn rol in de totstandkoming van de beruchte *Lancelotcompilatie* heeft veel aandacht gekregen, ook in het internationale onderzoek.[4] Deze rol zal hieronder nog uitgebreid aan bod zal komen. Dit hoofdstuk vangt echter aan met een bespreking van een ander handschrift, de Burgsteinfurtse codex. Aan de hand van deze casus zal ik een nieuwe stylometrische techniek voorstellen en kritisch de sterktes en zwaktes ervan toelichten. Vervolgens zal deze techniek worden toegepast op de *Lancelotcompilatie*.

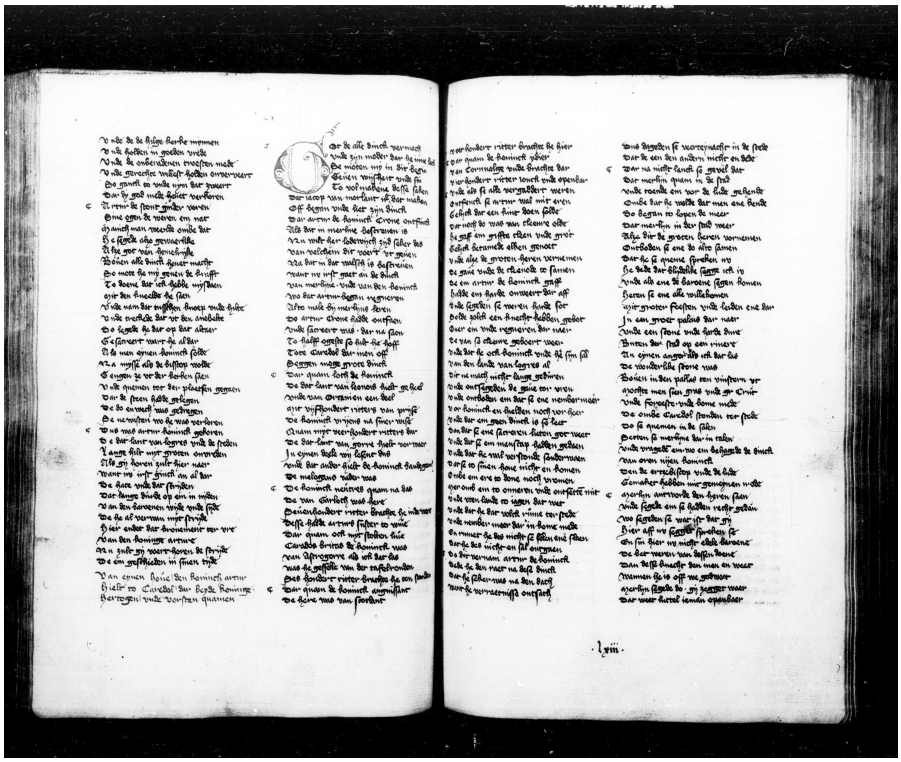
Op Witte Donderdag 1327 legde Lodewijk van Velthem de laatste hand aan zijn *Merlijn-continuatie*, een berijmde bewerking van de Oudfranse *Suite-Vulgate du Merlin* (Besamusca et al. 2009b, 27). Het is een van de jongste Middelnederlandse Arturromans en schijnt een relatief groot publiek te hebben gevonden. Er zijn fragmenten overgeleverd van drie verschillende Middelnederlandse afschriften, maar een Duits handschrift (ca. 1425) biedt de compleetste tekstgetuige (Kienhorst 1988, 140-143). In dit handschrift BURGSTEINFURT, FÜRST ZU BENTHEIMISCHE SCHLOSSBIBLIOTHEK, B37 vinden we de ca. 26.000 verzen van de *Merlijn-continuatie* in een Middelnederduitse *Umschreibung* (zie Fig. 50, p. 201).

Velthem vermeldt in de proloog dat hij de *Merline* wenste verder te zetten die zijn grote voorbeeld, Jacob van Maerlant, eerder uit het Frans had vertaald:[5]

3 Zie Besamusca & Postma (vgl. 1997, 9). Het is merkwaardig dat verschillende Franse proza-teksten 'herberijmd' werden in het Middelnederlands. Een notoire uitzondering vormen de (Rotterdamse) fragmenten van de Middelnederlandse *Proza-Lancelot* (Lie 1987). Van dit handschrift zijn recent nieuwe fragmenten opgedoken (Minnen & Claassens 2005).

4 Ook Velthems historiografische teksten hebben overigens de aandacht getrokken in het buitenland; zie bv. Bibble (2000, 378-386).

5 Kritisch weergegeven naar Fig. 51, p. 202, met dank aan M. Demol.



Figuur 50: Fol. 62v-63r van codex BURGSTEINFURT, FÜRST ZU BENTHEIMISCHE SCHLOSSBIBLIOTHEK, B37, het tweekolommige handschrift (ca. 1425) met een verduitste *Umschreibung* van de *Merlijn*romans van Jacob van Maerlant en Lodewijk van Velthem. Bovenaan de tweede kolom van het links afgebeelde folium grijpt de auteursovergang tussen beide dichters plaats (zie ook Fig. 51, p. 202).

Got, de alle dinck vermach,
Unde zijn moder, dar he inne lach,
Se moten my in dit begin
Geven wijsheit unde sin
To volmakene desse saken,
Dat Jacop van Morlant irst dat maken
Off began, unde liet zijn dinck
Dar Artur de koninckcrone ontfinck,
Als dat in *Merline* bescreven is.
Nu wilt her Lodewijch, zijd seker des,
Van Vel[t]hem dit voert utgeven
Nadat in dat Walsch is bescreven.

Velthems tekst wordt in de Burgsteinfurtse codex inderdaad voorafgegaan door Maerlants *Historie vanden Grale* en *Boec van Merline* (ca. 1260). Beide teksten zijn vertalingen van de Oudfranse prozaadaptaties van *Estoire dou Graal* en de *Merlin* door Robert de Boron. Er wordt vaak verondersteld dat Velthem de oorspronkelijke tekst van Maerlant tot op zekere hoogte bewerkt heeft. Dat is echter moeilijk

Figuur 51: Bovenkant van de tweede kolom op fol. 62v van de Burgsteunfurtse codex (vgl. Fig. 50, p. 201). Velthem beschrijft hoe hij Maerlants tekst wil verderzetten.

vast te stellen, omdat Maerlants oorspronkelijke tekst niet onafhankelijk van Velthem is overgeleverd (Kienhorst 1988, 140-143). Voor één passage in Maerlants dubbelroman over Maskeroen is bijvoorbeeld duidelijk dat die op Velthem kan teruggaan (Besamusca et al. 2009a, 12-13). De passage werd ingevoegd tussen Maerlants romans en vertoont onmiskenbaar invloed van het werk van Velthem (Warnar 2009). Deze *Maskeroen*-passage biedt een bewerking van de Latijnse *Processus Satanae*, waarin het personage Maskeroen de mensheid voor God roept en in een soort rechtsgeding de heerschappij van de Duivel over de zielen van de mensen tracht te verkrijgen. Maria komt echter tussenbeide en weet het pleit in het voordeel van God beslechten.

Het Burgsteinfurtse handschrift biedt een interessante casus voor een stylometrische analyse. Gegeven het feit dat Velthem Maerlants tekst herwerkt kan hebben en het feit dat de tekst slechts is overgeleverd in een veel jonger, sterk verduitst handschrift is het een boeiende vraag of stylometrische methodes nog in staat zijn om de auteursovergang in de codex bloot te leggen. Als referentiemateriaal kunnen we opnieuw een beroep doen op de *Spiegel historiael*, meerbepaald *P3* en *P4P5*. Voor de Burgsteinfurtse codex doen we een beroep

op de editie op de *Cd-rom Middelnederlands*. Die editie is niettemin erg kritisch, want de schrijfwijze van de teksten is door de editeur verregaand genormalizeerd.[6] Het gebruik van deze editie maakt het stilistisch onderscheiden van beide auteurs dus nog moeilijker, omdat de editeur – een beetje als een moderne kopiist – stijlverschillen tussen de auteurs mogelijk heeft uitgevlakt. Omdat Velthems aandeel in de codex veel groter is dan dat van Maerlant, beperk ik me voor de *Merlijn-continuatie* tot een sample uit de tekst.[7] Maerlants teksten tellen – zonder de *Maskeroen*-passage (974 verzen) – 9.113 verzen in het corpus.

Voor het samplen van de Burgsteinfurtse codex kunnen we een beroep doen op een segmentatieprocedure die gelijkt op deze die we voor Velthems Vijfde Partie gebruikten maar toch iets verder gaat. Voor Velthems Vijfde Partie maakten we gebruik van groepjes opeenvolgende, niet overlappende kapittels in de tekst (§ 5.2.4). Hier kunnen we – voor een iets fijnmaziger resultaat – gebruik maken van deels overlappende reeksen verzen van een gelijke grootte, i.e. *windows* (vgl. Van Dalen-Oskam & Van Zundert 2007, 351). De sample-grootte kunnen we aanduiden als de *window size* (*ws*); de afstand tussen de eerste verzen van twee opeenvolgende *windows* kunnen we aanduiden als de *step size* (*ss*). Merk op dat de *step size* groter moet zijn dan nul – want anders zouden we nooit vooruitgaan door de tekst – maar kleiner of gelijk aan de *window size* – omdat we anders verzen zouden overslaan. Met deze methode ($0 > ss >= ws$) kunnen we teksten nu verdelen in samples. Laat ons dit hier eerst doen voor het referentiemateriaal van Maerlant en Velthem uit de *Spiegel historiael*, zodat we een aantal representatieve samples verkrijgen van de schrijfstijl van beide auteurs in *P3* en *P4P5*.

Vervolgens gooien we de *windows* of samples van beide auteurs op één hoop en we bepalen de cumulatieve frequentie van elk rijmwoord. We beperken onze verdere metingen tot de *n* rijmwoorden die het frequentst voorkomen in deze samples (bv. 100 rijmwoorden). Voortbouwend op Burrows' Delta, definiëren we nu een soort centroïde of stilistisch prototype voor elke auteur, op basis van deze *n* rijmwoorden. Voor elk van deze *n* rijmwoorden wordt eerst de relatieve frequentie bepaald in elk voorbeeld-sample van de auteur. Vervolgens steken we in de centroïde van de auteur de gemiddelde frequentie

centroïde

6 De enige volledige – maar extreem kritische – editie van het handschrift is die van Van Vloten (1880). De veelbekritiseerde editeur (1818–1883) heeft ervoor gekozen om de spelling van de tekst te 'restaureren' en de Duitse *Umschreibung* opnieuw om te zetten naar een Middelnederlandse spelling. Dat gebeurde echter op een weinig inzichtelijke en inconsequente wijze (Lodder 2003). De jongere, meer betrouwbare editie door Sodmann (1980) blijft beperkt tot Maerlants aandeel en daarom zal ik toch gebruik maken van de editie-Van Vloten (vgl. ook Hogenbirk 2009b, 53).

7 Het gaat om de eerste 14.237 verzen uit de tekst. Dit specifieke aantal verzen is niet betekenisvol en komt voort uit een niet-gerapporteerd vooronderzoekje waarin ik Velthems *Merlijn-continuatie* heb vergeleken met Maerlants *Alexanders Geesten* die in mijn corpus 14.237 verzen telde.

van deze rijmwoorden, evenals de standaarddeviatie ten aanzien van dit gemiddelde. De centroïde voor elke auteur zal uiteindelijk dus voor n rijmwoorden de gemiddelde frequentie en standaarddeviatie in het referentiemateriaal bevatten. De centroïde biedt zo een vrij naïeve voorstelling van de schrijfstijl van een auteur: we weten hoe vaak een auteur een rijmwoord gemiddeld gebruikt, alsook welke afwijking zijn gebruik daarin kan vertonen.

Een dergelijk centroïde kunnen we vervolgens inzetten om de stijlschommelingen in een testtekst te meten. Als we de Burgsteinfurtse codex in *windows* verdelen, verkrijgen we een reeks samples die we individueel kunnen vergelijken met het stilistisch prototype of het centroïde van elke auteur. Achtereenvolgens kunnen we voor elk sample uit de testtekst een soort Delta berekenen ten aanzien van de centroïde van een auteur en nagaan of zich in de resulterende reeks waarden abrupte overgangen voordoen. Laat C het centroïde van een auteur in kwestie zijn en W een *window* dat we ermee willen vergelijken. Voor de vergelijking met n rijmwoorden kunnen we de volgende eenvoudige formule hanteren om de Delta tussen C en W te berekenen:

$$\Delta(C, W) = \sum_{i=1}^{n} \frac{1}{\sigma_i(C)} |\mu_i(C) - f_i(W))| \tag{8.1}$$

Voor elk rijmwoord berekenen we eerst het absolute verschil tussen de frequentie van het rijmwoord in het test-*window* en de gemiddelde frequentie voor hetzelfde rijmwoord in de centroïde. Vervolgens wegen we dit verschil door het te delen door de standaarddeviatie in de frequentie van dit rijmwoord in het referentiemateriaal. Deze gewogen verschillen worden tot slot samengeteld. De wegingsfactor heeft tot gevolg dat rijmwoorden waarvan het gebruik relatief grotere schommelingen vertoont in het referentiemateriaal voor de bewuste auteur relatief minder zwaar zullen doorwegen in de uiteindelijke afstandsbepaling. De achterliggende gedachte is dat rijmwoorden die een auteur inconsequent gebruikt, slechtere markeerders zijn van zijn gemiddelde schrijfstijl. Concreet betekent een hogere Delta-waarde dus een relatief grotere afstand in stijl tussen een test-*window* en een centroïde, terwijl een lagere Delta een grotere stijlgelijkenis impliceert.

Laat ons deze techniek nu eens toepassen op de casus van het Burgsteinfurtse handschrift. Eerst verdelen we het beschikbare referentiemateriaal van Maerlant (*P3*) in *windows*. Vervolgens doen we hetzelfde met het beschikbare testmateriaal uit de Burgsteinfurtse codex. Daarna berekenen we voor elk van de verkregen test-*windows* de Delta ten aanzien van Maerlants centroïde. We kunnen volgende hypothese formuleren: voorafgaand aan de auteursovergang in het handschrift zouden de *windows* relatief lagere Delta-waardes moeten vertonen ten aanzien van Maerlant's centroïde. Deze *windows* gaan oorspronkelijk immers terug op Maerlant en zouden daarom een rela-

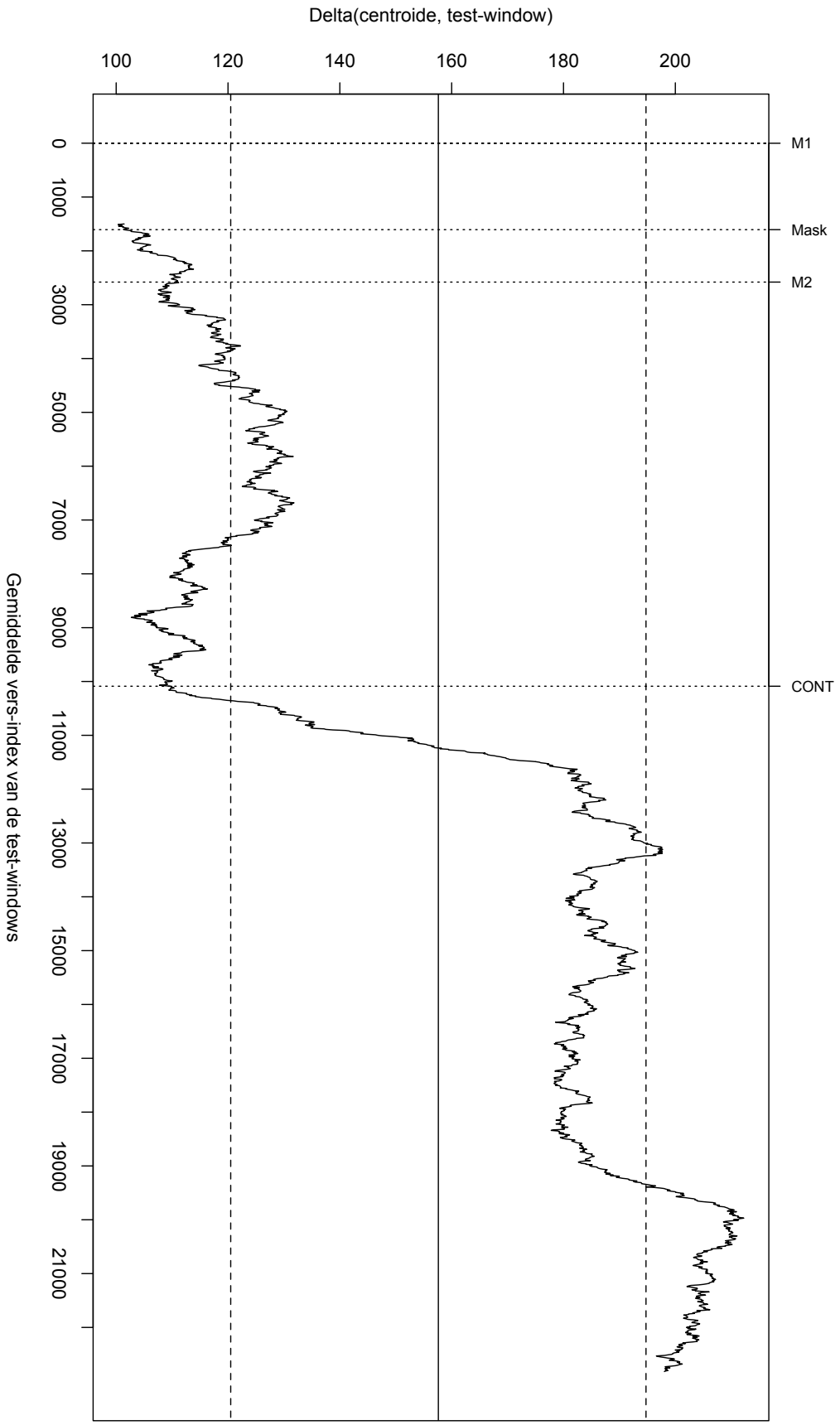
tief grotere stilistische gelijkenis moeten laten zien met Maerlants stijl dan de *windows* uit de *Merlijn-continuatie* die op Velthem teruggaan.

Onderstaande figuur geeft het resultaat weer van een experiment met Maerlants *P3* als referentiemateriaal voor de beschikbare teksten uit het Burgsteinfurtse handschrift (Fig. 53 (p. 207), *ws*=3.000, *ss*=10, *n*=100). De horizontale as geeft hierin de Burgsteinfurtse codex weer, verdeeld in een aantal opeenvolgende *windows* met dezelfde verslengte. De Delta-waardes worden hier op de horizontale as geprojecteerd op basis van de index van hun middelste vers. Op de verticale as wordt de Delta-waarde van ieder *window* weergegeven, ten aanzien van het centroïde dat werd samengesteld op basis van Maerlants *P3*. De horizontale lijnen in de grafiek bieden een leeshulp voor de interpretatie van de grafiek: de volle horizontale lijn geeft de gemiddelde delta-waarde weer in dit experiment; de horizontale streepjeslijnen geven dit gemiddelde weer, respectievelijk plus en min de standaarddeviatie. De verticale stippellijnen markeren verschillende 'gebeurtenissen' in de Burgsteinfurtse tekst: van links naar rechts gaat het om: het begin van Maerlants eerste roman (M1), het begin van de geïnterpoleerde *Maskeroen*-episode (MASK), het begin van Maerlants tweede roman (M2) en de start van Velthems *Merlijn-continuatie* (CONT).

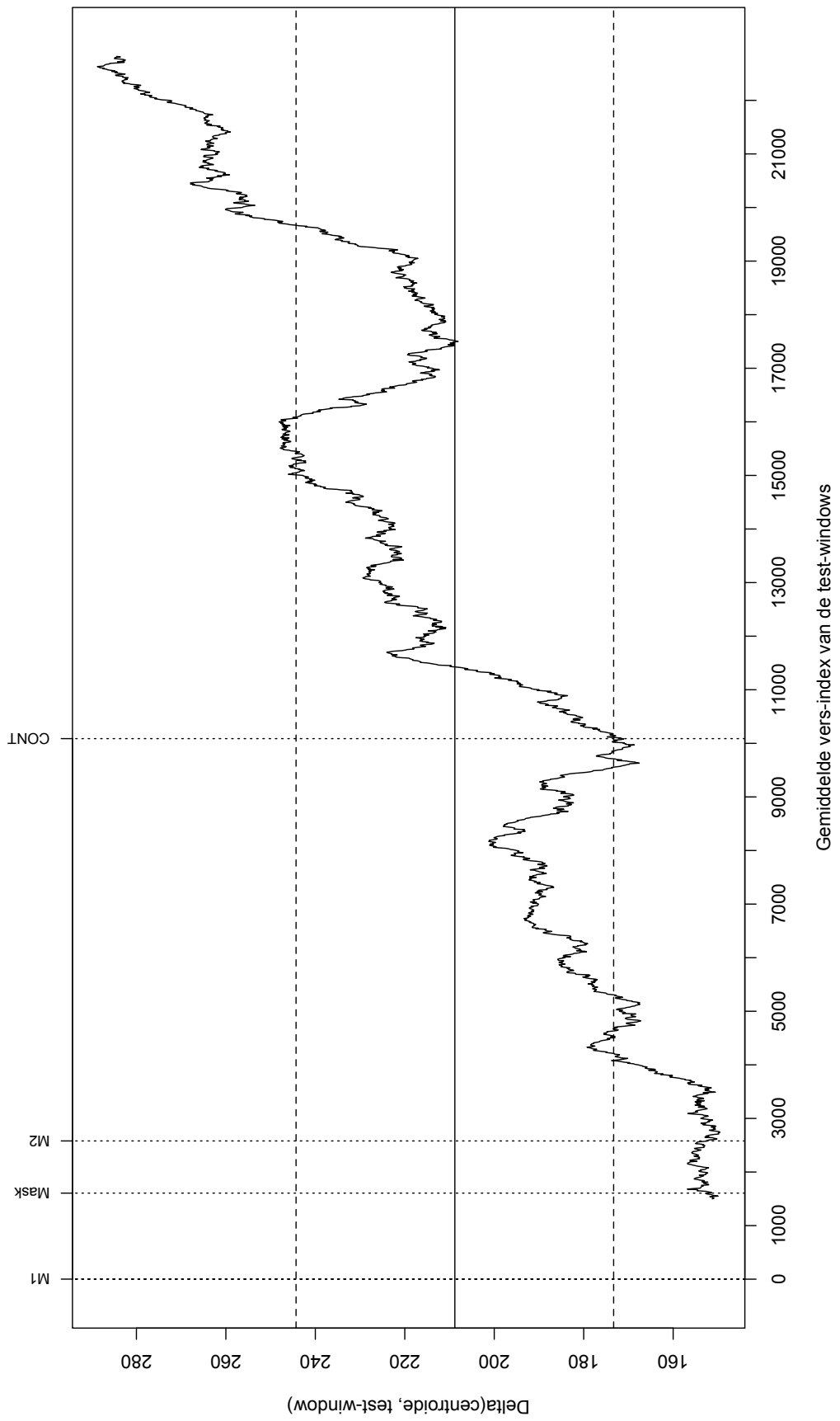
Het resultaat dat we uit deze grafiek kunnen aflezen, bevestigt grotendeels onze hypothese: vóór de auteursovergang (vgl. CONT) schijnt de afstand van de Burgsteinfurt-*windows* ten opzichte van Maerlants *P3*-centroïde veelal kleiner (< gemiddelde). Ná de auteursovergang (ca. vers 10.000) is die afstand duidelijk groter (> gemiddelde). Het verschil in Delta-waarden correspondeert duidelijk met de overname door Velthem in het handschrift. Hieruit zouden we voorzichtig kunnen concluderen dat deze stylometrische methode inderdaad geschikt is om de stijlschommelingen in dit handschrift te bestuderen.

Als we deze procedure op het Burgsteinfurtse handschrift nu toepassen op basis van Velthems centroïde (*P4P5*), verwachten we precies het tegenovergestelde resultaat: de *Merlijn-continuatie* zou immers dichter bij Velthems schrijfstijl moeten aanleunen, hetgeen zou moeten resulteren in relatief lagere Delta-waardes na de auteursovergang. Onderstaande figuur leert echter dat dit experiment allesbehalve de inverse van de vorige analyse oplevert en dus niet aan onze verwachtingen beantwoordt (Fig. 53, p. 207). Ook in deze analyse zijn de waardes na CONT immers relatief hoger dan ervoor. Hoe moeten we deze verrassende uitkomst verklaren?

Misschien kunnen we via dimensie-reductie een beter zicht krijgen op de zaak. De onderstaande figuur geeft het resultaat weer van een correspondentie-analyse (Fig. 54, p. 208), gebaseerd op de 100 frequentste rijmwoorden in (ditmaal niet-overlappende) samples van 3.000 rijmwoorden uit *P3*, *P4P5*, Maerlants dubbelroman zonder de *Maskeroen*-passage (*MerlM*) en Velthems continuatie (*MerlV*). Een eerste observatie is dat samples afkomstig uit dezelfde tekst clusteren:

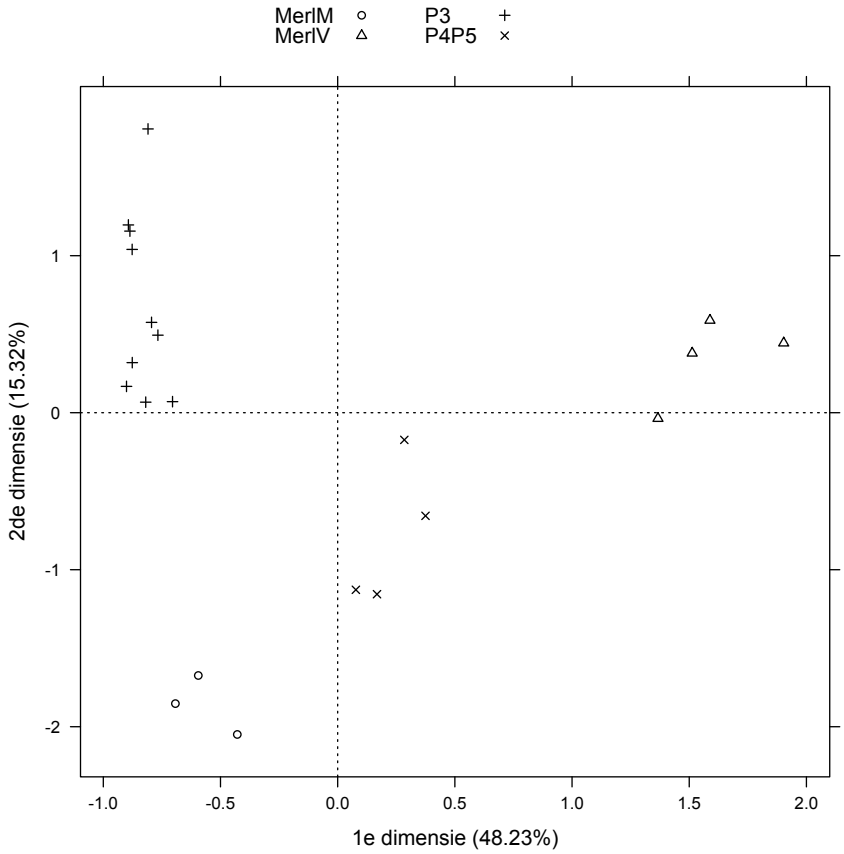


Figuur 52: Resultaat van een Delta-experiment met Maerlants P_3 als referentiemateriaal voor het Burgsteinfurtse handschrift (ws=3.000, ss=10, n=100). De horizontale as stelt de Burgsteinfurtse codex voor, verdeeld in een aantal opeenvolgende *windows*. Op de verticale as wordt de Delta-waarde van ieder *window* weergegeven, ten aanzien van het centroïde dat werd samengesteld op basis van Maerlants P_3.



Figuur 53: Resultaat van een Delta-experiment met Velthems P_4P_5 als referentiemateriaal voor het Burgsteinfurtse handschrift (ws=3.000, ss=10, n=100). De horizontale as stelt de Burgsteinfurtse codex voor, verdeeld in een aantal opeenvolgende *windows*. Op de verticale as wordt de Delta-waarde van ieder *window* weergegeven, ten aanzien van het centroïde dat werd samengesteld op basis van Velthems P_4P_5.

de sample uit *MerlM* vormen bijvoorbeeld een hecht groepje (drie cirkeltjes linksonder in de figuur). Een tweede belangrijke vaststelling is dat de samples van onze twee auteurs in de eerste dimensie foutloos gescheiden worden door de verticale stippellijn door de nulwaarde op de horizontale as. Het is in dat opzicht interessant dat de factoren 'kopiist', 'editeur' of 'onderwerp' onze analyse niet schijnen te verstoren want de samples van *MerlM* en *MerlV* zijn duidelijk gesitueerd in verschillende regio's van het strooidiagram.



Figuur 54: Correspondentie-analyse (eerste twee dimensies) van samples van 3.000 rijmwoorden uit Maerlants *P3* en *MerlM* en Velthems *P4P5* en *MerlV*. De analyse is gebaseerd op de relatieve frequenties van de 100 meest frequente rijmwoorden in deze samples.

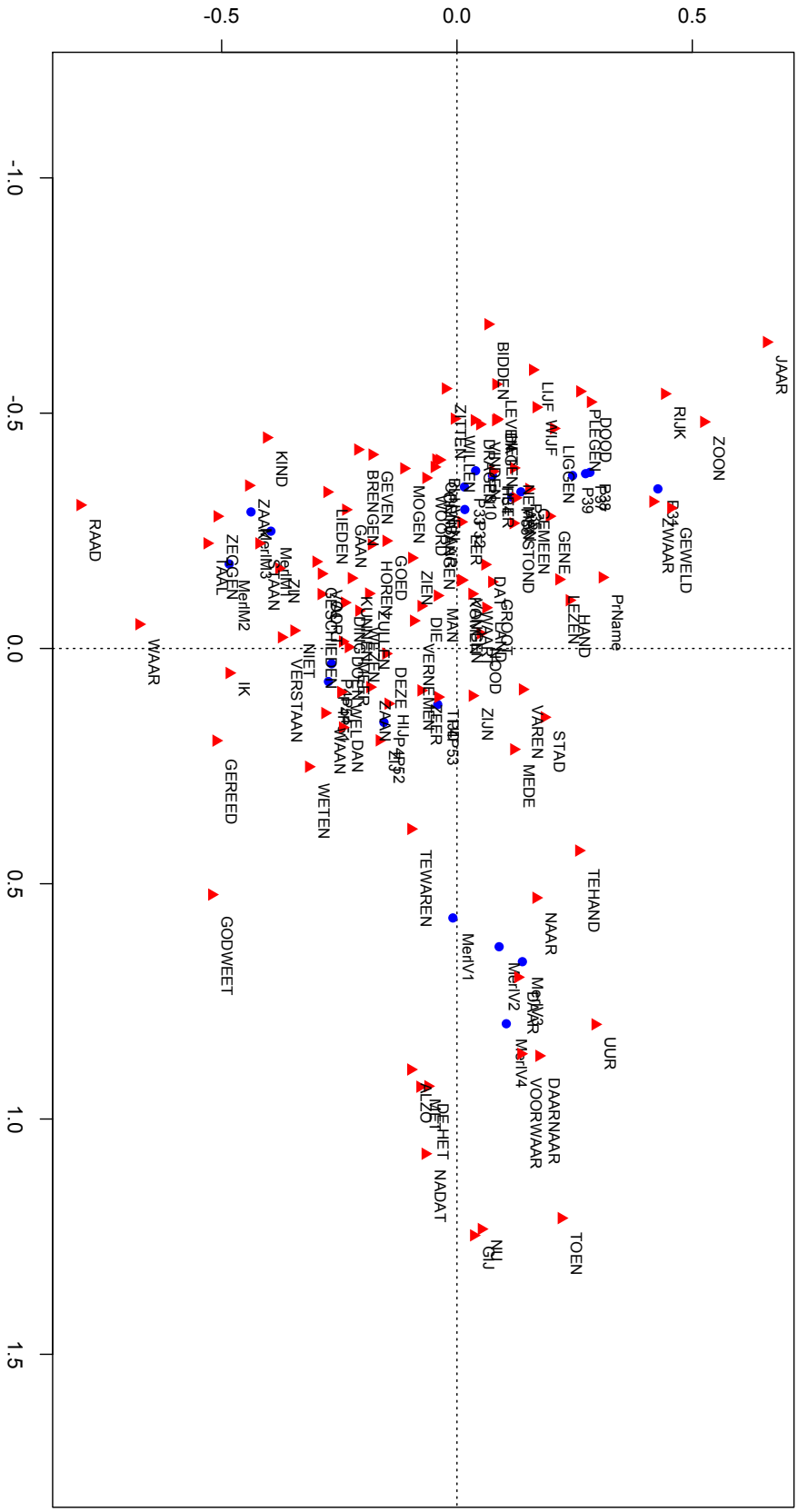
Ondanks de scheiding die wordt gerealiseerd in de eerste dimensie, valt het wel op dat de afstand tussen Maerlants teksten *MerlM* en *P3* veel kleiner is dan de afstand tussen Velthems teksten *P4P5* en *MerlV* – de samples uit deze laatste tekst worden duidelijk geïsoleerd uiterst rechts in de figuur. Als we de tekstscheiding in de eerste dimensie mogen in verband brengen met de factor 'auteurschap', zijn Velthems teksten dus verder van elkaar verwijderd dan we zouden verwachten op de basis van de onderlinge verwijdering tussen Maerlants teksten.

De tweede figuur hieronder, waarin ook de 'ladingen' van de individuele rijmwoorden worden getoond (Fig. 55, p. 210), is bijzonder slecht leesbaar maar de extreme positie van enkele individuele rijmwoorden uiterst rechts maakt wel duidelijk dat de uitgesproken verwijdering van Velthems *MerlV* wordt veroorzaakt door de hoge frequentie van een klein aantal rijmwoorden. Erg kenmerkend schijnt de combinatie van de lexemen NU en GIJ. Dat deze rijmwoorden zo dicht op elkaar worden afgebeeld, toont overigens aan dat deze rijmwoorden vaak binnen eenzelfde rijmpaar moeten worden gebruikt (*u* : *nu* of omgekeerd *nu* : *u*). Ook andere rijmlexemen die makkelijk binnen dezelfde rijmparen kunnen worden gebruikt, nemen in deze figuur een vergelijkbare positie in (NAAR, DAAR, VOORWAAR, DAARNAAR). De stoplap GODWEET schijnt eveneens typisch voor Velthems rijmstijl, maar omdat dit rijmlexeem iets lager wordt afgebeeld, schijnt Velthem er vooral in *P4P5* een beroep op te doen.

Mogelijk is een genre-gerelateerd verschil in rijmwoorden dus de oorzaak van de relatief grote verwijdering tussen Velthems teksten. Dan zouden wij met andere woorden stoten op de beperking die Brinkman reeds formuleerde in zijn Vijfde Gebod, namelijk dat een stylometrische analyse van auteurschap slechts mogelijk is binnen de grenzen van een enkel genre. Velthem gebruikt steeds veel stoplappen, maar kennelijk bestaat er een specifieke set rijmwoorden die hij nog vaker gebruikt in zijn Arturtekst dan in zijn historiografische tekst. Dat de stoplap GODWEET relatief frequenter is in het historiografische *P4P5*, kunnen we bijvoorbeeld opvatten als een indicatie dat Velthem het waarheidsgehalte van zijn verhaal in deze tekst sterker wilde benadrukken. Niettemin weet de eerste dimensie in deze correspondentie-analyse Velthems teksten toch nog van die van Maerlant te onderscheiden: zo bekeken is Velthems *P4P5* wat betreft bepaalde stijlkenmerken kennelijk toch nog meer gelijkend op *MerlV* dan op Maerlants teksten. Misschien kunnen we deze 'relatieve nabijheid' toch nog uitbuiten.

Vijfde Gebod: De eenheid van genre

Onderstaande figuur biedt een analyse waarin we opnieuw een beroep doen op de metriek van de 'Dubbele Delta', die we reeds gebruikten voor de analyse van Velthems Vijfde Partie hierboven (Fig. 56 (p. 212), vgl. § 5.2.2). Daarbij hebben we voor elk *window* uit Burgsteinfurt eerst de delta berekend tot Velthems centroïde (= $\delta 1$) en vervolgens de delta tot Maerlants centroïde (= $\delta 2$). Vervolgens kunnen we de Dubbele Delta ($\delta\delta$) gelijkstellen aan $(\delta 1/\delta 2) - 1$. Deze afstandsmaat zal mogelijk een meer fijnzinnige analyse opleveren: we erkennen het feit dat er een verschil zit op Velthems stijl in de test- en referentietekst maar we willen nagaan of de testtekst toch niet *relatief dichter* ligt bij de referentietekst van Velthem dan bij die van de controle-auteur Maerlant. In onderstaande figuur herhalen we de bovenstaande experimenten met dezelfde segmentatieparameters (*ws*=3.000, *ss*=10), maar we gebruiken de Dubbele Delta in plaats van


0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
1.5
JAAR
ZOON
RIJK
GEWELD
ZWAAR
PrName
TEHAND
UUR
DAARNAAR
VOORWAAR
TOEN
NU
NADAT
ALZO
DE
HET
MerIV1
MerIV2
MerIV3
MerIV4
NAAR
STAD
VAREN
MEDE
TEWAREN
GODWEET
GEREED
IK
WAAR
RAAD
WETEN
NIET
VERSTAAN
KIND
ZIJN
MerIM2
MerIM3
ZAAN
ZEGGEN
PAAL
GEVEN
BRENGEN
GAAN
LIEDEN
GOED
HOREN
MOGEN
ZIEN
DEZE
HIJ
ZAAN
ZUN
MAN
DIE
VERNEMEN
MEER
TRAP53
TRAP52
LIGGEN
GENE
HAND
LEZEN
LIJF
WIJF
BIDDEN
LEVEN
DAGEN
PEER
ZITTEN
WILLEN
WOORD
WAAR
P39
P38
P33
P10
P32
P31
P30
P29
P28
P27
P26
P25
P24
P23
P22
P21
P20
P19
P18
P17
P16
P15
P14
P13
P12
P11
P10
P9
P8
P7
P6
P5
P4
P3
P2
P1
P0
POOD
PLEGEN
GEMEEN
STOND
PROGRAM
GOD
ZUN
DAN
WAAN
WELDAN
PRAGEN
KLUNKEN
DOEN
HEEFT
GEDOCHTEN
LIEDEN
HUIS
HUIZEN
HUIZEN
HUIZEN

de gewone Delta (Fig. 56, p. 212). Deze analyse levert een betere benadering op van onze initiële hypothese, want zij suggereert dat Velthems bijdrage aan de Burgsteinfurt-tekst inderdaad relatief dichter ligt bij de stijl van deze auteur in *P4P5*. In het geval van een relatief lagere $\delta\delta$, is een *window* immers relatief meer gelijkend op Velthems stijl dan Maerlants stijl (immers, $\delta_1 < \delta_2$).

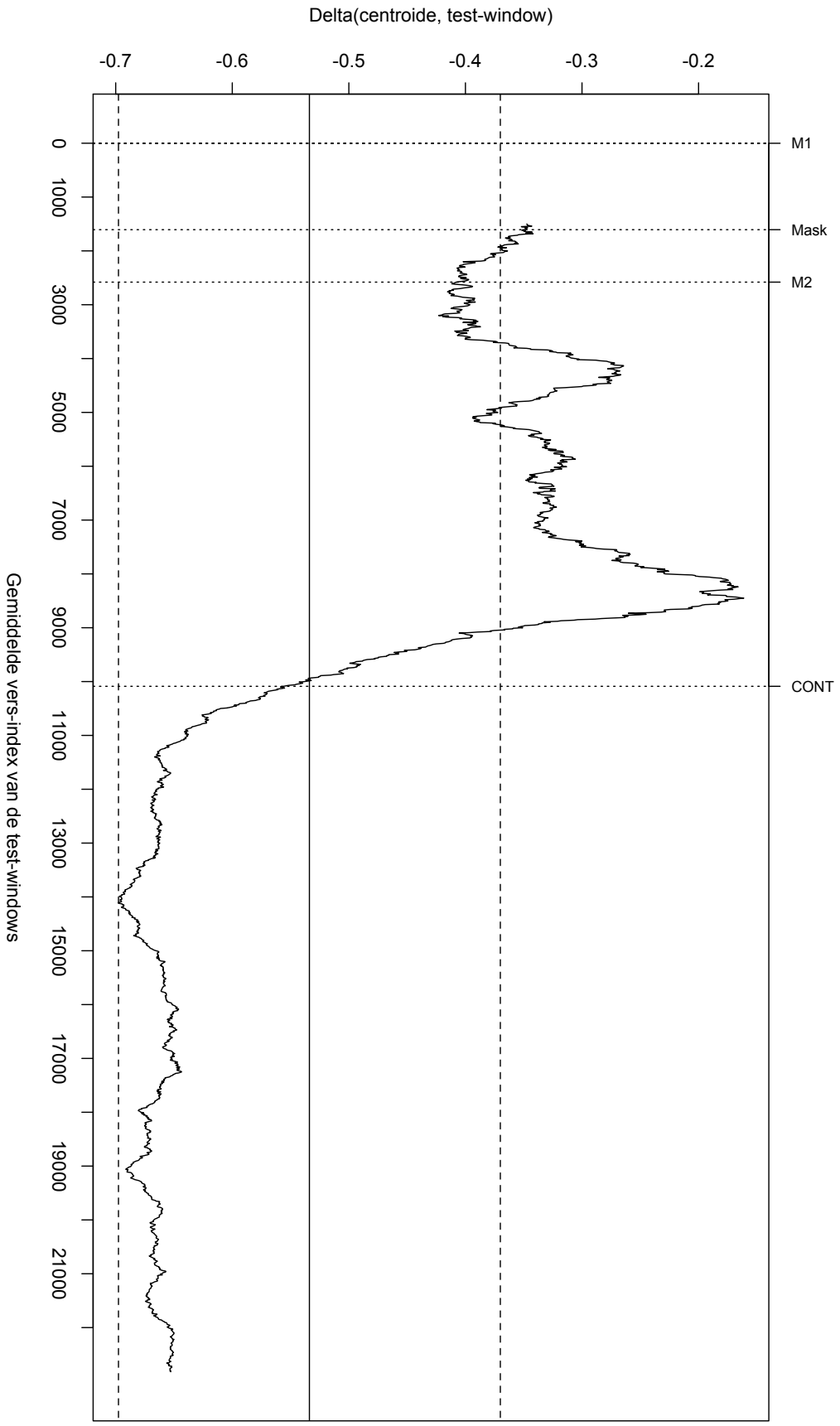
Toch is een over-enthousiaste interpretatie van dit resultaat niet aangewezen. In het geval van de Dubbele Delta-berekening is nul immers een kritische waarde: als een sample een negatieve $\delta\delta$ wordt toegewezen, betekent dit in onze berekeningswijze dat we het bewuste sample in principe aan Velthem moeten toeschrijven. Merk echter op dat geen van de Burgsteinfurt-*windows* uit het Maerlant-deel in onderstaande figuur een positieve waarde bereikt. Strikt genomen zouden we deze Maerlant-passages toch nog aan Velthem moeten toewijzen, wat een ongewenst resultaat lijkt.[8] Uit bovenstaande sectie kunnen we dus onthouden dat de centroïde-analyse een interessante techniek vormt om de stijlschommelingen binnen teksten bloot te leggen, maar dat we terughoudend moeten zijn in onze interpretatie van dergelijke resultaten. In deze experimenten werd immers duidelijk dat Velthems stijl deels kan correleren met het genre waarin hij schrijft, een gegeven dat reeds werd opgemerkt (vgl. Hogenbirk 2009b, 71, zie ook § 9.1.3).[9] Dan zouden we moeten benadrukken hoe gelukkig we mogen zijn dat Velthems *MerlV* in het Burgsteinfurtse handschrift is overgeleverd. Velthems eventuele bijdrage aan de *Lancelotcompilatie* moet immers hoe dan ook binnen het genre van de Arturepiek worden gesitueerd. Velthems Arturiaanse *Merlijn*-tekst biedt in dat opzicht een meer betrouwbaar uitgangspunt dan zijn historiografische werk.

8.2 HET BOEK VAN LANCELOT

In de Koninklijke Bibliotheek van Den Haag wordt een uniek manuscript uit de vroege veertiende eeuw bewaard (129 A 10). Dit Middelnederlandse handschrift staat in de medioneerlandistiek bekend als

8 Merk op dat de δ-waarden voor de *Maskeroen*-passage in geen van deze experimenten opvallende afwijkingen vertoonden. Dat hangt natuurlijk samen met het feit dat hier een relatief grote *ws* wordt gebruikt, zodat de afwijkende stijl van de kortere *Maskeroen*-passage niet kon worden opgepikt. Een correspondentie-analyse – die ik hier niet verder bespreek – met een kleinere *ws* kon echter Velthems auteurschap voor deze passage aannemelijk maken. Ik verwijs hiervoor naar de parallelle analyses van Brinkman (2010), die binnenkort gepubliceerd zullen worden. Merk trouwens op dat Velthem de *Maskeroen*-passage niet noodzakelijk zelf moet ingevoegd hebben: in principe kan een latere kopiist een passage uit Velthems vroegere tekst hebben ingebouwd. Dit is echter onwaarschijnlijk.

9 Afgezien van het genre-verschil zijn ook bijkomende verklaringen denkbaar voor de weinig geïnspireerde rijmtrant in *MerlV*. Een anonieme beoordelaar wees erop dat reeds de *Suite-Vulgate du Merlin*, de Franstalige brontekst van *MerlV*, een relatief saaie tekst is ('een uitermate vervelende aaneenschakeling van veldslagen'). Dit kan zeker een gedeeltelijke verklaring bieden voor de 'dichterlijke desinteresse' die uit Velthems nauwelijks gevarieerde rijmwoordenschat valt af te lezen.



Figuur 56: Resultaat van een 'Dubbele Delta'-experiment met Velthems P_4P_5 als referentiemateriaal voor het Burgsteinfurtse handschrift (ws=3.000, ss=10, n=100). Maerlants P_3 wordt gebruikt als controle-centroïde bij het berekenen van de 'Dubbele Delta'. Op de verticale as wordt de $\delta\delta$-waarde van ieder *window* weergegeven.

de Haagse *Lancelotcompilatie*.[10] Men duidt dit handschrift aan als een 'compilatie' omdat verscheidene, reeds bestaande romans in de codex tot een samenhangend geheel zijn samengevoegd. Het manuscript wordt trouwens de 'schatkamer' van de Middelnederlandse arturistiek genoemd, omdat maar liefst 10 van de 17 overgeleverde Middelnederlandse Arturromans uit deze compilatie bekend zijn (Besamusca 1985). Toch is het handschrift slechts het laatste deel van de oorspronkelijke compilatie: dit volume moet immers zijn voorafgegaan door (tenminste) een ander handschrift, dat verloren is gegaan.[11] In het overgeleverde handschrift treft men voortdurend terugwijzingen aan naar de inhoud van het verloren gegane eerste deel.

Codicologen hebben de genese van dit handschrift betrekkelijk goed kunnen reconstrueren. Aanvankelijk moet het plan zijn geweest om in de compilatie drie teksten op te nemen, namelijk de Middelnederlandse vertalingen van de drie delen van de Oudfranse *Lancelot en prose*. Deze prozateksten werden in hun Middelnederlandse pendanten opvallend genoeg 'herberijmd'. Deze drie omvangrijke 'kerndelen' uit de compilatie worden doorgaans aangeduid als de *Roman van Lanceloet* (hieronder afgekort tot LAN, in het corpus 36.900 rijmwoorden), de *Queeste van den Grale* (QUE, 11.158 rijmwoorden) en *Arturs doet* (DOE, 12.298 rijmwoorden).[12] Reeds in dit stadium werd in de *Lanceloet* de *Perchevael* ingevoegd, een Middelnederlandse versvertaling van de Oudfranse *Perceval* door Chrétien de Troyes (PER, 5.596 rijmwoorden). Maar liefst vijf kopiisten hebben tot op dit punt meegewerkt aan de compilatie: de kopiisten A, C, D en E hebben alle bijgedragen aan het afschrift van de *Lanceloet*. Zij werden waarschijnlijk gesuperviseerd door een soort 'werkleider', kopiist B, die zelf het afschrift van het einde van *Lanceloet* en de volledige *Perchevael* voor zijn rekening nam. Al snel werd besloten om de *Perchevael* opnieuw los te maken uit de

10 Een recente monografie biedt Besamusca (2003). Een codicologische behandeling vindt men in de bijdrage van Klein bij Besamusca & Postma (1997, 51-90) of bij Biemans (2009).

11 In theorie is het mogelijk dat het overgeleverde manuscript door meer dan één codex werd voorafgegaan, maar deze mogelijkheid acht men zelden waarschijnlijk. Een notitie achterin de Burgsteinfurt-codex leert dat de bibliotheek waarin dit handschrift ca. 1425 werd bewaard, eveneens *twe nye boke van lantslotte vnde eyn olt boek van lantslotte* bevatte. Mogelijk is dit een referentie aan het verloren compilatie-handschrift (Draak 1974). Zie ook onder.

12 Voor de *Lancelotcompilatie* maak ik in dit onderzoek gebruik van de recent gelanceerde online weergave of 'leeseditie' van de codex (`www.kb.nl/lancelot`) waarin verschillende, reeds bestaande edities zijn samengevoegd door H. Brinkman en J. Biemans. In de oudere editie-Jonckbloet zijn die delen vervangen waarvan een jongere versie beschikbaar was. Die vervangingen gaan terug op digitale documenten die door een groot aantal editeurs belangeloos zijn ter beschikking gesteld. De 'gulle schenkers' van edities waren in alfabetische volgorde: B. Besamusca, F. Brandsma, G. Claassens, W. Gerritsen, M. Hogenbirk, D. Johnson, W. Kuiper, S. Oppenhuis de Jong, A. Postma en R. van Eekelen. Zie voor gedetailleerde referenties de online verantwoording bij de 'leeseditie'. Deze 'compilatie' van edities en transcripties is sterk hybride maar biedt als geheel de recentste weergave van de *Lancelotcompilatie*.

Lanceloet, en de *Perceval*-vertaling aan het einde van *Lanceloet* in te bouwen.

Op dat moment gebeurde er iets merkwaardigs: de compilatie kwam in handen van een anonieme 'corrector' die de voorlopige compilatie van correcties en marginale aantekeningen voorzag.[13] Deze corrector heeft men niet kunnen identificeren als een van de bovengenoemde kopiisten, zodat er blijkbaar nog een zesde persoon in het spel was. Vervolgens kwam de gecorrigeerde compilatie terug in handen van kopiist B, die de compilatie nog gevoelig uitbreidde. Tussen de *Perchevael* en *Queeste*, werd de *Roman van Moriaen* ingevoegd (MOR). Tussen de *Queeste* en *Arturs doet* werden daarna nog eens vijf andere romans geïnterpoleerd, achtereenvolgens: *Wrake van Ragisel* (WRAK, 3.420 rijmwoorden), *Ridder metter mouwen* (MOU, 4.022 rijmwoorden), *Walewein ende Keye* (KEY, 3.668 rijmwoorden), *Lanceloet en het hert met de witte voet* (VOE, 856 rijmwoorden) en tot slot *Torec* (TOR, 3.854 rijmwoorden). Het is in de secundaire literatuur daarom gebruikelijk om de drie langere 'kernteksten' te onderscheiden van de zeven kortere 'geïnterpoleerde' romans.

Er zijn in de literatuur enkele verantwoorde gissingen gedaan naar de inhoud van het eerste deel van de compilatie. Het eerste deel moet tenminste tweederde van de oorspronkelijke *Lanceloet* hebben bevat, want de overgeleverde codex bevat eigenlijk slechts het laatste deel van *Lanceloet* en daarin wordt herhaaldelijk verwezen naar het begin van de roman (Besamusca & Postma 1997, 12-13). Als we kijken naar parallelle tekstcollecties in de Franse literatuur, kan het eerste compilatie-deel heel goed vertalingen hebben bevat van de Oudfranse *Estoire dou Graal*, *Merlin en prose* en *Suite-Vulgate du Merlin*. Deze drie teksten werden in Oudfranse handschriften met de zogenaamde 'Vulgaat-cyclus' geregeld gebruikt om de *Lancelot en prose* van een voorgeschiedenis te voorzien. De voormelde Burgsteinfurt-codex is natuurlijk erg interessant in deze context: zouden de romans in deze *Merlijn*-codex deel hebben uitgemaakt van de *prequel* voor de teksten in het Haagse handschrift?

Dat is bijzonder aannemelijk. Ook in deze *Merlijn*-teksten vinden we regelmatig vooruitwijzingen naar Arturiaanse stof die 'later' zal

13 Gerritsen (1976) heeft beargumenteerd dat de correcties werden aangebracht met het oog op een mondelinge voordracht: de corrector had bijvoorbeeld veel aandacht voor sprekerswisselingen, zodat het lijkt dat hij de intentie had om uit het handschrift voor te lezen. Biemans (2005) beargumenteerde bovendien dat de correcties blijk geven van een 'tendens tot prozaïsering': door het verbinden van verzen met o.m. voegwoorden kreeg de compilatie meer weg van een proza- dan een verstekst. Dat is eveneens een interessante interpretatie, zeker omdat de editeurs van de recent opgedoken fragmenten van de proza-Lancelot zich hebben afgevraagd of deze medio-veertiende eeuwse tekst niet werd (af?)geschreven in de omgeving van Aarschot/Wezemaal (Minnen & Claassens 2005, 170), op een boogscheut van Velthems parochie. De compilatie-correcties en (het afschrift van) de proza-Lancelot zouden daarom in nagenoeg dezelfde periode door streekgenoten kunnen zijn gemaakt. Tot op heden blijft mij trouwens onduidelijk waarom de hypotheses van Gerritsen en Biemans elkaar zouden moeten uitsluiten: men kan proza immers even goed hebben voorgedragen.

behandeld worden maar in de Burgsteinfurtse codex uiteindelijk nooit ter sprake komt. Hoewel enkele van deze vooruitwijzingen gewoon vertaald zijn uit het Oudfrans, zien we dat andere naar alle waarschijnlijkheid werden toegevoegd in het Middelnederlands (Besamusca et al. 2009b, 24-25). Dit bewijst dat deze Middelnederlandse *Merlijn*-compilatie blijkbaar van in het begin bestemd was om in een groter kader te functioneren. Als de compilator de ontbrekende 57.500 verzen van de *Lanceloet* zou hebben gecombineerd met de Burgsteinfurt-teksten (ca. 36.000) zou dat een handschrift hebben opgeleverd met een grootte (≈ 93.500 verzen, ca. 260 folia) die erg vergelijkbaar is met die van het overgeleverde compilatie-deel (≈ 87.000 verzen, ca. 241 folia).[14]

Hét belangrijkste argument is echter te vinden op het laatste beschreven folium van de compilatie (Fig. 57, p. 216). Dit bevat namelijk een colofon – in het handschrift van kopiist B – dat als volgt leest: *Hier indet boec van Lancelote dat heren Lodewijcs es van Velthem*. Blijkbaar is de naam van onze Brabantse dorpspriester ook met dit handschrift op de een of andere manier verbonden. De genitief *Lodewijcs* laat zich echter moeilijk doorgronden. Is het colofon een bezittersaantekening en wordt hier gewoon duidelijk gemaakt dat Velthem op een gegeven ogenblik de bezitter was van de *Lancelotcompilatie*? Of mogen we *Lodewijcs* als een onderwerpsgenitief interpreteren en was Velthem de geheimzinnige compilator die het handschrift samenstelde?

Deze vraag houdt de medioneerlandistiek al geruime tijd bezig. Om haar te kunnen beantwoorden, moet men rekening houden met de uiterst complexe 'rolverdeling' die aan de *Lancelotcompilatie* ten grondslag ligt. Als we abstractie maken van de vier minder belangrijke kopiisten (A, C, D en E) in het handschrift, kunnen we zes abstracte rollen onderscheiden in de totstandkoming van de *Lancelotcompilatie*. Als de hand van de corrector niet van kopiist B kan zijn zoals paleografen claimen, hebben we met tenminste twee verschillende personen te maken. Het is bovendien aannemelijk dat tenminste één van deze zes rollen aan Lodewijk van Velthem moet worden toebedeeld.

ROL 1 – de kopiist: diegene die de echte kopieerarbeid verrichte (in hoofdzaak kopiist B);

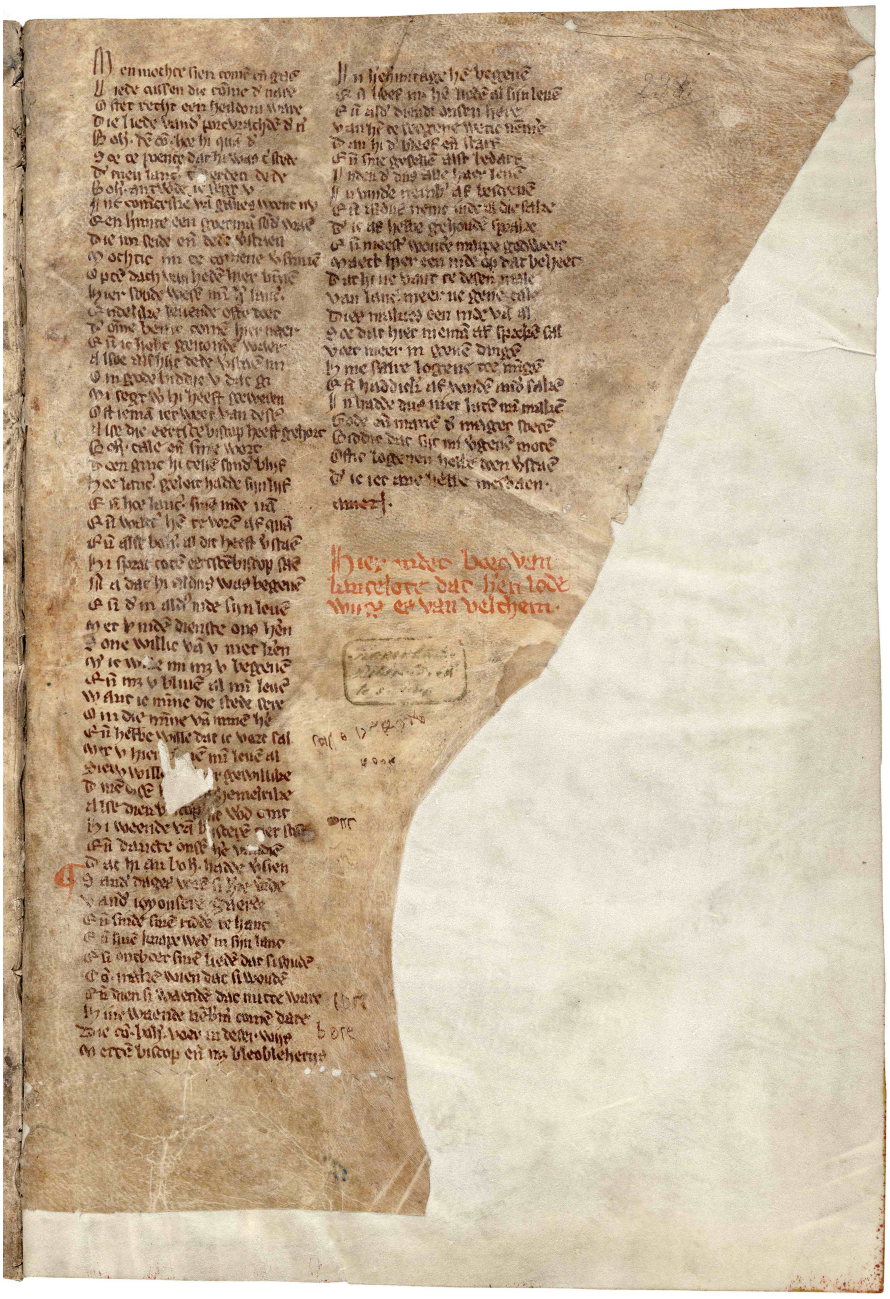
ROL 2 – de compilator: diegene die de bouwstenen van de compilatie bewerkte en aan elkaar verbond via overgangspassages;

ROL 3 – de corrector: diegene die de compilatie in een vroeg stadium van correcties voorzag;

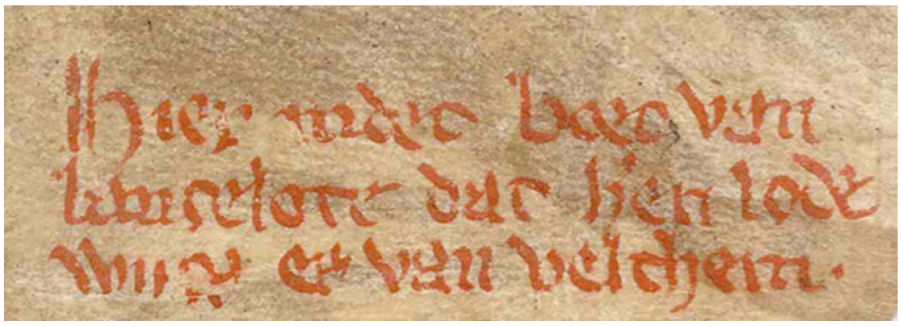
ROL 4 – de architect: diegene die de algemene opzet en samenstelling van de compilatie heeft bedacht;

ROL 5 – de opdrachtgever: diegene in wiens opdracht 129 A 10 oorspronkelijk werd vervaardigd.

14 Cijfers en berekeningen naar Besamusca & Postma (1997, 13).



Figuur 57: Het laatste beschreven folium uit de Haagse *Lancelotcompilatie* (Den Haag, Koninklijke Bibliotheek, 129 a 10, Brabant, ca. 1325); zie ook Fig. 58, p. 217.



Figuur 58: Het mysterieuze colofon van de Haagse *Lancelotcompilatie* (DEN HAAG, KONINKLIJKE BIBLIOTHEEK, 129 A 10), waarin duidelijk wordt dat Lodewijk van Velthem iets met de compilatie heeft te maken (zie ook Fig. 57, p. 216). In een diplomatische weergave leest het colofon: 'Hier indet boec van lancelote, dat h*er*en lodewijcs es van velthem'.

ROL 6 – de bezitter: de eerste bezitter van 129 A 10 en dus vermoedelijk de oorspronkelijke bestemmeling.

Laat ons onderaan deze rolverdeling beginnen, bij de rol van de bezitter. De meeste onderzoekers gaan ervan uit dat het colofon achterin de codex tenminste een bezittersaantekening was en dat hierin dus de geïntendeerde eerste bezitter van het handschrift kenbaar wordt gemaakt. Er bestaat eigenlijk geen gegronde reden om die aanname in twijfel te trekken.[15] Ook andere gegevens maken immers aannemelijk dat 129 A 10 in de onmiddellijke nabijheid van Velthem heeft gefunctioneerd (Biemans 2009). Dat Velthem de eerste bezitter van het handschrift was, suggereert bovendien dat hij de opdrachtgever was, want er zijn geen aanwijzingen (bv. vermelding van een gift of een dedicatie) dat iemand anders de codex voor Velthem liet vervaardigen. De meeste onderzoekers gaan er daarom van uit dat Lodewijk van Velthem zowel de eerste bezitter als opdrachtgever was van 129 A 10. Het is daarenboven plausibel dat de rol van architect (rol 3) door de opdrachtgever werd waargenomen: alles wijst erop dat Velthem zelf het initiatief voor de *Lancelotcompilatie* heeft genomen en het lijkt een terechte aanname dat het oorspronkelijke idee voor de compilatie van Velthem kwam. Het heeft er daarom alles van dat ook de samenstelling van de compilatie op Velthem teruggaat.

Tot op dit punt schijnt er min of meer een consensus te bestaan in de vakliteratuur. Of wij Velthem mogen gelijkstellen met de corrector (rol 3) is echter al veel minder duidelijk. Dat de corrector reeds tijdens de vervaardiging van de compilatie uitgebreid feedback heeft gegeven, suggereert alleszins dat deze nauw betrokken was bij het Lancelot-

15 Bij een gebrek aan gegevens laat ik hierbij wel in het midden of er later op basis van 129 A 10 nog een dedicatie-exemplaar werd vervaardigd, dat Velthem bijvoorbeeld kan hebben aangeboden een adellijke patroon. Het is immers slechts 129 A 10 dat hier aan de orde is.

project. Het is daarom verleidelijk om in de hand van de corrector die van de bezitter/opdrachtgever/architect te herkennen, *casu quo* Velthem. De corrector was in dat geval de ongeruste opdrachtgever-architect die het project dat hij had uitbesteed, van nabij probeerde op te volgen en bij te sturen. Uit deze aanname zou volgen dat Lodewijk van Velthem en kopiist B niet dezelfde persoon waren, want de corrector en kopiist B hebben een ander handschrift.

De eerste rol in het schema hierboven is voorbehouden voor B, de kopiist die het gros van de compilatie heeft afgeschreven en bovendien leiding heeft gegeven aan de vier andere afschrijvers in de vroegste fase. Traditioneel wordt de rol van B onderscheiden van die van de compilator. Deze compilator heeft de oorspronkelijke teksten voor hun opname in de compilatie bewerkt en verscheidene overgangspassages tussen teksten aangebracht. Daarbij is het niet onbelangrijk dat men B vaak slechts als een reproducerende kopiist beschouwde, en in de compilator een meer creatieve dichter herkende. De figuur van de compilator is daarom wel in contact gebracht met de rol van de architect. Men heeft zich afgevraagd of de architect van de compilatie, eventueel Velthem zelf, de compilator kan zijn geweest. Binnen die zienswijze zou de architect de teksten dan op voorhand in het klad bewerkt hebben. Kopiist B zou de teksten dan in een kladversie hebben aangeleverd gekregen en ze in de netversie hebben overgeschreven (vgl. Klein 1998, 114-115). Lodewijk van Velthem zou dan álle bovenstaande rollen kunnen hebben vervuld, behalve die van kopiist. Toch mag men niet al te snel in deze voor de hand liggende oplossing berusten.

Steeds vaker gaan er namelijk stemmen op om de creatieve inbreng van kopiist B niet te onderschatten. Het initiatief voor de compilatie mag dan mogelijk niet door hem zijn genomen, hij zou wel de praktische uitvoering ervan nogal soeverein op zich hebben genomen (vgl. Klein 1998, 114-115). Voor heel wat ingrepen is immers via codicologische weg betoogd dat kopiist B ze zelfstandig heeft bedacht en uitgevoerd, mogelijk zelfs zonder overleg met een opdrachtgever. De identificatie van Velthem als de Lancelot-compilator is daarom geen vaststaand gegeven. Sommige onderzoekers hebben zich trouwens afgevraagd of men B niet mag gelijkstellen met de compilator en of de hand van kopiist B niet gewoon die van Velthem kan zijn geweest. Binnen die zienswijze zou Velthem alle bovenstaande rollen kunnen hebben vervuld, behalve die van corrector. Toch heeft de identificatie van Velthem als kopiist B amper gehoor gevonden: het voornaamste probleem van deze identificatie is dat dan onduidelijk wordt wie de rol van corrector kan hebben vervuld.

Concluderend: Lodewijk van Velthem zou de eerste bezitter van het Haagse handschrift 129 A 10 zijn geweest. Het is aannemelijk dat het idee voor (de samenstelling van) deze *Lancelotcompilatie* van Velthem zelf kwam. Daarom heeft het er alles van dat Velthem zelf de opdracht heeft gegeven voor de vervaardiging van het handschrift. Mogelijk

was Velthem de Lancelot-compilator en heeft hij de oorspronkelijke teksten zelf bewerkt en de compilatie van overgangspassages voorzien. Verschillende onderzoekers hebben de identificatie van Velthem als de Lancelot-compilator verdedigd, ook op basis van stilistische argumenten (Besamusca 1991; Hogenbirk 2009b). Velthems teksten zouden dan in het net zijn gekopieerd door kopiist B. Mogelijk is Velthem zelfs identiek aan kopiist B maar deze hypothese heeft weinig bijval gevonden, omdat dan onduidelijk wordt wie de corrector was. Veel onderzoekers zien in Velthem daarom liever de corrector van de compilatie: als architect zou hij de vorderingsstaat van het project van nabij hebben gevolgd en bijgestuurd.

8.2.1 *Centroïdes*

De *Lancelotcompilatie* telt in ons corpus in totaal 87.230 rijmwoorden, waarop we de hierboven besproken centroïde-aanpak kunnen toepassen (§ 8.1). Vanzelfsprekend biedt het oeuvre van Lodewijk van Velthem het geijkte referentiepunt voor een dergelijke analyse. Verschillende onderzoekers hebben immers beargumenteerd dat wij Velthem mogen gelijkstellen met de mysterieuze compilator die de bouwstenen van de compilatie heeft bewerkt en gestroomlijnd. De meeste onderzoekers zijn er trouwens van overtuigd dat geen van de oorspronkelijke romans in de *Lancelotcompilatie* door de compilator werden geschreven. Verschillende teksten zijn namelijk overgeleverd in fragmenten die door codicologen vóór ca. 1325 worden gedateerd en de vervaardiging van 129 A 10 dus in de tijd zijn voorafgegaan. Dit bewijst dat deze bouwstenen reeds in omloop waren voor de compilator ze in handen kreeg. Bovendien wijst het dialect van de meeste grondteksten in de richting van Vlaamse auteurs, terwijl de taal van de Lancelot-kopiisten of Velthem eerder door een Brabantse tongval wordt gekenmerkt.

De centroïde-aanpak lijkt op het eerste zicht heel geschikt om te verkennen op welke delen de compilator het duidelijkst zijn stempel heeft gedrukt. Men zou bijvoorbeeld kunnen verwachten dat een overgangspassage die door de compilator tussen twee romans werd ingevoegd, zal aansluiten bij Velthems stijl. Vanzelfsprekend zou een overgangspassage daarvoor lang genoeg moeten zijn. Onderstaande figuren geven het resultaat weer van een aantal Delta-analyses waarbij we de *Lancelotcompilatie* eerst hebben gesegmenteerd met de parameters *sample size ss*=100 *window size ws*=2.000 – voor een goede granulariteit van de metingen. In de eerste figuur werkten we met een gewone Delta (voor *n*=100), waarbij Velthems *MerlV* als referentiemateriaal werd gekozen (Fig. 59, p. 221).[16] In de tweede figuur werkten we met de combinatie van Velthems *MerlV* en *P4P5*, maar gebruikten we het

16 Hetzelfde experiment werd uitgevoerd met het centroïde uit Velthems *P4P5* maar dat leverde geen wezenlijke verschillen op (idem dito voor de combinatie van *MerlV* en *P4P5*.

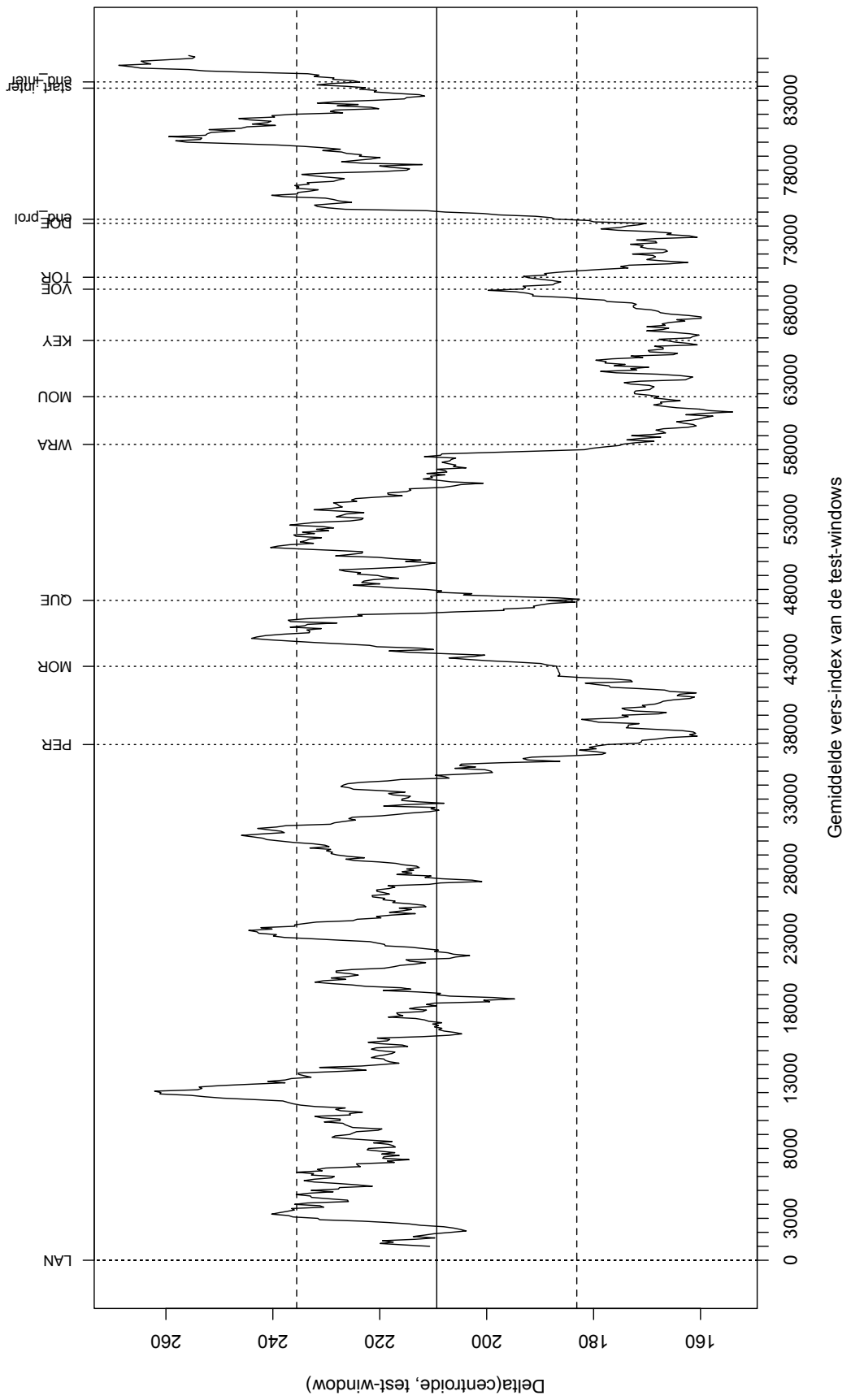
centroïde uit Maerlants *P3* als controle-materiaal voor een experiment met de Dubbele Delta (Fig. 60, p. 222). In deze figuren drukken de horizontale lijnen opnieuw de gemiddelde Delta-waardes uit (ook min en plus en de standaarddeviatie); de verticale lijnen drukken gebeurtenissen uit in de compilatie-tekst, zoals het begin van een nieuwe bouwsteen.

In deze figuren kunnen we een aantal interessante trends observeren, die we hieronder als hypotheses in het achterhoofd kunnen houden. De *windows* uit de kernteksten (LAN, QUE en DOE) vertonen een relatief stabiele Delta-waarde, meestal net boven de gemiddelde Delta-waarde. Dat deze teksten relatief verder verwijderd zijn van de stijl van Velthem, schijnt een eerste stylometrische bevestiging dat de compilator deze kernteksten inderdaad niet zelf heeft geschreven. In sommige gevallen is de overgang of 'naad' tussen twee teksten geassocieerd met plotse lokale daling in de Delta-waardes (bv. tussen MOR en QUE in de eerste figuur). De identificatie van Velthem als de compilator, zou verklaren waarom dergelijke secties 'reageren' op Velthems stijl.

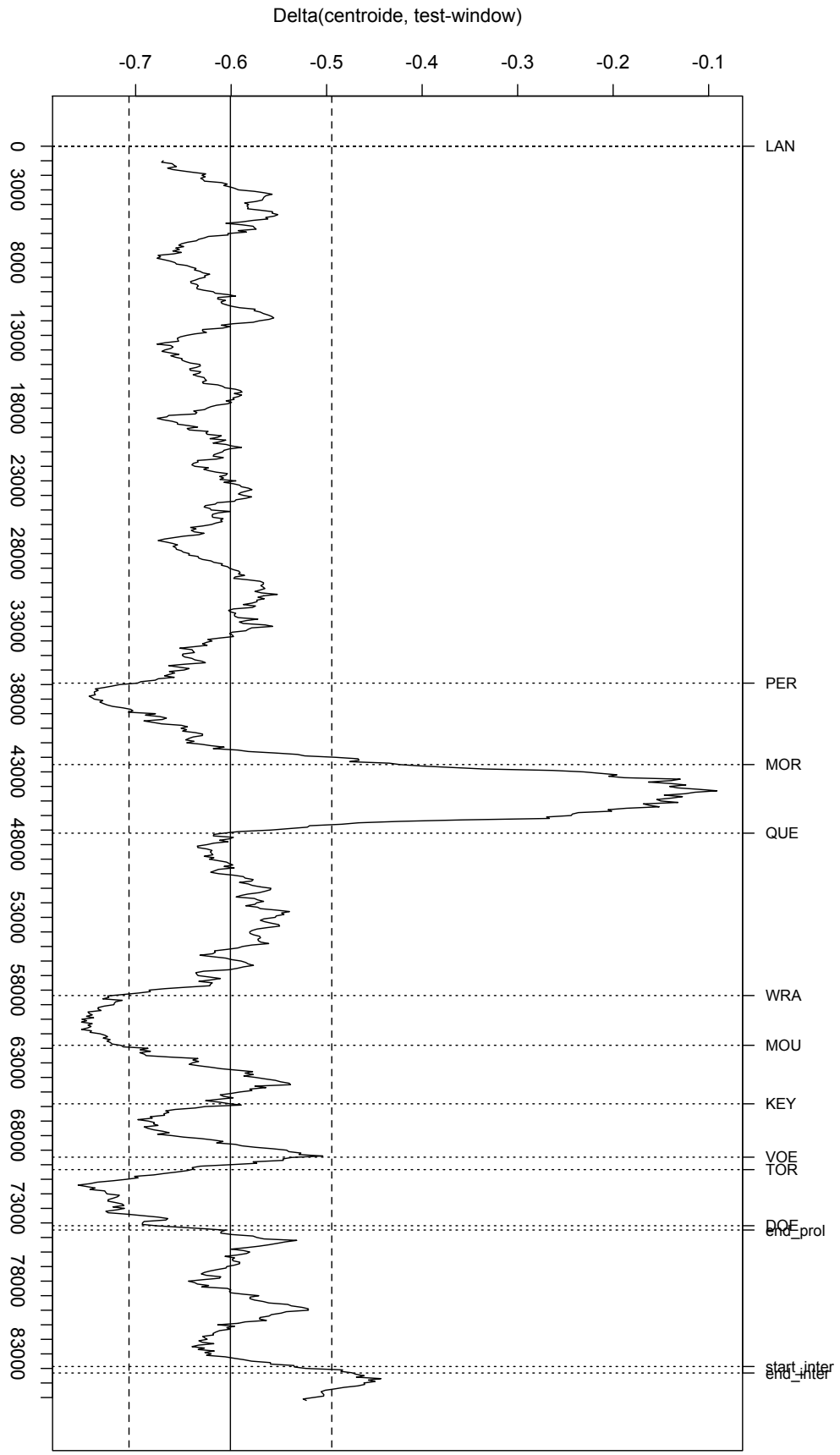
De resultaten voor de geïnterpoleerde romans zijn minder eenduidig. De *windows* van deze teksten zijn zeker in de eerste analyse geassocieerd met relatief lage Delta-waardes hetgeen een relatief grote stilistische gelijkenis met Velthems *MerlV* suggereert. Als Velthem inderdaad de compilator is, wijzen deze analyses uit dat de geïnterpoleerde romans heviger door de compilator zouden kunnen zijn bewerkt dan de kerndelen, een hypothese die ook in het traditionele onderzoek is verdedigd.[17] Zeker WRAK en TOR schijnen op het eerste zicht stevig bewerkt, alsook PER en KEY, zij het in iets mindere mate.

De geïnterpoleerde romans zijn natuurlijk korter, hetgeen voor deze analyses niet ideaal is, zeker als de compilator heeft ingegrepen aan het begin en einde van deze kortere teksten (vgl. de relatief grote *ws*=2.000). Toch rijst nu reeds de vraag in hoeverre de geïnterpoleerde romans nog een getrouwe afspiegeling vormen van de stijl van hun oorspronkelijke auteurs. Hoewel de interpolaties hierin ongetwijfeld onderling zullen verschillen, moeten wij nu reeds rekening houden met de mogelijkheid dat meeste romans duchtig door de compilator zijn bewerkt. Toch is er één duidelijke uitzondering: de Delta-waardes voor de *Moriaen* zijn in beide analyses opvallend hoog. Dat is vooral opvallend voor de steile piek die met MOR is geassocieerd in de tweede analyse (Fig. 60, p. 222). Kennelijk wijkt de stijl in deze tekst uitzonderlijk sterk af van Velthems stijl, want de Delta-waardes voor de andere teksten in de tweede analyse steken eerder flauw af tegen de piek van MOR. Dit is een teken aan de wand dat de compilator deze tekst mogelijk minder verregaand heeft bewerkt.

17 Zie Besamusca (2003, 30) of Oppenhuis de Jong (2003, 135-158). Vgl. echter de nuancering door Hogenbirk (2009b, 55).



Figuur 59: Resultaat van een Delta-experiment met Velthems *MerlV* als referentiemateriaal voor de *Lancelotcompilatie* (ws=2.000, ss=100, n=100). De horizontale as stelt de *Lancelotcompilatie* voor, verdeeld in een aantal opeenvolgende *windows*. Op de verticale as wordt de Delta-waarde van ieder *window* weergegeven, ten aanzien van het centroïde dat werd samengesteld op basis van Velthems referentietekst.



Figuur 60: Resultaat van een Dubbele Delta-experiment met Velthems *MerIV* en P_4P_5 als referentiemateriaal voor de *Lancelotcompilatie* (*ws*=2.000, *ss*=100, *n*=100). De horizontale as stelt de *Lancelotcompilatie* voor, verdeeld in een aantal opeenvolgende *windows*. Op de verticale as wordt de Dubbele Delta-waarde van ieder *window* weergegeven, ten aanzien van het centroïde dat werd samengesteld op basis van Velthems referentietekst. In deze grafiek werd het controle-centroïde gebruikt uit Maerlants P_3.

8.2.2 *De kernteksten*

Laat ons de kernteksten uit de compilatie eens van naderbij bekijken. Traditioneel gaat men ervan uit dat de compilator deze teksten niet zelf heeft gedicht, maar een beroep kon doen op reeds bestaande teksten (Besamusca 2003, 19-56). Deze hypothese lijkt zeker niet incongruent met de Delta-analyses hierboven. Omdat de kernteksten in veel opzichten een hechte drie-eenheid vormen, is het een interessante vraag of deze trilogie door slechts één auteur werd vertaald. Oorspronkelijk werd veel geloof gehecht aan een enkelvoudig auteurschap voor deze teksten. De teksten gaan immers terug op gelijkaardige Franse bronteksten en schijnen in dezelfde periode en regio te zijn vervaardigd (Vlaanderen, ca. 1280?).

Janssens was de eerste om deze aanname in vraag te stellen. Hij beschouwde de dichter van de *Lanceloet* als een virtuoos, maar vond dat de *Queeste*-dichter de diepere betekenis van zijn grondtekst niet begrepen had (bv. kleurensymboliek). De *Queeste*-vertaling zou volgens Janssens dan ook niet op de *Lanceloet*-dichter kunnen teruggaan. Deze visie werd vervolgens betwijfeld door Besamusca (2003, 33), omdat Janssens zich zou gebaseerd hebben op slechts een klein aantal, willekeurig gekozen samples. Om de rest van Besamusca's betoog te begrijpen, is het nodig om dieper in te gaan op een specifiek kenmerk van de kernteksten.

De oudste Franse Arturromans (bv. *Chevalier de la charette*) worden gekenmerkt door de aanwezigheid van een alwetende IK-verteller. Deze eerste-persoonsverteller werd in latere prozaromans (zoals de *Préparation à la Queste*) vervangen door een onpersoonlijke, abstracte verteller: *li contes* ('het verhaal'). Op die manier trachtten auteurs hun teksten een aura van objectiviteit te verlenen en werd de tussenkomst van een subjectieve, autoritaire IK vermeden (Besamusca & Postma 1997, 4, 9-10). *Li contes* treedt vaak op in 'formele overgangen' of passages die de sprong maken tussen verschillende verhaaldraden.

In Middelnederlands teksten zien we dat *davonture* – eveneens in de betekenis van 'het verhaal' – kon optreden als het equivalent van *li contes*. Toch is *davonture* niet de enige vertaling voor *li contes* die in Middelnederlandse bewerkingen kon worden gebruikt. In ongeveer de helft van de gevallen zien we dat *li contes* wordt vervangen door de oudere eerste persoonsverteller, in de vorm van het enkelvoudig pronomen *ik* of het meervoudige *wi*. Deze merkwaardige herinvoering van de oorspronkelijke eerste persoonsverteller is een bizar kenmerk van veel Middelnederlandse Arturromans, vooral omdat deze vrijelijk kon worden gecombineerd met het modernere *li contes* uit de brontekst.

De kernteksten uit de *Lancelotcompilatie* worden alle drie gekenmerkt door de aanwezigheid van deze merkwaardige vertellerscombinatie (Besamusca & Brandsma 1994). Besamusca benadrukte terecht dat

dit kenmerk niet noodzakelijk al in de brontekst van de compilator aanwezig was. De overgeleverde fragmenten van andere romans uit de compilatie tonen namelijk aan dat de compilator het kenmerk in veel teksten heeft ingevoegd (Besamusca 2003, 44-47). Daarom bleef het lang onduidelijk wie verantwoordelijk was voor de merkwaardige vertellerscombinatie in de kernteksten: de compilator of de oorspronkelijke auteur?

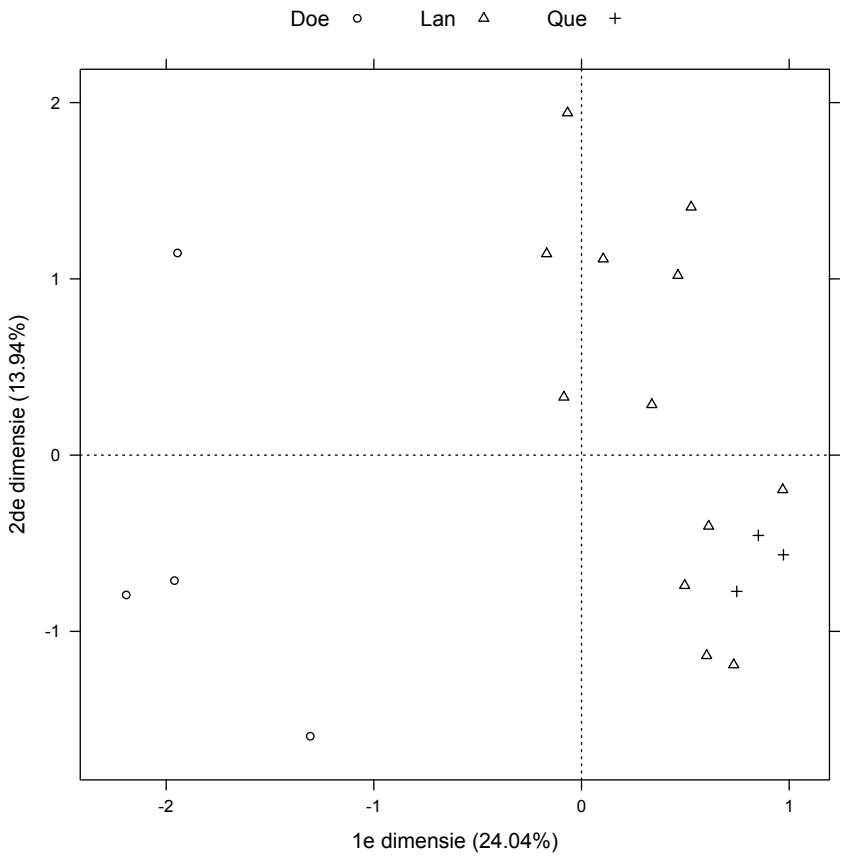
Dit veranderde echter in 1992, toen een minuscuul fragmentje werd ontdekt van de oorspronkelijke tekst van *Arturs doet*, tezamen nog geen 160 zwaar verminkte regels (Croenen & Janssens 1994a, b). Toch was duidelijk dat in dit fragment tot tweemaal toe een eerste persoonsverteller op de voorgrond treedt. Zo werd het onmiskenbaar dat de merkwaardige combinatie van een onpersoonlijke en een persoonlijke verteller reeds voorkwam in de grondtekst van de compilator. Besamusca (2003, 47) betoogde in zijn recente monografie dat deze vertellerscombinatie een uitvinding was van een Vlaamse dichter, want dit fragment leverde volgens hem het onomstotelijke bewijs ('conclusive evidence') dat de drie kernteksten door dezelfde auteur werden vertaald.

Besamusca's conclusie schijnt niettemin overhaast. Ten eerste zijn er er nog steeds geen fragmenten opgedoken die kunnen bewijzen dat de vertellerscombinatie ook optrad in de grondteksten van *Lanceloet* en de *Queeste*. Besamusca (2003, 47) vond het te ver gezocht om te veronderstellen dat de eerste twee kernteksten werden geschreven door dichters die de technieken van de vertaler *Arturs doet* hebben gekopieerd. Die aanname lijkt echter incongruent met de observatie dat de compilator het kenmerk wél uit andere teksten heeft overgenomen en zelf heeft ingevoegd in de compilatie-teksten. Toch is het inderdaad zeer aannemelijk dat de vertellerscombinatie ook in de grondtekst van de andere twee kerndelen voorkwam.

De vraag is echter of dit kenmerk alleen een onomstotelijk bewijs levert voor een enkelvoudig auteurschap van de kerndelen. De verteller van een tekst – zelfs een vrij unieke vertellersinstantie, zoals in de kerndelen – is natuurlijk een erg algemeen tekstkenmerk. Mogelijk is dit kenmerk zelfs zo algemeen dat het niet erg bewijskrachtig is voor het enkelvoudig auteurschap van teksten. Opnieuw kunnen de compilator-ingrepen immers illustreren dat een vertellersinstantie makkelijk kon worden geïmiteerd door andere auteurs. De combinatie was immers ongewoon en kan door latere imitators makkelijk zijn opgepikt.

Daarom is het nu interessant om een stylometrische analyse uit te voeren van de kerndelen, zonder daarbij uit te gaan van de voormelde hypotheses over het auteurschap van teksten. Het grote voordeel is dat de kerndelen erg omvangrijk zijn, zodat we zelfs met een luxueuze sample-grootte van 3.000 rijmwoorden (=*ws*=*ss*), een groot aantal samples per tekst verkrijgen. In onderstaande figuur is het

strooidiagram weergegeven waarin we deze samples uit LAN, QUE en DOE via een correspondentie-analyse hebben geanalyseerd (Fig. 61, p. 225). Daarvoor werden de 100 rijmwoorden gebruikt die cumulatief het meest frequent waren in deze teksten. In de eerste, horizontale dimensie wordt duidelijk een oppositie gerealiseerd tussen enerzijds de samples uit DOE uiterst links, en anderzijds de samples uit LAN en QUE die vermengd worden aan de rechterkant van het diagram.



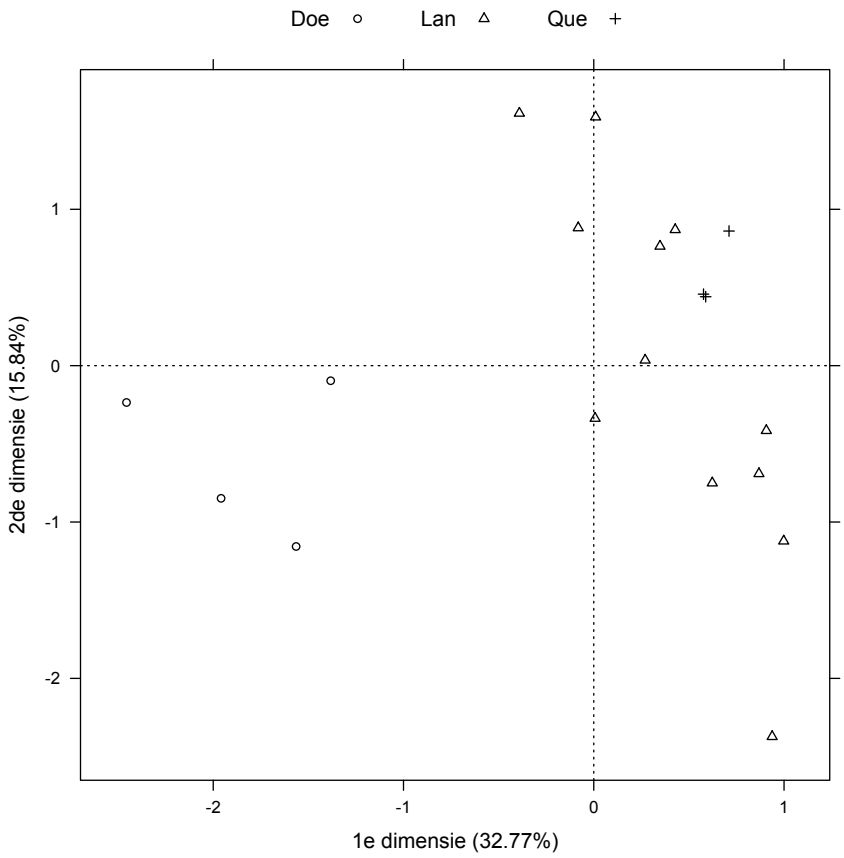
Figuur 61: Correspondentie-analyse (eerste twee dimensies) van samples van 3.000 rijmwoorden uit de kerndelen van de compilatie: LAN, QUE en DOE. De analyse is gebaseerd op de relatieve frequenties van de 100 meest frequente rijmwoorden in deze samples.

We mogen echter niet voortvarend zijn in onze interpretatie van dit resultaat: een nadere blik op de ladingen van de rijmwoorden in deze analyse toonde aan dat deze oppositie ten dele werd veroorzaakt door een aantal rijmen die met de inhoud van deze teksten samenhangen. Zo bleek het gebruik van de rijmwoorden KONING, KONINGIN en TROUW kenmerkend voor DOE. Dat is verontrustend: in *Arturs doet* treedt de vorst Artur immers veel sterker op de voorgrond dan in andere teksten, waar hij (net als Guinevere) veelal slechts een rol op de achtergrond van het verhaal speelt. De oppositie tussen de eerste twee

kernteksten en de derde kerntekst zou dus meer op de inhoud dan op het auteurschap van de kerndelen kunnen teruggaan.

Toch blijft de stilistische oppositie overeind als men ook een kleiner aantal rijmwoorden beschouwt. In onderstaande figuur wordt de analyse herhaald, maar ditmaal met slechts de 30 meest frequente rijmwoorden in deze teksten (Fig. 62, p. 226). De gebruikte rijmwoorden zijn (gerangschikt naar cumulatieve frequentie):

ZIJN, PRNAME, TOEN, DOEN, GIJ, ZIEN, DAARNAAR, DAAR, DEZE, NU, STOND, KOMEN, WEZEN, ZAAN, WAAN, TIJD, ZULLEN, HEER, STAD, DAT, ZIJ, IK, MEDE, MEDE, TOE, MAN, ZEER, AL, TEHAND, MOGEN



Figuur 62: Correspondentie-analyse (eerste twee dimensies) van samples van 3.000 rijmwoorden uit de kerndelen van de compilatie: LAN, QUE en DOE. De analyse is gebaseerd op de relatieve frequenties van de 30 meest frequente rijmwoorden in deze samples.

Het is duidelijk dat de spreiding van samples in de eerste, horizontale dimensie eigenlijk niet wezenlijk is gewijzigd ten opzichte van de vorige correspondentie-analyse, waar inhoudsgerelateerde factoren mogelijk een rol speelden. Experimenten met andere parameters konden dit globale beeld niet wijzigen. Als we een blik werpen op

bovenstaande lijst hoogfrequente rijmwoorden, lijkt het een te verdedigen aanname dat inhoudsgerelateerde rijmwoorden hier een veel minder sterke rol kunnen hebben gespeeld. Ook strooidiagrammen waarin de ladingen van de individuele rijmwoorden werden afgebeeld, brachten geen inhoudsgerelateerde afwijkingen aan het licht.

Het resultaat van deze analyse is dus verrassend: een statistische vergelijking van de hoogfrequente rijmwoordenschat in de kerndelen van de *Lancelotcompilatie* legt een merkwaardige oppositie bloot tussen, enerzijds de *Lanceloet* en de *Queeste* en anderzijds het laatste kerndeel, *Arturs doet*. Hoewel het nooit volledig valt uit te sluiten dat de afwijkende inhoud van *Arturs doet* in deze tegenstelling meespeelt, blijft het effect van verwijdering ook bij zeer kleine *n*-waardes overeind. Zou de compilator hier voor iets tussen zitten? Zou hij bijvoorbeeld sterker in *Arturs doet* dan in de andere kernteksten hebben ingegrepen? Ook dat is bij gebrek aan de authentieke bronteksten nooit volledig uit te sluiten.

Bekend is wél dat de compilator twee teksten zou hebben kunnen ingevoegd. Kapittel 19 van *Arturs doet*, bijvoorbeeld, handelt over Arturs oorlog tegen de Romeinen. Deze passage wijkt duidelijk af van de Franse brontekst en schijnt woordelijk vervangen te zijn door Maerlants behandeling van dezelfde stof in de *Spiegel historiael* (Besamusca 2003, 51-54). Natuurlijk kan Maerlants werk bekend zijn geweest bij de oorspronkelijke Vlaamse dichter van *Arturs doet*: daarom kan deze zelf de afwijkende passage hebben ingevoegd. Gezien Velthems vertrouwdheid met de *Spiegel* en zijn vermoedelijke compilatorschap, is het echter minstens zo aannemelijk dat Velthem voor de Romeinen-inlassing verantwoordelijk is geweest.[18] *Arturs doet* wordt trouwens voorafgegaan door een korte, traktaat-achtige proloog die we niet terugvinden in de Oudfranse tekst. Deze is geïnspireerd op het Latijnse *De modo orandi* van Hugo van Sint-Victor en behandelt de praktijk van het bidden. Hoewel de compilator ook hier niet volledig buiten verdenking valt, denken onderzoekers dat deze afwijkende proloog veeleer teruggaat op de oorspronkelijke auteur (Besamusca & Lie 1994).[19]

Zouden deze passages ten dele verantwoordelijk kunnen zijn voor de afwijkende stilistische rijmwoordenschat in *Arturs doet*? Neen, want zij werden volledig verwijderd uit *Arturs doet* in het corpus vooraleer ik bovenstaande analyses heb uitgevoerd. Mogelijk zijn deze passages wel illustratief voor het afwijkende karakter van DOE ten opzichte van de andere twee kernteksten. Het is trouwens interessant dat DOE in de

18 Hoewel de passage erg kort is (446 rijmwoorden), kon een CA de afwijkende stijl ervan oppikken (vgl. Brinkman 2010). Ik ga hier niet verder op in, omdat het onduidelijk is in hoeverre een genre-verschil hiervoor verantwoordelijk is.

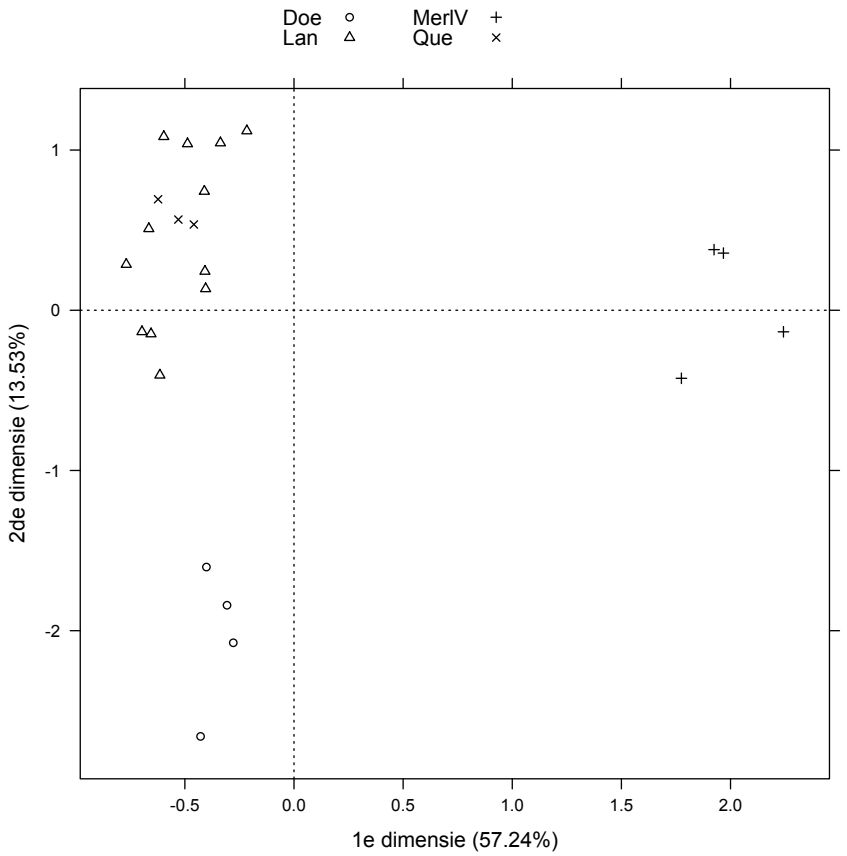
19 Warnar (2009) bracht hier wel tegen in dat deze gebedsproloog overeenkomsten vertoont met Velthems behandeling van Hugo van Sint-Victor in de Vierde Partie. Vanwege de extreem korte lengte (296 rijmwoorden) heb ik geen verdere analyse van deze passage uitgevoerd.

voorgaande Delta-experimenten consequent een iets hogere waarde uitlokte dan de andere twee teksten – vgl. Fig. 59 (p. 221) en Fig. 60 (p. 222). Ook uit verschillende analyses hieronder zal nog blijken dat de stijl van de drie kernteksten vanzelfsprekend dicht bij elkaar aanleunt, maar dat de rijmstijl in DOE consequent nét iets anders is dan die van de andere twee kerndelen. Mogelijk moeten we hier wel rekening houden met een diachroon verschil: staan we toch voor één dichter maar vertaalde die het laatste kerndeel gewoon veel later in zijn loopbaan?

In onderstaande analyse worden de drie kernteksten geconfronteerd met Velthems *MerlV* (Fig. 63, p. 229). Geen van de drie kernteksten vertoont een toenadering tot Velthems *Merlijn*-roman: indien wij Velthem mogen gelijkstellen met de Lancelot-compilator, toont deze analyse dat de romans alleszins niet significant verschillen inzake de mate van compilator-inmenging. In de tweede, verticale dimensie wordt DOE nog steeds netjes onderscheiden van de andere twee kernteksten. Het gebruik van een controle-auteur kan dus (ook in andere, niet-gerapporteerde experimenten) het afwijkende gedrag van het laatste kerndeel niet uitvagen. Een diachroon verschil is moeilijk uit te sluiten, maar wordt door deze analyse alleszins niet onderschreven.

Bovenstaande analyses bieden in isolatie vanzelfsprekend onvoldoende bewijs om het identieke auteurschap van de kernteksten definitief te verwerpen. Wel voeden zij de twijfel of een enkelvoudig auteurschap van de kerndelen zo vanzelfsprekend is als vroeger werd aangenomen.[20] Een analyse van de hoogfrequente rijmwoordenschat in deze teksten bracht immers een markante tegenstelling aan het licht tussen de eerste twee kerndelen en het laatste kerndeel. Hoewel andere factoren (zoals diachronie of inhoud) nooit volledig buiten spel zijn te zetten, kan men uit dit resultaat een argument afleiden voor een meervoudig auteurschap van de Lancelot-kerndelen. Hoewel slechts doorgedreven onderzoek over deze kwestie duidelijkheid zal kunnen brengen, illustreert het resultaat wel duidelijk de meerwaarde van de stylometrische benadering. De resultaten van deze analyses hoeven niet noodzakelijk als een bewijs te fungeren maar zij blijken wel waardevol om erg vertrouwde teksten met een nieuwe bril te benaderen. Onderzoekers kunnen bovenstaande stylometrische resultaten in de toekomst immers als een referentiepunt gebruiken bij hun lectuur van deze teksten.

20 Indien wij inderdaad met twee verschillende auteurs hebben te maken, is het een interessante vraag in hoeverre zij elkaar hebben gekend en vertaaltechnieken hebben uitgewisseld. We kunnen ons slechts afvragen of hun samenwerking in de context van een soort van dichtersatelier was georganiseerd, zoals dat bij Maerlant en Utenbroeke mogelijk het geval was (§ 4.1.1).



Figuur 63: Correspondentie-analyse (eerste twee dimensies) van samples van 3.000 rijmwoorden uit de kerndelen van de compilatie (LAN, QUE en DOE) en Velthems *MerlV*. De analyse is gebaseerd op de relatieve frequenties van de 30 meest frequente rijmwoorden in deze samples.

8.2.3 *De interpolaties*

De kernteksten zijn door hun grote omvang natuurlijk bijzonder aanlokkelijk voor een stylometrische analyse. Toch moet het niet onmogelijk zijn om ook de geïnterpoleerde romans aan een stylometrisch onderzoek te onderwerpen. Daarvoor is het nodig om deze interpolaties en hun onderlinge verschillen wat grondiger toe te lichten (een overzicht bij Hogenbirk 2004, 128-132). Ook voor deze teksten blijkt het steevast moeilijk om eventuele compilator-ingrepen af te bakenen. De schaarse fragmentjes die van de 'oerversies' van deze teksten zijn opgedoken, geven soms wel inzicht in de bekortingen van de compilator maar geven lang niet altijd uitsluitsel over mogelijke toevoegingen.

De eerste roman die reeds in een bijzonder vroeg stadium in de compilatie werd geïnterpoleerd is de *Perchevael* (Oppenhuis de Jong 2003), een Middelnederlandse bewerking van de Oudfranse *Perceval*

(ca. 1190). De grondtekst van de compilator schijnt afhankelijk te zijn geweest van *Première continuation* van de tekst (ca. 1215), hetgeen een terminus post quem biedt voor de Middelnederlandse bewerking (Caers & Kestemont 2011). De archaïsche spelling van het oudste bekende *Perchevael*-fragment (laatste kwart 13e eeuw) doet niettemin vermoeden dat de *Perchevael* al snel na de eerste voortzetting moet zijn gedicht (ca. 1225?). Hoewel de Oudfranse *Perceval* en *Première continuation* tezamen zowat twee derden dekken van de inhoud van de compilatie-*Perchevael*, is de origine van tenminste 40% van de compilatie-versie onduidelijk. Zeker één derde van de tekst werd in een recent onderzoek toegeschreven aan de Lancelot-compilator, die een groot aantal passages moet hebben toegevoegd die alle het personage van Walewein, de bekende neef van Koning Artur, naar de voorgrond halen.[21] De compilator schijnt zijn inspiratie voor deze toevoegingen in diverse bronnen te hebben gehaald om zo de lacunes op te vullen die hij in zijn brontekst zag.

De volgende interpolatie is de *Roman van Moriaen*, opvallend genoeg de eerste roman uit ons taalgebied met een zwarte titelheld. Ook van de oerversie van de *Moriaen* zijn enkele fragmenten opgedoken die het toelaten om de compilatie-versie iets beter te plaatsen.[22] Tot op heden heeft men geen Franse brontekst voor de *Moriaen* kunnen aanduiden en daarom gaat men ervan uit dat de *Moriaen* een oorspronkelijk Middelnederlands verhaal is. Daarin moet Moriaen zijn opgetreden als de biologische zoon van Perchevael, die Moriaen verwekte bij een 'Moorse' dame. De verhaallijn van de Moriaen staat volledig in het teken van Moriaens *Vatersuche*. In de compilatie-versie is Moriaens vader echter veranderd in Acglavael, Perchevaels broer (zie ook onder). Afgezien van deze markante wijziging (Besamusca 2003, 85-87), is men in het algemeen relatief optimistisch over de oorspronkelijkheid van de compilatie-*Moriaen*: de overgeleverde fragmenten van de oerversie doen uitschijnen dat de compilator deze teksten minder uitvoerig heeft bewerkt (Van Buuren 1973, 43).

Tussen de *Queeste* en *Arturs doet* werden nog vijf andere romans ingevoegd.[23] Deze hebben als gemeenschappelijk kenmerk dat zij de rol van Walewein in de verf zetten. Waarschijnlijk werden zij door de compilator ingevoegd – al dan niet op instigatie van de architect – om een tegenwicht te bieden voor de grote aandacht die de grondteksten aan de figuur van Lancelot besteedden.

---

21 Zie Oppenhuis de Jong (2003, bv. 9). Het is opgemerkt dat de titel *Perchevael* voor de compilatie-versie eigenlijk misleidend is, omdat er zoveel nadruk wordt gelegd op de figuur van Walewein. De overgang van *Lanceloet* naar *Perchevael* is jammer genoeg niet bewaard in 129 A 10. Zie voor de codicologische aspecten van de interpolatie de bijdrage van Klein bij Besamusca & Postma (1997, 51-90).

22 Een inleiding bij Paardekooper-van Buuren & Gysseling (1971). De tekst is recent in het modern Frans vertaald door Finet-van der Schaaf (2009).

23 Deze tweede reeks interpolaties werd recent heruitgegeven en vertaald in het modern Engels door Johnson & Claassens (2003).

De eerste interpolatie na de *Queeste* is de *Wrake van Ragisel*, een vroeg-dertiende-eeuwse (Brabantse?) bewerking op rijm van de Oudfranse *Vengeance Raguidel*. De Middelnederlandse bewerking valt op doordat deze – veel sterker dan de Franse brontekst – Walewein in een positief daglicht stelt, als een stereotiepe, hoofse minnaar. Een vergelijking met de fragmenten die van oorspronkelijke *Wrake* zijn overgeleverd, leert dat de compilator zijn grondtekst aanzienlijk heeft bekort. Daarbij moesten vooral overbodige passages die niet essentieel waren voor de voortgang van het verhaal het ontgelden. Verder schijnt de compilator een zogenaamde *entrelacement*-structuur aan de *Wrake* te hebben opgelegd. Verschillende verhaaldraden worden door de verteller – de combinatie van *ic* en *davonture* – in overgangspassages met andere verhaaldraden in de compilatie verbonden. De compilatie-*Wrake* schijnt daarom relatief duchtig bewerkt te zijn door de compilator (Besamusca 2003, 99-100).

De volgende interpolatie is de *Roman van de riddere metter mouwen* (De Haan & Jongen 1983). De grondtekst van de compilator was naar alle waarschijnlijkheid van Vlaamse origine en lijkt tot op zekere hoogte afhankelijk van de Oudfranse roman *Richard li biaus*, waarin eveneens een hoofdpersonage optreedt met een witte mouw. Toch schijnt de oorspronkelijke auteur ook bij veel andere teksten te rade te zijn gegaan. De *Riddere metter mouwen* schijnt overigens in een nauw intertekstueel verband te staan met de *Roman van Walewein* (zie onder) en de *Moriaen* (Besamusca 1993). Hoewel Smith (1989) heeft beargumenteerd dat de compilator zijn exemplaar drastisch moet hebben omgewerkt, betoogde Besamusca (2003, 114-115) dat een dergelijk drastische omwerking ongebruikelijk zou zijn voor de compilator. Volgens Besamusca is de stem van de compilator wel hier en daar merkbaar in bepaalde auctoriële interventies, maar zou de compilator zich toch vooral beperkt hebben tot een bekorting van zijn legger.

Walewein ende Keye is de volgende interpolatie (Hogenbirk 2011). Ook van deze roman is geen Franse brontekst bekend zodat we ook hier mogelijk met een oorspronkelijk Middelnederlands verhaal hebben te maken. Omdat van de eventuele brontekst van deze roman (nog?) geen fragmenten zijn opgedoken, is het niet absoluut zeker dat de compilator van een brontekst kon gebruik maken, hoewel dat bijzonder aannemelijk is, gegeven de origine van de andere compilatieteksten (Hogenbirk 1994). Het relatief klassieke verhaal gaat over een serie conflicten tussen de 'ideale schoonzoon' Walewein en zijn prototypische opponent, de verraderlijke seneschalk Keye. Van de oerversie van *Walewein ende Keye* zijn geen fragmenten bekend, maar ook hier kunnen verdachte passages worden aangewezen die hoogstwaarschijnlijk op de compilator teruggaan (Hogenbirk 2011, 46-49).

De volgende interpolatie is de relatief korte versnovelle *Lanceloet en het hert met de witte voet*. De tekst vertoont een merkwaardige over-

eenkomst met de Oudfranse *Lay de Tyolet*, alleen schijnt de naam van de protagonist veranderd van Tyolet in Lancelot en schijnt de figuur van Walewein – opnieuw – meer op de voorgrond te treden (Zemel & Besamusca 1999, 181-210). Van een eventuele grondtekst is nog geen spoor opgedoken, zodat het auteurschap van de compilator niet met absolute zekerheid is te verwerpen.

De laatste interpolatie voor *Arturs doet* is *Torec*. Zoals gezegd claimt Jacob van Maerlant in zijn *Historie van Troyen* dat hij een *Toerecke* heeft geschreven en daarom gaat men er meestal van uit dat de compilator zich van een grondtekst van de hand van Jacob van Maerlant kon bedienen. Deze grondtekst vormde naar alle waarschijnlijkheid een bewerking van de (niet-overgeleverde) Franse Arturroman *Torrez, le chevalier au cercle d'or* (Van Oostrom 1979). Onderzoekers hebben zich echter zeer pessimistisch uitgelaten over de oorspronkelijkheid van de *Torec* (Claassens 2009). Maerlants oorspronkelijke tekst is dan wel niet overgeleverd, maar alles wijst erop dat de compilator de *Toerecke* verregaand naar zich heeft toegetrokken.

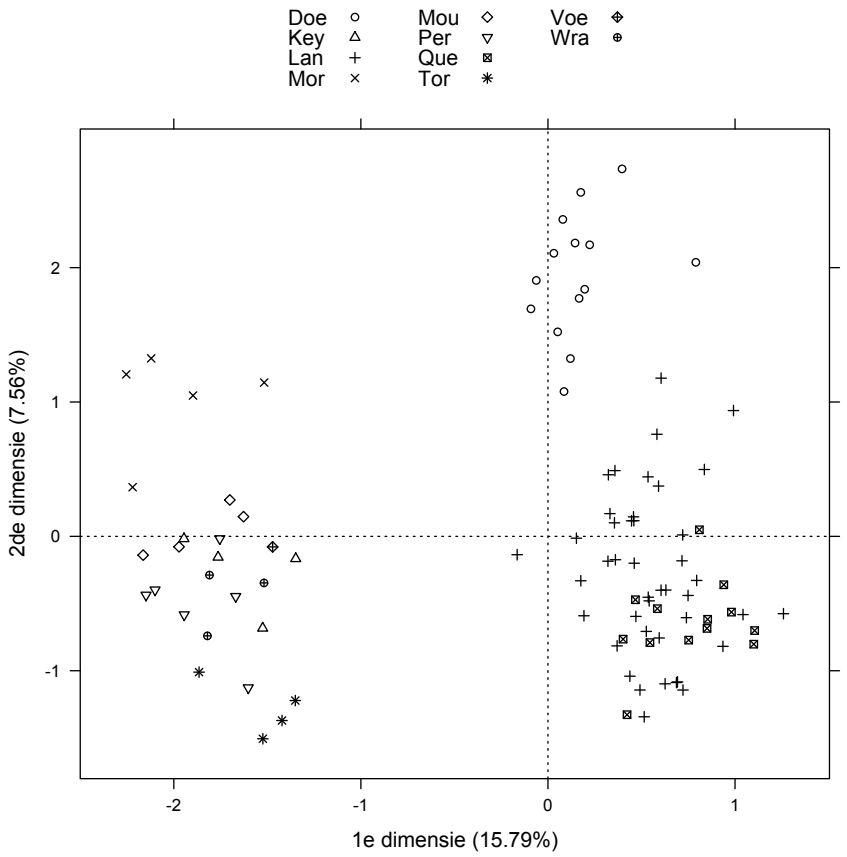
Men kan bovenstaand overzicht van de interpolaties en het onderzoek hiernaar al eens confronteren met de voorgaande Delta-analyses van de compilatie als geheel – Fig. 59 (p. 221) en Fig. 60 (p. 221). Daarbij valt op hoe nauw de waargenomen stijlschommelingen aansluiten bij het traditionele onderzoek naar de compilator-ingrepen – en zo onrechtstreeks Velthems compilatorschap kracht bijzetten. De lage Delta's voor *Perchevael* onderstrepen bijvoorbeeld de 'corrupte' aard van deze interpolatie. Dat maar liefst 40% van deze interpolatie op de compilator zou teruggaan laat zich duidelijk aflezen uit grote gelijkenis van *windows* uit de compilatie-*Perchevael* met Velthems werk. Vergelijkbaar zijn de lage Delta-scores voor *windows* uit de *Wrake van Ragisel* en de *Torec*: ook volgens deze stylometrische analyses mag men de oorspronkelijkheid van de rijmwoordenschat in deze interpolaties niet onderschatten.

Andere teksten lijken volgens deze analyses dan weer minder te hebben geleden onder de activiteiten van de compilator. De relatief hoge scores voor de *Riddere metter mouwen*, bijvoorbeeld, suggereren dat deze roman beter intact is gebleven. Hetzelfde lijkt op het eerste zicht waar voor de versnovelle over het hert met de witte voet, die ondanks haar bescheiden lengte nog een korstondig 'piekje' in de Delta-waarden weet uit te lokken. Verbluffend is niettemin het resultaat voor de *Moriaen*, die het verst van alle interpolaties verwijderd blijkt van de stijl van Velthem (Fig. 60, p. 222). Ook dit resultaat sluit niettemin aan bij het traditionele onderzoek waar is betoogd dat *Moriaen* niet al te verregaand zou zijn bewerkt.[24]

Laat ons eens een correspondentie-analyse uitvoeren van alle compilatieteksten, met kleine samples van 856 rijmwoorden, zodat we de korte versnovelle over het hert met de witte voet nog in een enkel sam-

24 Zie bijvoorbeeld Besamusca (2003, 114-115) of Hogenbirk (2009b, 64).

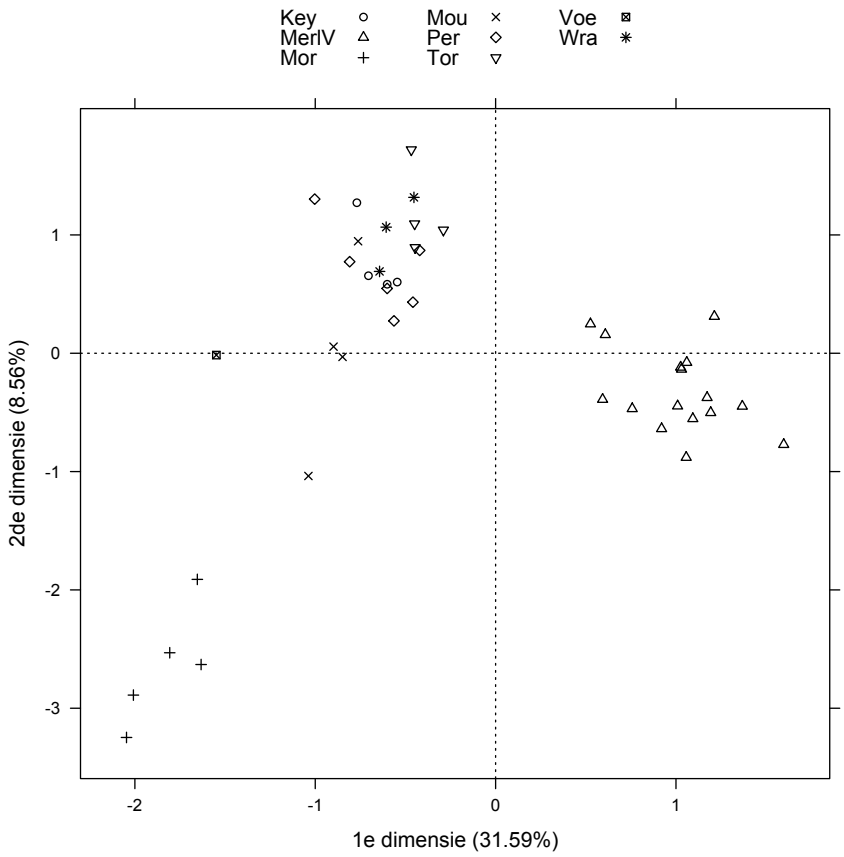
ple kunnen meenemen. Onderstaande analyse (voor n=100) bevestigt overduidelijk de relevantie van het traditionele onderscheid in de literatuur tussen de kerndelen en de interpolaties (Fig. 64, p. 233). Reeds in de eerste dimensie wordt een duidelijke links-rechts tegenstelling in het leven geroepen tussen beide. (Merk trouwens op hoe *Arturs doet* in de tweede, verticale dimensie van de eerste twee kerndelen opnieuw wordt onderscheiden.) De interpolaties vormen links een hechte stippenwolk, waaruit slechts *Moriaen* zich tot op zekere hoogte schijnt te kunnen losmaken. Een andere suggestie die we voorzichtig aan dit strooidiagram kunnen ontlenen, is dat geen van de auteurs van de kerndelen ook één van de geïnterpoleerde romans heeft geschreven, hetgeen een theoretische mogelijkheid was.



Figuur 64: Correspondentie-analyse (eerste twee dimensies) van samples van 856 rijmwoorden van alle compilatie-teksten. De analyse is gebaseerd op de relatieve frequenties van de 100 meest frequente rijmwoorden in deze samples.

Als we deze analyse herhalen zonder de kerndelen, maar Velthems *MerlV* toevoegen, verkrijgen we het resultaat dat hieronder is afgebeeld (Fig. 65, p. 234). Opnieuw worden onze initiële hypotheses bevestigd. Slechts de *Moriaen* weet zich los te maken van het interpolatiekluwen – dat Velthems *MerlV* opnieuw iets dichter schijnt te benade-

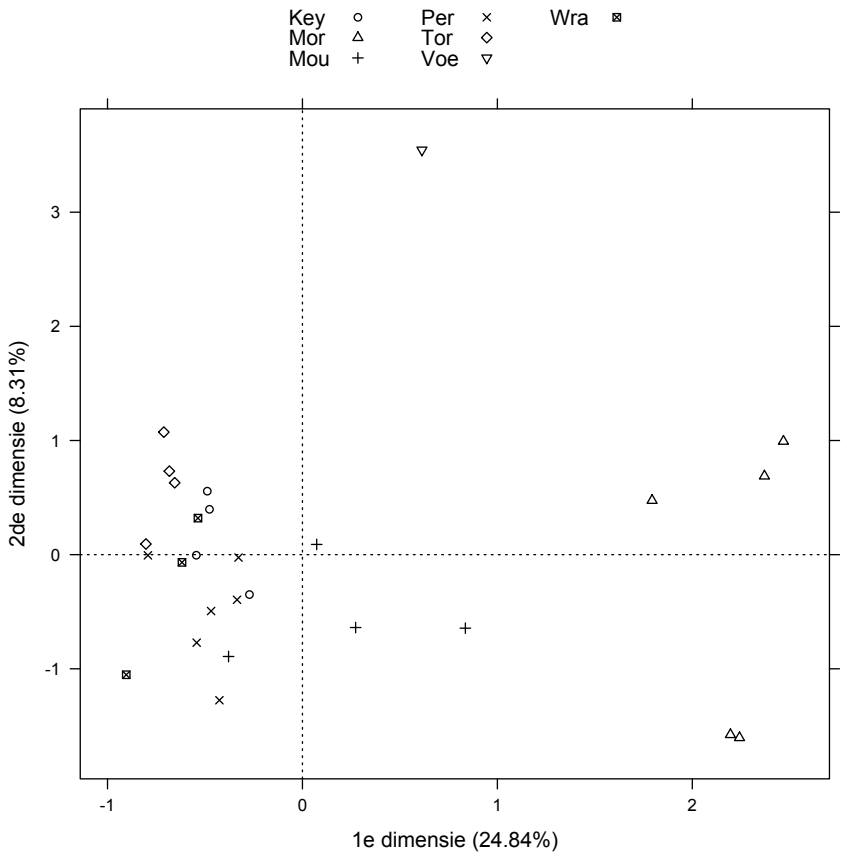
ren. Vooral de hechte cluster van de andere geïnterpoleerde romans is echter markant, omdat een identiek auteurschap voor deze teksten erg onaannemelijk is. Zo is Wra duidelijk te oud om door Maerlant te kunnen zijn geschreven (Gysseling 1980, 352), maar zou volgens deze analyse wel stilistisch verwant zijn aan Tor. Hoogstwaarschijnlijk worden wij hier geconfronteerd met het feit dat deze romans hier in hun gecompileerde versie worden bestudeerd: men zou uit dit resultaat namelijk kunnen opmaken dat deze romans in gelijklopende mate inmenging van de compilator hebben ondergaan. Die interpretatie strookt niet alleen met voorgaand onderzoek maar zou ook verklaren waarom de samples uit bv. Mor geen onmiddellijke verwantschap ten opzichte van de andere interpolaties vertonen.



Figuur 65: Correspondentie-analyse (eerste twee dimensies) van samples van 856 rijmwoorden uit samples van de compilatie-teksten, met uitsluiting van de kerndelen. Ook Velthems *MerlV* is ter vergelijking toegevoegd. De analyse is gebaseerd op de relatieve frequenties van de 100 meest frequente rijmwoorden in deze samples.

Een analyse waarin *MerlV* wordt weggelaten, wijzigt dit beeld niet (Fig. 66, p. 235). Opnieuw maken vooral Mor en in iets mindere mate Mou zich los van de sample-cluster met de andere compilatie-teksten. Een nadere analyse toonde dat de extreme positie van Voe deels werd

veroorzaakt door de hoge frequentie van het rijmwoord VOET maar andere analyses (o.m. met een kleinere *n*, zonder VOET) benadrukten dat ook de andere rijmwoorden aan het afwijkend karakter van VOE bijdragen. Als MOR en MOU niet werden geschreven door de compilator, Velthem of de auteur(s) van de kerndelen, rijst nu natuurlijk de prangende vraag wie dat wel heeft gedaan. Daarvoor kunnen we nu de blik wenden naar de Arturromans die buiten de *Lancelotcompilatie* zijn overgeleverd.



Figuur 66: Correspondentie-analyse (eerste twee dimensies) van samples van 856 rijmwoorden uit samples van de compilatie-teksten, met uitsluiting van de kerndelen. De analyse is gebaseerd op de relatieve frequenties van de 100 meest frequente rijmwoorden in deze samples.

8.2.4 *De wereld buiten de compilatie*

Het is reeds benadrukt dat slechts een klein aantal Middelnederlandse Arturromans buiten de *Lancelotcompilatie* zijn overgeleverd. Als we abstractie maken van het uitzonderlijk vroeg gedateerde, oostelijk *Tristant*-fragment (Caers & Kestemont 2011), resten er ons eigenlijk maar drie min of meer volledige teksten die geen onmiddellijke inmenging

van Velthem verraden. Toch behoren deze drie Arturromans tot de interessantste werken die ons uit de Middelnederlandse ridderepiek bekend zijn.

De *Roman van Ferguut* (FERG) is ons bekend uit een unieke Brusselse kopie (ca. 1350) die is overgeleverd in een bekend convoluut-handschrift dat in Leiden wordt bewaard.[25] De oorspronkelijke *Ferguut* schijnt ca. 1240 te zijn gedicht door een Oost-Vlaming en vormt een bewerking van de Oudfranse *Roman de Fergus* van Guillaume le Clerc (Kuiper 1989). Voortbordurend op de thematiek van de *Conte du Graal* behandelt de tekst de lotgevallen van de onervaren (en aanvankelijk nogal boerse) ridder Ferguut en zijn liefde voor de mooie Galiene. Slechts de eerste helft van de roman is getrouw aan de Oudfranse brontekst: halfweg de tekst schijnt de auteur zijn brontekst te hebben opgegeven. De auteur schijnt het tweede deel van de *Fergus* op uiterst creatieve wijze te hebben herwerkt. Kuiper (1989) heeft beargumenteerd dat deze inhoudelijke volta correleert met een auteursovergang: een eerste auteur zou de tekst onafgewerkt hebben gelaten en een tweede dichter – die geen toegang meer had tot de brontekst – zou de tekst enkele jaren later hebben voortgezet. Hoewel er enkele stilistische fenomenen in deze richting wijzen, heeft Kuipers betoog amper gehoor gevonden. Recent schijnen de meeste onderzoekers zelfs geen rekening meer te houden met een dubbel auteurschap voor de *Ferguut*. Het probleem is natuurlijk dat de meeste verschillen tussen beide helften makkelijk kunnen verklaard worden door het feit dat de dichter zijn grondtekst op een gegeven moment heeft verlaten, zodat de meeste stijlverschillen niet noodzakelijk op een auteursverschil teruggaan.[26]

Een tweede roman is *Lantsloot vander Haghedochte* (HAGH). Hoewel de roman niet volledig bewaard is, resten ons wel zeer omvangrijke fragmenten van een uniek handschrift van de tekst, die goed zijn voor 5.973 rijmwoorden in het corpus (Gerritsen et al. 1987). Van de talrijke Middelnederlandse bewerkingen van de *Lancelot en prose*, schijnt de *Lantsloot* een van de oudste te zijn. De *Lantsloot* (ca. 1250?) dient zeker niet verward te worden met de *Lanceloet* uit de compilatie. Hoewel beide teksten op dezelfde Franse brontekst teruggaan zijn de

25 De tekst is recent heruitgegeven en in het Engels vertaald door Johnson & Claassens (2000a).

26 Hieronder maak ik gebruik van een editie die mij werd bezorgd door Kuiper, waarvoor ik hem hartelijk wil danken. In de unieke *Ferguut*-kopie is een veertiende-eeuwse corrector aan het werk geweest die nogal wat aanpassingen in de (tegen dan verouderde) tekst heeft aangebracht. In het bestand dat hier wordt gebruikt heeft Kuiper waar mogelijk de oorspronkelijke lezing van de tekst hersteld. Rijmwoorden die niet restaureerbaar waren worden in deze analyse genegeerd. Hoewel ik hier niet dieper op inga, ga ik hieronder uit van een enkelvoudig auteurschap voor *Ferguut*. Geen van de analyses die ik tijdens mijn onderzoek heb uitgevoerd, leverde immers bijkomende indicaties op een meervoudig auteurschap. Natuurlijk is de hier gebruikte rijmwoorden-methode niet optimaal voor de analyse van een tamelijk korte tekst als de *Ferguut*, zodat toekomstig onderzoek hieromtrent uitsluitsel zal moeten brengen.

verschillen tussen beide opvallend. Gerritsen heeft in een invloedrijk artikel de *Lantsloot* voor de *Lanceloet* gedateerd. Hij beschouwde de *Lantsloot* als:

> een vrije bewerking-in-verzen die in esthetisch opzicht als conservatief kan worden getypeerd. De Middelnederlandse dichter nam de prozavorm niet over en zwakte het effect van een aantal nieuwerwetse stijlkenmerken – de kroniekachtige verteltrant zonder ik-verteller, de verbazingwekkende precisie in de aanduidingen van tijd en ruimte, het realisme, de overvloed aan details – in zijn bewerking bewust af (Gerritsen 1988, 203).

Gerritsen steunt in dit citaat op de erg belangrijke dissertatie van Van Oostrom (1981) over de bewerkingstechniek van de tekst. Van Oostrom kon overtuigend aantonen dat de *Lantsloot* inderdaad een relatief 'primitieve' indruk maakte, doordat de Franse brontekst verregaand werd geïdealiseerd. Waarschijnlijk was deze bewerking bedoeld voor een publiek dat nog niet erg vertrouwd was met de hoofse levensidealen die in deze periode opgang maakten. Dat de compilator van 129 A 10 de voorkeur gaf aan de latere bewerking in de *Lanceloet*, kwam waarschijnlijk doordat deze modernere versie beter aansloot bij de laatste ontwikkelingen in het genre in de vroege veertiende eeuw.

Een derde Arturroman die buiten de *Lancelotcompilatie* is overgeleverd, is de *Roman van Walewein*.[27] De roman is, buiten een klein fragment, slechts bekend uit een afschrift dat in het colofon in 1350 wordt gedateerd (Kienhorst 1988, 217-218). Dat deze klassieke Waleweinroman – althans, voor zover wij weten – niet werd opgenomen in de compilatie is verbazend. De roman zou immers op veel Arturteksten een belangrijke invloed hebben uitgeoefend, waaronder de *Moriaen* en *Riddere metter mouwen*, die beide wél in de compilatie werden opgenomen (Besamusca 1991). Bovendien had de compilator een grote voorliefde voor de figuur van Walewein.

De *Walewein* is een oorspronkelijke Middelnederlandse Arturroman, die opvalt door zijn ironische portrettering van Arturs neef, die in deze tekst met het bekende epitheton *der avonturen vader* werd bedacht. De epiloog van deze Vlaamse tekst (ca. 1250?) vermeldt dat de tekst door twee auteurs werd geschreven. Ene *Penninc* zou de tekst zijn begonnen, maar 'ongeveer 3.300 verzen voor het eind' zou *Pieter Vostaert* hem hebben afgelost, aldus het colofon. Het baanbrekend stylometrisch onderzoek van Van Dalen-Oskam & Van Zundert (2007) toonde inderdaad aan dat er zich 3.300 regels voor het einde van de tekst een markante stijlbreuk voordoet (zie ook § 2.2.2). Hoewel men rekening moet houden met de mogelijkheid dat Vostaert de tekst van Penninc (oppervlakkig) kan hebben bewerkt, zal ik de *Walewein*

27 Ook deze tekst is recent heruitgegeven en in het Engels vertaald door Johnson & Claassens (2000b).

voor dit onderzoek in twee delen opsplitsen: WalA (door Penninc) en WalB (door Vostaert).[28]

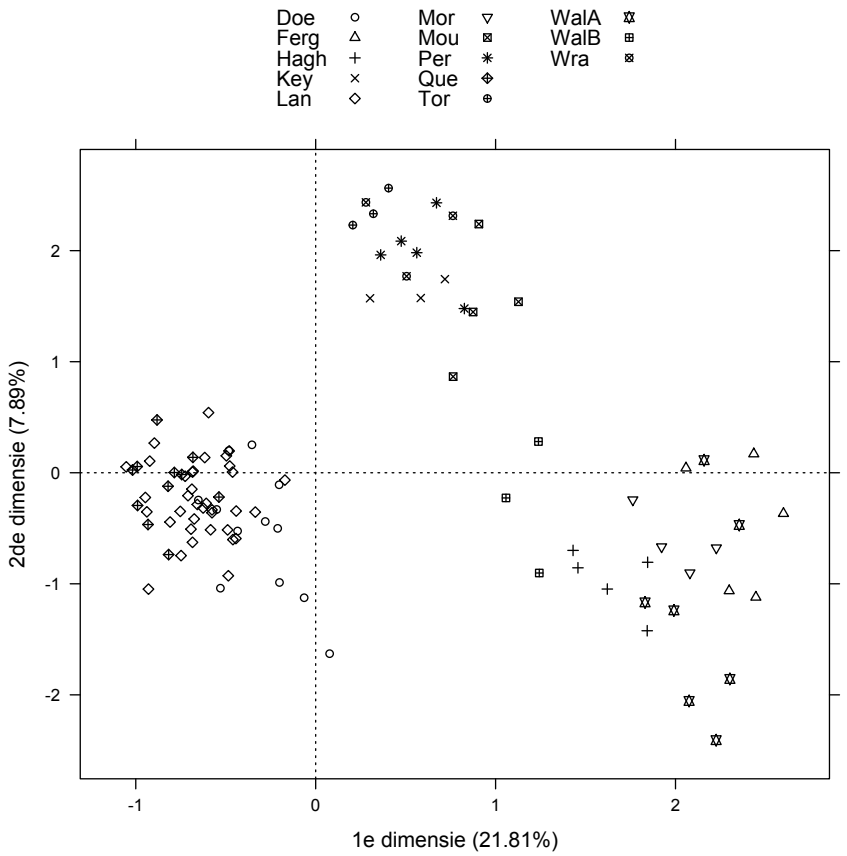
Laat ons deze drie – of liever: vier – Arturromans nu betrekken in onze correspondentie-analyses van de compilatie-teksten. In onderstaande figuur is een sample-grootte van 1.000 rijmwoorden gebruikt, zodat de kortere novelle Voe niet meer wordt meegenomen (Fig. 67, p. 239). Ook in deze analyse worden de langere kerndelen in een relatief hechte cluster duidelijk afgezet tegen alle andere teksten in de analyse. Enkele geïnterpoleerde romans vormen zoals in voorgaande experimenten een relatief dichte sample-wolk (Tor, Wrak, Per, Key). Ook de niet-compilatie-teksten vormen een relatief hechte groep. Merkwaardig genoeg treffen we de *Moriaen* – die ook hierboven reeds afwijkend gedrag vertoonde – niet aan bij de andere interpolaties maar in de groep teksten die buiten de compilatie zijn overgeleverd. Ook Mou schijnt in dit diagram te worden aangetrokken door de niet compilatie-teksten, al is deze toenadering minder uitgesproken.

Een correspondentie-analyse van de interpolaties en beide *Walewein*-delen (zie Fig. 68 (p. 240), vooral de eerste dimensie) suggereert inderdaad dat de relatief intacte rijmwoordenschat van Mou enkele gelijkenissen vertoont met de *Walewein*, en dan specifiek Vostaerts aandeel (WalB). Nadere experimenten met deze teksten leverden echter geen dwingende argumenten op om te speculeren over een identiek auteurschap van Mou en WalB. In toekomstig onderzoek zou men via 'filter-technieken' zoals *unmasking* (Koppel et al. 2009) kunnen trachten de compilator-invloed op Mou te neutraliseren. Het valt niet uit te sluiten dat Vostaerts rijmstijl in een dergelijke analyse meer uitgesproken gelijkenissen zou vertonen met Mou.

Een sterkere casus biedt de *Moriaen*: alle hierboven gerapporteerde experimenten bevestigen unaniem dat de stijl van deze roman in uitzonderlijke mate afwijkt van de rest van de compilatie-teksten en dus waarschijnlijk relatief intact is gebleven in de compilatie-versie. Onderstaande figuur biedt het resultaat van een correspondentie-analyse van samples van 1.000 rijmwoorden uit Hagh, Ferg, Mor en WalB (Fig. 69, p. 241). De samples van WalB en Ferg vormen beide relatief hechte groepjes samples per tekst. Opvallend genoeg zien we hoe de samples Mor en Hagh met elkaar vermengen, hoewel er op het eerste zicht geen verklaring voor deze toenadering bestaat. Talrijke andere experimenten met uiteenlopende parameters, bevestigden dat de stilistische cohesie tussen de *Moriaen* en *Lantsloot vander Haghedochte* vrijwel nooit is te doorbreken.

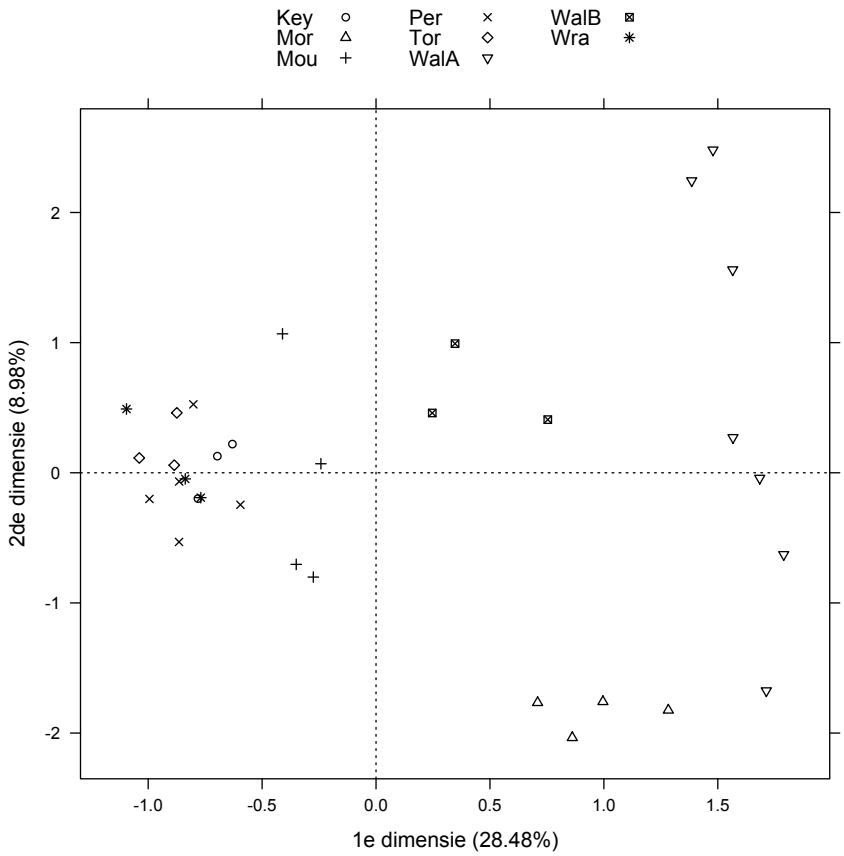
Maakt deze analyse een identiek auteurschap voor deze teksten aannemelijk? Mogelijk wel. Uit retorische overwegingen heb ik tot nu toe nagelaten te vermelden dat het identieke auteurschap van *Mo-*

28 Ik gebruik de standaardeditie op de *Cd-rom Middelnederlands* en ontleen mijn splitsingswijze aan Van Dalen-Oskam & Van Zundert (2007, 347-348). De tekst is dus gesplitst na vers 7.872 (*Salic sterven ic sal u gheven*), zodat WalA 7.856 rijmwoorden telt (met uitsluiting van de proloog), en WalB 3.323 rijmwoorden (zonder het colofon).



Figuur 67: Correspondentie-analyse (eerste twee dimensies) van samples van 1.000 rijmwoorden uit alle teksten uit de *Lancelotcompilatie*, alsook de Arturteksten die buiten dit handschrift zijn overgeleverd (WalA, WalB, Ferg, Hagh). De analyse is gebaseerd op de relatieve frequenties van de 100 meest frequente rijmwoorden in deze samples.

riaen en *Lantsloot vander Haghedochte* eigenlijk geen nieuwe hypothese is. Klaas Hanzen Heeroma (1909-1972) was een Nederlands filoloog en dichter die zich in zijn laatste levensjaren intensief bezig heeft gehouden met medioneerlandistiek (Van Oostrom 1998). In verschillende studies heeft Heeroma uitgebreid het identieke auteurschap voor deze teksten beargumenteerd, eerst voor *Moriaen* en *Karel ende Elegast*, vervolgens ook voor *Lantsloot vander Hagedochte*. Dat Heeroma ook de bekende Middelnederlandse roman over *Karel ende Elegast* in zijn betoog betrok is uiterst fascinerend, maar niet zonder problemen voor deze studie. De *Elegast* behoort immers tot het epische (sub)genre van de Karelepiek, terwijl de *Moriaen* en *Lantsloot* tot de Arturepiek behoren. In deze studie is reeds meermaals benadrukt dat het genre van een Middelnederlands tekst in sommige gevallen een bepalende invloed heeft gehad op de rijmwoordenschat van een tekst. Als we Heeroma's toeschrijving in een uitgebreider onderzoek willen toetsen,

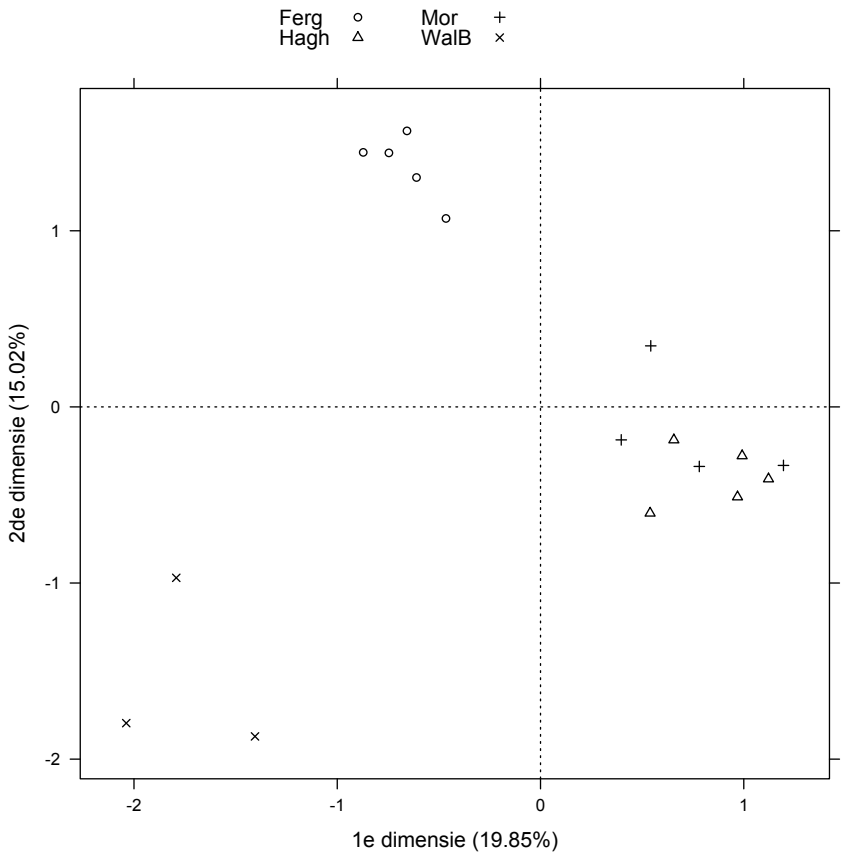


Figuur 68: Correspondentie-analyse (eerste twee dimensies) van samples van 1.000 rijmwoorden uit alle geïnterpoleerde teksten uit de *Lancelotcompilatie*, alsook beide *Walewein*-delen (WalA, WalB). De analyse is gebaseerd op de relatieve frequenties van de 100 meest frequente rijmwoorden in deze samples.

moeten we daarom eerst een beter zicht krijgen op de verschillen in rijmwoordenschat tussen de Middelnederlandse Artur- en Karelepiek. Daarvan zal ik in het volgende hoofdstuk werk maken.

8.3 CONCLUSIE

In dit hoofdstuk heb ik verslag gedaan van een stylometrisch onderzoek naar de Middelnederlandse Arturepiek, een genre dat in verschillende opzichten wordt gedomineerd door de figuur van Lodewijk van Velthem. Eerst werd een Delta-achtige procedure voorgesteld: deze gaat door de samples van een testtekst en meet telkens de afstand van een *window* uit de tekst ten opzichte van het centroïde van een auteur, gedefinieerd op basis van een relevante referentietekst van de auteur in kwestie (§ 8.1). Deze procedure werd toegepast op een verzameling *Merlijn-romans* (door Maerlant en Velthem) in het Burgsteinfurtse handschrift.



Figuur 69: Correspondentie-analyse (eerste twee dimensies) van samples van 1.000 rijmwoorden uit*WalB*, de *Roman van Moriaen*, *Lantsloot vander Haghedochte* en de *Ferguut*. De analyse is gebaseerd op de relatieve frequenties van de 100 meest frequente rijmwoorden in deze samples.

Wat betreft Maerlant leverde de centroïde-benadering het gewenste resultaat en werd de auteursovergang in de *Merlijn*-codex goed blootgelegd. Voor Lodewijk van Velthem werd echter een omgekeerd resultaat geboekt. Een nadere inspectie leerde dat er serieuze, mogelijk genre-gerelateerde verschillen bestaan tussen Velthems rijmwoordenschat in de *Merlijn-continuatie* en zijn *P4P5*. Slechts door het gebruik van de Dubbele Delta-metriek kon uiteindelijk een acceptabel resultaat worden bereikt.

De centroïde-aanpak werd in de volgende sectie toegepast op de Haagse *Lancelotcompilatie*, waarin een anonieme compilator (Lodewijk van Velthem?) verschillende reeds bestaande Arturromans in elkaar heeft vervlochten (§ 8.2). De stijlschommelingen die de Delta-benadering in de compilatie herkende kwamen verrassend nauwkeurig overeen met het traditionele onderzoek naar de compilatie (§ 8.2.1). Eerst werden de de zogenaamde 'kerndelen' onderzocht, die door de compilator relatief ongemoeid zouden zijn gelaten (§ 8.2.2). Hoewel

men meestal uitgaat van een enkelvoudig auteurschap van deze teksten, spraken verschillende correspondentie-analyses deze aanname tegen. De hoogfrequente rijmwoordenschat van *Arturs doet* wijkt namelijk af van die van de andere kernteksten. In de toekomst zal men daarom rekening moeten houden met de verrassende mogelijkheid dat meer dan één (Vlaamse?) dichter deze kernteksten kan hebben verdietst.

Een analyse van de geïnterpoleerde romans bevestigde de traditionele visie dat deze romans sterker zijn bewerkt door de compilator, hoewel de ingevoegde romans verschillen in de mate waarin zij zijn aangepast (§ 8.2.3). Teksten als de *Torec* en *Perchevael* lijken in het algemeen relatief sterk op Velthems *Merlijn-continuatie*, hetgeen de hypothese bevestigt dat de compilator – naar alle waarschijnlijkheid Velthem – ze duchtig heeft bewerkt. Andere romans, onder meer de *Riddere metter mouwen* lijken dan weer minder verregaand aangepast en weerspiegelen mogelijk meer getrouw de stijl van hun oorspronkelijke auteur.

Dé verrassing in dit hoofdstuk was echter de *Roman van Moriaen*, een interpolatie die al bij de eerste experimenten in het oog sprong door zijn afwijkende stijl. Omdat de rijmwoordenschat van deze roman mogelijk relatief intact is gebleven in de compilatie-versie, werd de *Moriaen* in de laatste sectie van dit hoofdstuk (§ 8.2.4) vergeleken met enkele romans die buiten de compilatie zijn overgeleverd. De hoogfrequente rijmwoordenschat van de *Moriaen* vertoonde frappante overeenkomsten met die van *Lantsloot vander Haghedochte*, een van de oudste bewerkingen van de Lancelot-stof in het Nederlandse taalgebied. Het identiek auteurschap voor deze teksten werd reeds jaren geleden door Heeroma betoogd en zal in het volgende hoofdstuk nader onderzocht worden.

9 KONING EN KEIZER

9.1 CONINC ENDE KEYSER MEDE

Het vorige hoofdstuk eindigde met een *cliffhanger*: een correspondentie-analyse bracht een grote overeenkomst aan het licht tussen de hoogfrequente woordenschat van de *Moriaen* en *Lantsloot vander Haghedochte* (Fig. 69, p. 241).[1] Dat deze opvallende overeenkomst indicatief is voor een identiek auteurschap van deze teksten is een fascinerende hypothese, vooral omdat Heeroma reeds jaren geleden dezelfde stelling heeft bepleit op basis van een traditioneel stilistisch onderzoek. Heeroma schreef dezelfde dichter ook het bekende werk *Karel ende Elegast* toe. Dat is merkwaardig omdat de *Elegast* traditioneel tot de Karelepiek wordt gerekend terwijl de andere twee teksten typische Arturromans zijn. In dit hoofdstuk zal ik Heeroma's hypothese verder toetsen. Daarvoor is het echter noodzakelijk om eerst de invloed van de tekstsoort op de rijmwoordenschat van beide types ridderromans beter in kaart te brengen. In eerdere hoofdstukken is namelijk gebleken dat het type tekst soms een bepalende invloed heeft gehad op – zelfs de meest frequente – rijmwoordenschat (§ 6.3 en § 8.1).

9.1.1 *Twee matières*

De Arturepiek en Karelepiek zijn de twee belangrijkste tekstsoorten die onderzoekers doorgaans onderscheiden binnen de Middelnederlandse ridderepiek (Janssens 1998b, 16-17). Het traditionele onderscheid tussen beide tekstsoorten is voornamelijk op inhoudelijke criteria gebaseerd en gaat terug op de verschillen in het stofcomplex waaruit zij putten. De Arturroman (*matière de Bretagne*) ontleent vooral aan de middeleeuwse legendevorming rond de legendarische Britse vorst Artur en zijn rondetafelridders. De Karelepiek (*matière de France*) behandelt daarentegen de epische stof rond Karel de Grote en zijn vazallen (Van den Berg 1992b, 406-408). Van den Berg (1988, 1992a, b) heeft in verschillende publicaties ook de stilistische verschillen tussen beide teksttypes onderzocht: teksten uit de Karelepiek geven bijvoorbeeld blijk van een formulair discours dat sterker lijkt ingebed in orale tradities.

Toch is een star onderscheid tussen de Middelnederlandse Artur- en Karelepiek moeilijk houdbaar. Op het gebied van versificatie vertonen zij bijvoorbeeld nauwelijks verschillen. In beide tekstsoorten zijn

1 Dit hoofdstuk bouwt in belangrijke mate voort op een eerder gepubliceerd artikel (Kestemont 2012c).

werken vormgegeven met behulp van volrijmende coupletten. Dat is opvallend omdat de meeste ridderromans omwerkingen vormen van Franstalige bronteksten die op dit vlak wel een duidelijk onderscheid vertonen: verschillende Franse Arturromans zijn geschreven in octosyllabische coupletten, terwijl de meeste *chansons de geste* uit de Karelepiek in assonerende strofevormen werden gedicht (Van Oostrom 2006, 233). Ook de datering en lokalisering van de teksten uit de Artur- en Karelepiek vertonen in ons taalgebied weinig verschillen.[2] Zelfs qua thematiek lijkt men het erover eens dat het onderscheid misschien niet zo eenduidig is als vroeger wel eens werd aangenomen (Janssens 1998b, 17). Zemel (2005, 2010) beargumenteerde dat de auteurs van de *Roman van Walewein* stilistisch aansluiten bij de Karelepiek en nog recent ontdekte Hogenbirk (2009a) aan het *chanson de geste* ontleende motieven in de Arturroman *Moriaen*. Dichters uit beide tekstsoorten kenden bovendien elkaars werk.[3] Besamusca besprak enkele jaren geleden nog een mooi voorbeeld daarvan: tenminste één van de dichters van de Karelroman *Ogier van Denemarken* verraadt kennis van de Middelnederlandse Arturroman *Lantsloot vander Haghedochte* (Besamusca 2005, 142). Ondanks de verschillen in stofcomplex bestaan er tussen de Middelnederlandse Artur- en Karelepiek dus ook markante overeenkomsten.

Gegeven het recente onderzoek is het dus ondenkbaar dat de Middelnederlandse Artur- en Karelepiek in gescheiden literaire circuits functioneerden.[4] Toch zijn momenteel geen gevallen bekend van dichters die zich met zekerheid in beide teksttypen bekwaamden. Alleen al daarom is het interessant om de Artur- en Karelepiek gezamenlijk aan een stylometrische studie te onderwerpen: mogelijk brengen kwantitatieve vergelijkingstechnieken nieuwe overeenkomsten aan het licht, hetgeen zeker interessant zou zijn vanuit het perspectief van Brinkmans Vijfde Gebod: DE EENHEID VAN GENRE.

Vijfde Gebod: De eenheid van genre

Een dergelijke onderneming gaat echter niet over rozen: in zowel de Artur- als Karelepiek is over het auteurschap van teksten bitter weinig bekend, zodat betrouwbare referentiepunten voor het ijken van technieken schaars zijn. Het leeuwendeel van de Middelnederlandse Artur- en Karelepiek is anoniem overgeleverd (Kienhorst 1988; Besamusca 1983, 1985). In de Arturepiek is zoals reeds gezegd het auteurschap van de *Merlijn*-romans van Maerlant en Velthem uit het Burgsteinfurtse handschrift bekend (Besamusca et al. 2009b, 12-13), evenals dat van de *Torec* in de *Lancelotcompilatie*, waarvan het oorspronkelijk auteurschap aan Maerlant wordt toegeschreven (Claassens 2009). Daarbuiten is slechts bekend dat ene Penninc en Vostaert beiden aan de *Walewein* hebben geschreven (Van Dalen-Oskam 2007).

2 Zie bv. Caers (2011); Caers & Kestemont (2011). Merk wel op dat de productie van de tekstdragers mogelijk diachrone schommelingen in literaire smaak verraadt (Klein 1995, 18-20).

3 Zie recent bv. Zemel (2010, 11-18) of Hogenbirk (2011, 169).

4 Vgl. Janssens (1998a) of Hogenbirk (2011, 157-159).

Wat betreft auteurschap is in de Karelepiek nog minder bekend: Maerlant vermeldt slechts in zijn *Spiegel historiael* ene *van Haerlem Clays, ver Brechten sone* als auteur van een tekst over *Willeme van Oringen*. Mogelijk doelde hij daarmee op een van de twee *Oringen*-teksten die zijn overgeleverd, maar zeker is dat niet (Caers & Kestemont 2011). Buiten Maerlant (en eventueel Velthem) zijn trouwens ook geen auteurs bekend die met zekerheid meer dan één ridderroman hebben geschreven, al werd vaak aangenomen dat de kerndelen van de *Lancelotcompilatie* (vgl. § 8.2.2) door één enkele Vlaming zouden zijn verdietst (Besamusca 2003, 44-47). Niettemin heb ik precies die stelling in het vorige hoofdstuk in twijfel getrokken (vgl. § 8.2.2).

9.1.2 *Eindjes aan elkaar knopen*

Zoals in dit hoofdstuk nog ruimschoots zal blijken, wordt het onderzoek naar de Middelnederlandse ridderroman door een aantal praktische factoren ernstig belemmerd. De grootste hindernis – en dat geldt natuurlijk niet alleen voor dit onderzoek – is de gebrekkige bronnenoverlevering: in statistische termen biedt de overlevering slechts een willekeurig sample van de oorspronkelijke 'populatie' aan ridderteksten. De betrouwbaarheid van deze noodgedwongen steekproef is moeilijk of niet te achterhalen (vgl. Van Oostrom 2006, 331). Hoewel ik in dit hoofdstuk een zo goed mogelijke dekking van het bewaarde materiaal nastreef, is dit streefdoel van exhaustiviteit dus hoe dan ook relatief. Alleszins valt op dat de Middelnederlandse Artur- en Karelepiek door een heel ander type van overlevering worden gekenmerkt. In de rest van deze sectie bespreek (en problematiseer) ik beknopt de materiaalverzameling die de basis van dit hoofdstuk vormt.

Voor deze studie heb ik gebruik gemaakt van een verzameling digitale edities die veelal beschikbaar waren op de *Cd-rom Middelnederlands*. Een bespreking van de geanalyseerde werken is te vinden in de repertoria van Besamusca (1983, 1985) en Kienhorst (1988) en enkele overzichtsartikelen (Van den Berg & Besamusca 1992; Klein 1995; Van der Have 2005). Over het algemeen zijn meer verschillende Karel- dan Arturromans overgeleverd, maar daar staat tegenover dat de Karelepiek veel fragmentarischer is. In het extreme voorbeeld van *Oringen*-2 (Kienhorst 1988) resten ons amper een 70-tal zwaar gehavende verzen. De beperkte omvang van deze fragmenten speelt het onderzoek vanzelfsprekend parten.

De *Elegast* is de enige Kareltekst die – naast enkele schamele restanten van handschriften – volledig is overgeleverd in jongere drukken (Fig. 70, p. 249). De langste Kareltekst biedt evenwel de *Roman van Lorreinen-I*, met meer dan 10.000 overgeleverde verzen – oorspronkelijk moeten dat er ruim 40.000 geweest zijn (Van der Have 2005, 83). Dat is voor de studie van deze tekst natuurlijk een gelukkig toeval. Toch is het belangrijk dat men deze omvangrijke tekst niet te zwaar laat

doorwegen in de algemene beeldvorming van de Karelepiek. Verschillende Karelteksten schijnen via relatief oppervlakkige *Umschreibungen* naar het oosten te zijn gemigreerd. Erg oostelijke of sterk verduitste tekstgetuigen zijn op voorhand uit deze studie geweerd omdat zij zich slecht laten vergelijken met de hoofdmoot van het corpus, onder meer de *Limburgse Aiol* (Goossens 2002) of handschrift H96 van de *Renout van Montalbaen* (Kienhorst 1988, 176-177).[5]

Onderstaande tabellen – naamgeving en siglen zoveel mogelijk volgens Kienhorst (1988) – bieden een overzicht van de in deze studie betrokken Karelteksten en het aantal beschikbare rijmwoorden per tekst – Tab. 20 (p. 247) en Tab. 21 (p. 248). Als van dezelfde tekst verschillende fragmenten bestaan, zijn deze samengevoegd onder één tekstnoemer, waarbij de (zeldzame) overlapping tussen tekstgetuigen werd vermeden. Alle digitale versies van deze teksten werden geoogst van de *Cd-rom Middelnederlands*, met uitzondering van de *Garijn van Montglavie*, die van de tekstdrager zelf werd getranscribeerd en *Oringen-2*, die rechtstreeks overgenomen is uit Kienhorst (1998).

Voor de *Elegast* is in navolging van Claassens gekozen voor de redactie in Incunabel A (Fig. 70, p. 249), met toevoeging van de regels 1226–1244 en 1378–1394 uit Claassens (2002). De zogenaamde 'oudere' en 'jongere' delen die zijn beargumenteerd voor *Madelgijs* (Duijvestein 2002, 23-24), *Ogier van Denemarken* (Besamusca 2005) en de *Roman van Lorreinen* (Van der Have 1990) zijn onderscheiden omdat hier meerdere auteurs in het spel kunnen zijn. Op die wijze wordt hier zoveel mogelijk gegarandeerd dat iedere afzonderlijke 'tekst' door slechts één auteur is geschreven. Ik gebruik hieronder zoveel mogelijk de traditionele nomenclatuur van Kienhorst (1988), al is voor verschillende teksten onduidelijk hoe zinvol deze is bij ontstentenis van een Franse brontekst (Besamusca 2008). Merk ook op dat de fragmenten van de zogeheten *Gwidekijn van Sassen* (H32) en het fragment over het verraad van Pepijn – *Ongeïdentificeerd (3)* of fragment H126 (Van Oostrom 1987) – uit dezelfde codex stammen (Kienhorst 1988, 68-69 & 232), zodat zij mogelijk tot dezelfde (auteurs?)tekst mogen gerekend worden. Met uitzondering van dit geval is voor geen enkele van de hier onderscheiden Karelteksten tot dusver een identiek auteurschap grondig beargumenteerd.[6]

De Middelnederlandse Arturepiek telt minder, maar wel beter overgeleverde teksten (Besamusca 1983). In totaal zijn een zeventiental Middelnederlandse Arturteksten overgeleverd, waarvan ik er twee

5 Van der Have (2005, 79) vermeldt een ongeïdentificeerd fragment met het sigle G2 (Gent, Universiteitsbibliotheek, 1647) maar dit fragment wordt door Kienhorst (1988, 64-65) vermeld als een restant (H31B) van een handschrift van de kruisvaartroman *Godevaerts kintshede / Roman van Antiochië*. De *Lion van Bourges* (H57, nog vermeld bij Besamusca en Kienhorst) zou volgens Claassens (1996) geen ridderroman, laat staan een Karelroman zijn. Beide fragmenten zijn hier daarom niet opgenomen.

6 Voor de *Renout van Montalbaen* is de mogelijkheid van een meervoudig auteurschap overwogen maar verworpen (Spijker 1990, 242-250).

Tekst	Afkorting of mix	Samengevoegde tekstgetuigen	Aantal rijmwoorden (aantal samples)
Vlaamse Aiol, Aiol (2)	VlaAiol	H2	1.198 (1x1.198)
Chanson d'Aspremont, olim *Ongeïdentificeerd (5)*	A	H128[a]	49
Aubri de Borgengoen	A	H7	350
Beerte metten breden voeten	A	H10	207
Fierabras	A	H20	360
Floovent	A	H23	604
Garijn van Montglavie	B	H29	150
Geraert van Viane	B	H30	192
Gwidekijn van Sassen	B	H32	200
Hughe van Bordeeus	Hughe	H33, H34, H35, H36	1.397 (1x1.397)
Karel ende Elegast	Elegast	Incunabel A[b]	1.400 (1x1.400)
Laidoen / Lorreinen-II	B	H45	310
Lorreinen / Lorreinen-I	Lorreinen	H58, H59, H60	10.632 (6x1.772)
Loyhier ende Malaert	B	H61, H62	605

Tabel 20: Overzicht van de gebruikte teksten uit de Karelepiek (deel 1).

a Momenteel: Antwerpen, Bibliotheek van het Ruusbroecgenootschap, Hs. Neerl. 15/1, olim Ibidem, Fragment-Averbode A.
b Den Haag, Koninklijke bibliotheek, 169 G 63.

Tekst	Afkorting of mix	Samengevoegde tekstgetuigen	Aantal rijmwoorden (samples)
Madelgijs-1 (eerste, korte versie)	MadelgijsKorte	H63B, H64, H66, H67, H68, H69, H70, H71	1.933 (1x1.933)
Madelgijs-2 (tweede, lange versie)	C	H63A, H72, H65	745
Ogier van Denemarken-1	C	H78, H79	184
Ogier van Denemarken-2	C	H80	389
Renout van Montalbaen	Renout	H92, H93, H94, H95	2.641 (1x1.320+1x1.321)
Roelantslied	Roelant	H98, H99, H100, H101, H102	1.306 (1x1.306)
Willem van Oringen-1	C	H120	362
Willem van Oringen-2	D	Oxford-fragment[a]	42
Van den bere Wisselau	D	H121	561
Ongeïdentificeerd (2)[b]	D	H125	48
Ongeïdentificeerd (3)[c]	D	H126	359
Ongeïdentificeerd (4)[d]	D	H127	156
Ongeídentificeerd (6)[e]	D	H129	157

Tabel 21: Overzicht van de gebruikte teksten uit de Karelepiek (deel 2).

a Oxford, Bodleian Library, Ms. Dutch B.2, fol. 6. Ik ga er in navolging van de recente literatuur (bv. Van der Have 2005) van uit dat de *Oringen-1* en *Oringen-2* tot verschillende teksten moeten gerekend worden.
b Fragment G1 bij Van der Have (2005).
c Ook bekend als 'Het verraad van Pepijn' (Van Oostrom 1987); fragment L bij Van der Have (2005).
d Fragment AL bij Van der Have (2005).
e Fragment AV bij Van der Have (2005).

Figuur 70: Afbeelding van de titelbladzijde uit Incunabel A van *Karel ende Elegast* (Den Haag, Koninklijke Bibliotheek, 169 g 63). Het gaat om een exemplaar van een van de oudste drukken (1486-1488) van de enige Middelnederlandse Kareltekst die naar alle waarschijnlijkheid volledig bewaard is gebleven. De tekst is recentelijk nog toegankelijk uitgegeven naar deze druk door Claassens (2002).

op voorhand heb uitgesloten voor het onderzoek in dit hoofdstuk, omwille van hun beperkte lengte (*Lanceloet en het hert met de witte voet* en de oostelijke *Tristan*). Het zij nogmaals benadrukt dat meer dan tien Arturromans zijn overgeleverd in twee compilatie-handschriften die in verband kunnen worden gebracht met Lodewijk van Velthem (vgl.

Draak 1974, 363). De compilatie-teksten worden ook in dit hoofdstuk in hun gecompileerde vorm bestudeerd: de overgeleverde fragmenten van de originele oerteksten zijn veelal té beperkt in omvang om voor dit type onderzoek in aanmerking te komen. Daarbij is geen poging ondernomen om Velthems stilistische inmenging te neutraliseren – de verantwoording van die ingreep alleen al zou een complete monografie vergen. Maar liefst 12 van de 15 hier bestudeerde Arturteksten hebben ons dus waarschijnlijk bereikt via de schrijftafel van één enkele beoefenaar van de Arturroman.

9.1.3 *Clusters*

Wanneer we de rijmwoordenschat van de Middelnederlandse Arturen Karelepiek aan een vergelijkend stylometrisch onderzoek willen onderwerpen, zijn we aangewezen op samples uit bovenstaande teksten. Samples moeten voor een stylometrisch onderzoek echter groot genoeg zijn: vooral bij de kortere Karelteksten wringt hier het schoentje. Als we uitgaan van een minimale sample-grootte van ca. 1200 woorden – en gegeven de onderzoeksresultaten in de vorige hoofdstukken is dat reeds krap bemeten – kunnen we voor sommige teksten tenminste één volledig sample extraheren. Voor de *Lorreinen-I* en de *Renout* levert dit zelfs respectievelijk 6 en 2 samples op.

Met deze ondergrens dreigen verschillende, korte maar interessante teksten evenwel uit de boot te vallen. Ik heb daarom op betrekkelijk willekeurige wijze enkele 'gemixte' samples samengesteld uit de fragmentarische restanten van teksten die te klein zijn om een volledig sample te vullen – vgl. Tab. 20 (p. 247) en Tab. 21 (p. 248). Deze samengestelde samples tellen respectievelijk 1.570 (mix A), 1.457 (mix B), 1.680 (mix C) en 1.323 (mix D) rijmwoorden, zodat zij de sample-grootte benaderen van de overige 'atomische' Karelepische samples.

Dit levert in totaal 17 bruikbare Karelsamples op (13 atomische en 4 samengestelde) met een gemiddelde grootte van ca. 1.561 rijmwoorden per sample. Ik heb er bij het ontwerp van mijn methodologie lang aan getwijfeld of het noodzakelijk was om de grootte van alle samples gelijk te trekken, bijvoorbeeld door het (eventueel willekeurig) verwijderen van rijmwoorden en er slechts 1.198 te bewaren per sample. Gezien het beperkte aantal rijmwoorden waarover dit onderzoek reeds beschikte, leek het verwijderen van zoveel waardevolle informatie uiteindelijk toch minder opportuun dan het werken met intacte samples van ongelijke maar toch erg vergelijkbare grootte. Het is evenwel duidelijk dat wij Brinkmans Derde Gebod – inzake tekstlengte – hier tot op zekere hoogte schenden.

Derde Gebod: Tekstlengte

Een ander probleem stelt de vergelijking met de Arturepiek, waar door het grote aantal goed overgeleverde teksten een veel groter aantal samples (per tekst) beschikbaar is. Middelnederlandse Karelepiek kan

men vanzelfsprekend niet meer bijmaken; wel kan men simuleren dat er minder aan Arturepiek is overgeleverd. De simulatie van een meer gebrekkige overleveringssituatie voor de Arturepiek is vanzelfsprekend erg artificieel, maar lijkt noodzakelijk voor het compenseren van het kwantitatieve overwicht van de Arturepiek ten opzichte van de veel minder bedeelde Karelepiek. Omdat de meeste Karelepische fragmenten niet uit het einde of het begin van een roman lijken overgeleverd, heb ik ervoor geopteerd om uit de beschikbare Arturteksten steeds een sample met de middelste 1.561 (de gemiddelde lengte van de Karelsamples) rijmwoorden te lichten. Op deze manier creëren we 15 atomische 'Artur-fragmentjes' van 1.571 rijmwoorden die in aantal en grootte vergelijkbaar zijn met wat ons rest aan Karelepiek.

Als we de rijmwoorden in deze samples voorstellen in een *bag-of-words*-model, is het een interessante vraag of er een duidelijk verschil in rijmwoordenschat valt waar te nemen tussen de samples uit beide tekstsoorten. Voor een dergelijke analyse kunnen we gebruik maken van een statistische techniek voor clustering. Net zoals technieken voor dimensie-reductie is clustering eveneens een methode om de belangrijkste structuur in een grote frequentie-tabel op inzichtelijke wijze te visualiseren (Hoover 2003).[7]

clustering

De specifieke clustertechniek die hieronder wordt toegepast is *divisive clustering*. De opzet van deze techniek noemt men 'hiërarchisch', want zij zal (steeds kleinere) groepen van gelijkaardige samples in de data proberen te herkennen. Er wordt daarbij *top down* gewerkt: eerst wordt gezocht naar grote, ruwe groepen samples in de data die vervolgens, op lagere niveaus in de analyse, worden opgedeeld in meer verfijnde groepjes. De 'familie-structuur' die zo uit de data wordt gedistilleerd, kan dan worden weergegeven als een boomdiagram, waarin in een oogopslag duidelijk wordt welke groepjes samples veel met elkaar gemeen hebben en welke helemaal niet. Samples die veel gelijkenissen met elkaar vertonen wat betreft hun rijmwoorden, zullen op een relatief laag niveau in de boom clusteren, terwijl sterk verschillende groepen teksten pas op een veel hoger niveau in de boom met elkaar zullen verbonden worden. Clustering levert voor auteursherkenning meestal minder genuanceerde resultaten op (Hoover 2003) maar is wel geschikt voor het opsporen van de algemene stilistische verwantschap binnen een grotere verzameling teksten.

divisive clustering

Het beste voorbeeld van de werking van een cluster-analyse kunnen we nog geven aan de hand van de ridderromans in het corpus zelf. *Divisive clustering* probeert immers eerst grote groepen in de data te onderscheiden, zodat de analyse misschien al op een hoog niveau in het boomdiagram een tweedeling zal voorstellen tussen de Artur- en Karelromans. Een dergelijk resultaat valt te verwachten op basis van

7 Ik ga hier voorbij aan de meetkundige formules waarop deze techniek stoelt. Dergelijke methodes zijn immers goed beschreven in publicaties die ook voor filologen met een minder kwantitatieve scholing erg toegankelijk zijn (Baayen 2008).

voorgaand stilistisch onderzoek, in het bijzonder dat van Van den Berg (1992a).

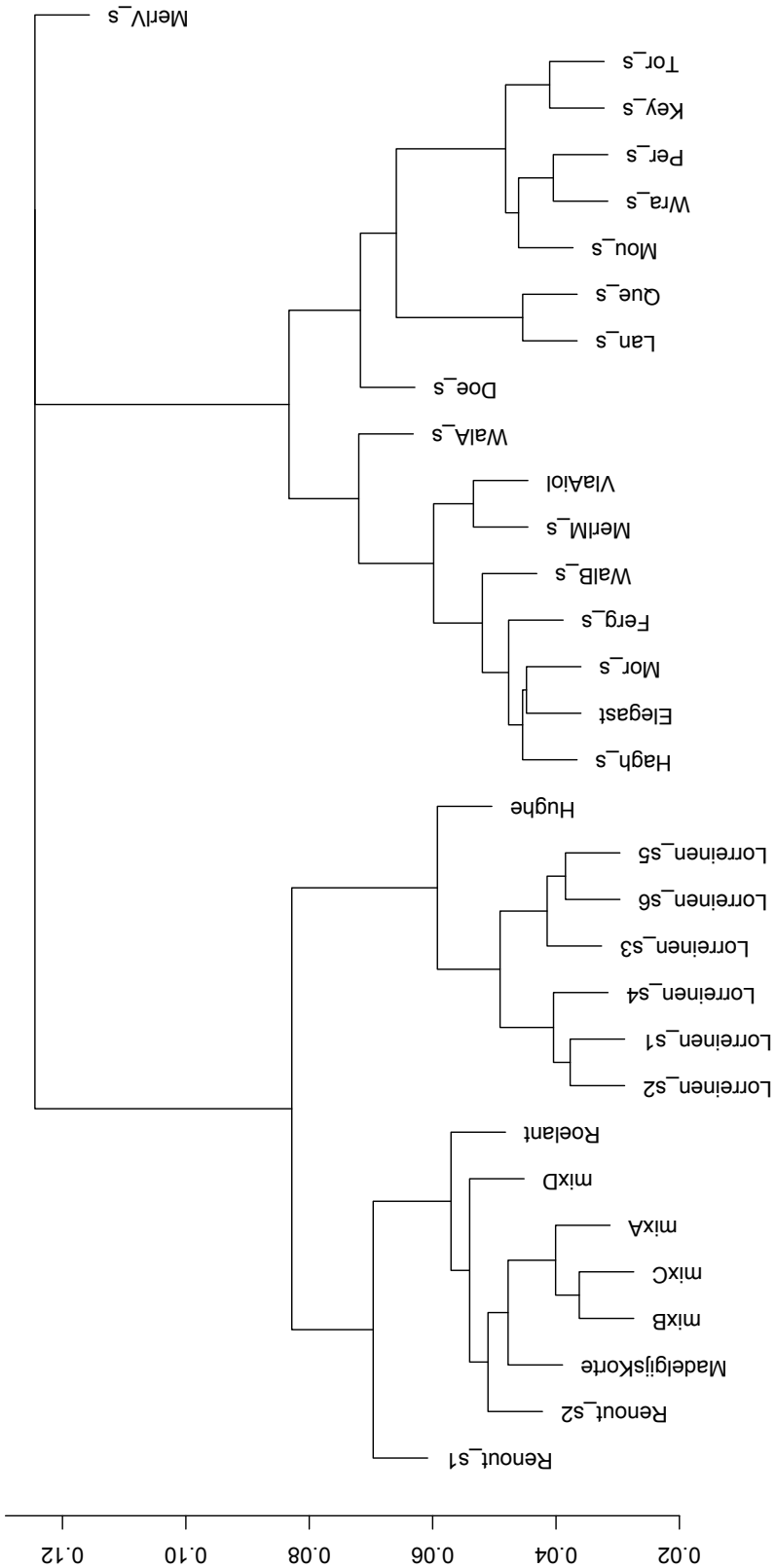
Indien we deze clustertechniek toepassen op onze Artur- en Karelepische samples (17+15 = 32 samples) voor de 1.000 meest frequente rijmwoorden in dit corpus, is het onderstaande boomdiagram het resultaat.[8] Een tekstnaam gevolgd door een '_s' in onderstaande figuur wijst erop dat het sample slechts een steekproef is (DOE_S is de steekproef uit *Arturs doet*). In deze naamgeving volgt op dit achtervoegsel het volgnummer van een sample als er meerdere samples voor een tekst beschikbaar zijn (RENOUT_S2 is het tweede sample uit de *Renout van Montalbaen*).

Onderstaande analyse illustreert treffend de werking van de clustering (Fig. 71, p. 253). Reeds op het hoogste niveau stelt de analyse een belangrijke driedeling voor die sterk gerelateerd blijkt aan het traditionele onderscheid tussen Artur- en Karelepiek: (1) links treffen we een hechte groep van Karelepische werken aan, (2) in het midden een groep samples waarin de Arturepiek domineert en (3) uiterst rechts een sterk geïsoleerd sample uit Velthems *Merlijncontinuatie*. Over die laatste tekst is men snel uitgepraat: de rijmstijl van deze tekst is in het verleden reeds gekarakteriseerd als bijzonder stoplaprijk en repetitief (Hogenbirk 2009b, 59-60). Deze analyse toont nog maar eens hoe extreem ver Velthems stijl in deze roman verwijderd is van ander ridderepische werken (zie ook § 8.1). Zo ver zelfs, dat de roman ontsnapt aan de belangrijke tweedeling die zich manifesteert in de cluster-analyse op het tweede hoogste niveau.

De Karelepische werken vormen links in de afbeelding een relatief dichte eenheid die duidelijk kan onderscheiden worden van de cluster met Arturepische werken. Merk op hoe dicht de diverse samples uit de *Lorreinen-I* met elkaar clusteren, hetgeen vanzelfsprekend een artefact is van het feit dat deze samples uit een enkele tekst zijn geëxtraheerd. Het sample van *Hughe van Bordeeus* (HUGHE) gelijkt volgens deze verkennende analyse meer op de *Lorreinen-I* dan het uiterst linkse groepje werken met de samengestelde samples en voorts onder meer het *Roelantslied*. Deze coherente groep Karelteksten vertoont onderling grote stilistische overeenkomsten maar blijkt dus amper verwant aan de groep vertakkingen in het midden van de figuur.

De middelste groep in het diagram wordt gedomineerd door Arturromans en wordt op het tweede hoogste niveau uitgesplitst in twee subgroepen. In de eerste daarvan (rechts in het diagram) vinden we

8 Een technische noot: gebruikt is de standaard *diana*-functie met *default options* (o.m. Euclidische afstand) uit de *open-source* software *R* (`http://www.R-project.org`). *Divisive clustering* is aangewezen omdat we geïnteresseerd zijn in grotere subgroepen (met name de tweespalt tussen Artur- en Karelepiek). Zie Baayen (2008, §5.1.5). Wat betreft het optimale aantal rijmwoorden in een dergelijke analyse bestaat overigens niet echt een consensus. De uitkomsten die hier gerapporteerd worden waren steeds stabiel en weken niet wezenlijk af van het resultaat met andere, vergelijkbare aantallen rijmwoorden.



Figuur 71: Cluster-analyse voor de hierboven gedefinieerde samples uit de Middelnederlandse Artur- en Karelepiek. Voor deze analyse werd een beroep gedaan op de 1.000 meest frequente rijmwoorden in deze samples.

een cluster van geïnterpoleerde romans uit de *Lancelotcompilatie*: WRA, PER, KEY en TOR. Op een iets hoger niveau worden deze onderscheiden van de 'kernteksten' uit de compilatie (LAN, QUE en DOE). Merk op hoe deze analyse – nog maar eens – het verschil beklemtoont tussen DOE en de twee eerste kerndelen.

Vooral het clusteren van de steekproeven uit de vier geïnterpoleerde romans is markant. Hoogstwaarschijnlijk worden wij hier opnieuw geconfronteerd met het feit dat deze romans hier in hun gecompileerde versie worden bestudeerd (vgl. § 8.2.3). Ook dit resultaat suggereert dat deze romans in gelijklopende mate inmenging van de compilator hebben ondergaan en opnieuw vertonen de steekproeven uit MOU en MOR geen onmiddellijke verwantschap met de andere interpolaties. De grote verwijdering tussen de linkse cluster (met o.m. *De roman van Lorreinen* en de rechtse cluster (met verschillende Lancelot-interpolaties) laat momenteel niet toe om tussen deze teksten inzake auteurschap een brug te slaan. Toekomstig onderzoek kan trachten de stilistische inmenging van de compilator in de interpolaties te isoleren om dit resultaat te verifiëren, zoals door Hogenbirk (2009b) is gedaan.

De situatie is evenwel anders voor de groep vertakkingen, waarin de geïnterpoleerde roman MOR zich wel weet te onttrekken aan het 'compilatie-kluwen' (ongeveer centraal in het diagram). Blijkbaar is de stilistische eigenheid van de oerversie van deze tekst wel voldoende gevrijwaard voor een rechtstreekse vergelijking met de rijmstijl van andere teksten. De centrale groep vertakkingen bevat immers naast een groot aantal Arturromans van buiten de *Lancelotcompilatie* (FERG, HAGH, WALA, WALB en MERLM), opvallend genoeg ook enkele Karelteksten (ELEGAST en VLAAIOL). De verrassing in deze macroanalyse wordt dus eigenlijk geleverd door twee Karelepische teksten. De boomstructuur reproduceert immers vrij getrouw het traditionele onderscheid dat in de vakliteratuur tussen de Artur- en Karelepiek wordt gehanteerd. Twee Karelepische teksten ontsnappen echter aan die regelmaat en zoeken het gezelschap op van Arturepische teksten (en zelfs specifiek die teksten die zich van het compilatie-kluwen onderscheiden). Blijkbaar zijn er in het rijmvocabulaire stilistische selectieprocessen aan het werk die van deze teksten atypische Karelepiek maken en zorgen voor een aansluiting bij de Arturepiek.

Het resultaat van deze clustering is op zich niet nieuw: men wist reeds dat er een Karelepische traditie bestond die is te onderscheiden van haar Arturepische pendant. Wel biedt dit resultaat een kwantitatieve, objectieve bevestiging van dat inzicht, want ook met cijfers valt het onderscheid tussen deze *matières* te motiveren. Een belangrijke nuance is dat een cluster-analyse onbevooroordeeld of *unsupervised* is: de analyse heeft geen toegang tot een voorhand gegeven indeling naar tekstsoort. Bijgevolg wordt er ook geen poging ondernomen om de data in een op voorhand gedefinieerd model te dwingen. Voorgaand onderzoek op dit gebied was meestal expliciet gericht een op voor-

hand geïsoleerde stilistische categorie (vgl. § 2.2), zoals epitheta of nadrukformules (Van den Berg 1988, 1992a; Van Driel 2007). Het is interessant dat de nieuwe techniek die hier is toegelicht de data tot op zekere hoogte voor zichzelf laat spreken en toch nog steeds het traditionele onderscheid uit de rijmtellingen kan distilleren.

9.1.4 *Rangsom*

In de vorige sectie is gebleken dat er vanuit een macro-perspectief een kwantitatief verschil bestaat tussen teksten uit de Karel- en Arturepiek inzake rijmvocabulaire. Aangezien deze bewering nogal abstract is, rijst de evidente vraag welke rijmwoorden dan precies aanzienlijk tot dit verschil bijdragen, oftewel: welke rijmwoorden zijn eerder typisch en onderscheidend voor het ene dan wel het andere deelcorpus? In een overzichtsartikel biedt de corpuslinguïst Kilgariff (2001) een kritisch overzicht van de tests die voor het beantwoorden van een dergelijke vraag geschikt zijn.

Kilgariff raadt aan gebruik te maken van een techniek die is gestoeld op de *Mann-Whitney ranks test* (ook bekend als de *Wilcoxon test* of *rank sums test*). Om na te gaan voor welke woorden het gebruik significant verschilt tussen corpus X en Y is het nodig beide corpora te verdelen in samples van een gelijkaardige grootte – het precieze aantal samples per deelcorpus hoeft niet gelijk te zijn. In het fictieve voorbeeld van Kilgariff (2001, 237-238) tellen twee corpora X en Y bijvoorbeeld elk 5 samples. Om na te gaan of het voorkomen van een bepaald woord verschilt in X en Y wordt eerst het voorkomen van dit woord in elk van de samples geteld. De samples worden vervolgens op basis van deze telling gerangschikt waardoor aan elk van de 10 samples een rangnummer wordt toegekend – zie onderstaande Tab. 22, p. 255).

Mann-Whitney ranks test, Wilcoxon test, rank sums test

Telling	3	3	12	13	15	18	24	27	33	88	**Rangsom**
Corpus	Y	Y	X	Y	X	X	X	Y	Y	X	
Rang(X)			3		5	6	7			10	31
Rang(Y)	1	2		4				8	9		24

Tabel 22: Illustratie van de *Wilcoxon ranks sums*-proef op basis van twee fictieve corpusjes. Dit voorbeeld is integraal overgenomen uit Kilgariff (2001, 237-238).

Vervolgens worden de eigenlijke waardes in deze pientere proef genegeerd (bv. 3, 12, . . . , 88) en wordt uitsluitend nog gekeken naar de rangorde van de samples. Het voordeel van deze (niet-parametrische) test is dat extreme waardes (zoals in dit geval 88) geen onrealistisch groot effect op de uitkomst hebben. Concreet worden de vijf rangordes die geassocieerd zijn met de X-samples samengeteld en hetzelfde gebeurt voor de vijf Y-samples ('Rangsom'). De resulterende waardes

(24 < 31) kunnen gebruikt worden voor het beschrijven van de mate van verschil tussen de betrokken deelcorpora – het is ook met deze waardes dat de statistische significantie kan worden berekend.

Deze proef wordt slechts uitgevoerd voor een aantal voldoende frequente woorden (om een teveel aan nulwaardes te vermijden). De resulterende waardes voor de rijmwoorden zullen een 'kop' en een 'staart' vertonen die geassocieerd zijn met de woorden waarvan het gebruik typisch is voor het eerste dan wel het tweede deelcorpus. Het zijn deze woorden (met extreme uitkomsten) die interessant kunnen zijn voor het beschrijven van de belangrijkste verschillen tussen de corpora X en Y.

Ik heb deze test uitgevoerd op de hierboven beschreven Artur- en Karelepische samples, met uitsluiting van Velthems *Merlijn*-sample dat zich afgaande op de cluster-analyse als een weinig representatieve *outlier* gedroeg. Bovendien heb ik slechts het eerste van de zes beschikbare samples uit *Lorreinen-I* benut, omdat de uitkomst al snel duidelijk maakte dat de analyse anders een al te groot gewicht toekende aan deze uitgebreide tekst. De Karelepische samples die in bovenstaande afbeelding onder de Arturepiek clusterden, heb ik echter niet verwijderd om de berekeningen niet al te zeer op voorhand bij te sturen. In onderstaande tabel wordt het resultaat van deze analyse getoond voor de 1.000 meest frequente rijmwoorden. Voor elk van deze rijmwoorden is de *rank sums*-test berekend en de rijmwoorden werden op basis van deze uitkomst gerangschikt.[9] Voor elk uiteinde van deze rangschikking worden hier de rijmwoorden met de meest extreme waardes gerapporteerd: tweemaal 5% van de 1.000 oorspronkelijke rijmwoorden oftewel 50 voor respectievelijk de Karel- en Arturepiek.

Reeds een eerste, vluchtige blik op bovenstaande tabellen maakt duidelijk dat hier geen volslagen nieuwe gegevens worden aangereikt – vgl. Tab. 23 (p. 257) en Tab. 24 (p. 258). Wel kan het interessant zijn om na te gaan hoe deze zuiver kwantitatief verzamelde gegevens zich verhouden tot de 'handmatige' karakterisering van beide tekstsoorten in vroeger, traditioneler onderzoek. Een interpretatie van het resultaat in deze tabellen helpt bovendien om zicht te krijgen op de verschillen in vocabularium die de clustering in de voorgaande sectie verklaren.

De rijmwoordenschat van de Karelepiek wordt duidelijk gekenschetst door een martiaal vocabulaire waarin de veldslag en de (lichamelijke) gevolgen daarvan centraal staan: DOOD, GESCHAL (van de strijdhoorn), KRACHT en MOED. Opvallend is het groot aantal rijmwoorden dat schijnt voort te komen uit stereotiepe epithetische persoonsaanduidingen van strijders (DEGEN, WIGANT) of de waarderende omschrijving van krijgers: FIER, FIJN, GOED, KOEN, SNEL en STERK (vgl. Van den Berg 1992b, 409-410). Een leuk detail is bovendien dat ei-

9 Technische noot: gebruikt is de *ranksums*-functie uit *SciPy* (`http://www.scipy.org`) (zonder *tie-handling*), waarbij gerankt is op de *z*-statistiek. In navolging van Kilgariff (2001, 246-248) is geen rekening gehouden met de *p*-waardes van de uitkomsten, omdat het verwerpen van de nulhypothese hier sowieso niet erg betekenisvol is.

Rangorde	Rijmwoord	Rangorde	Rijmwoord
1	PrName	26	DEGEN
2	MIJN	27	WIJL
3	FIJN	28	KRACHT
4	HANGEN	29	MINDER
5	MOED	30	PALEIS
6	GOED	31	OVERAL
7	WIJN	32	GEMOED
8	ACHTEN	33	GEBOD
9	GESCHAL	34	VERRADEN
10	VAST	35	BEKEND
11	HEMELRIJK	36	RIJK
12	WIGANT	37	LACHEN
13	SARACEEN	38	NEEF
14	MAAG	39	HEIDE
15	SULTAN	40	LIEF
16	GRAAF	41	ZWAAR
17	FIER	42	IN
18	CHRISTENHEID	43	ZENDEN
19	STERK	44	KROON
20	KIND	45	NET
21	CHRISTEN	46	GETELT
22	SNEL	47	KOND
23	DOOD	48	KOEN
24	WAAR	49	TEZAMEN
25	GINDER	50	POORT

Tabel 23: Resultaat van een *ranksums*-analyse (deel 1) van de 1.000 meest frequente rijmwoorden in de hierboven besproken sampleverzameling. Hier worden (in volgorde) de 50 rijmwoorden weergegeven die het meest typisch zouden zijn voor de rijmwoordenschat in de Middelnederlandse Karelepiek.

Rangorde	Rijmwoord	Rangorde	Rijmwoord
1	DEEL	26	DAAR
2	DEZE	27	UUR
3	WEER	28	DAARMEDE
4	AVONTUUR	29	WETEN
5	WEZEN	30	HOF
6	ALBLOOT	31	WONDER
7	KENNEN	32	WAAN
8	DAARNAAR	33	BEDIE
9	ZIJ	34	DAARBIJ
10	GENEZEN	35	DAN
11	SCHIJNEN	36	NU
12	BEZIEN	37	DAARAF
13	ZAAK	38	SLAG
14	NEDER	39	ALDAAR
15	BLIJ	40	OORLOF
16	SCHEIDEN	41	VOORWAAR
17	PLEIN	42	WAART
18	LIGGEN	43	TERNEDER
19	DAARBINNEN	44	TROUW
20	JONKVROUW	45	GEWOON
21	GIJ	46	MANIER
22	DE/HET	47	EEN
23	BIDDEN	48	RIJDEN
24	UITERMATE	49	KASTEEL
25	VINDEN	50	TOEN

Tabel 24: Resultaat van een *ranksums*-analyse (deel 2) van de 1.000 meest frequente rijmwoorden in de hierboven besproken sample-verzameling. Hier worden (in volgorde) de 50 rijmwoorden weergegeven die het meest typisch zouden zijn voor de rijmwoordenschat in de Middelnederlandse Arturepiek.

gennamen naar verhouding veel vaker voorkomen in Karel- dan in Arturromans (PRNAME). Dat kan samenhangen met het relatief groter aantal verschillende Karelepische personages, die bovendien vaak via complexe familierelaties aan elkaar gelieerd zijn (KIND, MAAG en NEEF). Hieraan gerelateerd treffen we ook rijmwoorden aan die verband schijnen te houden met een feodale thematiek (GRAAF, KROON) die eerder eigen is aan de Karel- dan aan de Arturepiek.

De Karelepiek schijnt bovendien gekarakteriseerd door een religieuze thematiek (HEMELRIJK en eventueel GEBOD). Er bestaat een scherpe oppositie tussen de Karolingische *milites christi* (CHRISTEN, CHRISTENHEID) en de ongelovigen waartegen een heilige strijd wordt geleverd (SARACEEN, SULTAN). Uit het rijmvocabulaire van de Karelepiek spreekt dus een grote thematische eenheid en focus (Van den Berg 1992a, 408).

Een dergelijke focus is iets minder makkelijk uit de rijmwoordenlijst voor de Arturepiek te distilleren. Zo is er een groot aantal rijmwoorden dat zich moeilijk rechtstreeks met de inhoud van teksten laat verbinden, zoals DAAR, DAARNAAR, NEDER, DE/HET en talloze andere eerder inhoudsarme rijmwoorden (Van den Berg 1992a, 410-411). Niettemin zijn er ook hier rijmwoorden die zich laten verbinden met strijdgewoel (GENEZEN, SLAG). Waarschijnlijk moeten we hier vooral denken aan (twee)gevechten op een kleinere schaal dan de grootschalige Karelepische veldslagen. Enkele rijmwoorden schijnen duidelijk samen te hangen met de klassieke hoofse thematiek van de Arturepiek: HOF, KASTEEL en TROUW aan de JONKVROUW. Ook de typische queestethematiek van de Arturroman schijnt door in enkele rijmwoorden die we kunnen in verband brengen met de (exotische) reizen van Arturs ridders: OORLOF, SCHEIDEN en RIJDEN, richtingaangevend WAART en eventueel ook AVONTUUR, PLEIN en WONDER.

9.2 DE ZAAK-HEEROMA

Laat ons even terugkeren naar de cluster-analyse hierboven (Fig. 71, p. 253). Hieruit bleek dat er duidelijk een macro-verschil bestaat tussen de rijmwoordenschat die door dichters werd gehanteerd in enerzijds de Middelnederlandse Karelromans en, anderzijds, de romans uit de Arturepiek. De daaropvolgende rangsom-analyse toonde aan dat de meest typerende rijmwoorden voor elke tekstsoort soms goed te interpreteren waren vanuit de dominante verhaalinhouden en thematiek in deze tekstsoorten – Tab. 23 (p. 257) en Tab. 24 (p. 258). Toch was de clustering hierboven niet absoluut: een tweetal Karelteksten maakten zich namelijk los van de andere Karelromans en clusterden in het boomdiagram met enkele Arturromans. Dat *Karel en Elegast* zich onder dit duo bevond is natuurlijk fascinerend, gezien de these van Heeroma dat deze Karelroman door dezelfde dichter werd geschreven als de *Moriaen* en de *Lantsloot*. Zelfs in het algemene boomdiagram

(Fig. 71, p. 253) was namelijk reeds sprake van een duidelijke stilistische toenadering tussen deze teksten, want zij werden op een laag niveau samengevoegd.

Laat ons eens inzoomen op de groep vertakkingen die naast een aantal Arturromans (o.m. FERG en de *Walewein*-delen) ook de Karelteksten ELEGAST en VLAAIOL bevatte. We kunnen nu ook de rest van de overgeleverde rijmwoorden uit deze teksten inschakelen, want tot nu toe werkten we slechts met steekproeven. Om een beter zicht te krijgen op de inhoudsonafhankelijke aspecten van de stijl in deze werken, beperken we ons tot de 100 meest frequente rijmen in deze samples.[10] Onderstaande figuur toont het resultaat van de cluster-analyse van dit materiaal (Fig. 72, p. 261). Het boomdiagram leert dat Maerlants MERLM stilistisch relatief ver staat van de andere werken in deze groep en samen met VLAAIOL reeds in de eerste vertakking wordt afgesplitst – op het afwijkende sample MERLM_S2 na.[11] Eenzelfde effect doet zich voor bij Pennincs aandeel in de *Walewein*: zijn samples (WALA) distanciëren zich bij de op een na hoogste vertakking van de rest van de teksten. Relatief hecht is de cluster (uiterst links) met HAGH, MOR en ELEGAST. Op een iets hoger niveau sluiten respectievelijk FERG en WALB zich bij die groep aan.

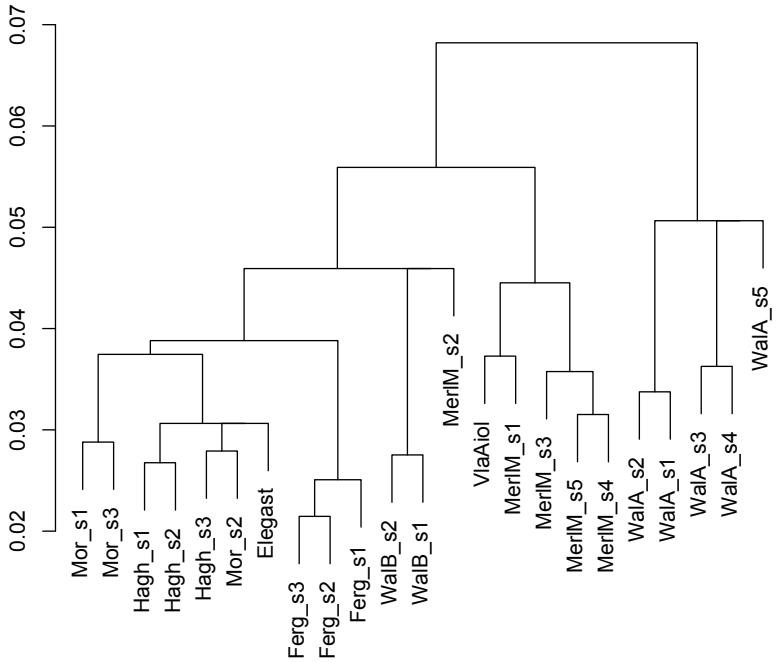
Voorgaand selectieproces levert zo een kleine groep van vijf ridderromans op die ten opzichte van de rest van het corpus stilistisch relatief coherent blijken: HAGH, MOR, ELEGAST, FERG en WALB.[12] Op dit moment loont het de moeite om met een correspondentie-analyse na te gaan of er binnen deze groep gelijkenissen tussen teksten optreden die eventueel indicatief kunnen zijn voor een enkelvoudig auteurschap.

De onderstaande figuur bevat het strooidiagram dat resulteert uit een correspondentie-analyse van HAGH, MOR, ELEGAST, FERG en WALB (Fig. 73, p. 262). Voor deze analyse heb ik de beschikbare samples nogmaals in tweeën opgedeeld (ca. 780 rijmwoorden) om na te gaan of we ook tussen dergelijk kleine samples interessante overeenkomsten zichtbaar kunnen maken. Deze correspondentie-analyse realiseert een duidelijk onderscheid tussen Vostaerts WALB (ruitjes linksboven) en FERG (driehoekjes linksbeneden). Beide teksten zijn bovendien duidelijk te onderscheiden van een sample-amalgaam centraal rechts in de figuur, bestaande uit de samples MOR, HAGH en ELEGAST. De overeenkomsten tussen de hoogfrequente rijmwoordenschat in deze teksten – niettegenstaande de kleine sample-grootte – zijn zo sterk dat zij nog

10 Ik laat MOU in dit hoofdstuk verder buiten beschouwing (zie ook vorige hoofdstuk) en doe liever geen expliciete uitspraken over het auteurschap of de oorspronkelijkheid van de tekst.

11 Wegens zijn beperkte lengte laat ik de *Vlaamse Aiol* hier verder buiten beschouwing. Verdere experimenten leerden dat de stilistische toenadering van deze tekst tot MERLM in bovenstaande figuur niet significant was.

12 Verdere experimenten zijn ook uitgevoerd met WALA maar het probleem was dat deze tekst door zijn relatieve grootte de andere teksten uit balans bracht in de analyse. Alleszins was geen sprake van een duidelijke toenadering van WALA tot de ELEGAST-cluster of andere teksten.

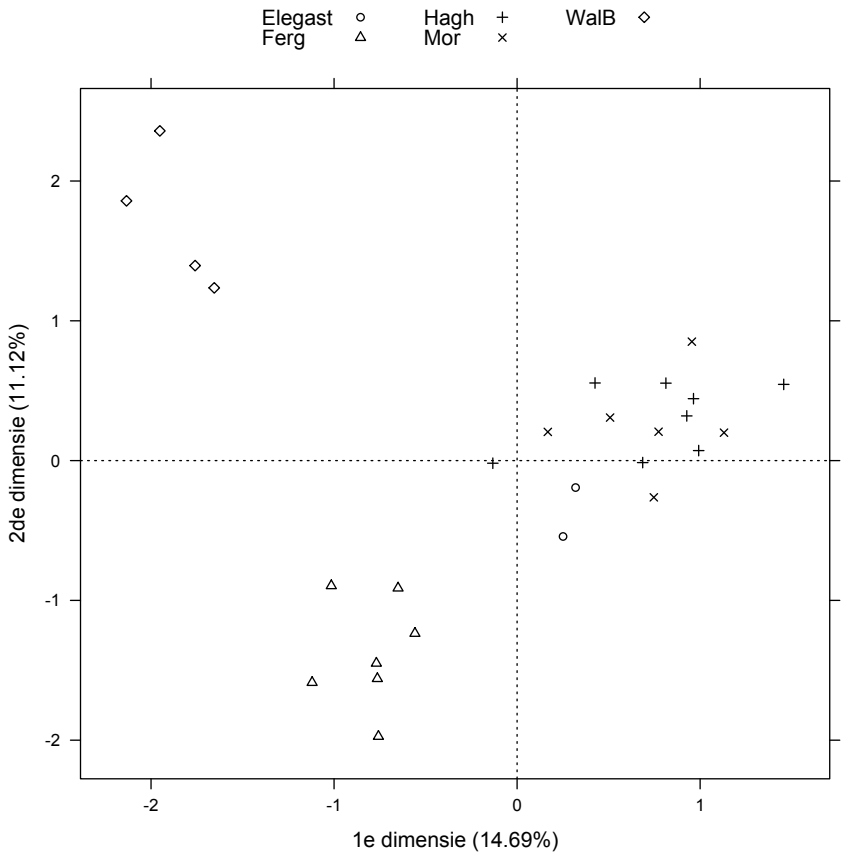


Figuur 72: Cluster-analyse van de samples (ca. 1500 rijmwoorden) uit MOR, HAGH, ELEGAST, FERG, WALA, WALB, VLAAIOL en MERL. Voor deze analyse werd een beroep gedaan op de 100 meest frequente rijmwoorden in deze samples.

nauwelijks van elkaar zijn te onderscheiden in het strooidiagram. De correspondentie-analyse oppert zo duidelijk de mogelijkheid van een identiek auteurschap voor de Middelnederlandse ridderromans *Karel ende Elegast*, *Roman van Moriaen* en *Lantsloot vander Haghedochte*.

Andere analyses (bijvoorbeeld met andere sample-groottes en vergelijkingsteksten) konden dit beeld niet wijzigen: de stilistische eenheid tussen deze teksten bleek zo goed als nooit te doorbreken. Zelfs wanneer men het aantal rijmwoorden in de tabel terugdringt tot 50, blijft de spreiding uit bovenstaande figuur behouden – vaak werd het effect nog sterker. De onderstaande opsomming bevat een lijst van de 50 meest frequente rijmwoorden uit deze teksten:

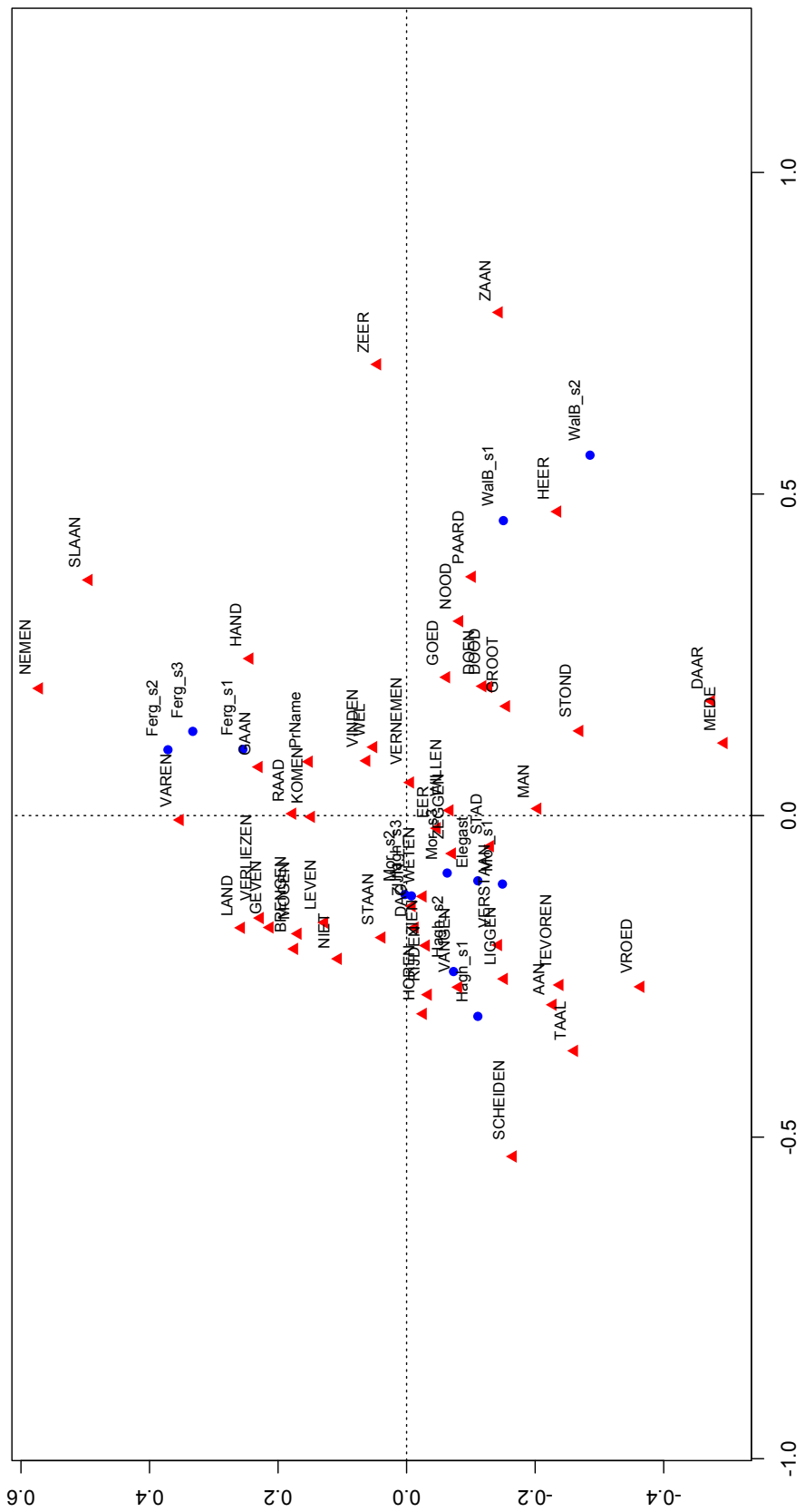
> PRNAME, ZIJN, DOEN, KOMEN, ZIEN, WEL, NIET, MOGEN, GROOT, GAAN, STAAN, HEER, MAN, STAD, ZEER, DAG, LAND, LEVEN, GOED, VINDEN, HOREN, DAAR, VAREN, LIGGEN, BRENGEN, RIJDEN, VERSTAAN, VERLIEZEN, VERNEMEN, STOND, VANGEN, HAND, GEVEN, NEMEN, ZAAN, SCHEIDEN, VROED, DOOD, WE-



Figuur 73: CA van relatief kleine samples (ca. 700 rijmwoorden, zie boven) uit MOR, HAGH, ELEGAST, FERG en WALB. Voor deze analyse werd een beroep gedaan op de 100 meest frequente rijmwoorden in deze samples.

TEN, TEVOREN, SLAAN, MEDE, TAAL, NOOD, AAN, EER, ZEGGEN, RAAD, PAARD, WILLEN

Deze rijmwoorden werden gebruikt voor de correspondentie-analyse waarvan het resultaat hieronder is afgebeeld (Fig. 74, p. 263). Het lijkt redelijk te claimen dat deze hoogfrequente rijmwoorden sterk inhoudsonafhankelijk zijn, want het optreden ervan lijkt nauwelijks gekoppeld aan de inhoud van deze teksten. Deze is echter niet erg overzichtelijk want naast de eigenlijke samples (bolletjes), zijn de ook de ladingen van de individuele rijmwoorden afgebeeld (driehoekjes). Een dergelijke visualisering is vanzelfsprekend chaotisch maar wel erg belangrijk om te verifiëren of de verschillen tussen groepen samples niet veroorzaakt worden door een te klein aantal, mogelijk te zeer inhoudsafhankelijke rijmwoorden. Dat deze figuur hier en daar moeilijk leesbaar is, lijkt in dit opzicht een goed teken, want de spreiding van de teksten blijkt veroorzaakt door een mengelmoes aan betekenisarme rijmwoorden.



Figuur 74: Correspondentie-analyse van de 50 vaakst voorkomende rijmwoorden (vgl. de lijst in de hoofdtekst) in Ferg, WalB, Hagh, Mor en Elegast (voor samples van ca. 1500 rijmwoorden, afgebeeld als bolletjes in het diagram). In dit diagram zijn ook de ladingen van de rijmwoorden (als driehoekjes) weergegeven. Uit de spreiding van deze driehoekjes kunnen we opmaken welke rijmwoorden typisch zijn voor welke (groep) werken.

Typische rijmwoorden voor FERG zijn NEMEN, SLAAN, VAREN of GAAN, terwijl Vostaert in WALB eerder toevlucht heeft gezocht tot rijmwoorden als ZAAN, HEER of ZEER. Voor de cluster met HAGH, ELEGAST en MOR blijken dan weer andere rijmwoorden typerend, zoals SCHEIDEN, TAAL, AAN en TEVOREN. Als deze teksten wat betreft dergelijke rijmwoordfrequenties sterke gelijkenissen vertonen, kan men zich de vraag stellen of deze niet door een identiek auteurschap zouden kunnen verklaard worden (vgl. Van Oostrom 1981, 57-62). Andere (bijvoorbeeld inhoudsgerelateerde) verklaringen lijken op het eerste zicht moeilijker te beargumenteren. De auteursvraag werpt zich niettemin vooral op omdat de hier geanalyseerde redacties van deze ridderromans erg diverse materiële origines hebben: incunabel A van *Karel ende Elegast* is een relatief jonge druk (Fig. 70, p. 249), *Lantsloot vander Haghedochte* is slechts bekend uit de fragmentarische resten van een enkel handschrift, terwijl de hier gebruikte Haagse *Moriaen* bekend staat als een door de compilator bewerkte versie. Voor de *Elegast* komt daar nog eens bij dat de tekst traditioneel tot een andere epische tekstsoort wordt gerekend dan de *Moriaen* en de *Lantsloot*.

Dat deze teksten, zo divers in oorsprong, toch zulk een opvallende stilistische verwantschap vertonen, pleit voor een toeschrijving aan één en dezelfde dichter. Het dialect van de drie teksten schijnt zich trouwens niet tegen een gelijkstelling van hun dichters te verzetten. Van Sterkenburg (1982, 28) karakteriseerde de taal van de *Lantsloot*-fragmenten als Zuidwestelijk Vlaams, mogelijk eerder West-Vlaams. Merk op dat de *Lantsloot* daarvoor lang in Holland is gesitueerd, hetgeen door Draak (1974, 373-374) nog werd aangevoerd als een argument tegen Heeroma's these. De oorspronkelijke *Moriaen* zou ontstaan zijn 'in noordelijk West-Vlaanderen, hoogstwaarschijnlijk te Brugge' (Paardekooper-van Buuren & Gysseling 1971, 37), al schijnt mij de bewijslast voor een specifiek Brugse lokalisering niet dwingend. Kienhorst (1988, 80) plaatste de *Elegast* in zijn repertorium op gezag van Van den Berg weinig specifiek in Vlaanderen.

9.3 EERHERSTEL VOOR ICARUS?

9.3.1 *Reminiscenties*

Het identieke auteurschap van *Moriaen*, *Karel ende Elegast* en *Lantsloot vander Haghedochte* is geen nieuwe hypothese. Daarom zal ik nu het werk bespreken van onderzoekers die zich eerder over deze auteurskwestie hebben gebogen. Het zal blijken dat verschillende mensen aan de hand van traditioneel onderzoek reeds hadden gepleit voor de attributie die hier via kwantitatieve weg is beargumenteerd. Belangrijk in deze context is een bundel postuum verschenen studies van K.H. Heeroma: *Moriaen, Lantsloot en Elegast. Luisterend gelezen* uit 1973, eerst verschenen in de 89e jaargang van het *Tijdschrift voor Nederlandse Taal-*

en Letterkunde. In deze studies heeft Heeroma uitgebreid het identieke auteurschap voor deze teksten beargumenteerd, eerst voor *Moriaen* en *Elegast*, vervolgens ook voor *Lantsloot* (vgl. Van Oostrom 1998). Zijn betoog stoelde in het bijzonder op gelijk(luidend)e versregels en overeenkomsten in vertel- en verstechniek. Het komt mij in deze studie vanzelfsprekend goed uit dat Heeroma de stilistische overeenkomsten tussen deze drie teksten al bijzonder gedetailleerd in kaart heeft gebracht – en dat anderen na hem belangrijke aanvullingen hebben geleverd.[13]

Dat Heeroma eenzelfde 'dichterlijke stem' het eerst herkende in *Moriaen* en *Elegast*, verrast bij nader inzien niet. In de *Elegast* wordt Karel de Grote in het holst van de nacht door een engel uit bed gehaald en verzocht uit stelen te gaan. Een verwarde Karel verlaat te paard heimelijk het slot Ingelheim; opvallend genoeg zijn de deuren en stalpoort geopend (*Noch poerte diene wederstoet*, v. 141) en kan hij onopgemerkt wegkomen. Onderweg wordt Karel plots door spijtgevoelens overvallen en denkt onwillekeurig aan Elegast, een edelman die hij om een kleinigheid verbannen heeft en uit zijn erfelijke landgoeden gezet. Er is duidelijk sprake van een feodale thematiek. Via een innerlijke monoloog verneemt het publiek dat Karel met spijt beseft dat hij zijn leenman zo heeft veroordeeld tot een bestaan als dief. Karel komt in een donker woud plots op een open plek en merkt dat een schrikwekkende verschijning zijn kant komt opgereden: een duistere ridder met *wapenen swart als colen*:

> Swart was helm ende schilt
> Die hi aenden hals hilt;
> Sinen halsberch mochtmen loven;
> Swart was den wapenroc daer boven;
> Swart was dors daer hi op sat;
> (Claassens 2002, vs. 277-281)

Karel *waende dat die duvel ware* (v. 286) want *Tis al swert, peert ende man*. De verteller verplaatst zich vervolgens in Elegasts gedachten: Wie is die ridder met die indrukwekkende uitrusting? Is hij verdwaald? Wanneer zij elkaar kruisen, weigeren zij ieder zichzelf voor stellen. Het komt tot een heftig tweegevecht, waarin Karel de onbekende ridder weet te ontwapenen. Daarop moet de zwarte ridder zijn identiteit prijsgeven: *Heere, ic hete Elegast* (v. 474). Elegast verhaalt zijn lotgevallen na zijn verbanning:

> Daer ic bi leve, moet ic stelen
> Mer so vele isser an:
> En steels geenen armen man
> Die bi sijnre pine leeft.

13 Alle citaten hier zijn naar Claassens (2002) (*Elegast*), Paardekooper-van Buuren & Gysseling (1971) (*Moriaen*) en Gerritsen et al. (1987) (*Lantsloot*).

Dat pelgrim ofte coopman heeft,
Dies laet ic hem ghebruken wel,
Mer ic en verseker niemant el.
(Claassens 2002, vs. 502-508)

Wanneer Karel zich op zijn beurt voorstelt, liegt hij echter over zijn identiteit en maakt zich aan Elegast bekend als Adelbrecht, een wrede roofridder die voor niets terugdeinst. Samen besluiten zij uit stelen te gaan. Belangrijk is dat Karel zijn identiteit tijdens de nachtelijke rooftocht weet te verhullen en dat een groot deel van de spanning in de roman hierdoor wordt gecreëerd. Er is sprake van een 'dubbele lens-effect' of dramatische ironie (Vekeman 1970–1971, 5-6): de verteller laat het publiek delen in de kennis over Karels identiteit maar laat Elegast hieromtrent in het ongewisse. Hierdoor kan het verhaalde steeds vanuit twee camera-standpunten belicht worden.

De *Roman van Moriaen* vertoont qua macro-structuur op het eerste gezicht weinig overeenkomsten met de *Elegast*. Toch wordt de moderne lezer die beide teksten naast elkaar legt door enkele micro-structurele gelijkenissen getroffen. Een zwaargewonde ridder bereikt Arturs hof in *Bertangen* en komt te paard de hofzaal ingestormd. Dat was mogelijk – zo merkt verteller terzijde op (v. 37) – omdat het daar de gewoonte was *datmen gene porte sloet*. Walewein snelt de ridder te hulp en heft hem uit het zadel. Arturs neef ontpopt zich prompt tot veldarts en dient de ridder in zijn mond een kruid toe, dat zo goed is *dattet stelpte mans bloet* (v. 114).[14] De ridder herstelt wonderwel onmiddellijk en vertelt beschaamd dat armoede hem na het verlies van zijn erfelijke landen tot een leven als dief had gedwongen. Hij lijkt eenzelfde lot als Elegast beschoren, al spaart deze ridder net geen pelgrims of kooplieden:

Wie so quam in min gemoet
Vordi penninge of ander goet
Waest pelegrim oft coepman
Ic dede den raet dat ict gewan
(Paardekooper-van Buuren & Gysseling 1971, vs. 161-164)

Op een dag ging hij echter een gevecht aan met een ridder die hem overwon en opdroeg zich aan Arturs hof te melden. Uit zijn beschrijving kan men opmaken dat het om Perchevael moet gaan. Na een discussie besluiten Lancelot en Walewein samen op zoek te gaan naar Perchevael. Op de negende dag van hun zoektocht ontmoeten zij plots een vervaarlijke verschijning: een zwaarbewapende ridder.

Hi was al sward ic segt v hoe
Sijn hoeft lichame ende hande

14 Onwillekeurig kan een lezer hier reeds denken aan het magische kruid dat Elegast in zijn mond steekt, waarmee hij op magische wijze het 'Latijn' van de dieren begrijpt (vs. 763ff).

Was al sward. sonder sine tande
Ende wapine ende scilt sekerlijc
Was al enen moer gelijc
Ende alse sward alse een rauen
(Paardekooper-van Buuren & Gysseling 1971, vs. 422-427)

Op zich is de parallelle beschrijving opvallend, maar veel te topisch om bewijskrachtig te zijn voor een identiek auteurschap. Interessanter in dat opzicht zijn de overeenkomsten in verteltechniek tussen beide romans. De zwarte verschijning uit de passage hierboven – slechts zijn tanden blijken wit – houdt halt en tracht Walewein en Lancelot te dwingen om zich bekend te maken. Hierdoor beledigd gaat Lancelot de strijd met hem aan. Walewein houdt afstand en via een interne monoloog vernemen we hoe Walewein denkt dat de zwarte ridder eerder *die duuel was dan een man* (v. 481). Er volgt een heftig tweegevecht waarin de frasering niet alleen sterk aan het tweegevecht van Elegast en 'Adelbrecht' doet denken, maar waarin de dichter zich een voortreffelijk stilist toont:

Al had middernacht gewesen
Ende alse donker alst noit was
Men mochte blomen ende gras
Hebben bekint vanden lichte
Die sparken vlogen so bedichte
Vten helmen ende vten swerden
Ende vilen neder opter eerden
(Paardekooper-van Buuren & Gysseling 1971, vs. 503-509)

Het gevecht wordt echter zo gevaarlijk dat Walewein de vechtjassen uit elkaar moet halen. Hij maakt de brutale zwarte ridder duidelijk dat hij gewoon hoofs om hun identiteit had moeten vragen. De zwarte ridder weet echter van geen ophouden en zegt dat hij geen angst heeft, al waren zij Lancelot en Walewein, twee rondetafelridders die hij nog nooit gezien heeft, maar waarvan hij weet dat men hen de grootste lof toezwaait. Walewein doorziet de ware toedracht van de situatie *in sinen moet*. De verteller stelt het prijsgeven van de identiteit van beide ridders echter nog even uit.

Daardoor ontstaat ook hier tijdelijk een dubbele lens-effect: het publiek deelt in de kennis van de verteller die niet alle personages hebben. De zwarte ridder verhaalt vervolgens zijn lotgevallen: hij is de zoon van Acglavael, de broer van Perchevael, die hem verwekt heeft bij een dame in *Moriane*. Omdat Acglavael door omstandigheden niet daar kon blijven, dreigen hij en zijn moeder uit hun erfgoed gezet te worden. Ook hier blijken feodale gevoeligheden mee te spelen. Omstandig vertelt Moriaen dat hij daarom op zoek is naar zijn vader. Pas honderden regels later wordt de identiteitsspanning (en het dubbele lens-effect) doorbroken en maken beide rondetafelridders

zich bekend (v. 881ff). Moriaen zakt door de grond van schaamte en verontschuldigt zich. Samen besluiten zij de zoektocht naar Acglavael verder te zetten. Natuurlijk is het dubbele vertelspectief als *fait divers* in de *Moriaen* veel minder ver doorgetrokken dan in de *Elegast*, waar het een van de belangrijkste elementen in de plot is. Toch roept deze subtiele narratieve overeenkomst vragen op over het verband tussen beide werken.

9.3.2 *Ohrenphilologie?*

Ongetwijfeld zijn het dergelijke soms kleine, maar steeds treffende gelijkenissen die Heeroma's scherpe intuïtie – oftewel *Ohrenphilologie* (Draak 1974, 373) – hebben aangewakkerd. In zijn zoektocht naar gelijkenissen zijn echter veel meer stilistische overeenkomsten naar boven gekomen, waaronder in tweede instantie ook gelijkenissen met *Lantsloot vander Haghedochte*. Deze Lancelot-roman vormt – in tegenstelling tot de naar alle waarschijnlijkheid oorspronkelijke *Moriaen* en *Elegast* – een bewerking van de Franse *Lancelot en prose*. Daardoor vertoont deze roman mogelijk minder narratieve of thematische gelijkenissen met de andere twee teksten. Toch verzamelde Heeroma een pagina's lange lijst van overeenkomstige verzen uit de drie werken (Heeroma 1973, 95-90 & 117-127) met zowel vrijwel identieke versregels als sterk gelijkluidende verzen. Ook in het gebruik van 'overglijding' – de plots overgang van indirecte naar directe rede – zouden de drie teksten grote overeenkomsten vertonen (Heeroma 1973, 136-140). Het interessante van veel van Heeroma's parallellen is dat zij sterk genoeg zijn om als duidelijke gelijkenis beschouwd te worden maar anderzijds zijn vele niet woordelijk genoeg om een naïeve of slaafse stilistische navolging te suggereren.[15]

Heeroma's soms nogal wild beargumenteerde toeschrijvingen werden aanvankelijk neergesabeld in de vakliteratuur want té subjectief.[16] Wat betreft zijn werk over het *Gruuthuse*-handschrift trok Gerritsen (1969, 214) zelfs een parallel tussen Heeroma en Icarus. In de latere literatuur vinden we gaandeweg echter een voorzichtige herwaardering van verschillende van zijn theses. Zo luidde de derde stelling bij Van Oostroms invloedrijke dissertatie over *Lantsloot*:

> Heeroma's stelling dat *Lantsloot vander Haghedochte* en *Moriaen* van dezelfde auteur zijn, is weliswaar niet bewezen doch wel zo aannemelijk dat het onderzoek van de Middelnederlandse Arturromans er terdege rekening mee moet houden (Van Oostrom 1998, 108, n. 3).

15 Besamusca (2003, 27-28) hield in navolging van Van Oostrom de mogelijkheid van een identiek auteurschap open maar herkende in deze gelijkluidende verzen eerder vrijblijvend een gemeenschappelijke stilistische traditie.

16 Vgl. Van Sterkenburg (1982, 14-15) of Draak (1974, 375).

Op verschillende plaatsen in zijn studie leverde Van Oostrom (1998) dan ook bijkomende argumenten voor Heeroma's these.[17] Hoewel hij steeds voorzichtig bleef, was Van Oostrom nog in 1998 benieuwd naar het mogelijke resultaat van kwantitatieve onderzoekstechnieken in dezen: 'Maar wie weet wat mede door de digitalisering op dit gebied nog eens aan frappante convergenties aan de dag zal komen' (Van Oostrom 1998, 109). Van Sterkenburg (1987) had reeds een belangrijke aanzet gegeven voor dat digitaal onderzoek: op basis van digitale lemmalijsten stelde hij vast dat *Lantsloot* en *Moriaen* respectievelijk 58% en 65% van elkaars woordenschat gemeen hebben.

De hypothese dat *Lantsloot* en *Moriaen* door dezelfde dichter werden geschreven, pleit voor een extra laag in de interpretatie van de proloog van *Moriaen* in de *Lancelotcompilatie*. Daarin is eerst (vs. 1-22) overduidelijk de compilator aan het woord die zijn interpolatie inleidt. Van de regels 23-29 is echter onduidelijk of zij op de compilator dan wel de oorspronkelijke dichter van de *Moriaen* teruggaan:

> Ic wane die gene di Lancelote maecte,
> Dat hem in sijn dichten vaecte,
> Dat hi vergat ende achterliet
> Van Moriane dat scone bediet.
> Mi wondert wies si hen onderwinden
> Die dichten wilt ende rimen vinden,
> Sine volbrachten daer af die tale.
> (Paardekooper-van Buuren & Gysseling 1971, vs. 23-29)

Traditioneel identificeert men de *ic* (v. 23) in deze verzen als de compilator, die in deze passage verwijst naar de Middelnederlandse *Lanceloet*-vertaler (*die gene di Lancelote maecte*). Deze referentie aan een Lancelot-dichter heeft immers veel betekenis binnen de *Lancelotcompilatie*, waar de Middelnederlandse *Lanceloet* inderdaad voor de *Moriaen* is opgenomen (Besamusca 2003, 85). Bekeken vanuit de context van de *Lancelotcompilatie* is deze lezing hoe dan ook correct, want de compilator heeft deze verzen handig aangewend om de *Moriaen*-interpolatie in te leiden.

Als deze verzen echter het begin vormden van de oorspronkelijke *Moriaen*, hadden zij in die context een heel andere betekenis. Dat zou zeker het geval zijn wanneer de *Moriaen*-dichter op dat moment reeds zijn *Lansloot vander Haghedochte* had afgewerkt. De oorspronkelijke dichter presenteerde zijn *Moriaen* dan als een vervolg op zijn Middelnederlandse *Lantsloot* en opende met een sneer aan het adres van de Franse (!) *Lancelote*-dichter (*dat hem in sijn dichten vaecte*). Bij zijn eerdere vertaling van de *Lancelot en prose* had de *Moriaen*-dichter

17 Van Oostrom (1998, 202) opperde in zijn studie dat ook de *Riddere metter mouwen* door de *Moriaen*-dichter kan zijn geschreven. Nieuw onderzoek kan zoals gezegd pogen de invloed van de compilator te neutraliseren. Voor mogelijke toevoegingen van de compilator, zie Besamusca (2003, 116-117).

namelijk gemerkt dat de Franse dichter het *scone bediet van Moriane* niet kende (*dat hi vergat ende achterliet*). Hij vraagt zich bovendien af waarom andere Middelnederlandse dichters (*wies si hen onderwinden / die dichten wilt ende rimen vinden*) het verhaal niet eerder hebben neergeschreven (*sine volbrachten daer af die tale*). Op het niveau van de oorspronkelijke *Moriaen* is dus een heel andere interpretatie van deze proloog mogelijk.[18] Een dergelijke lezing zou een nieuw licht werpen op de poetica van de anonieme *Moriaen*-dichter, maar benadrukt ook nog maar eens hoe sluw de compilator zijn bronnen gerecycleerd kan hebben.[19]

Met deze proloog zijn we genaderd tot een voor deze studie heikel punt: de oorspronkelijkheid van de *Moriaen*. De tekst is buiten enkele schamele fragmenten (Van Buuren 1973) slechts bekend in zijn compilatie-versie. De compilator schijnt ten minste één belangrijke aanpassing in zijn grondtekst te hebben doorgevoerd. De eerste 22 verzen van de compilatie-*Moriaen* zijn in dezen verhelderend:

Ons maect cont die aventure
Van enen riddere nu ter ure,
Die Moriaen was geheten.
Som die boeke doen ons weten
Dat hi Perchevals sone was,
Ende som boke secgen oec das,
Dat hi was Acglavaels soene,
Perchevals broder was die goene.
Ende so was hi Perchevaels broder kint,
Want men wel ter waerheit vint
Dat Percheval ende mede Galaet
Beide bleven, dat wel verstaet,
Maget doet biden grale.
Om dit secgic van Perchevale,
Dat sijn sone nine mach wesen;
Maer ic hebbe in boken gelesen

18 De *Moriaen* bevat overigens wel meer 'terugverwijzingen' naar Lancelot-stof (Besamusca 2003, 87-89). Merk echter op dat ook deze verwijzingen oorspronkelijk alle kunnen teruggaan op *Lantsloot* in plaats van *Lanceloet*.

19 Ik moet volledigheidshalve aangeven dat verschillende collega's – waaronder de anonieme beoordelaars van deze studie – vriendelijk maar kordaat hebben geprotesteerd tegen deze nieuwe lezing van de *Moriaen*-proloog. Zij hebben mij echter niet van ongeldigheid van mijn argumentatie kunnen overtuigen. Cruciaal is dat de Middelnederlandse dichter in mijn lezing niet noodzakelijk kritiek uit op zichzelf of op zijn eigen *Lantsloot*-vertaling, maar wél op de Franse dichter en brontekst omdat die het personage Moriaen niet vermelden. Bovendien kan hij met de frase *wies si hen onderwinden / die dichten wilt ende rimen vinden* slechts verwijzen naar Middelnederlandse dichters want de Franse dichters schreven proza. In deze laatste verzen geeft de dichter mogelijk dus kritiek op andere Middelnederlandse dichters, want hij verwijt hen dat zij de Moriaen-stof nog nooit hebben neergeschreven. Anderzijds wil ik wel graag erkennen dat er naast mijn specifieke lezing van deze proloog, nog andere interpretaties – waaronder de traditionele – mogelijk blijven.

Dat Moriaen was sijns broder sone.
Ende an ene Morinne was die gone
Gewonnen; ende oec ten selven tide
Dat Acglavael sochte wide en side
Lancelote, die verloren was,
Alsmen hier te voren las.
(Paardekooper-van Buuren & Gysseling 1971, vs. 1-22)

In deze – gewoonlijk aan hem toegeschreven – proloog (tenminste vs. 1-22) vermeldt de compilator het bestaan van andere verhalen waarin Perchevael, en niet zijn broer Acglavael als vader van Moriaen optreedt. Waarschijnlijk verwijst hij hiermee naar de oer-versie van de *Moriaen*, want reeds door Te Winkel (1878, 22-38) werd beargumenteerd dat de tekst op verschillende plaatsen verraadt dat Perchevael oorspronkelijk als Moriaens vader moet zijn opgetreden. Dat aspect van zijn grondtekst bracht de compilator echter in verlegenheid, want elders in zijn tekstverzameling verscheen Perchevael als een maagdelijke (en dus per definitie kinderloze) ridder. De compilator zag zich daarom gedwongen om Perchevael als vader te vervangen door Acglavael (vgl. De Haan 1972).

De complexiteit van die aanpassing hoeft men echter niet te overschatten.[20] Vanuit rijmtechnisch oogpunt stelde de aanpassing alvast weinig problemen, aangezien de inwisseling van Acglav*ael* en Perchev*ael* geen rijmparen verstoorde. Zo kan men zich de vraag stellen of de inwisselbaarheid van beide eigennamen zelfs ten dele verklaart waarom precies Acglavael Moriaens nieuwe vader werd. Bij een nadere studie van de enige fragmenten buiten de compilatie (die sterk parallel lopen met de compilatie-versie) vroeg ook Van Buuren (1973, 43) zich reeds af 'of de in de compilatie geïnterpoleerde roman – anders dan bijvoorbeeld de geïnterpoleerde *Wrake van Ragisel* of *Riddere metter mouwen* – niet of minder ingrijpend is gewijzigd t.o.v. zijn voorbeeld'. Mogelijk moet men het oorspronkelijk karakter van de compilatie-*Moriaen* in het algemeen niet al te somber inschatten en is vooral in de slotpassages van de roman gesnoeid.[21] In het geval van een identiek auteurschap voor *Lantsloot*, *Elegast* en *Moriaen*, zou

20 Heeroma (1973, 180, n. 1) veronderstelde bijvoorbeeld reeds dat Acglavael al in Velthems grondtekst optrad, 'maar dan alleen als begeleider van Perchevael en niet als vader van Moriaen. De compilator had dan ook maar weinig te verschuiven om de 'maagdelijkheid' van Perchevael te redden'. Dat blijft een theoretische mogelijkheid maar bij gebrek aan internationale parallellen hiervoor – Perchevael heeft zelden 'begeleiders', tenzij Walewein – is deze niet erg overtuigend (met dank aan Marjolein Hogenbirk voor deze opmerking). Zie ook De Haan (1972, 211) en Besamusca (1991, 91-93) of Besamusca (2003, 85-90), ook voor de grote onzekerheid over het al dan niet optreden van andere personages (zoals Lancelot) in de oorspronkelijke *Moriaen*. Gezien de opmerkelijke stilistische coherentie van de *Moriaen*, komt het mij voor dat bijvoorbeeld de Lancelot-verhaaldraad helemaal niet door de compilator hoeft te zijn toegevoegd.

21 Vgl. De Haan (1972, 213) en Hogenbirk (2011, 42).

dat kunnen verklaren waarom nog zoveel stilistische overeenkomsten zichtbaar zijn.

9.4 CONCLUSIE

Dit hoofdstuk bood een verkennende studie van de rijmstijl in de Middelnederlandse Artur- en Karelepiek (§ 9.1.1). Na een overzicht van het beschikbare materiaal voor beide tekstsoorten van de Middelnederlandse ridderroman (§ 9.1.2), werd een cluster-analyse op deze teksten toegepast. Zonder enige moeite bracht de cluster-analyse een tweedeling aan in de teksten die grotendeels overeenkomt met het traditionele onderscheid dat men tussen Artur- en Karelteksten maakt (§ 9.1.3). Vervolgens werd met een rangsom-test nagegaan welk rijmwoorden typisch waren voor beide tekstsoorten (§ 9.1.4). Daaruit bleek dat meest typische rijmwoorden heel goed kunnen geïnterpreteerd worden vanuit de dominante thematiek in beide types epiek.

Het hierboven beschreven onderzoek bracht één geval aan het licht waarin een auteur mogelijk zowel Arturepiek als Karelepiek heeft beoefend (§ 9.3). Stylometrische analyses van de *Roman van Moriaen*, *Lantsloot vander Haghedochte* en *Karel ende Elegast* wezen op frappante overeenkomsten tussen de rijmwoordenschat van deze werken. Het is opvallend dat deze kwantitieve resultaten naadloos aansluiten bij de bevindingen van wijlen K.H. Heeroma. Op basis van een, hoewel sterk bekritiseerde, kwalitatieve analyse herkende Heeroma – en anderen in zijn kielzog – eenzelfde dichterlijke stem in deze ridderromans. Een grondige herwaardering van Heeroma's these lijkt dan ook nodig.

Natuurlijk staat het Heeroma's critici – *niders* noemde hij ze zelf spottend (vgl. Van Oostrom 1998, 107) – nog vrij om voor een meer behoedzame benadering te opteren: in theorie bestaat de mogelijkheid dat wij te maken hebben met drie contemporaine dichters uit eenzelfde 'Vlaamse school' die uit een vergelijkbaar register putten (vgl. Besamusca 1993, 27-28). Wie doorgaat op Heeroma's *élan* kan zich echter aan een hernieuwde, avontuurlijke lectuur van deze teksten wagen: bijvoorbeeld in de proloog van de compilatie-*Moriaen* vinden wij mogelijk een verwijzing naar de eerdere *Lantsloot*-vertaling van de dichter. Wat betreft de chronologie van dit oeuvre, kan men daarom rekening houden met een ontstaansdatum van *Moriaen* na *Lantsloot vander Haghedochte*, terwijl men de onmiskenbare literaire kwaliteiten van de *Elegast* mogelijk met een ervaren dichter op het toppunt van zijn kunnen in verband kan brengen. Men zou zich dus sterk kunnen maken dat de Nederlandse literatuurgeschiedenis hiermee een formidabel oeuvre rijker is, maar zoals steeds kunnen slechts doorgedreven onderzoek en nieuwe materiële vondsten hieromtrent uitsluitsel brengen.

Deel IV

TOT SLOT

10 BESLUIT

Tot slot van deze studie is het zinvol om in dit hoofdstuk terug te blikken op de resultaten van dit onderzoek (§ 10.1). Veeleer dan een opsommerige samenvatting te bieden van het voorgaande, zal ik mijn onderzoeksresultaten kritisch evalueren, onder meer tegen het achterdoek van Brinkmans Tien Geboden, die helemaal in het begin van deze studie aan bod kwamen (§ 2.2.3). In dit hoofdstuk ga ik vervolgens ook in op de lacunes van mijn onderzoek en stip ik interessante richtingen aan voor toekomstig onderzoek (§ 10.2).

10.1 WAT IS ONDERZOCHT?

10.1.1 *Hoogfrequente rijmwoorden*

Volgens Brinkmans Eerste Gebod – DE EENHEID VAN KOPIIST – kan men het auteurschap van middeleeuwse teksten slechts aan een stylometrische studie onderwerpen als men beschikt over afschriften van de bewuste teksten van dezelfde kopiist. Het onderzoek gerapporteerd in hoofdstuk 3 benadrukte inderdaad hoe groot het probleem van kopiistenvariatie is voor de stylometrische studie van middeleeuwse literatuur. Kopiisten genoten een relatief grote vrijheid om op verschillende niveaus aanpassingen door te voeren in afschriften, zodat er hoe dan ook geen sprake kan zijn van een-op-een-relatie tussen de taal van een afschrift en de taal van een autograaf. Deze observatie biedt wel interessante perspectieven voor de studie van stijl- of toch taalgebaseerde kopiistenidentificatie. Zoals gedemonstreerd in § 3.3 blijken stylometrische technieken die ontwikkeld zijn voor moderne teksten, soms geschikt om de kopieerstijl van middeleeuwse afschrijvers te modelleren. Verder onderzoek in deze richting loont de moeite, omdat computationele middelen ook in de paleografie een vernieuwend onderzoeksinstrument kunnen vormen (zie § 3.5).

Toch is Brinkmans Eerste Gebod in voorgaande studie op grote schaal geschonden. De basisaanname was daarbij dat de rijmwoordenschat van Middelnederlandse teksten in grote mate stabiel is gebleven in de overlevering, zoals wordt geïllustreerd door de parallelle tekstgetuigen die van werken zijn overgeleverd. Het rijmwoord kan binnen die visie worden opgevat als een 'reddingsboei' voor het stylometrisch auteursonderzoek naar Middelnederlandse rijmteksten (§ 4.1). Indien men de oppervlakkige spellingvariatie in Middelnederlandse rijmwoorden normaliseert via lemmatisering (§ 4.1.2), vormen rijmlexemen een veelbelovende uitvalsbasis voor het stijlonderzoek naar teksten

waarvan slechts veel jongere tekstgetuigen (mogelijk van verschillende kopiisten) en geen autografen zijn overgeleverd (vgl. Brinkmans Eerste Gebod, maar ook het Achtste Gebod, MINIMALE AFSTAND TOT DE AUTOGRAAF).

In hoofdstuk 2 heb ik toegelicht welke methodologische inzichten uit het internationaal stylometrisch onderzoek aan de basis lagen van deze studie. Sinds het pioniersonderzoek van Mosteller & Wallace (1964) bestaat er een brede consensus in de vakliteratuur dat 'functoren' erg aantrekkelijk zijn voor de kwantitatieve studie naar auteurschap (§ 2.1). Hoogfrequente morfemen zijn een goed uitgangspunt voor stylometrisch onderzoek, onder meer door hun goede spreiding en inhoudsonafhankelijkheid. Dat moderne lezers relatief onoplettend met dergelijke morfemen omspringen, draagt bij aan hun populariteit in het onderzoek. Langzaam maar zeker zijn deze opvattingen ook binnengesijpeld in de medioneerlandistiek, zoals bijvoorbeeld in Brinkmans Tweede Gebod inzake de INHOUDSONAFHANKELIJKHEID van de geanalyseerde tekstkenmerken (§ 2.2).

Vooral in hoofdstuk 2 is beargumenteerd dat functoren verschillende eigenschappen gemeen hebben met de meest frequente rijmwoorden in Middelnederlandse teksten. Verschillende Middelnederlandse rijmwoorden zijn namelijk zo frequent dat zij eveneens gekenmerkt worden door een goede spreiding en inhoudsonafhankelijkheid (§ 4.1.3). Mijn argumentatie bouwde hier in belangrijke mate voort op vroeger stilistisch onderzoek naar formulair taalgebruik, waaronder dat naar stoplappen. De hierboven voorgestelde selectiemethode voor rijmwoorden onderscheidt zich echter van dit voorgaand onderzoek doordat deze *data driven* is en niet afhankelijk van de menselijke interpretatie van bepaalde stijlfenomenen.

In hoofdstuk 4 is via een relevante testcasus aangetoond dat stylometrische technieken valide resultaten opleveren wanneer zij worden toegepast op de hoogfrequente rijmwoordenschat van teksten. Jacob van Maerlant en Filip Utenbroeke hebben in elkaars onmiddellijke omgeving een omvangrijke bijdrage geleverd aan de rijmkroniek *Spiegel historiael*. Omdat deze auteurs bijvoorbeeld qua beroep, dialect, scholing en brontekst zo dicht bij elkaar gestaan moeten hebben, was dit een ideale testcasus voor auteursherkenning. Het auteurschap van 'ongeziene' samples (vgl. Tiende Gebod, SAMPLES) uit deze teksten bleek met een eenvoudige techniek voor tekstclassificatie niettemin opvallend goed te achterhalen (§ 4.2). Daarbij kwamen we ook tegemoet aan Brinkmans richtlijnen inzake DE EENHEID VAN GENRE, EENHEID VAN ONTSTAANSDATUM en DE EENHEID VAN ONTSTAANSREGIO. Deze experimenten toonden wel aan dat er geen reden is tot overdreven optimisme, want – in overeenstemming met het Derde Gebod inzake TEKSTLENGTE en veel ander internationaal onderzoek – werd duidelijk dat (samples uit) teksten lang genoeg moeten zijn vooraleer ze betrouwbaar kunnen worden toegeschreven (§ 4.2.3).

In hoofdstuk 4 werd wel afgeweken van het criterium van de gelijke sample-grootte (Vierde Gebod). Daar werd immers gewerkt met samples die een gelijk aantal kapittels bevatten maar daarom geen gelijk aantal verzen (vgl. ook § 9.1.3). Op die manier stelden de samples geen willekeurige reeks opeenvolgende verzen voor maar vielen zij samen met goed te interpreteren literair-historische grootheden, zoals een reeks kapittels of een boek uit de *Spiegel historiael*. Experimenten toonden aan dat het ook mogelijk is om met samples van een verschillende maar vergelijkbare grootte te werken voor auteursherkenning. Doorgedreven onderzoek naar het effect van tekst- en sample-grootte in auteursherkenning is daarom zeker gewenst.[1]

De resultaten van hoofdstuk 4 waren bemoedigend maar behandelden slechts twee auteurs, terwijl uit het internationale onderzoek bekend is dat auteursherkenning moeilijker wordt als men het aantal verschillende auteurs in experimenten opvoert (Luyckx & Daelemans 2011). In hoofdstuk 5 werd daarom ook Lodewijk van Velthem in het onderzoek betrokken, de Brabantse priester-dichter die de *Spiegel historiael* in het begin van de veertiende eeuw vervolledigde (§ 5.1). Door de veertiende-eeuwse Brabander Velthem te vergelijken met Utenbroeke traden we wel enkele geboden met de voet. Het Zesde en Zevende Gebod adviseren immers een eenheid van ontstaansregio en ontstaansdatum voor teksten waarvan het auteurschap wordt geanalyseerd. Deze Geboden zijn natuurlijk heel zinvol vanuit een theoretisch oogpunt: indien men dergelijke factoren zo constant mogelijk houdt, kan men immers veel sterker claimen dat men énkel auteurschap onderzoekt. Als men in de praktijk echter een anonieme tekst aan een bekende auteur wil toeschrijven, kan informatie omtrent de ontstaansregio en ontstaansdatum van de anonieme tekst anderzijds wel heel waardevol zijn.

De mogelijkheid bestond dat Velthem als vroegveertiende-eeuwse Brabander stilistisch iets makkelijker zou zijn te onderscheiden van Utenbroeke en Maerlant dan Velthems voorgangers onderling. Verkennende experimenten toonden echter aan dat dit op het eerste zicht niet zo was, want ons attributie-algoritme schreef in testscenario's geregeld Velthem-kapittels aan een verkeerde auteur toe (§ 5.1.2). Een doorgedreven foutenanalyse toonde zelfs dat de betreffende toeschrijvingstechniek vaak bij dezelfde kapittels – uit het vierde boek van de Vijfde Partie (§ 5.2.1) – de mist inging. Daarbij rees de vraag of dit vierde boek over de bekende Guldensporenslag eigenlijk wel van

1 Stamatatos (2009) onderscheidt in zijn overzichtsartikel twee benaderingen in computationele auteursherkenning: *instance-based approaches* en *profile-based approaches*. In het eerste geval wordt een anonieme tekst vergeleken met het oeuvre van een kandidaat-auteur dat is verdeeld in samples van dezelfde of een vergelijkbare grootte als de anonieme tekst. In het tweede geval wordt een anonieme tekst rechtstreeks met het hele (onverdeelde) oeuvre van een auteur. Bij een recente vergelijking van attributie-methodes constateerde Luyckx (2010, 112-113) dat de *profile-based approaches* – die dus (de profielen van) teksten van een heel andere omvang vergelijken – vaak betere resultaten opleverden (Luyckx 2010, 119).

Velthem was. Mogelijk kon een antwoord op deze vraag de foute toeschrijvingen van deze kapittels verklaren. Velthem integreerde immers geregeld materiaal van anderen in zijn eigen werk zonder bronvermelding.

Via een uitbreiding van ons 'geheugen-gebaseerd' attributie-algoritme (§ 5.2.2) konden we de afwijking van het vierde boek scherper in beeld brengen: in het vierde boek wijkt Velthems rijmwoordenschat consequent op dezelfde manier af van zijn gangbare stijl als in passages in de Vijfde Partie die aan andere bronnen zijn ontleend (§ 5.2.4). Dit resultaat was problematisch want er moest of iets niet kloppen met de meettechnieken, of er moest iets schelen aan het vierde boek. Daarom was het belangrijk om deze uitkomst terug te koppelen naar de tekst zelf en het vroegere onderzoek ernaar, of in de woorden van De Baere (1993, 48):

> Om de statistische methode veilig te hanteren is een eerste vereiste dat men voortdurend lezenderwijs contact houdt met de teksten. Anders dreigt men in abstracties te verzeilen die nauwelijks nog iets zeggen over de teksten, waarvan men is uitgegaan.

Dit literatuuronderzoek leerde dat het niet de eerste keer was dat men (onrechtstreeks) vraagtekens had geplaatst bij Velthems auteurschap voor het vierde boek (§ 5.2.5). E. van den Berg had eerder al de stilistische onregelmatigheden in dit boek opgemerkt en ook historici hadden zich al afgevraagd hoe Velthem aan dergelijk gedetailleerde informatie over de slag was geraakt. De kans is daarom meer dan reëel dat Velthem voor het vierde boek kon beschikken over een reeds bestaande brontekst, waarschijnlijk van een andere (Vlaamse?) auteur die een ooggetuige kan zijn geweest van het gewapend treffen op 11 juli 1302. Wat hier ook van waar is, dit hoofdstuk bracht een belangrijk probleem onder de aandacht: het gebrek aan absolute zekerheid over het auteurschap van veel middeleeuwse teksten. De *ground truth* of *gold standard* die men in het auteurschap van sommige historische teksten herkent, kan daardoor bijzonder verraderlijk zijn.

Velthems stijl kon dan nog afwijken in het vierde boek, mogelijk heeft Velthem gewoon een oudere tekst van zichzelf in zijn Vijfde Partie geïnterpoleerd en kan dat de afwijkende stijl van het vierde boek verklaren. Het is immers amper onderzocht of Middelnederlandse auteurs als Velthem diachrone of genre-gerelateerde stijlschommelingen vertonen in hun teksten. Wie geïnteresseerd is in de oeuvres van wereldlijke Middelnederlandse auteurs komt automatisch uit bij de veelschrijver Maerlant. Het oeuvre van deze dichter is niet alleen divers en omvangrijk maar geniet bovendien het voordeel dat het auteurschap van deze tekstverzameling – anders dan bij Velthem of Boendale – zelden of nooit ter discussie staat (§ 6.1).

In hoofdstuk 6 gingen we eerst in op de spreiding van Maerlants hoogfrequente rijmwoordenschat. Verschillende meettechnieken toon-

den aan dat de meest frequente rijmwoorden in Maerlants oeuvre als geheel inderdaad goed gespreid zijn over (samples uit) de individuele teksten (§ 6.2). Gedurende zijn carrière heeft Maerlant zich dus bediend van een relatief stabiele set hoogfrequente rijmwoorden. Toch kan deze stilistische stabiliteit niet al te ver worden doorgetrokken. Een correspondentie-analyse toonde immers aan dat Maerlants epische teksten – zelfs wat betreft de meest frequente rijmwoordenschat – uiteenvallen in drie tekstgroepen: de ridderepiek, de didactische teksten en zijn meer historiografische werk (een structuur die overigens al veel eerder in dit oeuvre was herkend). Deze groepen correleren dus met het genre van teksten, maar ook met de diachrone structuur van Maerlants werken, twee factoren die in dit oeuvre hoe dan ook moeilijk te ontwarren zijn (§ 6.3). Vanuit een poëticaal standpunt was dit een interessante observatie maar voor auteursherkenning hield zij een belangrijke waarschuwing in: de schrijfstijl van auteurs kan onmogelijk volledig los worden gezien van factoren als genre en ontstaansdatum, zelfs wanneer men metingen beperkt tot relatief inhoudsonafhankelijke teksteigenschappen (§ 6.4). Dat benadrukt het belang van het Tweede Gebod.

10.1.2 *Boendale*

Het tweede deel van deze studie is in zekere zin op de vlakte gebleven, omdat daarin slechts werd gewerkt met teksten waarvan de oorsprong relatief vaststond. (Velthems vierde boek heeft hier natuurlijk wel roet in het eten gestrooid.) Hoewel die stelling reeds op verschillende niveaus was genuanceerd, toonden verschillende experimenten wel aan dat stylometrische technieken op een zinvolle manier kunnen bijdragen aan de studie van middeleeuws auteurschap, in het bijzonder dat van de vele anonieme Middelnederlandse teksten die zijn overgeleverd. Gewapend met het inzicht in de mogelijkheden en beperkingen van de stylometrie uit het vorige deel, ging het derde deel van deze studie daarom in op enkele bekende casussen van betwist auteurschap in de medioneerlandistiek.

Hoofdstuk 7 behandelde de casus van de beruchte Antwerpse School, een coherente verzameling grotendeels anonieme geschriften uit de Scheldestad (eerste helft van de veertiende eeuw). De schepenklerk Jan van Boendale heeft hoe dan ook enkele van deze teksten geschreven – onweerlegbaar is zijn auteurschap voor bijvoorbeeld de *Lekenspiegel* – maar hoeveel meer van deze teksten aan Boendales bibliografie toegevoegd mogen worden, is tot op heden onduidelijk (§ 7.1.1). In dit hoofdstuk maakte ik gebruik van *Principal components analysis* maar ook van Burrows' invloedrijke Delta-procedure voor auteursherkenning (§ 7.1.2). Een driemensionele visualisering werd aangewend om het effect van enkele experimentele parameters te illustreren. De drie auteurs van de *Spiegel historiael* werden gebruikt

als *tertium comparationis* in deze experimenten, een aanbeveling uit Brinkmans Negende Gebod.

Wat betreft de didactische teksten uit de 'Antwerpse School' konden wij met de hier gebruikte onderzoekstechnieken geen markante stijlbreuken detecteren binnen het Antwerpse corpus (§ 7.3). De *Lekenspiegel* zou volgens de resultaten van deze studie qua rijmstijl hoe dan ook een erg centrale rol in dit corpus innemen, hetgeen niet onaannemelijk zou zijn als de auteur van dit omvangrijke werk, *casu quo* Boendale, ook de andere teksten in dit corpus heeft geschreven. De *Melibeus* en *Dietsche doctrinale* verwijderden zich wel tot op zekere hoogte van de rest van deze teksten, maar een analyse met de *tertium comparationis* relativeerde het belang van deze verwijdering.

In dit hoofdstuk kon ik zo geen sluitende argumenten leveren vóór Boendales auteurschap in dit corpus, want het blijft moeilijk om in te schatten hoe dicht de stijlen van twee teksten nu eigenlijk bij elkaar moeten liggen om tot een toeschrijving aan dezelfde dichter over te mogen gaan. Dit is een belangrijke vraag, ook voor toekomstig onderzoek, waarop deze studie niet kan claimen (een aanzet tot) een antwoord te hebben geformuleerd. Bij de huidige stand van het onderzoek lijkt het vooralsnog onmogelijk om voor dergelijke toeschrijvingen een universele constante of drempelwaarde te definiëren. Ieder computationeel onderzoek naar het auteurschap van anonieme teksten behoudt daarmee een groot *ad hoc*-karakter. Anderzijds kon dit onderzoek wel bijkomende stylometrische argumenten aandragen tégen het verwerpen van een enkelvoudig auteurschap binnen het didactische deel van dit corpus. Deze uitkomst biedt daarmee tot op zekere hoogte een 'bewijs uit het ongerijmde' voor de identificatie van Boendale als de auteur van de didactische geschriften uit het Antwerpse corpus.

Wat betreft de historiografische geschriften uit de 'Antwerpse School' had ik aanvankelijk gevreesd voor genre-gerelateerde artefacten (§ 7.4). Afgaande op het resultaat van het Maerlant-onderzoek in hoofdstuk 6 bestond immers de mogelijkheid dat wij op een verwijdering zouden stoten tussen de didactische en historiografische component binnen het Antwerpse corpus. Van een dergelijke scheiding zou onduidelijk zijn of die met een niet-identiek auteurschap kon samenhangen (§ 7.4.4). Verscheidene analyses demonstreerden dat de geanticipeerde verwijdering zich tot op zekere hoogte voordeed. Niettemin leek deze scheiding niet erg uitgesproken in vergelijking met de verwijdering van het corpus als geheel ten opzichte van het werk van andere auteurs.

Daarop bestond echter één uitzondering: de hoogfrequente rijmwoordenschat in een van de eerste delen van de *Brabantse yeesten* week duidelijk af van die van de rest van het Antwerpse corpus. Dat bleek ook het geval bij het gebruik van erg kleine samples en een erg beperkte set hoogfrequente rijmwoorden, zodat Boendales auteurschap

voor de bewuste passages onverwacht op de helling kwam te staan. In het verleden is het auteurschap van de anonieme *Brabantse yeesten* fel bediscussieerd, omdat de tekst slechts in jongere para-teksten aan Boendale is toegeschreven (§ 7.4.2). Zelfs de vijftiende-eeuwse voortzetter van de *Yeesten* had de mogelijkheid open gehouden dat meer dan één auteur aan de *Yeesten* had meegewerkt. Gedurende enkele eeuwen had men de *Yeesten* zelfs toegeschreven aan Nicolaas Noese, een relatief onbekende stadsklerk en tijdgenoot van Boendale. Willems, die het pleit uiteindelijk in Boendales voordeel heeft beslecht, heeft zelfs nooit definitief uitgesloten dat er sprake was van een meervoudig auteurschap.

Toch kon de uitkomst van het stylometrisch onderzoek Boendales auteurschap voor de gehele *Yeesten* maar moeilijk doen wankelen (§ 7.4.5). Een nadere blik op de uitkomst van de correspondentie-analyses toonde namelijk dat de geobserveerde stijlverschillen heel nauw samenhingen met de toon en inhoud van het 'verdachte' tekstdeel, zodat een meervoudig auteurschap zeker niet de enig mogelijke verklaring biedt voor de stijlverschillen. Een kritische analyse van de recente vakliteratuur over dit onderwerp bood bovendien ook bitter weinig aanknopingspunten om een meervoudig auteurschap voor de *Yeesten* te postuleren. Dit resultaat bracht mij daarmee in verlegenheid. Waar ik een afwijkend 'styloom' in hoofdstuk 5 nog had gebruikt om het auteurschap van Velthem voor zijn vierde boek in vraag te stellen, zag ik mij verplicht om het resultaat van de stylometrische analyse van een deel van de *Brabantse yeesten* te ontkrachten.

Toch lijkt het duidelijk dat de Niclaas-toeschrijving bij de huidige stand van het onderzoek niet zwaar genoeg kan doorwegen om Boendale het auteurschap voor de *Yeesten* (zelfs gedeeltelijk) te ontzeggen. De Niclaas-toeschrijving is uiteindelijk volledig terug te leiden tot één merkwaardige correctie in één handschrift. Wie meer duidelijkheid kan scheppen over de beweegredenen achter deze apocrief aandoende correctie, zou de studie van de Brabantse geschiedschrijving een grote dienst bewijzen.[2]

10.1.3 *De Artur- en Karelepiek*

In het daaropvolgende hoofdstuk 8 kwam de Middelnederlandse Arturepiek aan bod, een boeiende verzameling teksten waarover veel is gespeculeerd wat auteurschap betreft. Het uitgangspunt was de

2 Ik hou op dit moment ook rekening met een relatief prozaïsche verklaring: mogelijk bezat Nicolaas als Boendales naaste collega (en vriend?) gewoon een persoonlijk afschrift van de tekst met een bezittersaantekening die Nicolaas' naam vermeldde. Als de humanist Gevartius een dergelijk handschrift onder ogen heeft gekregen, kon hij Nicolaas op basis van een dergelijke aantekening makkelijk voor de auteur van de tekst hebben gehouden, met alle gevolgen vandien. Hoe dan ook biedt de handschriftelijke overlevering van de *Brabantse yeesten* nog een rijke voedingsbodem voor toekomstig onderzoek.

codex-Burgsteinfurt waarin twee *Merlijn*-teksten zijn overgeleverd van Jacob van Maerlant en van Lodewijk van Velthem (§ 8.1). Laatstgenoemde is waarschijnlijk de belangrijkste protagonist binnen de Middelnederlandse Arturepiek. Uitgaande van het materiaal door Maerlant en Velthem in de *Spiegel historiael* werd geprobeerd om de auteursovergang in het Burgsteinfurtse handschrift zichtbaar te maken. Daarvoor maakten we gebruik van een *windowing procedure* op basis van stilistische centroïdes, die schatplichtig was aan het werk van Van Dalen-Oskam & Van Zundert (2007) maar ook de Delta-procedure van Burrows (2002). Het beoogde effect werd in deze testcase tot op zekere hoogte bereikt, al was daarvoor de optimalisatie van technieken nodig. Velthems rijmstijl in zijn historiografische tekst bleek namelijk relatief uitgesproken verschillen te vertonen met de hoogfrequente rijmwoorden in zijn *Merlijncontinuatie*.[3]

Vervolgens werd deze *windowing procedure* ook toegepast op de *Lancelotcompilatie*, het bekende Haagse handschrift waarin een compilator via allerlei ingrepen een reeks bestaande Arturromans tot één geheel heeft verwerkt (§ 8.2). Veel van het onderzoek naar deze codex heeft zich toegelegd op de identiteit van de onbekende compilator: heel waarschijnlijk was dit Lodewijk van Velthem, aangezien deze in het colofon wordt vermeld als eerste bezitter van deze Arturcompilatie. Wanneer de *Lancelotcompilatie* werd geanalyseerd met werken van Velthem als referentiepunt konden de blootgelegde stijlschommelingen nagenoeg naadloos worden ingepast in het vroegere onderzoek naar de vermoedelijke ingrepen van de compilator. Eigenlijk resten er zo nog maar weinig argumenten die tégen Velthems compilatorschap pleiten, of in de woorden van Ockham: *entia non sunt multiplicanda*.

Een interessante vraag blijft of Velthem óók gelijk mag worden gesteld met kopiist B, de voornaamste afschrijver van het handschrift. De activiteiten van B vertonen immers verdacht veel overlappingen met die van de compilator: de hypothese dat B slechts lijdzaam de teksten overschreef die de compilator hem in het klad bezorgde, vloekt toch enigzins met het paleografische onderzoek dat B's actieve rol in de genese van het handschrift net heeft benadrukt (vgl. Klein 1998, 114-115). Toegegeven, zo wordt onduidelijk welke mysterieuze derde partij de rol van 'corrector' op zich heeft genomen, maar de vraag is of Velthem wel zo makkelijk in de rol van corrector kan worden gedwongen. Dat de experimentele correcties in het handschrift bijvoorbeeld een 'neiging tot prozaïsering' zouden verraden (Biemans

3 Ik permitteer mij nog een speculatieve opmerking: de Duitse kopiist van de proloog van Velthems *Merlijncontinuatie* in de codex-Burgsteinfurt was duidelijk niet vertrouwd met de plaatsnaam Velthem en noteert 'velchem' (vgl. Fig. 51, p. 202). Deze verschrijving zou men kunnen verbinden met kopiist B van de *Lancelotcompilatie*, in wiens handschrift de letter 't' makkelijk met de letter 'c' is te verwarren. Vgl. het uitzicht van het woord 'velthem' in Fig. 57, p. 216. Ik sluit daarom niet uit dat de schrijfwijze 'velchem' in Burgsteinfurt (mogelijk onrechtstreeks) teruggaat op een kopie van de bewuste *Merlijn*-teksten door kopiist B.

2005), pleit bijvoorbeeld precies tegen de gelijkstelling van de corrector met Velthem, een vrij conservatief auteur van wie veel rijmteksten maar geen prozateksten bekend zijn.

Overigens is in de vakliteratuur opvallend weinig gezegd over de historische situering van B's 'hulpkopiisten' (A, C, D en E). Veel onderzoekers hebben het schrift van deze hulpkopiisten bekritiseerd als slordig, onervaren en dus weinig professioneel (Biemans 2009, 243). Toch schijnt men er nog steeds van uit te gaan dat deze kopiisten door B werden aangestuurd in een soort economisch dienstverband (Klein 1995; Biemans 2009). Dat is verbazend, aangezien het amper voorstelbaar is dat Velthem voor een inferieur handschrift als 129 A 10 bereid was goed geld te betalen, laat staan dat met dat geld de schrijfarbeid van vier andere kopiisten in onderaanneming kon worden bekostigd. Veeleer dan in een professioneel samenwerkingsverband, kan de *Lancelotcompilatie* volgens mij zijn ontstaan in een milieu van onbezoldigde Arturliefhebbers. Mogelijk is de *Lancelotcompilatie* het resultaat van een soort gezamenlijk Arturproject door een groep 'amateurs' – in alle betekenissen van het woord.[4] Veel meer dan in Antwerpen rond de figuur van Boendale, bespeur ik hier concrete aanwijzingen voor een soort 'school' van letterlievenden rond de figuur van Velthem.

Wel is het hoe dan ook onaannemelijk dat dit eventueel 'verbond van letterlievenden' de (grotendeels Vlaamse) grondteksten van de compilatie zelf heeft geschreven. De stylometrische analyses in hoofdstuk 8 bevestigden immers de traditionele visie dat de compilator slechts reeds bestaande grondteksten heeft herwerkt (§ 8.2.1). Een roman als de *Torec* vertoont bijvoorbeeld nog maar bitter weinig overeenkomsten met de stijl van Maerlant, vermoedelijk de oorspronkelijke auteur van de tekst. Wat betreft de kerndelen van de compilatie werd trouwens een verrassend resultaat geboekt (§ 8.2.2): hoewel men tot op heden vaak is uitgegaan van een enkelvoudig auteurschap voor deze trilogie, vertoont de meest frequente rijmwoordenschat van *Arturs doet* vrij uitgesproken verschillen ten opzichte van de eerste twee kerndelen. Gezien de grote omvang van dit laatste (nog minder onderzochte) kerndeel is het echter niet aangewezen om de stylometrie hier te laten vooruitlopen op het traditioneel onderzoek. Wel kan het voor onderzoekers de moeite lonen om in hun toekomstige lectuur met dit resultaat rekening te houden.

De grote verrassing in dit hoofdstuk kwam echter van de geïnterpoleerde *Moriaen*. Zoals reeds beargumenteerd in het verleden, maakt deze roman in vergelijking met de andere interpolaties een vrij authentieke indruk. De compilator moet slechts oppervlakkig in deze roman hebben ingegrepen. Dat was vooral boeiend omdat Heeroma op

4 Dat er van de hulpkopiisten van de *Lancelotcompilatie* ook afschriften van andere teksten zijn overgeleverd (Klein 1995), kan eventueel worden opgevat aan een aanwijzing voor een zeker professionalisme. Toch betreft het hier haast nooit afschriften van *dezelfde* tekst, zoals dat bij de *Limborch*- en *Ferguut*-kopiisten net wel het geval is (Kwakkel & Mulder 2001).

basis van een traditioneel stilistisch onderzoek had betoogd dat deze roman door dezelfde auteur moest zijn geschreven als *Lantsloot vander Haghedochte*, een van de weinige Middelnederlandse Arturromans die buiten de compilatie is overgeleverd (zij het fragmentarisch). De vergelijking van de Arturromans buiten de compilatie met de reeds aan bod gekomen werken leerde inderdaad dat er verregaande stilistische overeenkomsten bestaan tussen de meest frequente rijmwoordenschat van beide teksten (§ 8.2.4).

Omdat Heeroma het mogelijk bij het rechte eind heeft gehad, was het noodzakelijk om ook een andere tekst in het onderzoek te betrekken die Heeroma aan dezelfde anonieme Vlaamse meester had toegeschreven: *Karel ende Elegast*. De *Elegast* wordt traditioneel echter tot de Karelepiek gerekend, een andere epische tekstsoort dan de Arturroman. De vergelijking van werken uit verschillende tradities was echter niet vanzelfsprekend, omdat de relatie tussen tekstsoort en rijmvocabulaire al vaker een problematische factor was gebleken in dit onderzoek. In het hoofdstuk 9 van deze studie heb ik daarom werk gemaakt van een stylometrische vergelijking van de Middelnederlandse Artur- en Karelepiek.

De Middelnederlandse Artur- en Karelepiek worden al van oudsher onderscheiden, omdat deze tekstsoorten uit verschillende verhaaltradities putten (§ 9.1.1). Recent hebben onderzoekers niettemin beklemtoond dat er ook heel wat raakpunten tussen deze tekstverzamelingen bestonden, zodat het onwaarschijnlijk is dat zij in gescheiden literaire circuits functioneerden. Een 'onbevooroordeelde' cluster-analyse van het rijmvocabulaire in deze ridderepische teksten leverde niettemin een resultaat waarin de Artur- en Karelepiek relatief netjes uit elkaar werden gehouden (§ 9.1.3). De daaropvolgende rangsom-analyse (§ 9.1.4) toonde dat zeker de Karelepiek gekenmerkt wordt door een relatief coherente rijmwoordenschat die goed aansluit bij vroegere zowel inhoudelijke als stijlkundige karakteriseringen. Een dergelijke thematische focus was minder aanwezig in de voor de Arturepiek typerende lijst rijmwoorden.

Opvallend genoeg ontsnapte *Karel ende Elegast* aan de tweedeling inzake *matière* in de cluster-analyse: de tekst zocht temidden van de andere Arturromans het gezelschap op van *Moriaen* en *Lantsloot*. Om na te gaan of er daadwerkelijk sprake was van een stilistische toenadering tussen deze teksten, werd via een cluster-analyse nagegaan welke andere teksten in het corpus qua rijmstijl dicht genoeg bij het bewuste trio lagen om als relevant controle-materiaal te dienen (§ 9.2). Uiteindelijk werden daarbij de *Ferguut* gekozen en Vostaerts bijdrage aan de *Walewein*.

Verschillende correspondentie-analyses benadrukten de frappante overeenkomsten in de hoogfrequente rijmwoordenschat tussen de *Elegast*, de *Moriaen* en *Lantsloot vander Haghedochte* – ondanks de grote verschillen in deze teksten wat betreft hun overlevering en verhaalin-

houd. Wie deze teksten herleest met dit resultaat in het achterhoofd, ziet interessante gelijkenissen verschijnen die onze lectuur van deze belangrijke werken in belangrijke mate zou kunnen verrijken (§ 9.3). Wie had vermoed dat de bedenker van het eerste zwarte hoofdpersonage uit de Nederlandse literatuur ook de uitdagende dichter zou zijn die Karel de Grote op zijn legendarische rooftocht uitstuurde? Dit resultaat nodigt de medioneerlandistiek met andere woorden uit tot een herwaardering van Heeroma's these die door dit kwantitatieve onderzoek duidelijk kracht wordt bijgezet. Deze intrigerende casus vormde het gewaagde slotakkoord van deze studie; gewaagd, want wie uitspraken doet over een kroonjuweel als de *Elegast* kan gegarandeerd tegenwind verwachten. Maar misschien is daar precies een kerntaak voor de stylometrie weggelegd: het beredeneerd opwerpen van uitdagende hypotheses voor toekomstig onderzoek.

10.2 WAT IS (NOG) NIET ONDERZOCHT?

De kritische nabeschouwing bij deze studie leert dat een groot aantal onderwerpen buiten het bereik van dit onderzoek zijn gevallen maar in de toekomst zeker nog aandacht verdienen. Een eerste lacune betreft enkele bekende gevallen van betwist auteurschap in de Middelnederlandse letterkunde, die ik zuiver omwille van tijdsgebrek niet heb kunnen behandelen. Een goed voorbeeld daarvan is de mysterieuze dichter die Jan van Boendale in zijn *Lekenspiegel* eervol vermeldt (zie ook § 7.1.1):

> Van Bruecele Heine van Aken
> Die wel dichte conste maken
> God hebbe die ziele sine
> Maecte dese twee versekine:
> "Vrient die werdet langhe ghesocht
> Selden vonden ende saen verwrocht."
> (De Vries 1844–1848, 183 boek III, kap. 17, vs. 91-96)

Boendale vermeldt hier ene Hein van Aken (in Duitsland?) en schrijft aan deze Brusselse dichter een spreuk toe die ons verder uit geen enkel Middelnederlands werk bekend is. Wel kennen we één werk dat met zekerheid aan de bewuste dichter kan worden toegeschreven: een laat veertiende-eeuwse codex van Brabantse origine (OXFORD, BODLEIAN LIBRARY, MARSHALL 29) bevat een berijmde verdietsing van *De l'ordene de chevalerie*, getiteld *Van den coninc Saladijn ende van Hughen van Tabaryen*. Deze korte verstekst eindigt met de verzen: *Dit heeft gedicht, te lovene en te eren / Allen ridderen, Hein van Aken* (Willems 1842, 103, vs. 274-275).

In de loop der jaren hebben filologen het oeuvre van deze Brusselse dichter gevoelig trachten uit te breiden, onder meer met de bekende *Roman van Limborch* en de *Brabantse Rose* (Lievens 1960). De secun-

daire literatuur over dit onderwerp is echter bijzonder omvangrijk en de handschriftelijke tradities en de onderlinge afhankelijkheid van deze teksten is bovendien complex. Hoewel de stylometrie een vernieuwende invalshoek voor deze fascinerende casus kan bieden, heb ik deze casus niet afdoende kunnen behandelen in de mij toegemeten tijd. Wellicht kan aan deze auteurskwestie in de toekomst nog een zelfstandige monografie worden gewijd. Hetzelfde geldt voor de vele 'oosterse' of 'klassieke' romans (zoals *Floris ende Blancefloer* en de *Parthenopeus van Bloys*), die hierboven schromelijk onderbelicht zijn gebleven.

In het algemeen vertoont deze studie overigens een duidelijke voorkeur voor langere teksten. Uit het internationale onderzoek in de stylometrie is bekend dat een tekst lang genoeg moet zijn om deze aan een kwantitatief onderzoek te kunnen onderwerpen. Aangezien dit onderzoek om methodologische redenen bovendien beperkt bleef tot de rijmwoordenschat van teksten, zijn veel kortere teksten uit de boot gevallen. Vanuit een kwantitatief standpunt dringt zich daarom de noodzaak op om ook niet-rijmwoorden in het onderzoek te betrekken, zodat ook minder omvangrijke teksten in het onderzoek kunnen behandeld worden.

Dat brengt ons bij het belangrijkste tekort van deze studie: een techniek voor auteursherkenning die gebaseerd is op rijmwoorden, negeert niet alleen alle stilistische informatie buiten de rijmpositie, maar is vanzelfsprekend ook ongeschikt om onberijmde en eventueel kortere prozateksten te onderzoeken. Die beperking laat zich niet zo sterk voelen in de oudste Middelnederlandse literatuur omdat deze nog relatief consequent berijmd is (Lie 1994). Toch moet men beseffen dat deze situatie al snel verandert in de veertiende eeuw, wanneer versteksten in stijgende mate concurrentie krijgen van volkstalige prozateksten. Zeker in bijvoorbeeld de Brabantse religieuze literatuur, zien we hoe proza-vormen in de veertiende eeuw snel aan populariteit winnen. Exemplarisch is de casus van de 'Bijbelvertaler van 1360', een anonymus uit het kartuizerklooster van Herne die een veelheid aan Latijnse teksten in Middelnederlands proza heeft omgezet (Kors 2007). De exacte omvang van dit merkwaardige oeuvre is echter nog betwist, zodat stylometrische technieken voor auteursherkenning in proza hier een belangrijk winstpunt zouden kunnen betekenen voor de literatuurgeschiedenis.

Verder mogen wij de Middelnederlandse situatie niet extrapoleren naar het buitenland. In de ons omringende cultuurgebieden, komen versteksten al veel vroeger in de verdrukking, zodat het aandeel prozateksten daar in het algemeen veel groter is. Ook in de Europese *lingua franca*, het Latijn, is het aantal proza-teksten verhoudingsgewijs natuurlijk bijzonder groot. De belangrijkste uitdaging voor toekomstig stylometrisch onderzoek zal dan ook zijn om een betrouwbaar instrumentarium te ontwikkelen voor de analyse van middeleeuwse

proza-teksten. Aangezien men zich in proza-teksten niet meer kan verlaten op rijmwoorden, zal dat hoe dan ook geen evidentie vormen.

Toch bestaan er experimentele technieken die voor de analyse van middeleeuwse prozateksten in aanmerking komen en nader onderzoek verdienen. Zo kan gewezen worden op *Unmasking*, een complexe maar doeltreffende techniek die is voorgesteld om oppervlakkige stijlverschillen tussen teksten – bijvoorbeeld gerelateerd aan vertelperspectief – via computationele weg op te sporen en uit te vagen (Koppel et al. 2009). In enkele voorbereidende experimenten heb ik deze techniek toegepast op diplomatische afschriften van middeleeuwse teksten. Hierin bleek duidelijk dat deze techniek in staat was om 'verdachte' ingrepen (bijvoorbeeld inzake een afwijkende spelling) van kopiisten te detecteren en te verwijderen, hetgeen perspectieven biedt voor de vergelijking van afschriften van teksten die door verschillende kopiisten werden vervaardigd.

Unmasking

Een laatste, weinig chauvinistische kanttekening daarbij is de volgende. De Middelnederlandse literatuur is rijk aan casussen van betwist auteurschap en vormt daarom een interessant domein om stylometrische technieken in te zetten. Het blijft echter de vraag of het relatief kleinschalige domein van de medioneerlandistiek wel geschikt is om betrouwbare technieken te ontwikkelen, laat staan te ijken. De Middelnederlandse literatuur is bijvoorbeeld relatief arm aan oeuvres of verzamelingen van parallelle kopieën van auteursteksten waarvan de teksttraditie betrouwbaar in kaart is gebracht. Andere talen – zoals het Latijn of Frans – zijn rijker aan omvangrijke teksttradities en oeuvres die als uitvalsbasis kunnen dienen voor het ontwikkelen van technieken. Het lijkt aangewezen om deze literatuur in het onderzoek niet links te laten liggen.

Tot slot zij opgemerkt dat de stylometrie niet beperkt is tot de studie van auteurschap maar veel meer facetten heeft. Hierboven werd reeds gewezen op de mogelijkheid van stylometrische 'kopiistherkenning' (§ 3.5), een aantrekkelijk onderzoeksspoor waarover zeker nog niet het laatste woord is gezegd. Een andere interessante toepassing is de 'stylochronometrie', waarin men onderzoekt of het mogelijk is om literaire teksten op basis van stijlverschijnselen te dateren, eventueel relatief ten opzichte van elkaar (Stamou 2008). Voorgaand onderzoek naar Oudfranse literatuur heeft aangetoond dat deze benadering voor middeleeuwse teksten vruchtbare resultaten kan opleveren (Ernestus et al. 2007). De stylometrie en computationele tekstanalyse is met bovenstaande studie duidelijk nog lang niet 'uitgezocht'.

stylochronometrie

11 APPENDIX

11.1 ANNOTATIEPRAKTIJK

Voor dit onderzoek maakte ik gebruik van een grotendeels geautomatiseerde technologie voor de linguïstische verrijking van Middelnederlandse teksten. In deze sectie licht ik de technische kant van mijn algemene werkwijze toe. Iedere tekst werd in drie opeenvolgende stadia verwerkt: voorbewerking, tokenisering en lemmatisering.

11.1.1 *Voorbewerking*

Voor dit onderzoek maakte ik gebruik van digitale edities van teksten. De meeste standaardedities van de geanalyseerde teksten waren reeds in een digitale versie beschikbaar op de *Cd-rom Middelnederlands*. Tenzij anders aangegeven in de hoofdtekst, heb ik dankbaar van deze *Cd-rom*-versies gebruik gemaakt. Slechts in een klein aantal gevallen steunde ik op andere edities, bijvoorbeeld voor de niet-uitgegeven *Garijn van Montglavie*, die ik zelf diplomatisch heb getranscribeerd.

In het geval dat meer dan één tekstgetuige van een Middelnederlandse tekst is overgeleverd, heb ik steeds een zo verantwoord mogelijke keuze gemaakt voor één versie of editie. De keuze voor een specifieke tekstversie is in de desbetreffende secties hierboven vaak reeds toegelicht. Voor enkele teksten met een zeer grote overlevering (bv. de *Spiegel historiael* of de *Lekenspiegel*), heb ik gewoon de hoofdtekst van de standaardedities gebruikt, omdat het om praktische redenen niet mogelijk was om het (zelden gedigitaliseerde en complexe) variantenapparaat van deze teksttradities te beschouwen. Deze keuze steunt op de aanname dat de hier bestudeerde rijmwoordenschat in Middelnederlandse teksten over het algemeen slechts aan een beperkte variatie onderhevig was.

Indien er meerdere versies of edities van een tekst beschikbaar waren, heb ik mijn keuze steeds gebaseerd op een combinatie van de volgende factoren: het digitaal beschikbaar zijn van een editie, de betrouwbaarheid van een editie, de lengte van de bewuste versie en de voorkeur in de vakbeoefening voor een bepaalde tekstgetuige. Voor *Karel ende Elegast*, heb ik in navolging van Claassens (2002) bijvoorbeeld voor een (aangevulde) editie van incunabel A gekozen, aangezien de oudere tekstgetuigen fragmentarisch zijn en nog niet gedigitaliseerd.

Ik heb – met uitzondering van de *Rijmbijbel*-analyses (§ 3.1) – geen rekening gehouden met het onderscheid tussen diplomatische en kriti-

sche edities: dit onderscheid was voor dit onderzoek minder relevant omdat de orthografie van de handschriften in de lemmatiseringsfase verregaand zou worden genormaliseerd. Sommige teksten (bv. *Der naturen bloeme*) waren reeds in een verrijkt formaat beschikbaar in het digitale *Corpus-Gysseling* op het INL (zie onder). Vanzelfsprekend heb ik in deze gevallen de voorkeur gegeven aan de reeds geannoteerde bestanden.

Alle digitale edities werden ontdaan van meta-informatie zoals (niet in het rijmschema betrokken) kapittelopschriften, paragraaftekens en een moderne regelnummering. Niet-alfabetische symbolen (bv. paragraaftekens en moderne interpunctie) werden verwijderd of in een te verwaarlozen aantal gevallen vervangen door hun ASCII-pendant (bv. ů > u in de *Limburgse Aiol*). Kleinschalige conjecturen door editeurs (bv. in het geval een onduidelijke lezing) werden in de regel aanvaard om het tekstverlies waar mogelijk te beperken. Verzen werden echter niet in het corpus opgenomen, indien het materieel tekstverlies te groot was om het rijmwoord nog van een betrouwbaar lemma te kunnen voorzien. Omwille van die selectie ligt het aantal vermelde verzen of rijmwoorden voor een tekst in deze studie vaak lager dan het verzenaantal dat elders wordt vermeld. Hoofdletters werden in kleine letters omgezet (bv. XI > xi), met uitzondering van rijmwoorden met een hoofdletter in kritische edities. In deze gevallen kon aan de hand van deze hoofdletter het optreden van eigennamen (aangeduid met de tag PRNAME) namelijk makkelijk gedetecteerd worden.

Omwille van het auteursrecht dat nog op verschillende edities rust, kan ik deze teksten jammer genoeg niet in hun verrijkte vorm beschikbaar maken. Aan het INL bestaan wel plannen om het geannoteerde *Corpus-Gysseling* op korte termijn openbaar te maken, hetgeen reeds een belangrijke hiaat zal opvullen. Vanzelfsprekend kunnen geïnteresseerde onderzoekers de gelemmatiseerde rijmwoorden en de scripts die ik voor deze studie gebruikte op aanvraag verkrijgen om ze voor onderzoeksdoeleinden aan te wenden.

11.1.2 *Tokenisering*

De metingen in deze studie werden – opnieuw met uitzondering van de *Rijmbijbel*-analyses (§ 3.1) – beperkt tot de gelemmatiseerde rijmwoorden in teksten. De woordscheiding en –aaneenschrijving waren in de middeleeuwen echter nog niet strikt beregeld. De oppervlaktevorm *daer ave* correspondeert bijvoorbeeld met de schrijfwijze *daerave* in andere handschriften. Daarom moest er aan de lemmatiseringsfase een tokeniseringsfase voorafgaan. Dergelijke verschillen inzake witruimte moeten immers worden genormaliseerd om abstractie te kunnen maken van mogelijk door een kopiist geïntroduceerde en daarom voor dit onderzoek minder relevante verschillen tussen teksten. Een *tokeniser* is een software-toepassing uit de computerlinguïstiek die automatisch

teksten 'tokeniseert', d.w.z. de scheiding en aaneenschrijving van de oppervlaktevorm van woorden (i.e. een reeks letters aan weerszijden door witruimte begrensd) normaliseert.

Als voorbeeldmateriaal deed ik voor deze studie een beroep op het literaire deel van het digitale *Corpus-Gysseling*, een omvangrijke collectie Middelnederlandse teksten die aan het INL wordt geannoteerd en onderhouden. De teksten in dit corpus zijn semi-automatisch getokeniseerd en voorzien van een uniforme woordsoort-code en een modern lemma. Voor de verrijking van teksten heb ik mij zoveel mogelijk gericht op de annotatie-praktijk in dit corpus. Dat was echter niet altijd vanzelfsprekend omdat de annotatie-principes nog niet expliciet zijn beschreven en daardoor in de meeste gevallen uit geannoteerde voorbeelden moeten worden afgeleid. Zo wordt de woordgroep *daer ave* in dit corpus consequent in een eenheid ondergebracht (*daer+ave*), waaraan één lemma wordt toegekend (DAARAF). In sommige gevallen lijkt echter nog sprake van schijnbare inconsequenties, die bij gebrek aan een meer uitgebreide documentatie moeilijk te doorgronden zijn. De woordgroep *al bloot* wordt bijvoorbeeld consequent tot *al+bloot* herleid, terwijl andere andere combinaties van *al*+ADJECTIEF niet worden geconcateneerd. Ik heb evenwel zo goed mogelijk de gehanteerde tokeniseringspraktijk in het digitale *Corpus-Gysseling* trachten te volgen. Bij twijfel consulteerde ik het online Middelnederlandse woordenboek (`http://www.inl.nl/`), dat deels op basis van dit corpus is vervaardigd.

Het tokeniseren werd hier geïmplementeerd als een classificatie-taak waarbij voor iedere oppervlaktevorm in een vers (behalve het eerste woord) werd voorspeld of die vorm moest verbonden worden met de vorige oppervlaktevorm in het vers. De classificatie-taak was dus binair (wél of niet concateneren). Voor de tokenisering gebruikte ik de TIMBL-software oftewel de *Tilburg Memory-Based Learner* (Daelemans et al. 2010b). De classifier werd getraind op voorbeeld-instanties uit het digitale *Corpus-Gysseling*. Naar analogie met de *sliding window*-aanpak in bijvoorbeeld de *Memory-Based Tagger* werd ieder woord in een te tokeniseren vers als een *focus token* voorgesteld, omringd door twee *context tokens* aan weerszijden (Daelemans et al. 2010a).

Bekijken we bijvoorbeeld het willekeurige vers *so dat riepen alle te samen*, dat in het digitale *Corpus-Gysseling* als volgt werd getokeniseerd: *so+dat riepen alle te+samen*. In dit vers worden dus tweemaal twee opeenvolgende woorden verbonden: *so* en *dat* worden samengenomen in *so+dat* en *te* en *samen* in *te+samen*. Als we dit voorstellen via een *sliding window* geeft dat 6 vectoren voor iedere oppervlaktevorm (waarbij een '=' uitdrukt dat er op de bewuste positie geen *context token* meer beschikbaar is; zie Tab. 25, p. 292). Op deze vectoren werd vervolgens een TIMBL-*tree* getraind via het efficiënte *IgTree*-algoritme (verder met *default options*). Alle hier geanalyseerde teksten werden via deze TIMBL-*tree* getokeniseerd.

Focus-2	Focus-1	Focus	Focus+1	Focus+2	Verbinden?
=	=	**so**	dat	riepen	NEE
=	so	**dat**	riepen	alle	JA
so	dat	**riepen**	alle	te	NEE
dat	riepen	**alle**	te	samen	NEE
riepen	alle	**te**	samen	=	NEE
alle	te	**samen**	=	=	JA

Tabel 25: Voorbeeld van trainingsvectoren voor de context-gevoelige tokenisering van Middelnederlandse oppervlaktevormen. Deze zes vectoren zijn ontleend aan het getokeniseerde voorbeeldvers *so+dat riepen alle te+samen* uit het digitale *Corpus-Gysseling*.

Een heikel punt bij het tokeniseren vormen de scheidbare werkwoorden: zoals in het modern Nederlands kunnen bepaalde Middelnederlandse werkwoorden in veel constructies een afgesplitst woord (bv. een partikel) regeren. Dit fenomeen zorgt ook in taaltechnologie voor het moderne Nederlands nog voor problemen (Chrupała 2006, 124). In het willekeurige voorbeeldvers *ende roofde den tempel ende sloten toe* worden de laatste twee oppervlaktevormen in het digitale *Corpus-Gysseling* samengenomen tot *sloten+toe* en wordt het lemma TOESLUITEN aan deze combinatie toegekend. Dergelijke partikels hoeven echter niet in de onmiddellijke omgeving van de persoonsvorm op te treden. In het voorbeeldvers *doe sprac hem dingel aldus toe*, worden de componenten *sprac* en *toe* door drie tussenliggende vormen gescheiden. Dergelijke fenomenen zullen in de hier gehanteerde vector-voorstellingen – beperkt tot twee *context tokens* aan weerszijden van het *focus token* – natuurlijk niet worden opgemerkt. In andere gevallen is het bovendien onduidelijk of men met een werkwoordelijk partikel heeft te maken, dan wel met een gewoon voorzetsel. In het vers *doe sprac die ingel nog voert* is het voor interpretatie vatbaar of wij *voert* als een partikel bij *sprac* mogen opvatten (cf. VOORTSPREKEN), dan wel als een zelfstandige eenheid (cf. SPREKEN ... VOORT).

Om deze reden heb ik ervoor gekozen om dergelijke partikels slechts met een vervoegd werkwoord te combineren, als beide componenten onmiddellijk op elkaar volgden. In het voorbeeld *ende roofde den tempel ende sloten toe* zou dus met het lemma TOESLUITEN worden gewerkt voor de combinatie *sloten+toe*, terwijl *voert* in het vers *doe sprac die ingel nog voert* een zelfstandig lemma wordt toegewezen (VOORT). Dergelijke partikels werden in de lemmatiseringsfase bovendien niet op een afwijkende wijze gemarkeerd.

In tegenstelling tot het verbinden van woorden werd woordscheiding (bijvoorbeeld in het geval van clisis) nog niet in dit stadium aangepakt. Soms bevat een Middelnederlandse oppervlaktevorm namelijk meerdere samenstellende delen (bv. *daers* uit *daer es*) die moeten

gescheiden worden voor de lemmatisering (cf. *daers* > DAAR+ZIJN). In dergelijke gevallen treden er niettemin vaak uitspraak-gerelateerde assimilatieverschijnselen in de spelling op. Deze maken het in veel gevallen moeilijk om de exacte overgang tussen de samenstellende eenheden aan te duiden. In deze gevallen wordt de scheiding tussen de twee samenstellende tokens in het digitale *Corpus-Gysseling* meestal niet expliciet aangegeven in de oorspronkelijke oppervlaktevorm, maar worden er in de lemmatiseringsfase wel twee lemmata aan het oorspronkelijke token toegekend (bv. *daers* > DAAR+ZIJN, *tpeert* > DAT+PAARD, *ics* > IK+DE/HET).

11.1.3 *Lemmatisering*

Zoals reeds toegelicht in § 4.1.2 werd voor het auteursonderzoek hierboven slechts gebruik gemaakt van tellingen op basis van gelemmatiseerde rijmwoorden. Hoewel rijmwoorden over het algemeen relatief stabiele elementen waren in de overlevering van Middelnederlandse teksten, waren ook zij aan een oppervlakkige spellingvariatie onderhevig. Daarom is het in verschillende opzichten raadzaam om stylometrische tellingen uit te voeren op de onderliggende lemma's of lexemen van rijmwoorden (Knowles & Mohd Don 2004, 70). Lemmatisering is een linguïstisch verrijkingsproces waarbij men een token – een getokeniseerd woord – met een lemma verbindt, i.e. de genormaliseerde hoofdvorm van een woord, zoals we die aantreffen in een woordenboek (Francis & Kucera 1982, 1). Aan tokens die onderling slechts verschillen inzake spelling, vervoeging of verbuiging, wordt hetzelfde lemma toegekend (Lyras et al. 2008, 1043ff). De bedoeling is om abstractie te kunnen maken van oppervlakkige verschillen tussen rijmwoorden (Crystal 1997). Door lemmatisering kunnen we immers op zinnige wijze generaliseren over de verschillende oppervlaktevormen van dezelfde lexemen in rijmpositie.

De lemmatiseringstechniek die voor deze studie is gehanteerd, wordt elders uitgebreider beschreven (Kestemont et al. 2010). Ik beperk mij hier daarom tot de hoofdlijnen. De ontwikkelde *lemmatiser* is net als de *tokeniser* getraind op het literaire deel van het digitale *Corpus-Gysseling*. In dit corpus zijn alle getokeniseerde woordvormen verbonden aan een lemma in een moderne spelling, bv. *peert* > PAARD, *sach* > ZIEN, *mine* > MIJN (Van der Voort van der Kleij 2005). Dergelijke lemmata kunnen worden teruggevonden in de recentste versie van het online Middelnederlands woordenboek. Het digitale *Corpus-Gysseling* bijna alle Middelnederlandse teksten (27 in totaal) die in tekstdragers zijn overgeleverd die vóór 1300 kunnen worden gedateerd. In totaal bevat het corpus ca. 573.000 tokens, waarvan er ca. 40.000 uniek zijn. Deze tokens zijn verrijkt met ca. 15.000 verschillende lexemen.

Voor het tokeniseren van een nieuwe tekst wordt een verschil gemaakt tussen 'geziene' woorden en 'ongeziene woorden'. Geziene

woorden zijn woorden die de *lemmatiser* reeds letterlijk is tegengekomen in het trainingsmateriaal: in dat geval is de kans groot dat één van de vormen uit het trainingsmateriaal verbonden is met het lemma dat we willen extrapoleren naar het nieuwe woord in de te lemmatiseren tekst. In sommige gevallen kan er sprake zijn van ambiguïteit: de relatief frequente oppervlaktevorm *doet* kan afhankelijk van de onmiddellijke context immers zowel een verschijningsvorm zijn van het adjectief DOOD, het substantief DOOD, dan wel een werkwoordsvorm van DOEN of DODEN. Voor de disambiguering van deze vormen doe ik een beroep op de *Memory-Based Tagger* (met *default options*: Daelemans et al. 2010a). Dit TIMBL-gebaseerd softwarepakket werd oorspronkelijk ontwikkeld voor automatische woordsoorttoekenning (*part-of-speech tagging*) maar kan ook probleemloos worden gebruikt voor lemmatisering, een probleem dat eveneens een sequentiële classificatietaak betreft. Na kruisvalidering (*ten-fold cross validation*) bleek deze aanpak voor de 'geziene' woorden in het trainingsmateriaal een gemiddelde score van iets minder dan 94% op te leveren. Voor het taggen van de rijmwoordenschat in de teksten voor deze studie scheen deze score echter hoger te liggen, doordat de woordenschat in rijmpositie veelal erg stereotiep was en de *lemmatiser* dus zelden voor problemen stelde.

Een groter probleem stelden de vele 'ongeziene' woorden die een *lemmatiser* in teksten aantreft, als een gevolg van de aanzienlijke spellingvariatie in middeleeuwse handschriften. Hiervoor werd beroep gedaan op een geheugengebaseerde *string distance*-metriek: een afstandsmaat die men op basis van het beschikbare trainingsmateriaal leert welke woorden meer op elkaar lijken dan andere en daarom met een grotere waarschijnlijkheid onder hetzelfde lemma mogen worden ondergebracht (Kestemont et al. 2010). In experimenten via kruisvalidering op basis van het digitale *Corpus-Gysseling* bleek dat deze aanpak de foutenmarge van de *lemmatiser* voor ongeziene woorden gevoelig kon terugdringen. Deze aanpak werd dan ook gebruikt voor het lemmatiseren van de ongeziene woorden in de teksten die voor deze studie werden verrijkt.

Het lemmatiseren van de teksten voor dit onderzoek gebeurde zoveel mogelijk op analoge wijze aan de verrijking van het digitale *Corpus-Gysseling*. In het geval van twijfel heb ik steeds de (op dat moment) recentste versie van dit woordenboek geconsulteerd. Ik herneem kort de basisprincipes (zie § 4.1.2) die ik heb gehanteerd. Homonieme lemmata werden in principe niet onderscheiden: dat wil bijvoorbeeld zeggen dat HEER in de betekenis van 'meester' niet formeel wordt onderscheiden van HEER in de betekenis van 'leger'.

Alle rijmwoorden en slechts de rijmwoorden zijn gelemmatiseerd: slechts die woorden die door de dichter gebruikt worden om tussen twee opeenvolgende verzen een rijm tot stand te brengen, kregen een lemma toegewezen. In de praktijk gaat het meestal gewoon om het laatste woord van een regel, maar niet altijd. In het geval van

proclisis bijvoorbeeld wordt het enclitische token meestal niet in de lemmatisering betrokken. Het versfinale token *tsweert* (in een rijm met *eert*) bijvoorbeeld, krijgt slechts het lemma ZWAARD en niet het lemma DAT+ZWAARD omdat het clitische *t-* geen rol speelt in de totstandkoming van het rijm en kan teruggaan op een kopiist. In principe kan een auteur oorspronkelijk immers *dat sweert* gebruikt hebben. In enkele gevallen kan toch sprake zijn van een combinatie van lemmata: in een stilistisch verfijnd rijmpaar als *vicaris* : *waer is* is ook het voorlaatste woord *waer* in het tweede vers cruciaal voor het rijm en wordt in dit geval het samengestelde lemma-label WAAR+ZIJN gebruikt.

In technische termen kregen dus alle woorden een lemma die de eerste beklemtoonde vocaal van een rijm bevatten, alsook alle woorden die daarop volgen. In drie gevallen is van deze annotatie afgeweken: om de inhoud van deze teksten van meet af aan zoveel mogelijk buiten spel te zetten, zijn drie soorten woorden afwijkend gecodeerd. Zij kregen geen echt inhoudelijk lemma maar eerder een vage aanduiding van hun woordsoort:

EIGENNAMEN – alle eigennamen (persoonsnamen, geografische aanduidingen, boektitels, ...) kregen het lemma PRNAME. Een uitzondering vormen enkele hoogfrequente eigennamen als GOD en CHRISTUS die zo algemeen zijn in Middelnederlandse teksten dat ze amper met de inhoud van teksten samenhangen;[1]

TELWOORDEN – alle hoofd- en rangtelwoorden (buiten 1, 2, 3 en afgeleiden) kregen het label NUMB. Deze betekenissen werden verbleekt omdat zij vaak inhoudsgevoelig zijn (bijvoorbeeld jaartallen) en omdat de numerieke inhoud die zij uitdrukken daarom in de regel minder van belang lijkt voor het stijleigen van auteurs;

BASTAARDWOORDEN – alle expliciet anderstalige woorden, zoals enkele Latijnse substantieven (bv. *deus*) of Franstalige woordgroepen, kregen het lemma FOREIGN. Dat gold niet voor erg gangbare leenwoorden.

In een klein aantal gevallen ben ik afgeweken van de oorspronkelijke lemmatiseringspraktijk van het digitale *Corpus-Gysseling*. Deze zijn onder te brengen in de volgende categorieën:

- In sommige gevallen was de oorspronkelijke annotatie-praktijk in het *Corpus-Gysseling* inconsequent. (Soms bleek deze later gecorrigeerd in het online woordenboek.) Zo waren de vormen *baroen, barone, baroenen* oorspronkelijk als BAROEN gelemmatiseerd, waar men eigenlijk de vorm BARON verwacht. Zelf heb ik daarom consequent het lemma BARON gehanteerd.

1 Naar analogie met het lemma CHRISTEN zijn ook de vormen van het rijmwoord SARACEEN niet door de tag PRNAME vervangen.

- In sommige gevallen werden reeksen van oppervlaktevormen (o.m. voegwoorden) als een *multiword-unit* beschouwd. Het gaat bijvoorbeeld om reeksen als *als ende als* (*als+ende+als*), *toter wilen dat* (*toter+wilen+dat*) of *ut ende ut* (*ut+ende+ut*). Ik ben de oorspronkelijke annotators niet gevolgd in het concateneren van dergelijke woordgroepen omdat deze soms ver terugreikten in het vers en daarmee de claim bemoeilijkten dat enkel de rijmwoordenschat van woorden hier werd geanalyseerd. Slechts het laatste element in deze reeksen werd daarom gelemmatiseerd (bv. *als ende als* > ALS).

- In een klein aantal gevallen heb ik de annotatie aangepast voor dit specifieke onderzoek. Een belangrijke aanpassing was het lemmatiseren van de bijzonder frequente demonstratieve voornaamwoorden *die(s)*, *dit*, *dat, das* en *des* in rijmpositie. Om het gebruik van deze rijmwoorden te normalizeren, heb ik voor deze rijmwoorden consequent de volgende labels gebruikt: *die(s)*: DIE, *dit*: DIT, *dat, das*: DAT en *des*: DE/HET. Die beslissing is linguïstisch gezien niet helemaal verantwoord, omdat bijvoorbeeld de vorm *dat* in het *Corpus-Gysseling* wel eens wordt geïnterpreteerd als een lidwoord (*de, het*) in plaats van als een voornaamwoord. De bedoeling was in deze gevallen echter niet om tot een correcte linguïstische analyse te komen, maar wel om de rijmwoorden op een consequente manier te normaliseren.

- Sommige specifieke woorden bleken voor de *lemmatiser* té moeilijk om te disambigueren. De oppervlaktevorm *bloot* kan afhankelijk van de context namelijk zowel een adjectivale verschijningsvorm zijn van het lemma BLOOT (in de betekenis van 'naakt') als van het lemma BLODE (in de betekenis van 'laf', 'niet strijdvaardig'). Omdat de tagger er – bij gebrek aan meer contextuele informatie – zelden in slaagde deze vormen correct te onderscheiden, heb ik de verschijningsvormen van bijvoorbeeld deze woorden consequent tot het lemma BLOOT herleid.

De hier bestudeerde teksten zijn dus automatisch geannoteerd via software. Na het verrijken van de teksten heb ik slechts een oppervlakkige controle van het resultaat uitgevoerd en de meest markante, consequent gemaakte fouten verbeterd. In het algemeen heb ik wel getracht om zo weinig mogelijk te interfereren met het resultaat van de automatische analyse. De verrijking van teksten is daarmee verre van foutloos: wie de data bekijkt waarop deze studie gestoeld is, zal merken dat de annotatiekwaliteit voor individuele passages soms laag kan zijn. Voor een traditionele filoloog kan het daarom vreemd lijken dat ik geen doorgedreven poging heb ondernomen om de slordige verrijking van bepaalde teksten te verbeteren. Deze kwestie vereist enige toelichting.

Ten eerste was er een praktische overweging: binnen de korte tijdsspanne van dit eenmansproject was het nauwelijks haalbaar om de verrijking van iedere tekst nauwgezet te corrigeren. Bovendien kon ik niet over fondsen beschikken om deze verbetering aan een geschoolde expert uit te besteden. Tevens behoorde het zeker niet tot de doelstellingen van dit project om een omvangrijk en nauwgezet geannoteerd corpus van Middelnederlandse teksten voor een breder publiek aan te leggen – een taak die overigens adequater vervuld kan worden door instituten die hiervoor de nodige expertise in huis hebben.

Veel belangrijker is echter de volgende overweging. Binnen de kwantitatieve filologie – en de *Digital Humanities* in het algemeen – speelt men sterk in op de recente ontwikkelingen waarbij een immens grote hoeveelheid tekst relatief snel in een digitaal formaat beschikbaar is geworden. Dergelijke, omvangrijke corpora kunnen simpelweg niet meer door een individuele onderzoeker onderzocht worden. Daarom is het een grote uitdaging om computationele methodes te ontwikkelen die onderzoekers een deel van de 'manuele' analyse uit handen kunnen nemen.

In de internationale stylometrie is het dan ook eerder ongebruikelijk om een *close reading* uit te voeren van een volledige data-set, laat staan om een verregaande correctie uit te voeren van de eventuele linguïstische verrijking van het materiaal. De bedoeling is net dat dit type verrijking een gedetailleerde inspectie van het materiaal overbodig maakt. Bovendien kunnen de meeste attributie-technieken een zekere foutenmarge of *noise* in de data tolereren: deze algoritmes gaan immers op zoek naar grotere patronen in woordfrequenties die het belang van individuele uitschuivers in de annotatie overstijgen.

Belangrijk is daarom dat deze op het eerste zicht slordige werkwijze wel degelijk goede resultaten oplevert voor auteursherkenning, zoals hierboven op verschillende plaatsen proefondervindelijk is aangetoond. Ook in toekomstig stylometrisch onderzoek zal men bovendien met *ruis* in de data te maken krijgen, zodat men niet de onrealistische verwachting mag koesteren dat men steeds met volledig correcte data zal kunnen werken. Een laatste overweging betreft consistentie. Aangezien ik amper ben tussengekomen in het resultaat van de tagger, zal dit resultaat vaak dezelfde fouten bevatten. Als er fouten zijn gemaakt, moeten deze door de tagger dus relatief consistent zijn gemaakt. Het feit dat de berekeningen teruggaan op hoogfrequente, en daarom meestal makkelijk te interpreteren fenomenen leidt er eveneens toe dat er slechts in een enkel geval aanleiding tot discussie zal zijn.

BIBLIOGRAFIE

Cd-rom Middelnederlands. Woordenboek en teksten. Sdu, 1998.

Aha, D.; Kibler, D. & Albert, M. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

Allen, J. On the authenticity of the Baligant episode in the *Chanson de Roland*. In Mitchel, J. (ed.), *Computers in the Humanities*, pp. 65–72. Edinburgh University Press, 1974.

Alpaydin, E. *Introduction to Machine Learning*. Adaptive computation and machine learning. The MIT Press, 2004.

Appelmans, J. Van Thomas' bijen tot *Jans Teesteye*. De Brabantse historiografen en hun didactische bekommernis (dertiende en veertiende eeuw). In Bauer, R. (ed.), *In de voetsporen van Jacob van Maerlant: liber amicorum Raf de Keyser: verzameling opstellen over middeleeuwse geschiedenis en geschiedenisdidactiek*, Symbolae Facultatis Litterarum Lovaniensis, Series A 30, pp. 258–283. Universitaire Pers Leuven, 2002.

Appelmans, J. The abbey of Affligem and the emergence of a historiographic tradition in Brabant (1268-1322). In Verbeke, W.; Milis, L. & Goossens, J. (ed.), *Medieval narrative sources: a gateway into the medieval mind*, pp. 163–180. Leuven University Press, 2005.

Argamon, S. Interpreting Burrows's Delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2):131–147, 2008.

Argamon, S. & Levitan, S. Measuring the usefulness of function words for authorship attribution. In *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing (2005)*, 2005.

Aronoff, M. & Fudeman, K. *What is Morphology?* Blackwell, 2005.

Asaert, G. *Honderd huizen aan de Grote markt van Antwerpen. Vijf eeuwen bewoningsgeschiedenis*. Waanders, 2005.

Baayen, R. *Word Frequency Distributions*. Text, speech and language technology 18. Springer, 2002.

Baayen, R. *Analyzing linguistic data. A practical introduction to Statistics using R*. Cambridge University Press, 2008.

Barthes, R. La mort de l'auteur. *Manteia*, 5:12–17, 1968. [Oorspronkelijk gepubliceerd in het Engels in *Aspen* (1967) 5/6].

Bender, P. Naar aanleiding van *Jan, u arme clerc* in den *Lekenspieghel* (deel III bl. 278). *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 18 (4):313–314, 1899.

Berry, D. The computational turn: Thinking about the Digital Humanities. *Culture machine*, 12, 2011. URL http://www.culturemachine.net/index.php/cm/article/view/440/470.

Berteloot, A. öf ghi yet wilt spreect ieghen mi in dietsche dat ict mach verstaen: welke taal sprak reynaert? In Jansen-Sieben, R.; Janssens, J. & Willaert, F. (ed.), *Medioneerlandistiek: een inleiding tot de Middelnederlandse letterkunde*, Middeleeuwse studies en bronnen 69, pp. 23–37. Uitgeverij Verloren, 2000.

Besamusca, B. *Repertorium van de handschriften van de Middelnederlandse Karelepiek. Een beknopte beschrijving van de handschriftelijke en gedrukte overlevering*. HES, 1983.

Besamusca, B. *Repertorium van de handschriften van de Middelnederlandse Arturepiek. Een beknopte beschrijving van de handschriftelijke en gedrukte overlevering*. HES, 1985.

Besamusca, B. (ed.). Lanceloet. *De Middelnederlandse vertaling van de* Lancelot en prose *overgeleverd in de Lancelotcompilatie. Pars 2 (vs. 5531-10740). Met een inleidende studie over de vertaaltechniek*. Middelnederlandse Lancelotromans 5. Van Gorcum, 1991.

Besamusca, B. Walewein, Moriaen *en de* Ridder metter mouwen. *Intertekstualiteit in drie Middelnederlandse Arturromans*. Middeleeuwse studies en bronnen 39. Uitgeverij Verloren, 1993.

Besamusca, B. The medieval Dutch Arthurian material. In Jackson, W. & Ranawake, S. (ed.), *The Arthur of the Germans. The Arthurian Legend in Medieval German and Dutch Literature*, pp. 187–228. University of Wales Press, 2000.

Besamusca, B. *The Book of Lancelot. The Middle Dutch* Lancelot *Compilation and the Medieval Tradition of Narrative Cycles*. Arthurian Studies 53. D.S. Brewer, 2003.

Besamusca, B. Zingende minstrelen, een jonge dichter en een voordracht in Vlaanderen. Over de twee prologen van *Ogier von Dänemark*. In Besamusca, B. & Tigelaar, J. (ed.), *Karolus Rex. Studies over de middeleeuwse verhaaltraditie rond Karel de Grote*, Middeleeuwse studies en bronnen 83, pp. 141–152. Uitgeverij Verloren, 2005.

Besamusca, B. Der Reiz der Versform. In Ridder, K. & Huber, C. (ed.), *Lancelot. Der mittelhochdeutsche Roman im europäischen Kontext*. Niemeyer, 2007.

Besamusca, B. Unidentifizierte mittelniederländische Karlsepen. Die idée fixe der mittelniederländischen Literatur als Übersetzungsliteratur. In Ziegler, H. (ed.), *Chanson de geste im europaïschen Kontext. Ergebnisse der Tagung der Deutschen Sektion der ICLS am 23. und 24.4.2004 in Köln*, Encomia Deutsch 1, pp. 26–35. V&R Unipress, 2008.

Besamusca, B. Het bibliografisch ganzenbord van Lodewijk van Velthems. In Besamusca, B.; Sleiderink, R. & Warnar, G. (ed.), *De boeken van Velthem. Auteur, oeuvre en overlevering*, Middeleeuwse studies en bronnen 119, pp. 73–87. Uitgeverij Verloren, 2009.

Besamusca, B. & Brandsma, F. Between audience and source: the first-person narrator in the Middle Dutch *Lanceloet*. In Busby, K. & Lacy, N. (ed.), *Conjunctures. Medieval Studies in Honor of Douglas Kelly*, Faux titres 83, pp. 15–19. Rodopi, 1994.

Besamusca, B. & Lie, O. The prologue to *Arturs doet*, the Middle Dutch translation of *La mort le Roi Artu* in the *Lancelot compilation*. In Kooper, E. (ed.), *Medieval Dutch Literature in its European context*, Cambridge studies in medieval literature 21, pp. 96–112. Cambridge University Press, 1994.

Besamusca, B. & Postma, A. (ed.). Lanceloet. *De Middelnederlandse vertaling van de* Lancelot en prose *overgeleverd in de* Lancelotcompilatie. *Pars 1 (vs.1-5530, voorafgegaan door de verzen van het Brusselse fragment).* Middelnederladse Lancelotromans 4. Uitgeverij Verloren, 1997.

Besamusca, B.; Sleiderink, R. & Warnar, G. (ed.). *De boeken van Velthem. Auteur, oeuvre en overlevering*. Middeleeuwse studies en bronnen 119. Uitgeverij Verloren, 2009a.

Besamusca, B.; Sleiderink, R. & Warnar, G. Lodewijk van Velthem. Ter inleiding. In Besamusca, B.; Sleiderink, R. & Warnar, G. (ed.), *De boeken van Velthem. Auteur, oeuvre en overlevering*, Middeleeuwse studies en bronnen 119, pp. 7–30. Uitgeverij Verloren, 2009b.

Bibble, M. The making of the Round Table. In Bibble, M. (ed.), *King Arthur's Round Table. An archaeological investigation*, pp. 337–392. Boydell & Brewer, 2000.

Biemans, J. Onsen Speghele Ystoriale in Vlaemsche. *Codicologisch onderzoek naar de overlevering van de* Spiegel historiael *van Jacob van Maerlant, Philip Utenbroeke en Lodewijk van Velthem, met een beschrijving van de handschriften en fragmenten.* Schrift en schriftdragers in de Nederlanden in de middeleeuwen 2. Peeters, 1997. Twee delen.

Biemans, J. De 'corrector' in de Lancelotcodex. Een hachelijk avontuur, waarbij aan het eind toch een nieuw perspectief gloort. In

Besamusca, B.; Sleiderink, R. & Uyttersprot, V. (ed.), *Maar er is meer. Avontuurlijk lezen in de epiek van de Lage Landen: studies voor Jozef D. Janssens*, Davidsfonds Literair, pp. 358–375. Davidsfonds, 2005.

Biemans, J. No miniatures, not even decoration, yet extraordinarily fascinating. New hypotheses concerning the *Lancelot Compilation* and related manuscripts. *Quaerendo*, 39(3–4):225–256, 2009.

Binongo, J. Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, 16(2):9–17, 2003.

Binongo, J. & Smith, W. The application of principal components analysis to stylometry. *Literary and Linguistic Computing*, 14(4):445–466, 1999.

Brinkman, H. 1330. Jan van Boendale wordt berispt wegens passages in *Der leken spieghel*. Een wereldbeeld in verzen. In Schenkeveld-van der Dussen, M. (ed.), *Nederlandse literatuur, een geschiedenis*, pp. 53–58. Nijhoff, 1993.

Brinkman, H. Als de nachtuil. Auteurschap en overlevering van middeleeuwse teksten. [Rede uitgesproken te Amsterdam op 19 maart 2010, ter gelegenheid van Brinkmans aanstelling als hoogleraar aan de Universiteit van Amsterdam], 2010.

Brinkman, H. & Mulder, H. Recht, historie en schone letteren: het arbeidsterrein van een Gents kopiistencollectief. Hs. Brussel KB 16.762-75 en het Comburgse handschrift. *Queeste. Tijdschrift over middeleeuwse letterkunde in de Nederlanden*, 10(1):27–78, 2003.

Brown, R. *A First Language*. Harvard University Press, 1973.

Bulacu, M. & Schomaker, L. Text-independent writer identification and verification using textural and allographic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):701–717, 2007.

Burgers, J. *De rijmkroniek van Holland en zijn auteurs. Historiografie in Holland door de Anonymus (1280-1282) en de grafelijke klerk Melis Stoke (begin veertiende eeuw)*. Historische Vereniging Holland, 1999.

Burrows, J. *Computation into criticism. A Study of Jane Austen's novels and an experiment in methods*. Clarendon Press, 1987.

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

Caers, B. Een *buchelin inn flemische*. Over ontstaan en verspreiding van de ridderepiek in de Nederlanden (ca. 1150-1450). *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 127(3):223–251, 2011.

Caers, B. & Kestemont, M. Over de datering van de Middelnederlandse ridderepiek. *Verslagen en Mededelingen van de Koninklijke Academie voor Nederlandse Taal- en Letterkunde*, 121(1):1–59, 2011.

Chiflet, J. *Le faux Childebrand relegué aux fables [...]*. s.n., 1659.

Chrupała, G. Simple data-driven context-sensitive lemmatization. *Procesamiento del Lenguaje Natural*, 37:121–127, 2006.

Chung, C. & Pennebaker, J. The psychological functions of function words. In Fiedler, K. (ed.), *Social Communication*, pp. 343–359. Psychology Press, 2007.

Claassen, E. *The author's footprints in the garden of fiction. Readers' generation of author inferences in literary reading*. Doctorale dissertatie, Vrije Universiteit Amsterdam, 2009. URL http://dare.ubvu.vu.nl/bitstream/1871/13123/5/5544.pdf.

Claassens, G. Eén ridder, twee vrouwen. Amsterdam, U.B., hs. I A 24k ontraadseld? *Spiegel der Letteren*, 38(4):239–257, 1996.

Claassens, G. Jan van Boendale (ca. 1280-1351). In Jeep, J. (ed.), *Medieval Germany. An encyclopaedia*, Routledge Encyclopaedias of the Middle Ages 6, pp. 405–407. Routledge, 2001.

Claassens, G. (ed.). Karel ende Elegast. Alfa. Literaire teksten uit de Nederlanden. Amsterdam University Press, 2002.

Claassens, G. De *Torrez* à *Torec*. Un roman Arthurien en moyen néerlandais et sa source inconnue en ancien français. In Goyens, M. & Verbeke, W. (ed.), Lors est ce jour grant joie née. *Essais de langue et de littérature françaises du Moyen Âge*, Mediaevalia Lovaniensia Series I, Studia 41, pp. 159–175. Leuven University Press, 2009.

Clark, H. & Clark, E. *Psychology and language. An introduction to psycholinguistics*. Harcourt, Brace & Jovanovich, 1977.

Craig, H. Stylistic analysis and authorship studies. In Schreibman, S.; Siemens, R. & Unsworth, J. (ed.), *A Companion to Digital Humanities*. Backwell, 2004. URL http://www.digitalhumanities.org/companion/.

Croenen, G. Het dubbele auteurschap van de *Grimbergsche oorlog*. In Sleiderink, R.; Uyttersprot, V. & Besamusca, B. (ed.), *Maar er is meer. Avontuurlijk lezen in de epiek van de Lage Landen. Studies voor Jozef D. Janssens*, Davidsfonds Literair, pp. 131–151. Davidsfonds, 2005.

Croenen, G. & Janssens, J. Een nieuw licht op de Lancelotcompilatie? De betekenis van het pas gevonden fragmentje van *Arturs doet* [(1)]. *Queeste. Tijdschrift over middeleeuwse letterkunde in de Nederlanden*, 1 (1):3–11, 1994a.

Croenen, G. & Janssens, J. Een nieuw licht op de Lancelotcompilatie? De betekenis van het pas gevonden fragmentje van *Arturs doet* (2). *Queeste. Tijdschrift over middeleeuwse letterkunde in de Nederlanden*, 1 (2):108–125, 1994b.

Crystal, D. *A Dictionary of Linguistics and Phonetics.* Oxford University Press, 1997.

Daelemans, W. & Van den Bosch, A. *Memory-Based Language Processing*. Studies in Natural Language Processing. Oxford University Press, 2005.

Daelemans, W.; Van den Bosch, A. & Zavrel, J. Forgetting exceptions is harmful in language learning. *Machine Learning*, 34(1–3):11–41, 1999. [Themanummer 'Natural Language Learning', C. Cardie & R. Mooney (ed.)].

Daelemans, W.; Zavrel, J.; Van den Bosch, A. & van der Sloot, K. MBT: Memory-Based Tagger, version 3.2, Reference guide. Technical report, University of Tilburg / University of Antwerp, 2010a. URL http://ilk.uvt.nl/downloads/pub/papers/Mbt_3.2_Manual.pdf. [ILK Research Group Technical Report Series 10-04].

Daelemans, W.; Zavrel, J.; Van der Sloot, K. & Van den Bosch, A. TiMBL: Tilburg Memory-Based Learner, version 6.3, Reference guide. Technical report, Tilburg University / University of Antwerp, 2010b. [ILK Research Group Technical Report Series 10-03].

De Baere, G. De Mengeldichten of Hadewijch met een janusgezicht. *Millennium. Tijschrift voor middeleeuwse studies*, 7(1):40–51, 1993.

De Bundel, K. Van woerde tot woerde oft van synne te sinne. *Petrus Naghel en het* translatorium *van de kartuis te Herne (ca. 1350-1400)*. Doctorale dissertatie, Katholieke Universiteit Leuven, 2009.

De Haan, M. Het bediet van moriane? *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 88(3–4):200–215, 1972.

De Haan, M. & Jongen, L. (ed.). Roman van de riddere metter mouwen. *Opnieuw naar de bewaarde bronnen uitgegeven*. Publikaties van de Vakgroep Nederlandse taal- & letterkunde 11. HES, 1983.

De Vreese, W. Boendales sterfjaar. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 13:157–158, 1894.

De Vreese, W. Boendale (Jan), ook geheten Jan de Clerc. *Zuid-Nederlandsche Taal- en Letterkundige Wetenschap*, 4:83–86, 1900.

De Vries, M. (ed.). *Der leken spieghel. Leerdicht van den jare 1330*. D. du Mortier en zoon, 1844–1848.

De Vries, M. & Verwijs, E. (ed.). *Jacob van Maerlant's* Spiegel historiael, *met de fragmenten der later toegevoegde gedeelten, bewerkt door Philip Utenbroeke en Lodewijc van Velthem.* H&S Publishers, 1982. [Herdruk van de Leidense uitgave uit 1861].

Deschamps, J. *Middelnederlandse handschriften uit Europese en Amerikaanse bibliotheken. Tweede herziene druk.* Brill, 1972.

Dewhurst, Jane. Generic hybridity in Hartmann von Aue's *Der arme Heinrich*. In Busby, Keith & Dalrymple, Roger (ed.), *Arthurian literature XX*, pp. 43–84. D.S. Brewer, 2003.

Diercxsens, J. *Antverpia Christo nascens et crescens [...]. Tomus II.* s.n., 1773.

Dimpel, F. *Computergestützte textstatistische Untersuchungen an mittelhochdeutschen Texten*. A. Francke Verlag, 2004.

Dimpel, F. Der Verlust der *Eneas*-Handschrift als Fiktion. Eine computergestützte, textstatistische Untersuchung. *Amsterdamer Beiträge zur älteren Germanistik*, 61(1):87–102, 2006.

Divaeus, P. *Rerum Brabanticarum libri XIX.* s.n., 1610.

Draak, M. Oude en nieuwe Lancelot-problemen, en de noodzakelijkheid van lezen. *Mededelingen der Koninklijke Nederlandse Academie van Wetenschappen, Afd. Letterkunde*, 39(8):363–378, 1974.

Draux, R. *Beatrijs' biecht. Stilistisch en semiotisch onderzoek*. Lambert Academic Publishing, 2009.

Drewnowki, A. & Healy, A. Detection errors on *the* and *and*: Evidence for reading units larger than the word. *Memory & Cognition*, 5(6): 636–647, 1978.

Duijvestein, B. Madelgijs. Zwerftocht van een epische stof. In De Schutter, G. & Goossens, J. (ed.), *Van* Madelgijs *tot* Malagis. *Een bundel opstellen verzameld naar aanleiding van de tachtigste verjaardag van Gilbert de Smet*, Studies op het gebied van de oudere Nederlandse letterkunde 1, pp. 23–34. Koninklijke Academie voor Nederlandse Taal- en Letterkunde, 2002.

Dunteman, G. *Principal Components Analysis*. Quantitative applications in the social sciences 69. Sage Publications, 1989.

Ernestus, M.; Van Mulken, M. & Baayen, H. Ridders en heiligen en tijd en ruimte: moderne stylometrische technieken toegepast op Oud-Franse teksten. *Taal en Tongval*, 58:1–83, 2007.

Fegley, R. *The golden spurs of Kortrijk: how the knights of France fell to the foot soldiers of Flanders in 1302*. McFarland, 2002.

Finet-van der Schaaf, B. (ed.). *Le Roman de Moriaen. Texte présenté, traduit et annoté*. Collection Moyen Âge européen. ELLUG, Université Stendhal, 2009.

Foucault, Michel. Qu'est-ce qu'un auteur? *Bulletin de la Société française de philosophie*, 3:73–104, 1969.

Fraeters, V. Medioneerlandistiek in context. Literair-historici op zoek naar Hermes en Philologia. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 120(4):298–308, 2004.

Francis, N. & Kucera, H. *Frequency Analysis of English Usage. Lexicon and Grammar*. Houghton Mifflin, 1982.

García, A. & Martín, J. Function words in authorship attribution studies. *Literary and Linguistic Computing*, 22(1):49–66, 2007.

Geirnaert, D. 'Membra disiecta'. Banden met het versneden verleden. In Jansen-Sieben, R.; Janssens, J. & Willaert, F. (ed.), *Medioneerlandistiek: een inleiding tot de Middelnederlandse letterkunde*, Middeleeuwse studies en bronnen 69, pp. 85–101. Verloren, 2000.

Geirnaert, D. & Geirnaert, N. Een onbekend gebleven *Spiegel historiael*-fragment. Nieuw materiaal van Lodewijk van Velthems aandeel in de Vierde Partie. *Verslagen en Mededelingen van de Koninklijke Academie voor Nederlandse Taal- en Letterkunde*, 116(1):147–172, 2006.

Génard, P. Jan van boendale. *Het Taelverbond*, pp. 161–163, 1853.

Génard, P. *Antwerpsch archievenblad. Eerste deel*. Van Merlen, 1864.

Gerritsen, W. Kritische kanttekeningen bij de inleiding tot Heeroma's editie van het Gruuthuse-liedboek. *De nieuwe taalgids*, 62(3):187–215, 1969.

Gerritsen, W. Geschiedverhaal of schetskaart. *De nieuwe taalgids*, 68 (2):89–109, 1975.

Gerritsen, W. Corrections and indications for oral delivery in the Middle Dutch *Lancelot* manuscript, The Hague K.B. 129 A 10. In Gumbert, J. & De Haan, M. (ed.), *Essays presented to G.I. Lieftinck. Deel 3: Neerlandica manuscripta*, Litterae textuales, pp. 39–59. Van Ghendt, 1976.

Gerritsen, W. Wat voor boeken zou Floris V gelezen hebben? In *Floris V. Leven, wonen en werken in Holland aan het einde van de dertiende eeuw. Acht voordrachten gehouden tijdens een symposium te Muiderberg, 29 en 30 september 1978*, Stichting Comité Oud-Muiderberg 14, pp. 71–86. Nijhoff, 1979.

Gerritsen, W. Vertalingen van Oudfranse litteraire werken in het Middelnederlands. In Stuip, R. (ed.), *Franse literatuur van de Middeleeuwen*, pp. 184–207. Coutinho, 1988.

Gerritsen et al., W. (ed.). Lantsloot vander Haghedochte. *Fragmenten van een Middelnederlandse bewerking van de* Lancelot en prose. Middelnederlandse Lancelotromans 2. Noord-Hollandsche Uitgevers Maatschappij, 1987.

Goossens, J. De jambische viervoeters van de Limburgse *Aiol*. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 118(4):317–334, 2002.

Govers, M.; Goris, M.; Van Heusden, P.; Schippers, A.; Sonnemans, G.; Wissink, W. & Kienhorst, H. (ed.). *Het Geraardsbergse handschrift. Hs. Brussel, Koninklijke Bibliotheek Albert I, 837-845*. Middelnederlandse verzamelhandschriften uit de Nederlanden 1. Uitgeverij Verloren, 1994.

Greene, V. What happened to medievalists after the death of the author? In Greene, V. (ed.), *The medieval author in medieval French literature*, Studies in Arthurian and courtly cultures, pp. 205–227. Palgrave Macmillan, 2006.

Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

Gysseling, M. (ed.). *Corpus van Middelnederlandse teksten (tot en met het jaar 1300). Reeks* II, *Literaire handschriften. Deel 1*. Bouwstoffen voor een woordarchief van de Nederlandse taal. Nijhoff, 1980.

Hage, A. Sonder favele, sonder lieghen. *Onderzoek naar vorm en functie van de Middelnederlandse rijmkroniek als historiografisch genre*. Historische Studies XLVIII / Bijdragen van het instituut voor Middeleeuwse Geschiedenis XXXXIX. Wolters-Noordhoff/Forsten, 1989.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P. & Witten, E. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18, 2009.

Heeroma, K. Elegast, Moriaen *en* Lantsloot. *Luisterend gelezen*. Brill, 1973. [Overdruk uit *Tijdschrift voor Nederlandse Taal- en Letterkunde* 89 (1973)].

Heymans, J. (ed.). Van den derden Eduwaert, coninc van Ingelant hoe hij van over die zee is comen in meyningen Vrancrijc te winnen ende hoe hij Doernic belach. *Uitgegeven met een inleiding over de Brabantse historiografie tussen ca. 1270 en ca. 1350*. Tekst en tijd 10. Alpha, 1989a.

Heymans, J. Geschiedenis in *Der Leken Spiegel*. In Dibbets, G. & Wackers, P. (ed.), *Wat duikers vent is dit! Opstellen voor W.M.H. Hummelen*, pp. 25–40. Quarto, 1989b.

Hinskens, F. & Van Dalen-Oskam, K. Kwantitatieve benaderingen in het taal- en letterkundig onderzoek. Een ruwe schets. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 123(1):1–21, 2007. [Themanummer 'Kwantitatieve benaderingen in de taal- en letterkunde', F. Hinskens & K. van Dalen-Oskam (ed.)].

Hogenbirk, M. *Die coenste die ie werd geboren*. Over *Walewein ende Keye*. *De nieuwe taalgids*, 87(1):57–75, 1994.

Hogenbirk, M. *Avontuur en anti-avontuur: een onderzoek naar* Walewein ende Keye, *een Arturroman uit de* Lancelotcompilatie. Uitgaven Stichting Neerlandistiek VU 42. Stichting Neerlandistiek VU, 2004.

Hogenbirk, M. De carrière van de zwarte ridder Moriaen: tussen *Conte du Graal* en *Aliscans*. *Queeste. Tijdschrift over middeleeuwse letterkunde in de Nederlanden*, 16(1):51–73, 2009a.

Hogenbirk, M. Is hij het? Lodewijk van Velthem en de compilator. In Besamusca, B.; Sleiderink, R. & Warnar, G. (ed.), *De boeken van Velthem. Auteur, oeuvre en overlevering*, Middeleeuwse studies en bronnen 119, pp. 47–72. Uitgeverij Verloren, 2009b.

Hogenbirk, M. The "I–word" and genre: Merging epic and romance in the *Roman van Walewein*. In Jones, C. & Whalen, L. (ed.), Li Premerains Vers. *Essays in Honor of Keith Busby*, Faux Titre 361, pp. 157–170. Rodopi, 2011.

Holmes, D. Authorship attribution. *Computers and the Humanities*, 28 (2):87–106, 1994.

Holmes, D. The evolution of stylometry in Humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117, 1998.

Hoover, D. Multivariate analysis and the study of literary style. *Literary and Linguistic Computing*, 18(4):341–360, 2003.

Hoover, D. Testing Burrows's Delta. *Literary and Linguistic Computing*, 19(4):453–475, 2004a.

Hoover, D. Delta prime. *Literary and Linguistic Computing*, 19(4):477–495, 2004b.

Hoover, D. Quantitative analysis and literary studies. In Schreibman, S. & Siemens, R. (ed.), *A companion to digital literary studies*, pp. 517–533. Blackwell, 2007.

Houthuys, A. *Middeleeuws kladwerk. De autograaf van de* Brabantse yeesten, *boek VI (vijftiende eeuw)*. Middeleeuwse studies en bronnen 117 & Schrift en schriftdragers in de Nederlanden in de middeleeuwen 4. Uitgeverij Verloren, 2009.

Houvardas, J. & Stamatatos, E. N-gram feature selection for authorship identification. In Euzenat, J. & Domingue, J. (ed.), *Proceedings of Artificial Intelligence: Methodologies, Systems, and Applications*, pp. 77–86. Springer Verlag, 2006.

Janssens, J. *Dichter en publiek in creatief samenspel. Over interpretatie van Middelnederlandse ridderromans*. Leuvense studieën en tekstuitgaven (nieuwe reeks) 7. Acco, 1988.

Janssens, J. Een geschiedenis van de Middelnederlandse epiek. Van publiek naar dichter: terug naar af...? In Van Oostrom, F. & Willaert, F. (ed.), *Grote lijnen. Syntheses over Middelnederlandse letterkunde*, Nederlandse literatuur en cultuur in de middeleeuwen 11. Prometheus, 1995.

Janssens, J. Van Mierlo of Van Oostrom? Literaire netwerken in de Middeleeuwen. In Janssens, J. (ed.), *Op avontuur. Middeleeuwse epiek in de Lage Landen*, Nederlandse literatuur en cultuur in de middeleeuwen 18, pp. 261–272 & 367–369. Prometheus, 1998a.

Janssens, J. Subtiel vertellen. Middeleeuwse epiek in de Lage Landen. In Janssens, J. (ed.), *Op avontuur. Middeleeuwse epiek in de Lage Landen*, Nederlandse literatuur en cultuur in de middeleeuwen 18, pp. 9–35 & 273–285. Prometheus, 1998b.

Janssens, J. Een kwarteeuw middelneerlandistiek. Weifelen tussen algemene literatuurwetenschap en cultuurgeschiedenis: een subjectieve geschiedenis. In de Clercq, M.; de Vin, D.; Janssens, J. & Mattheeusen, C. (ed.), *Verzoenende veelzijdigheid. Huldealbum opgedragen aan Prof. Dr. H. van Gorp*, pp. 109–126. Katholieke Universiteit Brussel, 2000.

Janssens, J.; Van Daele, R. & Uyttersprot, V. (ed.). Van den vos Reynaerde. *Het Comburgse handschrift*. Davidsfonds, 1998.

Johnson, D. Middle Dutch Arthurian romances: What are they and why should we read them? *Arthuriana*, 15(2):1–2, 2005.

Johnson, D. & Claassens, G. (ed.). *Dutch romances 2:* Ferguut. Arthurian Archives 7. D.S. Brewer, 2000a.

Johnson, D. & Claassens, G. (ed.). *Dutch romances 1.* Roman van Walewein. Arthurian Archives 6. D.S. Brewer, 2000b.

Johnson, D. & Claassens, G. (ed.). *Dutch romances 3. Five interpolated romances from the* Lancelot compilation. Arthurian Archives 10. D.S. Brewer, 2003.

Jonckbloet, W. (ed.). Die Dietsche doctrinale, *leerdicht van den jare 1345, toegekend aan Jan Deckers, clerk der stad Antwerpen*. s.n., 1842.

Jones, Oliphant T. E. & Peterson et al., P. Open source scientific tools for Python, 2001. URL http://www.scipy.org.

Jongen, L. & Piters, M. (ed.). Ghi Fransoyse sijt hier onteert. *De Guldensporenslag. Lodewijk van Velthem. Kritische editie van de Middelnederlandse tekst uit de* Voortzetting *van de* Spiegel historiael *met inleiding en vertaling*. Davidsfonds, 2002.

Juola, P. Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334, 2006.

Juola, P. Becoming Jack London. *Journal of Quantitative Linguistics*, 14 (2):145–147, 2007.

Kestemont, M. Een onderzoek met stijl. *Queeste. Tijdschrift over middeleeuwse letterkunde in de Nederlanden*, 14(2):174–182, 2007. [Recensie van Van Driel 2007].

Kestemont, M. Bloemlezen uit Velthem. Handschrift Wenen, Ö.N.B., Cod. 13.708 in Herne ten tijde van het Westers Schisma. In Besamusca, B.; Sleiderink, R. & Warnar, G. (ed.), *De boeken van Velthem. Auteur, oeuvre en overlevering*, Middeleeuwse studies en bronnen 119, pp. 251–266. Uitgeverij Verloren, 2009.

Kestemont, M. Velthem *et al*. A stylometric analysis of the rhyme words in the account of the Battle of the Golden Spurs in the fifth part of the *Spiegel historiael*. *Queeste. Tijdschrift over middeleeuwse letterkunde in de Nederlanden*, 17(1):1–34, 2010a.

Kestemont, M. Seghers wapenfeiten. Oude en nieuwe hypotheses omtrent de *Trojeroman*, het huis van Gaasbeek en het handschrift–Van Hulthem. *Spiegel der Letteren*, 52(3):249–275, 2010b.

Kestemont, M. Een stylometrisch onderzoek naar Jan van Boendales auteurschap voor de *Brabantse yeesten*. *Revue Belge de Philologie et d'Histoire – Belgisch Tijdschrift voor Filologie en Geschiedenis*, 89(4): 1019–1048, 2011a.

Kestemont, M. Eind goed, al goed? *Queeste. Tijdschrift over middeleeuwse letterkunde in de Nederlanden*, 18(1):81–84, 2011b. [Recensie van Draux 2009].

Kestemont, M. Het gewicht van de auteur. Reflecties over auteursattributie in samenleving en wetenschap. *Streven. Cultureel maatschappelijk maandblad*, 78(3):234–244, 2011c.

Kestemont, M. De meesters van de *Spiegel*. Auteursonderscheiding op basis van het frequente rijmwoord in het aandeel van Utenbroeke en Maerlant in de *Spiegel historiael*. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 127(2):97–121, 2011d.

Kestemont, M. Stylometry for medieval authorship studies: An application to rhyme words. *Digital Philology. A Journal of Medieval Cultures*, 1(1):42–72, 2012a.

Kestemont, M. What can stylometry learn from its application to Middle Dutch literature? *Journal of Dutch Literature*, 2(2):46–65, 2012b.

Kestemont, M. Auteursherkenning met rijmwoorden in de Middelnederlandse Artur- en Karelepiek. Eerherstel voor Icarus? *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 128(2):135–159, 2012c.

Kestemont, M. Arthur's authors. A quantitative study into the rhyme words of the Middle Dutch Arthurian epics. *Zeitschrift für deutsches Altertum und deutsche Literatur*, 142:1–33, 2013.

Kestemont, M. & van Dalen-Oskam, K. Predicting the past. Memory-based copyist and author discrimination in medieval epics. In Calders, T.; Tuyls, K. & Pechenizkiy, M. (ed.), *Proceedings of the twenty-first Benelux conference on artificial intelligence*, pp. 121–128, 2009.

Kestemont, M.; Daelemans, W. & De Pauw, G. Weigh your words – Memory-based lemmatization for Middle Dutch. *Literary and Linguistic Computing*, 25(3):287–301, 2010.

Kestemont, M.; Daelemans, W. & Sandra, D. Robust rhymes? The stability of authorial style in medieval narratives. *Journal of Quantitative Linguistics*, 19(1):54–76, 2012.

Kienhorst, H. *De handschriften van de Middelnederlandse ridderepiek. Een codicologische beschrijving. Deel 1*. Deventer studieën 9. Sub Rosa, 1988.

Kienhorst, H. Fragment van een onbekende Middelnederlandse ridderroman over Willem van Oringen: Oxford, Bodleian Library, Ms. Dutch b. 2, fol. 6. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 114 (2):125–137, 1998.

Kienhorst, H. De *Wrake van Ragisel*-fragmenten. Lay-out en opkomst van het literaire handschrift in de dertiende eeuw. *Jaarboek voor Nederlandse Boekgeschiedenis*, 6:49–66, 1999.

Kilgariff, A. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133, 2001.

Kinable, D. Geïntendeerde publieksgroepen in Boendales *Lekenspiegel* en *Jans Teesteye*. In Pleij, H. (ed.), *Op belofte van profijt. Stadsliteratuur en burgermoraal in de Nederlandse letterkunde van de middeleeuwen*, Nederlandse literatuur en cultuur in de middeleeuwen 4, pp. 69–100 & 357–362. Prometheus, 1991.

Kinable, D. *Facetten van Boendale. Literair-historische verkenningen van* Jans Teesteye *en de* Lekenspiegel. Leidse opstellen 31. Dimensie, 1997.

Kjell, B. Discrimination of authorship using visualization. *Information Processing and Management*, 30(1):141–50, 1994.

Klein, J.-W. *Het getal zijner jaren is onnaspeurlijk*. Een herijking van de dateringen van de handschriften en fragmenten met Middelnederlandse ridderepiek. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 111(1):1–23, 1995.

Klein, J.-W. De status van de *Lancelotcompilatie*: handschrift, fragmenten en personen. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 114 (2):105–124, 1998.

Klein, T. Zur Sprache des Münchener Prosa-Lancelot-Fragments. *Amsterdamer Beiträge zur älteren Germanistik*, 38/39:223–240, 1994.

Knowles, G. & Mohd Don, Z. The notion of a "lemma". Headwords, roots and lexical sets. *International Journal of Corpus Linguistics*, 9(1): 69–81, 2004.

Köhler, R. *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Brockmeyer, 1986.

Köhler, R. Synergetic linguistics. In Köhler, R.; Altmann, G. & Piotrowski, R. (ed.), *Quantitative Linguistik. Ein internationales Handbuch – Quantitative Linguistics. An International Handbook*, pp. 760–775. Walter de Gruyter, 2005.

Kooper, E. Introduction. In Kooper, E. (ed.), *Medieval Dutch literature in its European context*, Cambridge studies in medieval literature 21, pp. 1–8. Cambridge University Press, 1994.

Koppel, M.; Schler, J.; Argamon, S. & Messeri, E. Authorship attribution with thousands of candidate authors. In Efthimiadis, E.; Dumais, S.; Hawking, D. & Järvelin, K. (ed.), *Proceedings of the 29th International Conference of the Special Interest Group on Information Retrieval*, pp. 659–660, 2006.

Koppel, M.; Schler, J. & Argamon, S. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26, 2009.

Kors, M. *De bijbel voor leken: studies over Petrus Naghel en de Historiebijbel van 1361; ingeleid door Geert H.M. Claassens*. Publicaties van de Stichting Encyclopédie Bénédictine 4. Encyclopédie Bénédictine, 2007.

Kuiper, W. (ed.). Die riddere metten witten scilde. *Oorsprong, overlevering en auteurschap van de Middelnederlandse* Ferguut, *gevolgd door een diplomatische editie en een diplomatisch glossarium*. Thesaurus 2. Schipbouwer & Brinkman, 1989.

Kuiper, W. De filoloog als patholoog-anatoom. *Queeste. Tijdschrift over middeleeuwse letterkunde in de Nederlanden*, 5(2):172–180, 1998. [Themanummer over het Comburgse handschrift].

Kuiper, W. Een digitale knekelput. *Neder-L*, column 75, 2010. URL http://nederl.blogspot.com/2010/06/een-digitale-knekelput.html.

Kwakkel, E. Die Dietsche boeke die ons toebehoeren. *De kartuizers van Herne en de productie van Middelnederlandse handschriften in de regio Brussel (1350-1400)*. Miscellanea neerlandica 27. Peeters, 2002.

Kwakkel, E. & Mulder, H. *Quidam sermones*. Mystiek proza van de *Ferguut*-kopiist. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 117 (2):151–165, 2001.

Kwon, E.-Y. The "Natural Order" of morpheme acquisition. A historical survey and discussion of three putative determinants. *Teachers' College Columbia Working Papers in TESOL and Applied Linguistics*, 5 (1):1–21, 2005.

Lagast, A. *A la recherche de l'oeuvre perdue*. Kritische status quaestionis van het onderzoek naar de *Madoc*. *Millenium. Tijdschrift voor middeleeuwse studies*, 24(1):19–33, 2010.

Lewontin, R. On the measurement of relative variability. *Systematic Biology*, 15(2):141–142, 1966.

Lie, O. *The Middle Dutch Prose Lancelot. A study of the Rotterdam fragments and their place in the French, German, and Dutch* Lancelot en prose *tradition*. Middelnederlandse Lancelotromans 3. Noord-Hollandsche Uitgevers Maatschappij, 1987.

Lie, O. What is truth? The verse-prose debate in medieval Dutch literature. *Queeste. Tijdschrift over middeleeuwse letterkunde in de Nederlanden*, 1(1):34–65, 1994.

Lie, O. Middelnederlandse literatuur vanuit genderperspectief. Een verkenning. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 117(3): 246–267, 2001.

Lievens, R. De dichter Hein van Aken. *Spiegel der Letteren*, 4(1):57–74, 1960.

Lodder, F. Strijd tegen de bent van letterknechtjes. Johannes van Vloten (1818-1883). In Van Anrooij, W.; Hogenelst, D. & Warnar, G. (ed.),

Der vaderen boek. *Beoefenaren van de studie der Middelnederlandse letterkunde: studies voor Frits van Oostrom ter gelegenheid van diens vijftigste verjaardag*, pp. 77–90. Amsterdam University Press, 2003.

Love, H. *Attributing authorship. An introduction*. Cambridge University Press, 2002.

Lucas, H. Edward III and the poet chronicler John Boendale. *Speculum*, 12(3):367–369, 1937.

Luyckx, K. *Scalability Issues in Authorship Attribution*. Doctorale dissertatie, Universiteit Antwerpen, 2010.

Luyckx, K. & Daelemans, W. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, 26 (1):35–55, 2011.

Lyras, D.; Sgarbas, K. & Fakotakis, N. Applying similarity measures for automatic lemmatization. A case-study for modern Greek and English. *International Journal on Artificial Intelligence Tools*, 17(5): 1043–1064, 2008.

Manning, D.; Raghavan, P. & Schütze, H. *An introduction to Information Retrieval*. Cambridge University Press, 2008. URL http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf.

Mantingh, E. *Deen voren ende dander na*. De compositie van de Vijfde Partie en Velthems bewerking van Heelu's *Slag bij Woeringen*. In Besamusca, B.; Sleiderink, R. & Warnar, G. (ed.), *De boeken van Velthem. Auteur, oeuvre en overlevering*, Middeleeuwse studies en bronnen 119, pp. 207–230. Uitgeverij Verloren, 2009.

McCusker, L.; Gough, P. & Bias, R. Word recognition inside out and outside in. *Journal of Experimental Psychology: Human Perception and Performance*, 7(3):538–551, 1981.

Minnen, B. & Claassens, G. De *Roman van Lancelot* in Middelnederlands proza: het fragment-Wezemaal. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 121(2):169–183, 2005.

Moore, R. There's no data like more data: but when will enough be enough? In *Proceedings of the IEEE international workshop on intelligent signal processing*, 2001.

Morrow, D. Grammatical morphemes and conceptual structure in discourse processing. *Cognitive Science*, 10(4):423–455, 1986.

Mosteller, F. & Wallace, D. *Inference and disputed authorship: The Federalist*. Addison-Wesley, 1964.

Murk Jansen, S. Hadewijch's *Mengeldichten*. An experiment in statistical methods. *Dutch crossing*, 35:26–39, 1988.

Murk Jansen, S. *The measure of mystic thought. A study of Hadewijch's* Mengeldichten. Göppinger Arbeiten zur Germanistik 536. Kümmerle Verlag, 1991.

Nenadić, O. & Greenacre, M. Correspondence Analysis in R, with two- and three-dimensional graphics. The CA-package. *Journal of Statistical Software*, 20(3):1–13, 2007.

Nichols, S. Why material philology? Some thoughts. *Zeitschrift für deutsche Philologie*, 116:10–30, 1997. [Themanummer 'Philologie als Textwissenschaft. Alte und neue Horizonte', H. Tervooren & H. Wenzel (ed.)].

Oppenhuis de Jong, S. (ed.). *De Middelnederlandse* Perceval-*traditie. Inleiding en editie van de bewaarde fragmenten van een Middelnederlandse vertaling van de* Perceval *of* Conte du Graal *van Chrétien de Troyes, en de* Perchevael *in de* Lancelotcompilatie. Middelnederlandse Lancelotromans 9. Uitgeverij Verloren, 2003.

Paardekooper-van Buuren, H. & Gysseling, M. (ed.). Moriaen. *Opnieuw uitgegeven en geannoteerd*. Klassiek Letterkundig Panthéon 183. Thieme, 1971.

Pearson, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

Peng, F.; Schuurmans, D.; Keselj, V. & Wang, S. Language independent authorship attribution using character level language models. In EACL (ed.), *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 267–274, 2003.

Postma, A. Overzicht van *Scolastica*-handschriften. In Van Moolenbroek, J. & Mulder, M. (ed.), Scolastica willic ontbinden. *Over de* Rijmbijbel *van Jacob van Maerlant*, Middeleeuwse studies en bronnen 25, p. 145. Uitgeverij Verloren, 1991.

Prims, F. De stadsrekening van 1313. Het kapittel der lijftochten. *Antwerpsch Archievenblad*, 2e reeks, 6(4):228–245, 1931.

Prims, F. *Geschiedenis van Antwerpen. IV, 2*. Standaard Uitgeverij, 1933.

Prims, F. *Geschiedenis van Antwerpen. V, 3*. Standaard Boekhandel, 1935.

Quinlan, J. Induction of decision trees. *Machine Learning*, 1:81–206, 1986.

Rawlinson, G. *The significance of letter position in word recognition*. Doctorale dissertatie, University of Nottingham, 1976.

Rem, M. Middelnederlands met de computer. *Madoc*, 19(1):24–32, 2005.

Reynaert, J. De 10e Brief van Hadewijch en het 41ste der Limburgse Sermoenen. *Leuvense bijdragen*, 63(2):137–149, 1974.

Reynaert, J. Boendale of 'Antwerpse school'? Over het auteurschap van *Melibeus* en *Dietsche doctrinale*. In van Anrooij, W. (ed.), Al t'Antwerpen in die stad. *Jan van Boendale en de literaire cultuur van zijn tijd*, Nederlandse literatuur en cultuur in de middeleeuwen 24, pp. 127–157 & 177–182. Prometheus, 2002.

Reynaert, J. *Ene suptile clergie, daer groet soetheit in leit*. Filosofie en Middelnederlandse literatuur omstreeks het begin van de veertiende eeuw. *Queeste. Tijdschrift over middeleeuwse letterkunde in de Nederlanden*, 10(1):1–14, 2003.

Riedel-Bierschwale, H. *Das* Laiendoctrinal *des Erhart Groß. Edition und Untersuchung*. Studien und Texte zum Mittelalter und zur frühen Neuzeit 15. Waxmann Verlag, 2009.

Rudman, J. Non-traditional authorship studies in the *Historia Augusta*: Some caveats. *Literary and Linguistic Computing*, 13(3):152–157, 1998a.

Rudman, J. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31(4):351–365, 1998b.

Rybicki, J. & Eder, M. Deeper Delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, 26(3):315–321, 2011.

Salemans, B. *Building Stemmas with the Computer in a Cladistic, Neo-Lachmannian, Way. The Case of Fourteen Text Versions of* Lanseloet van Denemerken. Doctorale dissertatie, Katholieke Universiteit Nijmegen, 2000.

Sapir, E. *Language. An Introduction to the Study of Speech*. Harcourt, Brace & Co, 1921.

Schindler, R. The effect of prose context on visual search for letters. *Memory & Cognition*, 6(2):124–130, 1978.

Schlens, J. A tutorial on Principal Components Analysis, April 2009. URL http://www.snl.salk.edu/~shlens/pca.pdf. [Version 3.01].

Schmidt, K. *The authorship of* The Equatorie of the planets. D.S. Brewer, 1993.

Schoenaers, D. *My es leet dat ics vergat*. De Vijfde Partie van Lodewijk van Velthem en het vierde boek van Jan van Boendales *Brabantse yeesten*. In Besamusca, B.; Sleiderink, R. & Warnar, G. (ed.), *De boeken van Velthem. Auteur, oeuvre en overlevering*, Middeleeuwse studies en bronnen 119, pp. 231–250. Uitgeverij Verloren, 2009.

Schreibman, S.; Siemens, R. & Unsworth, J. (ed.). *A companion to Digital Humanities*. Blackwell, 2004. URL http://www.digitalhumanities.org/companion/.

Sebastiani, F. Machine Learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

Sleiderink, R. *De stem van de meester. De hertogen van Brabant en hun rol in het literaire leven (1106–1430)*. Nederlandse literatuur en cultuur in de middeleeuwen 25. Prometheus, 2003.

Sleiderink, R. Lodewijk van Velthem. De onnoemelijke ambitie van een Brabantse dorpspastoor. In Sleiderink, R.; Uyttersprot, V. & Besamusca, B. (ed.), *Maar er is meer. Avontuurlijk lezen in de epiek van de Lage Landen. Studies voor Jozef D. Janssens.*, Davidsfonds Literair, pp. 109–130. Davidsfonds, 2005.

Sleiderink, R. De verscheurde lelie. Velthem en het toekomstvisioen van Merlijn in de Vijfde Partie. In Besamusca, B.; Sleiderink, R. & Warnar, G. (ed.), *De boeken van Velthem. Auteur, oeuvre en overlevering*, Middeleeuwse studies en bronnen 119, pp. 161–182. Uitgeverij Verloren, 2009.

Smith, S. Van koning tot kroonprins. Over de structuur van de *Roman van den riddere metter mouwen*. In de Bree, F. & Zemel, R. (ed.), In onse scole. *Opstellen over Middeleeuwse letterkunde voor Prof. Dr. Margaretha H. Schenkeveld*, pp. 109–141. Stichting Neerlandistiek VU, 1989.

Snellaert, F. (ed.). *Nederlandsche gedichten uit de veertiende eeuw van Jan van Boendale, Hein van Aken en anderen*. Hayez, 1869.

Sodmann, T. (ed.). *Jacob van Maerlant*, Historie van den Grale *und* Boek van Merline*; nach der Steinfurter Handschrift herausgegeben*. Niederdeutsche Studien 26. Böhlau, 1980.

Sonnemans, G. Middelnederlandse versprologen. De oorspronkelijkheid van de overgeleverde redacties. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 108(4):324–353, 1992.

Sonnemans, G. *Functionele aspecten van Middelnederlandse versprologen*. Doctorale dissertatie, Radboud Universiteit Nijmegen, 1995.

Spijker, I. *Aymijns kinderen hoog te paard. Een studie over* Renout van Montalbaen *en de Franse* Renaut-*traditie*. Middeleeuwse studies en bronnen 22. Uitgeverij Verloren, 1990.

Stamatatos, E. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

Stamatatos, E.; Fakotakis, N. & Kokkinakis, G. Automatic text categorization in terms of genre and author. *Computational linguistics*, 26 (4):471–495, 2000.

Stamou, C. Stylochronometry: stylistic development, sequence of composition, and relative dating. *Literary and Linguistic Computing*, 23(2):181–199, 2008.

Stapel, R. Layer on layer. 'Computational archaeology' in 15th-century Middle Dutch historiography. *Literary and Linguistic Computing*, advance access, December 2012. doi: 10.1093/llc/fqs046.

Stein, R. Wanneer schreef Jan van Boendale zijn *Brabantsche yeesten*? *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 106(4):262–280, 1990.

Stein, R. Jan van Boendales *Brabantsche yeesten*: antithese of synthese? *Bijdragen en mededelingen betreffende de geschiedenis der Nederlanden*, 106(2):185–197, 1991.

Stein, R. *Politiek en historiografie. Het ontstaansmilieu van Brabantse kronieken in de eerste helft van de vijftiende eeuw*. Miscellanea neerlandica 10. Peeters, 1994.

Stella, F. Analisi informatiche del lessico e individuazione degli autori nelle *Epistolae duorum amantium* (XII secolo). In Wright, R. (ed.), *Latin vulgaire – latin tardif VIII. Actes du VIIIe colloque international sur le latin vulgaire et tardif, Oxford, 6-9 septembre 2006*, pp. 560–569. Olms-Weidmann, 2008.

Stoop, P. *Schrijven in commissie. Middelnederlandse biechtvaderpreken uit het Brusselse regularissenklooster Jericho in hun literaire context*. Doctorale dissertatie, Universiteit Antwerpen, 2009.

Summerfield, T. Schrijven in Maerlants schaduw. Lodewijk van Velthem, Edward I en koning Artur. In Besamusca, B.; Sleiderink, R. & Warnar, G. (ed.), *De boeken van Velthem. Auteur, oeuvre en overlevering*, Middeleeuwse studies en bronnen 119, pp. 183–206. Uitgeverij Verloren, 2009.

Talmy, L. *Towards a Cognitive Semantics*. The MIT Press, 2000.

Te Winkel, J. (ed.). Roman van Moriaen. Bibliotheek van Middelnederlandsche letterkunde 20. J.B. Wolters, 1878.

Te Winkel, J. *De ontwikkelingsgang der Nederlandsche letterkunde*. De Erven F. Bohn, 1922.

Team, R Development Core. R: A language and environment for statistical computing, 2011. URL http://www.R-project.org.

Thaisen, J. Probabilistic modeling of Middle English scribal orthographies. In *Digital Humanities 2010. Conference Abstracts*, pp. 37–39, 2010.

Tigelaar, J. *Brabants historie ontvouwd.* Die alder excellentste cronyke van Brabant *en het Brabantse geschiedbeeld anno 1500.* Middeleeuwse studies en bronnen 98. Uitgeverij Verloren, 2006.

Treschow, M.; Gill, P. & Swartz, T. King Alfred's scholarly writings and the authorship of the first fifty prose psalms. *The Heroic Age. A Journal of Early Medieval Northwestern Europe*, 12, May 2009. URL http://www.mun.ca/mst/heroicage/issues/12/treschowgillswartz.php.

Trio, P. Vlaanderen de Leeuw. Het relaas van de slag bij Kortrijk opnieuw aan de bronnen getoetst. In Trio, P.; Heirbaut, D. & Auweele, D. (ed.), *Omtrent 1302*, Alfred Cauchie reeks [4], pp. 81–110. Universitaire Pers Leuven, 2002.

Trio, P.; Heirbaut, D. & Auweele, D. (ed.). *Omtrent 1302*. Alfred Cauchie reeks [4]. Universitaire Pers Leuven, 2002.

Vaessens, T. *De revanche van de roman. Literatuur, autoriteit en engagement*. Vantilt, 2009.

Van Anrooij, W. Recht en rechtvaardigheid binnen de Antwerpse school. In Reynaert, J. (ed.), *Wat is wijsheid? Lekenethiek in de Middelnederlandse letterkunde*, Nederlandse literatuur en cultuur in de middeleeuwen 9, pp. 149–163 & 399–405. Prometheus, 1994a.

Van Anrooij, W. *Boek van de wraak Gods*. Griffioen. Querido, 1994b.

Van Anrooij, W. Boendales *Boec van der wraken*: datering en ontstaansgeschiedenis. *Queeste. Tijdschrift over middeleeuwse letterkunde in de Nederlanden*, 2(1):40–53, 1995.

Van Anrooij, W. Literatuur in Antwerpen in de periode ca. 1315-1350. In Van Anrooij, W. (ed.), Al t'Antwerpen in die stad. *Jan van Boendale en de literaire cultuur van zijn tijd*, Nederlandse literatuur en cultuur in de middeleeuwen 24, pp. 9–16. Prometheus, 2002.

Van Bossele, J. Een oud Engels liedje over de Guldensporenslag. *De Gids. Tijdschrift van de West-Vlaamse gidsenkring Kortrijk–Menen vzw*, 92:181–186, 2002.

Van Buuren, A. Een *Moriaen*-fragment geïdentificeerd. *De nieuwe taalgids*, 67(1):41–52, 1973.

Van Daele, R. De robotfoto van de Reynaertdichter. Bricoleren met de overgeleverde wrakstukken: "cisterciënzers", "grafelijk hof" en "Reynaertmaterie"'. *Tiecelijn. Tijdschrift voor reynaerdofielen*, 18(3): 179–205, 2005.

Van Dalen-Oskam, K. *Studies over Jacob van Maerlants* Rijmbijbel. Middeleeuwse studies en bronnen 57. Uitgeverij Verloren, 1997.

Van Dalen-Oskam, K. Kwantificeren van stijl. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 123(1):37–54, 2007. [Themanummer 'Kwantitatieve benaderingen in de taal- en letterkunde', F. Hinskens & K. van Dalen-Oskam (ed.)].

Van Dalen-Oskam, K. Digital resources and computational methods. *Zeitschrift für deutsche Philologie*, 130:375–390, 2011. [Themanummer 'Dialog mit den Nachbarn. Mittelniederländische Literatur zwischen dem 12. und 16. Jahrhundert', B. Bastert, H. Tervooren & F. Willaert (ed.)].

Van Dalen-Oskam, K. The secret life of scribes. Exploring fifteen manuscripts of Jacob van MaerlantÕs *Scolastica*. *Literary and Linguistic Computing*, 27(4):355–372, 2012.

Van Dalen-Oskam, K. & Van Zundert, J. Delta for Middle Dutch – Author and copyist distinction in *Walewein*. *Literary and Linguistic Computing*, 22(3):345–362, 2007.

Van den Berg, E. *Middelnederlandse versbouw en syntaxis. Ontwikkelingen in de versifikatie van verhalende poëzie ca. 1200 – ca. 1400*. H&S, 1983.

Van den Berg, E. Over het lokaliseren van Middelnederlandse rijmteksten. *Verslagen en Mededelingen van de Koninklijke Academie voor Nederlandse Taal- en Letterkunde*, s.n.(3):305–322, 1986.

Van den Berg, E. Van wiganten, onvervaerde helden en fiere ridders. Epithetische persoonsaanduidingen in de Middelnederlandse ridderepiek. *De nieuwe taalgids*, 81(2):97–110, 1988.

Van den Berg, E. Nadrukformules in Middelnederlandse ridderepiek. *De nieuwe taalgids*, 85(3):205–214, 1992a.

Van den Berg, E. Vorm en inhoud: ontwikkelingen binnen de ridderepiek ca. 1200 – ca. 1350. *De nieuwe taalgids*, 85(5):405–421, 1992b.

Van den Berg, E. & Besamusca, B. De Karelepiek in vogelvlucht. In Van den Berg, E. & Besamusca, B. (ed.), *De epische wereld. Middelnederlandse Karelromans in wisselend perspectief*, pp. 9–23. Countinho, 1992.

Van der Have, B. Roman der Lorreinen. *De fragmenten en het geheel*. Scriptum, 1990.

Van der Have, B. De Middelnederlandse Karelepiek. De fragmenten en het geheel. In Besamusca, B. & Tigelaar, J. (ed.), *Karolus Rex. Studies over de middeleeuwse verhaaltraditie rond Karel de Grote*, Middeleeuwse studies en bronnen 83, pp. 77–92. Uitgeverij Verloren, 2005.

Van der Voort van der Kleij, J. Reverse lemmatization of the dictionary of Middle Dutch (1885-1929) using pattern matching. In Kiefer, F.; Kiss, G. & Pajzs, J. (ed.), *Papers in Computational Lexicography*, pp. 203–210. Hungarian Academy of Sciences, 2005.

Van Driel, J. Stijlen van onderzoek. Anderhalve eeuw medioneerlandistiek van stilistisch perspectief. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 119(2):97–117, 2003.

Van Driel, J. *Prikkeling der zinnen. De stilistische diversiteit van de Middelnederlandse epische poëzie.* Walburg Pers, 2007.

Van Halteren, H.; Baayen, H.; Tweedie, F.; Haverkort, M. & Neijt, A. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1):65–77, 2005.

Van Hemelryck, T. Oublier l'auteur au moyen âge? Réflexions sur l'*ars oblivionis* des copistes. *Etudes de lettres*, 1-2:273–289, 2007.

Van Loey, A. Rijmonderzoek en Middelnederlandse dialectologie. *Handelingen van de Koninklijke Commissie voor Toponymie en Dialectologie*, 20:41–48, 1946.

Van Mierlo, J. *Geschiedenis van de letterkunde der Nederlanden. Deel 2.* Standaard Uitgeverij, 1940.

Van Mierlo, J. (ed.). *Hadewijch.* Mengeldichten. Leuvense studiën en tekstuitgaven 15. Standaard-Boekhandel, 1952.

Van Mierlo, J. De echtheid van de *Heimelichede der Heimelicheit* als werk van J. van Maerlant. In Draye, H. (ed.), *Feestbundel H.J. van de Wijer, den jubilaris aangeboden ter gelegenheid van zijn vijfentwintigjarig hoogleeraarschap aan de R.K. Universiteit te Leuven 1919-1943*, pp. 235–246. Instituut voor Vlaamsche toponymie, 1994.

Van Moolenbroek, J. & Mulder, M. (ed.). Scolastica willic ontbinden. *Over de* Rijmbijbel *van Jacob van Maerlant*. Middeleeuwse studies en bronnen 25. Uitgeverij Verloren, 1991.

Van Oostrom, F. De oorspronkelijkheid van de Torec, of De vrije val van een detail door de Nederlandse litteratuurgeschiedenis. *Spiegel der Letteren*, 21(3):197–201, 1979.

Van Oostrom, F. Lantsloot vander Haghedochte. *Onderzoekingen over een Middelnederlandse bewerking van de* Lancelot en prose. Middelnederlandse Lancelotromans 1. Noord-Hollandse Uitgevers Maatschappij, 1981.

Van Oostrom, F. Maecenaat en Middelnederlandse letterkunde. In Janssens, J. (ed.), *Hoofsheid en devotie in de middeleeuwse maatschappij:*

De Nederlanden van de 12e tot de 15e eeuw. Handelingen van het wetenschappelijk colloquium te Brussel 21–24 oktober 1981, pp. 21–40. s.n. [Brussel], 1982.

Van Oostrom, F. Hoe snel dichtten middeleeuwse dichters? *Literatuur*, 1(6):327–335, 1984.

Van Oostrom, F. Het verraad van Pepijn (Leiden, U.B., Ltk. 1199). In Buitink, S.; Van Buuren, A. & Spijker, I. (ed.), Rapiarijs. *Een afscheidsbundel voor Hans van Dijk*, Ruygh-bewerp 16, pp. 83–84. Instituut De Vooys voor Nederlandse Taal- en Letterkunde, 1987.

Van Oostrom, F. Maerlant tussen Noord en Zuid. Contouren van een biografie. In *Aanvaard dit werk. Over Middelnederlandse dichters en hun publiek*, Nederlandse literatuur en cultuur in de middeleeuwen 6, pp. 185–216 & 299–306. Prometheus, 1992.

Van Oostrom, F. *De waarde van het boek*. Amsterdam University Press, 1994.

Van Oostrom, F. *Maerlants wereld*. Prometheus, 1996.

Van Oostrom, F. Heeroma als mediëvist. *Driemaandelijkse bladen*, 50 (1–4):105–111, 1998. [Themanummer 'Balans in duplo: over persoon en werk van K.H. Heeroma – Muus Jacobse', R. Ebeling et al. (ed.)].

Van Oostrom, F. *Stemmen op schrift. Geschiedenis van de Nederlandse literatuur van het begin tot 1300*. Prometheus, 2006.

Van Sterkenburg, P. De lokalisering van de Middelnederlandse Arturroman *Lantsloot vander Haghedochte*: een voorstudie. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 98(1):1–29, 1982.

Van Sterkenburg, P. Een identiek auteurschap van *Lantsloot* en *Moriaen*? (samenvatting). *Nieuw Letterkundig Magazijn*, 5(1):21–22, 1987.

Van Uytven, R. Het Antwerpen van Jan van Boendales. In Van Anrooij, W. (ed.), Al t'Antwerpen in die stad. *Jan van Boendale en de literaire cultuur van zijn tijd*, Nederlandse literatuur en cultuur in de middeleeuwen 24, pp. 17–29. Prometheus, 2002.

Van Vloten, J. (ed.). *Jacob van Maerlant:* Merlijn. *Naar het enig bekende Steinforter handschrift uitgegeven*. E.J. Brill, 1880.

Vekeman, H. De verhaaltechniek in *Karel ende Elegast*. *Spiegel der Letteren*, 13(1):1–9, 1970–1971.

Verbruggen, J. *De Slag der Guldensporen. Bijdrage tot de geschiedenis van Vlaanderens vrijheidsoorlog. 1297–1305*. Standaard, 1952.

Vermassen, V. [Recensie van Houthuys 2009]. *Queeste. Tijdschrift over middeleeuwse letterkunde in de Nederlanden*, 18(1):84–87, 2011.

Warnar, G. Mysticus aan de lessenaar. Jan van Ruusbroec en de middeleeuwse literatuurtheorie. *Literatuur*, 14(3):140–145, 1997.

Warnar, G. Velthem in de Vierde Partie. De verwerking van de geleerdentraditie in het Middelnederlands. In Besamusca, B.; Sleiderink, R. & Warnar, G. (ed.), *De boeken van Velthem. Auteur, oeuvre en overlevering*, Middeleeuwse studies en bronnen 119, pp. 119–142. Uitgeverij Verloren, 2009.

Waterschoot, W. (ed.). *De Guldensporenslag. Een fragment uit de Voortzetting van de* Spiegel historiael. Klassieken Nederlandse letterkunde. Nijhoff, 1979.

Westgeest, H. De Leidse lapidariumfragmenten: delen van Maerlants *cortten lapydarys*? *Queeste. Tijdschrift over middeleeuwse letterkunde in de Nederlanden*, 8(1):1–26, 2001.

Willaert, F. [Recensie van Besamusca 1993]. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 111(3):258–261, 1995.

Willaert, F. *De ruimte van het boek. Literaire regio's in de Lage Landen tijdens de Middeleeuwen*. Bert van Selm-lezing 19. Stichting Nederlandistiek, 2010.

Willems, J. Berichten wegens oude Nederduitsche dichters. *Belgisch museum voor de Nederduitsche tael- en letterkunde en de geschiedenis des vaderlands*, 1:340–380, 1837.

Willems, J. (ed.). *De Brabantsche yeesten, of Rymkronyk van Braband, door Jan de Klerk, van Antwerpen*. Hayez, 1839.

Willems, J. Hugo van Tabarie door Hein van Aken. *Belgisch Museum voor de Nederduitsche tael- en letterkunde en de geschiedenis des vaderlands*, 6:94–104, 1842.

Windelberg, M. & Miller, D. How (not) to define the epic formula. *Olifant*, 8(1):29–50, 1980.

Wollheim, R. *On Art and the Mind. Essays and Lectures*. Allen Lane, 1972.

Wynants, M. De *Brabantse Yeesten* van Jan van Boendale en zijn navolger. *De Horen*, 21(1):39–41, 1994.

Zemel, R. De *Roman van Walewein* en het heldenlied. In *Maar er is meer. Avontuurlijk lezen in de epiek van de Lage Landen: studies voor Jozef D. Janssens*, Davidsfonds/Literair, pp. 27–44. Davidsfonds, 2005.

Zemel, R. Walewein en Ysabele in Endi. *Nederlandse letterkunde*, 15(1): 1–28, 2010.

Zemel, R. & Besamusca, B. *Lanceloet en het hert met de witte voet*. In Besamusca, B. (ed.), Jeesten van rouwen ende van feesten. *Een bloemlezing uit de* Lancelotcompilatie, Middelnederlandse tekstedities 6, pp. 181–210. Uitgeverij Verloren, 1999.

Ziolkowski, J. Lost and not yet found: Heloise, Abelard and the *Epistolae duorum amantium*. *Journal of Medieval Latin*, 14:171–202, 2004.

Zipf, G. *Human behavior and the Principle of Least Effort*. Addison-Wesley, 1949.