

MIKE KESTEMONT

De meesters van de *Spiegel*

Auteursonderscheiding op basis van het frequente rijmwoord in het aandeel van Utenbroeke en Maerlant in de *Spiegel historiael*

Abstract – Because of the poor quality of original matter transmitted to our times many Middle Dutch studies lack information on their authors. Hence, there is the need for a methodology to attribute and verify the authorship of Middle Dutch texts. Scholars are increasingly concerned with experiments in authorship attribution, based on insights from computational philology. In this paradigm a lot of attention is being paid to high-frequency words. This paper will research whether it is possible to stylistically verify medieval authorship. By means of *Machine Learning* we shall assess the verification of authorship in the Middle Dutch adaptation of the *Speculum historiale* on the basis of rhyme words.

1 Kop noch staart

In de studie van Middelnederlandse letterkunde kampen onderzoekers niet zelden met een gebrek aan feitelijke gegevens over teksten.¹ Waar en wanneer een tekst geschreven werd, door wie en voor wie, zijn vragen waarop de medioneerlandistiek vaak het antwoord moet schuldig blijven (Van Oostrom 2006: 233). Deze toestand komt in hoofdzaak voort uit de schamele overlevering: niet heel veel Middelnederlandse teksten zijn bewaard gebleven en als zij al gespaard werden, moet men zich meestal tevreden stellen met fragmentarische tekstgetuigen, die bovendien vaak kopieën (in het ‘kwadraat’) blijken van een veel latere datum (Geirnaert 2000). Pro- en epilogon met informatie over de ontstaanscontext van het literaire werk zijn dun gezaaid (Sonnemans 1995). Middelnederlandse literatuur is immers vaak ‘kop- en staartloos’ overgeleverd, aangezien latere afschrijvers niet altijd even geïnteresseerd leken in de herkomst van een tekst en er veeleer op gericht waren de tekst voor het functioneren in een nieuwe omgeving geschikt te maken. Auteurschap is wellicht datgene waarover men nog het slechtst geïnformeerd is: slechts een klein aantal auteursnamen is overgeleverd en nog een kleiner aantal auteursnamen valt met concrete, al dan niet overgeleverde werken te verbinden (Van Driel 2007: 163ff).

Het hoeft daarom niet te verbazen hoe vaak het auteurschap van Middelnederlandse werken onderwerp is geweest van wetenschappelijk onderzoek, maar ook van fascinatie en speculatie. Recent lijkt deze aandacht voor de auteur zelfs toe te nemen: volgens vele onderzoekers staat de gebrekkige kennis omtrent auteurs een beter literair-historisch inzicht in Middelnederlandse literatuur in de weg. Vaak is

1 Ik dank mijn promotoren Frank Willaert en Walter Daelemans en mijn goede collega's Karina van Dalen-Oskam, Herman Brinkman en Elisabeth de Bruijn, die dit stuk, soms meermaals, enthousiast van commentaar hebben voorzien. Verder dank ik mijn vader, René Kestemont, die buiten zijn drukke werkuren nog steeds de tijd vindt om als finale revisor op te treden voor mijn stukken. Vanzelfsprekend ligt de verantwoordelijkheid voor eventuele tekortkomingen geheel bij mijzelf.

daarbij niet zozeer de auteursnaam de inzet van het onderzoek. Immers, wat doet het ertoe dat ene Willem de *Reynaert* schreef en ene Diederik *Floris ende Blancefloer*? Het wordt pas interessant als we deze auteurs respectievelijk mogen gelijkstellen met Willem van Boudelo en Diederik van Hassenede, beiden in dezelfde periode werkzaam als klerk voor het grafelijk hof van Vlaanderen.² Dat zou immers betekenen dat beide dichters elkaar en elkaars werk gekend moeten hebben, wat intrigerende vragen oproept over de onderlinge relatie tussen deze werken. De auteur zelf is dus niet het doel van het onderzoek maar slechts een middel om inzicht te krijgen in literaire netwerken, schoolvorming, mecenaat, ...

Een methodologie voor het herkennen van Middelnederlandse auteurs is daarom hoognodig (Van Driel 2007: 166). Onderzoekers hebben vaak sterke intuïties over het auteurschap van teksten maar kunnen die vermoedens zelden sluitend staven. Vooral het literair taalgebruik of de *stijl* van auteurs is meer dan eens aangeduid als een betrouwbare indicator van auteurschap (Van Dalen-Oskam 2007). Onderzoekers hebben inhoudelijk onderzoek naar de verwantschap van teksten dan ook vaak aangevuld met formeel onderzoek, waarbij de toeschrijving van een tekst aan een auteur kracht werd bijgezet op stilistische gronden. Steeds vaker worden ook computationele middelen ingeschakeld om grotere corpora te doorzoeken en hypothesen een bredere, ook statistische basis te verlenen.

In het hier gepresenteerde onderzoek zoek ik expliciet aansluiting bij het opkomende onderzoek naar auteursherkenning in de *digital humanities* oftewel de 'stylometrie'. In dit kwantitatieve paradigma zijn interessante inzichten verworven omtrent de mogelijkheden tot het beschrijven van het stijleigen of 'styloom' van auteurs. Steeds vaker experimenteren ook medioneerlandici met de toepassing van stylometrische methodes op Middelnederlandse teksten. In dit artikel wil ik een bijdrage leveren aan dit opkomende paradigma en nagaan of auteurschap op basis van stilistische gronden kan worden vastgesteld. Vreemd genoeg is er wel veel onderzoek geweest naar het auteurschap van anoniem overgeleverde teksten maar is amper onderzocht of het auteurschap van teksten van een gekend auteur stilistisch geverifieerd *kan* worden (Van Dalen-Oskam 2007: 37). Als dat zou blijken, zou dit het attributie-onderzoek naar teksten van betwiste signatuur aanzienlijk meer slagkracht verlenen. De casus die ik hier zou willen behandelen, is het rijmwoord in de Tweede en Derde Partij van de *Spiegel historiael*, respectievelijk gedicht door Filip Utenbroeke en Jacob van Maerlant in het laatste kwart van de dertiende eeuw. Casus en methodologie zijn hieronder zo eenvoudig mogelijk gehouden. De bedoeling is voor een breed, ook minder gespecialiseerd publiek aan te tonen dat men op kwantitatieve wijze de stilistische vingerafdruk van een Middelnederlands auteur kan zichtbaar maken.

2 Met handen en voeten

Binnen de *digital humanities* verschijnt sinds enkele jaren een ware stortvloed aan publicaties over de stilistische bestudering van auteurschap (Holmes 1998;

² Voor de eventuele identificatie van Willem, zie Van Daele 2005, waar ook gewezen wordt op de nabijheid van Diederik (vgl. Van Oostrom 2006: 225).

Stamatatos 2009). Ik maak hier ruimte om enkele van de belangrijkste methodes en verworvenheden uit dit studiegebied te introduceren.³ In verschillende deelgebieden van de computationele filologie (*information retrieval, stylometry, Machine Learning, ...*) wordt de studie van auteurschap vaak 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 e.a. 2000: 472). Tekstclassificatie is niet de enige methode in dit paradigma, maar wel een dominante. Tekstclassificatie kent momenteel veel toepassingen, ook buiten de stijlstudie (Sebastiani 2002). Het bekendst is wellicht *spam filtering* waarbij een computer aan een emailbericht 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.⁴ De notie van het *trainen* is ontleend aan *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 gelijkenissen tussen de tweede en de eerste lucifer. Voorbeeld- of *training*-materiaal 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*).

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 *training*-instanties) is de sleutel tot succes. De bedoeling is dat om het even welke email kan worden voorgesteld aan de hand van een beperkt aantal kenmerken (*features*) die relevant kunnen zijn voor de classificatie. Enkel op basis van een consequente voorstelling van relevante features zal een classifier goed kunnen trainen. In tekstclassificatie – bijvoorbeeld het automatisch onderbrengen van nieuwsitems in de categorie *sport* of *economie* – wordt vaak gewerkt met een soort tabel waarin voor elk voorbeeld wordt aangegeven of een bepaald woord in de te labelen tekst voorkomt. Een voorbeeld van een dergelijke voorstelling wordt hieronder geboden (tabel 1). Op basis van een software-gestuurd leer-

3 Een nog grotendeels representatieve status quaestionis voor de neerlandistiek is te vinden in Hinskens & Van Dalen-Oskam 2007 en Van Dalen-Oskam 2007. De bondige en noodzakelijkerwijs onvolledige inleiding die hier tot het fenomeen tekstclassificatie wordt geboden, is specifiek gericht op niet-geïnitieerden. Om deze lezers niet af te schrikken met technische termen en een veelheid aan publicaties of problematiseringen, beperk ik mij tot de hoofdzaken en verwijs naar een klein aantal betrouwbare overzichtswerken in plaats van naar detailstudies.

4 Een introductie tot de algemene gegevens in deze alinea is te vinden in bijvoorbeeld Alpaydin 2004.

TABEL 1

	'loterij'	'X?x!'	'tijdschrift'	Klasse
Instantie 1	Ja	Nee	Nee	spam
Instantie 2	Nee	Nee	Ja	geen spam
Instantie 3	Nee	Nee	Nee	geen spam
Instantie 4	Ja	Ja	Nee	spam

Deze tabel biedt een fictief voorbeeld van hoe email-berichten kunnen worden voorgesteld in een classificatiesysteem voor spam filtering. In deze tabel bevat iedere rij een voorbeeld-email (instanties 1 tot 4), voorgesteld aan de hand van de aan- of afwezigheid van drie specifieke woorden (de kolommen 2 tot 4). Van iedere instantie wordt in de kolom uiterst rechts aangegeven tot welke klasse het bericht behoort (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 ongezien email-bericht voorkomt, zal de classifier het nieuwe bericht waarschijnlijk als spam beschouwen.

proces kan de classifier dan kennis opbouwen over hoe het best een nieuwe, ongeziene instantie wordt gelabeld. Een voorbeeld van een dergelijke leermethode wordt hieronder in detail besproken.

Ook in het onderzoek naar auteurs wordt vaak tekstclassificatie toegepast (Holmes 1998; Luyckx & Daelemans 2008; Stamatatos 2009). Men gaat bijvoorbeeld na of op basis van trainingmateriaal in de vorm van het *oeuvre* van een gekend 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 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* (Alpaydin 2004: 327ff; 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 'onbekende' testexemplaar zou toeschrijven. Dit proces 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 hoe betrouwbaar de classifier zou zijn als hij een label moet toekennen aan een werk van betwiste signatuur.

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 – men spreekt over het ‘styloom’ van een auteur (Van Halteren e.a. 2005). 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 (Luyckx & Daelemans 2008; Van Halteren e.a. 2005). In het onderzoek is gebleken dat één categorie kenmerken het bijzonder goed doet in het onderscheiden van auteurs: de kleine groep meest voorkomende elementen in een taal (Holmes 1998; Stamatatos e.a. 2000; Luyckx & Daelemans 2008; Stamatatos 2009). Het gaat om functiewoorden als lidwoorden, voornaamwoorden, voegwoorden, voorzetsels, ... Studies tonen aan dat verschillen tussen auteurs goed gemeten kunnen worden aan de hand van de hogere frequentieregioenen in een taal (Stamatatos 2009: 5-6). Auteurs verschillen dus niet zozeer in *welke* hoogfrequente woorden zij gebruiken – iedereen gebruikt lidwoorden – maar wel in welke *mate* zij die gebruiken – sommige auteurs gebruiken bepaalde lidwoorden meer dan andere auteurs (Burrows 2002; 2007). Belangrijk is dat men dus niet zozeer op zoek gaat naar enkele individueel heel kenmerkende teksteigenschappen (de aan- of afwezigheid van een kleine groep zeldzame woorden) maar naar een *combinatie* van verschillende redelijk goede kenmerken (de frequentie van een veel omvangrijkere groep alledaagse woorden).

Hoogfrequente functiewoorden zijn daarom methodologisch interessant (Stamatatos 2009: 5). Ten eerste komen functiewoorden frequent voor in *alle* teksten en bieden zij door hun goede spreiding een statistisch houvast. Laagfrequente woorden zoals hapax legomena of woorden die slechts schijnen voor te komen bij één auteur – ‘auteurshapaxen’ – hebben dat voordeel niet. Als een auteurshapax in het werk van één auteur al laagfrequent is, is de kans groot dat het woord in andere teksten van die auteur amper voorkomt. Een ander voordeel is dat hoogfrequente elementen in een taal bij alle auteurs voorkomen: als men twee auteurs vergelijkt, is de kans groot – hoe klein het onderzochte corpus ook is – dat zij beiden lidwoorden gebruiken. Functiewoorden zijn ook in dat opzicht aantrekkelijk, als een vergelijkingsbasis waarop auteurs effectief *kunnen* vergeleken worden. Het laatste – en misschien voornaamste – voordeel is dat functiewoorden grotendeels inhoudsonafhankelijk zijn: het onderwerp van een tekst beïnvloedt in wezen niet de frequentie van, bijvoorbeeld, lidwoorden. Dit aspect is zeker interessant voor tekstclassificatie, aangezien functiewoorden dan ook kunnen gebruikt worden voor auteursherkenning over de grenzen van genres heen. In het algemeen verklaart men de meerwaarde van functiewoorden door het feit dat ze niet bewust gecontroleerd kunnen worden door een auteur. Wie een auteur wil imiteren, zal inhoudsgerelateerde woordenschat of laagfrequente woordkeuzes makkelijk kunnen nabootsen. Moeilijker is het om de frequentie van bijvoorbeeld een lidwoord te imiteren.

De meerwaarde van functiewoorden is te illustreren aan de hand van een parallel in de schilderkunst, meer bepaald in de theorie van Giovanni Morelli (Wollheim 1972: 177ff). Ook veel schilderijen zijn anoniem overgeleverd, wat heeft geleid tot bloeiend attributie-onderzoek in de kunstgeschiedenis. 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. Zowat 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 over 'genrengrenzen' heen mogelijk maakt.

3 De medioneerlandistiek

Hieronder beschrijf ik een toepassing van de inzichten uit de niet-traditionele auteursstudies op Middelnederlandse letterkunde. Hoewel de aandacht voor het auteurschap van Middelnederlandse teksten een lange traditie kent, heeft het onderzoek ernaar slechts recent een kwantitatieve en computerondersteunde dimensie gekregen. Die ontwikkeling hangt samen met de groeiende beschikbaarheid van teksten in digitale vorm (Hinskens & Van Dalen-Oskam 2007: 15), bijvoorbeeld op de *Cd-rom Middelnederlands* (1998). Voor het verschijnen van de *Cd-rom* was computerondersteund onderzoek naar het auteurschap van teksten eerder zeldzaam (Hadewijch: Murk-Jansen 1988; *Ferguut*: Kuiper 1989; *Lancelotcompilatie*: Besamusca 1991). Daarna hebben verschillende onderzoekers de *Cd-rom* als instrument gehanteerd in de studie van auteursattributie op basis van stijl (Westgeest 2001; Reynaert 2002; Hogenbirk 2009).⁵ Niettemin is er ook kwantitatief auteursonderzoek geweest dat minder van de *Cd-rom* afhankelijk was (Burgers 1999; Croenen 2005). Opvallend is dat verschillende van deze onderzoekers vooral hebben gefocust op laagfrequente fenomenen, zoals *hapax legomena* of zeldzame syntactische constructies die eigen lijken aan de schrijfstijl van een bepaald auteur. Veel onderzoekers zijn ervan uitgegaan dat dergelijke stijlkenmerken resoluut aan een en dezelfde auteur mogen verbonden worden en gaan voorbij aan de mogelijkheid van schoolvorming of wederzijdse literaire beïnvloeding. Net omdat deze woorden zo opvallend zijn, kunnen zij door andere dichters al dan niet bewust makkelijk uit het werk van andere dichters zijn overgenomen (Van Driel 2007: 164).

Een baanbrekend onderzoek is dat naar het dubbel auteurschap van de *Walewein* door Karina Van Dalen-Oskam en Joris van Zundert (2007). Zij gebruikten een niet-traditionele onderzoeksmethode, namelijk de 'Delta-metrick' van Burrows die steunt op de inzichten uit de computationele filologie, in het bijzon-

⁵ Katty de Bundel promoveerde recent aan de Katholieke Universiteit Leuven op een proefschrift getiteld '*Van woerde tot woerde ofte van synne te sinne*'. *Petrus Naghel en het translatorium van de kartuis te Herne (ca. 1350-1400)*. In dit op het moment van schrijven nog ongepubliceerd onderzoek plaatst zij verschillende teksten op het conto plaats van de Bijbelvertaler van 1360 via een methodologie die doet denken aan die van Reynaert 2002.

der het belang van hoogfrequente woorden. Zij stelden vast dat de Delta-metrick ambigue resultaten opleverde: beperkte men zich tot de vijftig meest frequente woorden in de *Walewein*, dan was hun techniek succesvol in het detecteren van de *kopiïsten*wissel in het gebruikte handschrift maar werd de *auteurs*wissel aan het zicht onttrokken. Om de auteurswissel zichtbaar te maken, moesten zij uitwijken naar de frequenties van de honderd tot hondervijftig meest frequente woorden. In een recente paper is de problematiek van de *Walewein* hernomen (Kestemont & Van Dalen-Oskam 2009). De aanleiding werd gevormd door onderzoek naar een corpus van vijftien kopieën van enkele parallelle passages uit Jacob van Maerlants *Rijmbijbel*. De vijftien kopiïsten in dit corpus bleken met een relatief hoge slaagkans van elkaar te onderscheiden, onder meer op basis van de meest frequente functiewoorden, en hetzelfde gold voor de twee kopiïsten van het *Waleweinhandschrift*.⁶ De vergelijking leerde niettemin dat de meest frequente functiewoorden waardeloos waren in het tegen elkaar afzetten van auteurs. Dat was teleurstellend: die aspecten die typerend lijken voor hedendaagse auteurs, zouden volgens deze experimenten weinig informatie bevatten over het auteurschap van Middelnederlandse teksten en leken te zeer beïnvloed door de afschrijvers van een tekst. Toch bleek de situatie niet hopeloos. In onze experimenten op de *Walewein* konden wij aantonen dat de oorspronkelijke auteurs van de roman aan de oppervlakte van de tekst stilistisch weliswaar concurreerden met de kopiïsten, maar dat hun styloom intact was gebleven op een dieper niveau. Door abstractie te maken van oppervlakkige variatie – door het lemmatiseren van teksten, waarover hieronder meer – bleek het mogelijk om de kopiïsten buiten spel te zetten.

In deze bijdrage wil ik daarom aansluiting zoeken bij de gestaag groeiende interesse in de medioneerlandistiek voor niet-traditionele auteursattributie. In het bijzonder wil ik aandacht besteden aan een belangrijke handicap van veel van het voorgaande onderzoek: tot op heden is namelijk nooit geverifieerd of een Middelnederlands auteur wel herkend *kan* worden aan zijn stijl (vgl. Van Driel 2007: 166). Besamusca gaf bijvoorbeeld toe dat hij werkte vanuit ‘de overtuiging – meer is het niet’ dat een Middelnederlands dichter kan herkend worden aan zijn stijl (Besamusca 1991: 165-166). In het recente proefschrift van Joost van Driel wordt eveneens een belangrijke rol aan de auteur toegedicht als verklarende factor voor de grote stilistische diversiteit van Middelnederlandse epische poëzie (Van Driel 2007: 159ff). Die claim is niettemin opvallend omdat Van Driel in zijn studie geen enkel oeuvre heeft onderzocht. Ook hij gaat zo impliciet uit van de stilistische homogeniteit van het oeuvre van een dichter (Kestemont 2007: 179ff). Maar heeft een middeleeuwse auteur wel een unieke en constante stijl waarmee hij of zij zich onderscheidde van alle andere auteurs? De stilistische homogeniteit van een oeuvre wordt zelden in vraag gesteld, hoewel verscheidene onderzoekers in het verleden gegevens hebben aangereikt die de aanleiding voor een dergelijke problematisering hadden kunnen vormen. Velthems stijl, bijvoorbeeld, kan wel eens sterker dan vermoed samenhangen met het genre waarin hij dichtte (Hogenbirk 2009): is de stijl van een dichter die in verschillende genres dichtte dan wel zo homogeen als wij aannemen? Van Maerlant is beweerd dat die op de stijl van zijn

6 De methodologische kwesties van dit onderzoek worden hier omwille van de techniciteit en plaatsgebrek niet hernomen.

navolgers een grote invloed heeft uitgeoefend (Van Loey 1946: 42-43): is het dan wel mogelijk een stilistisch onderscheid te maken tussen Maerlant en zijn epigonen? De stijl van Middelnederlandse epische dichters is in hoge mate stereotiep (Kestemont 2007: 179ff): is het dan überhaupt wel mogelijk dat elke dichter een unieke stijl heeft? In verschillende opzichten kan een advocaat van de duivel dus betwijfelen of het mogelijk is om Middelnederlandse dichters te onderscheiden. Ik wil dit artikel daarom wijden aan een *proof of concept*: aan de hand van een overzichtelijke casus hoop ik aan te tonen dat een Middelnederlands dichter wel degelijk een stijleigen heeft. Mijn bedoelingen zijn vanzelfsprekend niet praktisch maar eerder theoretisch, aangezien gewerkt zal worden met teksten waarvan het auteurschap al vast staat.

Een groot probleem is echter de overlevering. Middeleeuwse teksten zijn ons meestal slechts uit latere afschriften (van eerdere afschriften) bekend. Het is algemeen geweten dat middeleeuwse kopiïsten een minder exacte opvatting van kopiëren hadden dan wij: door instabiele spellingsconventies en de afwezigheid van een standaardtaal werden de spelling en het dialect van teksten bij iedere kopie grondig aangepast. Deze vormen van tekstcorruptie blijven soms oppervlakkig en onschuldig (bv. dialectale of allografische variatie) maar in het geval van veel kopiëren blijkt de tekst ook op grotere schaal aangetast te worden. Van veel laat overgeleverde teksten kan daarom betwijfeld worden of zij nog de stijl weerspiegelen van de oorspronkelijke auteur, nadat zoveel opeenvolgende kopiïsten in de tekst ‘een hand’ hebben gehad. Deze schijnbaar onschuldige ingrepen hebben zwaarwichtige gevolgen voor het auteursonderzoek. Zoals recent onderzoek aantoonde, blijkt dat middeleeuwse kopiïsten een invloed hebben gehad op de hogere frequentiestrategie van een tekst (Van Dalen-Oskam & Van Zundert 2007; Kestemont & Van Dalen-Oskam 2009). Klaarblijkelijk genoten zij de vrijheid om net op de functiewoorden van een tekst een eigen stempel te drukken. Een klein bijwoordje bijvoorbeeld kan inderdaad makkelijk in een tekst worden toegevoegd of verwijderd. De vraag is dan welke woorden in de Middelnederlandse tekst nog in aanmerking komen voor auteursattributie, als zowel de hoog- als laagfrequente woorden beter uitgesloten worden.

Eén categorie woorden is in het verleden aangeduid als betrouwbaar voor het herkennen van Middelnederlandse auteurs: het rijmwoord (Besamusca 1991: 165-166; Westgeest 2001: 15-16; Van Driel 2007: 164-167). Zeker wat de Middelnederlandse epiek betreft – die voor het leeuwendeel paarsgewijs berijmd is – zou het rijmwoord een merkwaardig taai en stabiel element zijn, erg robuust ten aanzien van het overleveringsproces (Van den Berg 1983: 200ff; Van den Berg 1985; Van den Berg 1986: 305-306). Bekend is de visie dat wie kijkt naar de rijmen van een epische tekst, de oorspronkelijke dichter recht in het aangezicht kijkt. De eindeloze ketting van coupletten ligt structureel aan de basis van de berijmde Middelnederlandse tekst (Van Driel 2007: 37). Een kopiïst kon dan wel makkelijk aan de woorden morrelen binnenin het vers maar wat het rijmwoord betreft, zat een afschrijver redelijk vast aan de grondtekst. Immers, als hij een rijm wou aanpassen, zou hij ook op omslachtige wijze een deel van de grondtekst moeten herwerken. De spelling van rijmwoorden mocht dan nog soms aangepast worden, het onderliggende ‘woord’ werd intact gelaten (Van den Berg 1986: 305-306). Rijmwoorden lijken op die manier eilandjes van stabiliteit in de overlevering van Middelneder-

delandse teksten. Als er van de oorspronkelijke auteurstekst nog iets overblijft in de kopieën moet dat in de eerste plaats in het rijm zijn terug te vinden.

4 Twee meesters

Hieronder wil ik nagaan of het met een eenvoudige classifer mogelijk is twee Middelnederlandse auteurs te onderscheiden op basis van hun rijmvocabulaire. De casus waarmee gewerkt zal worden, is het aandeel van Filip Utenbroeke en Jacob van Maerlant in de *Spiegel historiael* (respectievelijk de Tweede en Derde Partie). De *Spiegel historiael* is een kolossale bewerking in Middelnederlandse verzen van de Latijnse wereldgeschiedenis *Speculum historiale* van Vincentius van Beauvais. Het initiatief voor dit project werd ca. 1280 genomen door Jacob van Maerlant, die vier grote tekstblokken of *Partieën* voorzag (Biemans 1997: 19ff; Van Oostrom 1996: 307ff). Deze *Partieën* werden onderverdeeld in *boeken*, die op hun beurt weer bestonden uit kleine hoofdstukjes, *kapittels* genaamd. Maerlant was wel de architect 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 wel de Derde Partie schreef (de geschiedenis tot net voor Karel de Grote), wist hij dus dat Utenbroeke de tweede voor zijn rekening zou nemen, hoewel dat in de tekst niet 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 priester Lodewijk van Velthem die op eigen initiatief ook een Vijfde Partie aan het geheel toevoegde (Besamusa, Sleiderink & Warnar 2009: 10ff). Het is overigens uit Velthems werk dat we Utenbroekes naam kennen, want die wordt bij zijn voorgangers verzwegen (Biemans 1997: 23).

De relatie tussen Maerlant en Utenbroeke is intrigerend: Maerlant en Utenbroeke moeten elkaar van nabij gekend hebben.⁷ 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: 23-34). Een dergelijk samenwerkingsverband kennen we uit de schilderkunst en ook van Vincentius weten we dat die hulp kreeg van een gelijkaardig type loop- en leerjongens. 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.

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

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 (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 casus: ze vertrokken vanuit dezelfde brontekst en hadden duidelijk de bedoeling om in samenwerking tot een coherent geheel te komen. Hier worden de Tweede en Derde Partie 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 beide als een organische eenheid voor tekstvergelijking kunnen dienen. Bovendien is beweerd dat van Maerlants rijmgedrag een dwingende invloed is uitgegaan op zijn navolgers. 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, š.n.b., Cod. 13.708, dat zijn toenaam – Tweede Partie-handschrift – aan de aanwezigheid van deze tekst dankt.⁸ Dit handschrift kwam aan het einde van de veertiende eeuw tot stand in de kartuis te Herne. Hoewel het handschrift niet volledig is, biedt 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.⁹ Het handschrift bevat 366 van de ca. 460 oorspronkelijke kapittels verspreid over zeven boeken. Maerlants Derde Partie is volledig overgeleverd in handschrift Den Haag, KA XX, een prachtig geillumineerd handschrift dat Maerlants aandeel in de *Spiegel historicael* volledig dekt.¹⁰ 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* (1998) en het zijn dan ook deze bestanden die hieronder worden gebruikt.

De rijmwoorden van het onderzochte corpus zijn integraal gelemmatiseerd op basis van een recente taaltechnologie die een groot deel van het lemmatiserings-

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

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

10 Handschrift 1 in Biemans 1997 (via de index, maar ook p. 197 e.v.).

werk automatiseert.¹¹ Lemmatiseren betekent dat aan een woord-in-context een uniform label of lemma wordt toegekend. Veelal gaat het om een genormaliseerde vorm, vergelijkbaar met de hoofdvorm waaronder het woord in een woordenboek is terug te vinden. De bedoeling is abstractie te maken van de spelling en flectie (verbuiging en vervoeging) van woorden zodat op zinnige wijze kan generaliseerd worden over het optreden van groepen woorden wier onderlinge verschillen voor een bepaalde taak irrelevant zijn. Voor auteursherkenning heeft deze normalisatie het voordeel dat men abstractie kan maken van spellingsverschillen die niet op de auteur teruggaan maar door een kopiïst zijn geïntroduceerd. Wat betreft flectie biedt deze aanpak bijvoorbeeld ook de mogelijkheid om te generaliseren over het optreden van de enkelvoudige dan wel meervoudige vorm bij een substantief. De gebruikte *lemmatizer* is getraind op het literaire deel van het digitale *Corpus-Gysseling* (geannoteerd en onderhouden door het Instituut voor Nederlandse Lexicologie). In dit corpus zijn de Middelnederlandse woorden verbonden met een modern lemma, of toch een lemma in een moderne spelling in het geval dat er geen moderne pendant meer voorhanden is. Ook in deze bijdrage wordt daarom gewerkt met dergelijke labels.

De principes die zijn gehanteerd bij het lemmatiseren van deze teksten worden kort toegelicht. Homonieme lemmata werden niet 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 stilzwijgend uit het corpus verwijderd. Alle rijmwoorden en slechts de rijmwoorden zijn gelemmatiseerd: slechts die woorden zijn gelemmatiseerd die door de dichter gebruikt worden om tussen twee opeenvolgende verzen een rijm tot stand te brengen. 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. De groep *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 kopiïst – in principe kan Maerlant 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 label *WAAR+ZIJN* gebruikt. In technische termen krijgen 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: – alle eigennamen (persoonsnamen, geografische aanduidingen, boektitels, ...) kregen het lemma *PrName*. Een uitzondering vormen enkele hoogfrequente eigennamen als *GOD* en *CHRISTUS* die zo algemeen zijn dat ze niet met de inhoud van teksten samenhangen;

11 Voor deze alinea volsta ik met een verwijzing naar Kestemont, Daelemans & De Pauw 2010, waarin de technische kant van het lemmatiseren (en het gebruikte trainingsmateriaal) grondig wordt toegelicht.

- alle hoofd- en rangtelwoorden (buiten 1, 2, 3 en afgeleiden) kregen het label *Numb*;¹²
- alle anderstalige woorden, zoals enkele Latijnse substantieven (bv. *deus*) kregen het lemma *Foreign*. Dat geldt niet voor 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: *daer ave* wordt daarom samengenomen in *daer+ave*; de onstane combinatie krijgt het lemma DAARAF. De richtlijnen die daarbij worden gehanteerd zijn dezelfde als in het verrijkte *Corpus-Gysseling*.

Het corpus werd automatisch geannoteerd en vervolgens manueel gecorrigeerd. Dit is slechts gebeurd door één persoon en aangezien het om een aanzienlijke hoeveelheid data gaat, wordt niet gegarandeerd dat het corpus geen fouten meer bevat. Bovendien zijn verschillende fenomenen voor interpretatie vatbaar. In het algemeen wordt er echter een zekere consistentie gegarandeerd door het feit dat de *lemmatizer* teruggaat op het *Corpus-Gysseling* dat in grote mate consistent geannoteerd is. Ook het feit dat de berekeningen hieronder teruggaan op hoogfrequente, meestal makkelijk te interpreteren fenomenen leidt ertoe dat er slechts in een enkel geval aanleiding tot discussie zal zijn. De data worden na de publicatie van deze bijdrage publiekelijk toegankelijk gemaakt via een website.¹³ Deze data zullen voor elk kapittel de lemmata voor de rijmwoorden bevatten maar niet de oorspronkelijke tekst aangezien deze door auteursrecht is beschermd. Wel kan wie ook beschikking heeft over de tekst, de oorspronkelijke tekst en de annotatie makkelijk opnieuw samenvoegen. Hieronder gaan bij wijze van voorbeeld enkele geannoteerde verzen (tabel 2).

TABEL 2

Van Nerons goeden beghinne (Utenbroeke)	Hoe Rome eerst dalen began (Maerlant)
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 bedwon- gen wesen wezen octavien siere dochter man man was nero ende also dan dan	nu helpt moeder ende maghet vrie vrij ic beginne die derde paertie partij vanden spieghale Ystoriale PrName ende hevet in bi ghetale getal viii bouke ende tog- het al clare klaar van cccc ende xx jare jaar wat overginc den roemscen rike rijk dat teersten so mogendelike mogend- lijck al meest aldie werelt dwanc dwin- gen nu ginc dalen an sinen ganc gang

Deze tabel illustreert op welke wijze het corpus is verrijkt en biedt een weergave van de eerste tien verzen (van het eerste kapittel van het eerste boek) van respectievelijk Utenbroekes Tweede Partie en Maerlants Derde Partie. De lemmata van de rijmwoorden zijn op het eind van elke regel in kapitalen weergegeven.

¹² De betekenis van hoofd- rangtelwoorden werd verbleekt omdat zij vaak inhoudsgevoelig zijn (bijvoorbeeld jaartallen) en omdat de numerieke inhoud die zij uitdrukken daarom in de regel niet van belang lijkt voor het stijleigen van een auteur.

¹³ Via <http://www.mike-kestemont.org>.

5 Experimentele setting

Hieronder wordt verslag gedaan van een serie experimenten met betrekking tot het onderscheiden van beide auteurs. De methodologie die hiervoor gebruikt wordt, gaat terug op de *leave one out evaluation*. In dit scenario wordt 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 (het aantal correct toegeschreven samples gedeeld door het totale aantal samples).

Hieronder zal geëxperimenteerd worden met de wijze van *sampling*. Hoe men het materiaal aan de classifier voert, is immers van groot belang. Hier wordt gewerkt met twee 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 kapitels bevatten. Zo kan flexibel gewerkt worden met ofwel een groot aantal samples die een klein aantal kapitels bevatten, ofwel een klein aantal samples die een groot aantal kapitels 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 kapitels. Een kapittel in het corpus telt gemiddeld 86 verzen.

De wijze waarop het rijmgebruik van een dichter in een sample wordt voorgesteld, is als volgt. Vooreerst wordt door alle samples gegaan en bijgehouden welke rijmwoorden erin optreden. Op basis van deze lijst wordt vervolgens een tabel aangelegd, waarin voor elk sample (de rij) wordt aangegeven hoe vaak (de cel) een bepaald rijmwoord (de kolom) in het bewuste sample voorkomt. Er wordt gewerkt met relatieve frequenties: het absolute aantal voorkomens van een bepaald rijmwoord in het sample wordt gedeeld door het totale aantal rijmwoorden in het sample. Hieronder zal enkel gewerkt worden met hoogfrequente rijmwoorden – de verantwoording van deze werkwijze komt hieronder. Ook hier is ruimte voor experimenteren: men kan zich beperken tot bijvoorbeeld de 10 meest frequente rijmwoorden, maar ook de 125 meest frequente.

Cruciaal in de experimenten is de classifier. Omwille van de duidelijkheid wordt hier gewerkt met een erg intuïtieve classificatiemethode. In *memory-based learning* staat het geheugen centraal (Daelemans & Van den Bosch 2005).¹⁴ 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

¹⁴ De software die voor het hier beschreven onderzoek werd gebruikt, is *imbi* of de *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 een inleiding tot het geheugen-gebaseerd leren. Wie meer wil lezen, kan terecht bij het boek *Memory-Based Language Processing* (Daelemans & Van den Bosch 2005).

TABEL 3

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

Deze tabel biedt een illustratie van de werking van het nearest neighbour-algoritme, toegepast op de volledige boeken van de Tweede Partij en Derde Partij. In deze tabel stelt elke rij een volledig boek voor aan de hand van de relatieve frequentie van de vijf meest frequente rijmwoorden in het onderzochte corpus (kolommen 1 tot 5). In de kolom 'Boek' wordt gespecificeerd om welk boek het in de bewuste rij gaat – de afkorting 'P3B6' slaat bijvoorbeeld op het zesde boek ('B6') van de Derde Partij ('P3'). In de kolom 'Nearest neighbour' wordt voor elk boek aangeduid welk ander boek uit het training-materiaal er het meest op gelijkt tijdens leave one out validatie. Het test-sample wordt toegeschreven aan de auteur van de nearest neighbour. De attributie is slechts één keer fout, in het geval van P2B7. De mate van gelijkheid wordt bepaald door de distance (kolom uiterst rechts) die op basis van de relatieve frequenties van de vijf rijmwoorden en de hierboven besproken afstandsmaat kan berekend worden. Alle getallen worden weergegeven tot op vier cijfers na de komma.

maken tussen voorbeelden. 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 een ongezien sample wordt bepaald op basis van de *nearest neighbour* in het geheugen. Belangrijk is hoe wordt bepaald hoe ongelijkend samples zijn: dat gebeurt via de afstand (*distance*) tussen beide (Daelemans & Van den Bosch 2005: 28ff). Een *distance* kan in deze experimen-

ten heel makkelijk worden geïmplementeerd aangezien met numerieke, continue waarden wordt gewerkt, in dit geval de relatieve frequentie van rijmwoorden (een eerder klein decimaal getal tussen 0 en 1). Als twee samples worden vergeleken, wordt het rijtje rijmwoorden afgegaan en berekent de classifier het (geschaalde) verschil voor de frequentie van ieder rijmwoord in beide samples.¹⁵ Vervolgens worden deze verschillen gewoon bij elkaar geteld. De resulterende waarde, de zogenoemde *distance*, zal hoog zijn in gevallen van weinig gelijkende samples maar klein in het geval van sterk gelijkende samples. Het sample in het geheugen op een minimale afstand van het ongeziene sample zal als *nearest neighbour* worden aangewezen. De classifier zal voorspellen dat aan het ongeziene sample dezelfde klasse moet worden toegewezen als die van de *nearest neighbour*.

In tabel 3 wordt weergegeven hoe de training-samples in het geheugen worden voorgesteld aan de hand van de relatieve frequentie van rijmwoorden – hier de vijf meest frequente. 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).¹⁶ 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).

6 Experimenten: de stoplap

Zoals hierboven toegelicht, is het om verscheidene redenen nuttig om in auteursonderscheiding te werken met hoofdfrequente items. Deze idee wordt hier toegepast op het rijmvocabulaire. De 100 meest frequente rijmwoorden in het hele corpus (dus Maerlant en Utenbroeke tezamen) zijn in volgorde:¹⁷

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 'NE

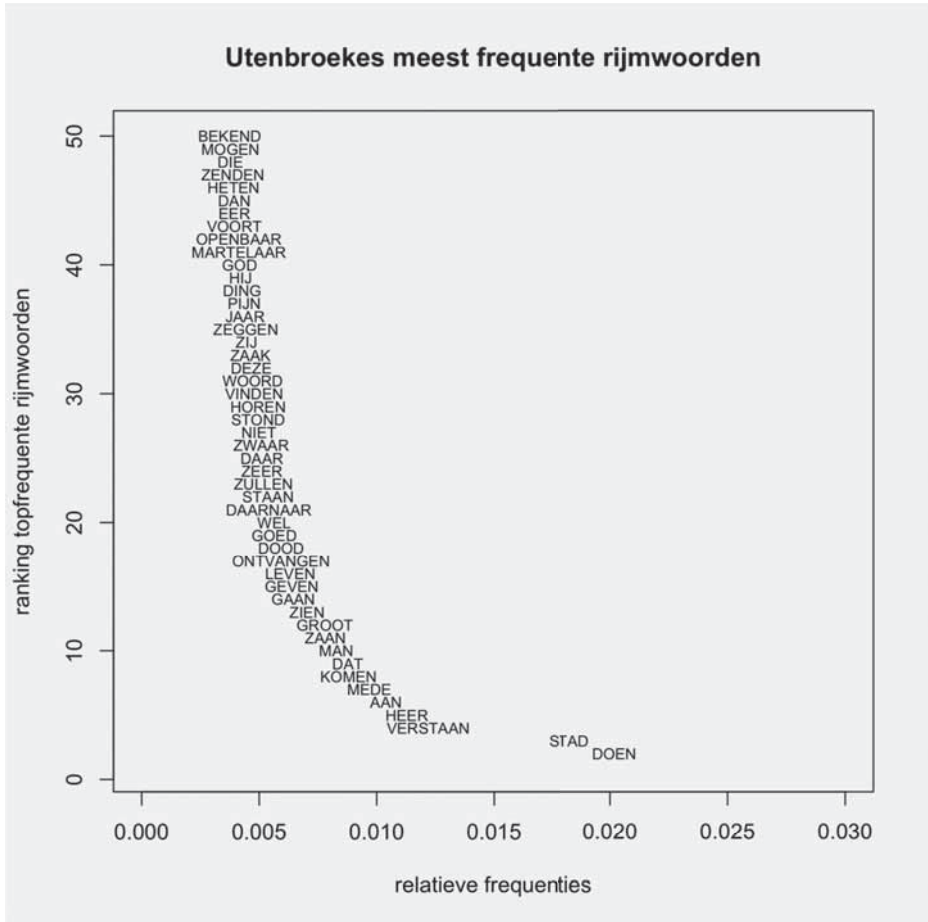
15 In tabel 3 wordt een voorbeeld gegeven van deze berekeningswijze (*scaled Manhattan distance without feature weighting*) met concrete getallen. Let wel: *timbl* zal de verschillen per feature-kolom 'schalen' tussen de hoogste en laagste waarde binnen de feature-kolom. De exacte berekeningswijze wordt wiskundig geformaliseerd in Daelemans & Van den Bosch 2005, 28-29. Wie de *distances* in tabel 3 zelf wil narekenen moet zich natuurlijk rekenschap geven van deze manier van schalen.

16 Merk op dat in deze setting (zie tabel 3) geen sprake hoeft te zijn van 'transiviteit': 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. Zie voor de exacte berekeningswijze in tabel 3 ook de vorige noot.

17 Stamatas (2009: 5) merkt op dat de keuze omtrent hoeveel of welke functiewoorden gebruikt worden vaak nog een arbitraire en taalspecifieke keuze is. Vaak zou een honderdtal woorden al ruimschoots volstaan; vaak beperkt men zich tot de 50 of zelfs 30 meest frequente woorden (Stamatas o.e.a. 2000: 747-475). Hieronder wordt gewerkt met maximum 75 rijmwoorden.

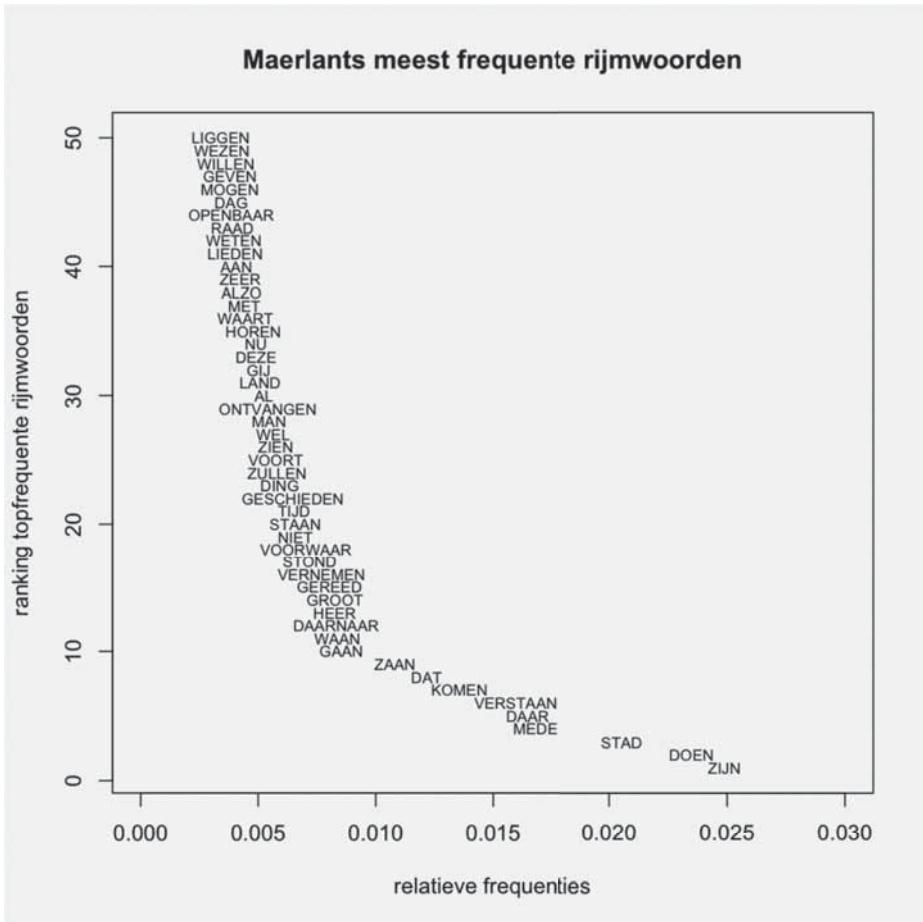
MEN', 48 'PIJN', 49 'AL', 50 'WEZEN', 51 'VOORT', 52 'RIJK', 53 'BLIJVEN', 54 'DAN', 55 'ZEGGEN', 56 'LIGGEN', 57 'KUNNEN', 58 'LIEDEN', 59 'GENE', 60 'ZIJ', 61 'PLEGEN', 62 'GOD', 63 'SCHARE', 64 'BEDE', 65 'DIE', 66 'TEHAND', 67 'HAND', 68 'ZOON', 69 'BRENGEN', 70 'LEZEN', 71 'BIDDEN', 72 'HIJ', 73 'ZENDEN', 74 'LIJF', 75 'VANGEN', 76 'KLAAR', 77 'GEWELD', 78 'LEREN', 79 'GEMEEN', 80 'NOOD', 81 'IK', 82 'BEKEND', 83 'WAART', 84 'RAAD', 85 'WIJF', 86 'ZIN', 87 'MARTELAAR', 88 'MEER', 89 'GEREED, 90 'POORT', 91 'ZITTEN', 92 'VAREN', 93 'NAAM', 94 'MAKEN', 95 'LATEN', 96 'KERK', 97 'DRAGEN', 98 'KWAAD', 99 'KIND', 100 'KEREN'

Wanneer we per auteur de frequentie van de vijftig vaakst voorkomende rijmwoorden uitzetten geeft dat de figuren 1a en 1b.



FIGUREN 1a en 1b De Boven- en onderstaande figuren zijn een voorstelling van de vijftig meest frequente rijmwoorden, respectievelijk in het aandeel van Utenbroeke (1a) en Maerlant (1b). Let wel: het meest frequente rijmwoord is PrName en valt in deze weergave buiten de grafiek; bij Utenbroeke valt ook ZIJN buiten de grafiek. In deze voorstelling zijn de rijmwoorden verticaal (langs de y-as) gerangschikt naargelang hun relatieve frequentie; hun positie op de x-as drukt hun effectieve frequentie uit. Deze rijmwoorden vertonen in beide grafieken een kromming die suggereert dat slechts een klein aantal rijmwoorden heel frequent voorkomt. Een veel groter aantal rijmwoorden komt veel minder vaak voor.

De curves van de rijmwoorden in deze figuren suggereren dat hoogfrequente rijmwoorden zich op dezelfde manier gedragen als functiewoorden: er blijkt enerzijds een *klein* aantal rijmwoorden te zijn dat *heel vaak* voorkomt en anderzijds een *groot* aantal rijmwoorden dat *heel zelden* voorkomt. Bijgevolg lijkt er ook bij rijmwoorden een kleine kruin van hoogfrequente rijmwoorden of ‘functionele rijmwoorden’ te bestaan. Deze zijn zo frequent dat zij niet inhoudsgebonden kunnen zijn want zij treden te vaak en in te diverse contexten op. Deze rijmwoorden dragen natuurlijk wel nog iets van betekenis in zich – het blijven woorden – maar die betekenis is zo algemeen dat het voorkomen van een rijm-



woord als HEER, ZIJN of GROOT nauwelijks kan verraden in welke inhoudelijke context het woord wordt gebruikt. Deze categorie rijmwoorden is trouwens ook echt ‘functioneel’: zoals reeds aangegeven door Van Driel, is het hoogfrequente rijmwoord in Middelnederlandse teksten zelden semantisch geladen (Van Driel 2007: 19ff & 37ff). Rijmwoorden dragen doorgaans weinig bij tot de voortgang van een verhaal, aangezien hun nut meestal beperkt blijft tot het tegemoet ko-

men aan de vormelijke eis van het rijm.¹⁸ In enkele extreme gevallen heeft men het daarom zelfs over ‘stoplappen’. Ook in dit opzicht gaat het hier dus om stoplappen als *functiewoorden*, aangezien hun nut of functie in de eerste plaats vormelijk is. Bovendien is wel beweerd dat verschillende hoogfrequente rijmparen niet eigen zijn aan het taalgebruik van één dichter (Van Driel 2007: 37). Jef Janssens stelde dat veel dichters putten uit éénzelfde pool of één beperkt register van algemeen bekende rijmwoorden: hij had het over een ‘pre-tekstuele potentie’ (Janssens 1988: 97). Zo lijkt het dat het hoogfrequente rijmwoord – alias de stoplap – inderdaad dezelfde voordelen heeft als het functiewoord: het is functioneel, niet gebonden aan het taalgebruik van individuele dichters, frequent en inhoudsonafhankelijk. Koppelen wij daaraan de robuustheid van het rijmwoord in de overlevering, dan wordt deze woordcategorie erg aantrekkelijk voor auteursonderzoek. Rest nog de vraag of het hoogfrequente rijmwoord daadwerkelijk *effectief* is in auteursattributie.¹⁹

Tabel 4 hieronder geeft experimenten weer op boekniveau waarbij stelselmatig het aantal hoogfrequente woorden in het experiment wordt opgevoerd. 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 kapitels 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 in een boek) heeft het algoritme genoeg om beter als kans te presteren (78%). Vanaf 8 rijmwoorden blijkt de classificatie foutloos (100%) en wordt de auteur feilloos herkend.

Hier werken we op boekniveau met een weliswaar klein aantal samples maar wel samples die een grote hap tekst voorstellen en waaruit dus relatief makkelijk algemene tendenzen kunnen worden opgemaakt. Hoewel het met dergelijke grote tekstblokken dus relatief makkelijk blijkt beide 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.²⁰ Tabel 5 beschrijft hieronder een experiment waarbij de sample-grootte stelselmatig artificieel wordt opgevoerd. Het vertrekpunt is 732 samples (366 per auteur) die elk één kapittel bevatten. Vervolgens 366 samples met elk twee kapitels, daarna 244 samples van drie kapitels, enzovoorts. De accuraatheid (aantal correct toegeschreven samples) wordt weergegeven naargelang het aantal topfrequente rijmwoorden dat wordt

18 Zie zeker de sectie ‘Het epische rijm’ in Van Driel 2007: 37ff.

19 Het hier gepresenteerde onderzoek vertoont qua methodologie overigens veel overeenkomsten met Van den Berg 1992, hoewel deze niet zozeer focust op auteursidentiteit.

20 Vergelijkbaar zijn de experimenten met datagrootte in Luyckx & Daelemans 2008. 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). Hier zal in verder onderzoek dieper op worden ingegaan aangezien de problematiek betrekkelijk complex is.

TABEL 4

Aantal rijmwoorden	Rijmwoord toegevoegd	Aantal correcte toeschrijvingen (%)
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%

Deze tabel geeft de resultaten weer van een leave one out-experiment op boekniveau. De werkwijze is exact hetzelfde als in tabel 3, alleen wordt hier gewerkt met een variabel aantal topfrequente 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.

meegenomen in ieder experiment. Figuur 2 biedt een grafische voorstelling van de curves voor de experimenten met respectievelijk 5, 30 en 75 topfrequente rijmwoorden.

Deze experimenten (tabel 5 en figuur 2) tonen dat de classifier aan een sample van slechts enkele kapitels weinig heeft: zeker tot een samplegrootte van tien kapitels (ca. 860 rijmwoorden) is de gemiddelde accuraatheid van de meeste toeschrijvingen eerder laag (< 90%). Toch presteert de classifier van meet af aan boven kans (> 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 samplegrootte. De accuraatheid stijgt vervolgens gevoelig als ook het aantal kapitels dat een sample voorstelt, wordt opgedreven: zoals bijvoorbeeld af te lezen uit figuur 2 vertonen alle curves de tendens om te stijgen naarmate ook de samplegrootte toeneemt. Afgezien van enkele uitschieters, blijkt de

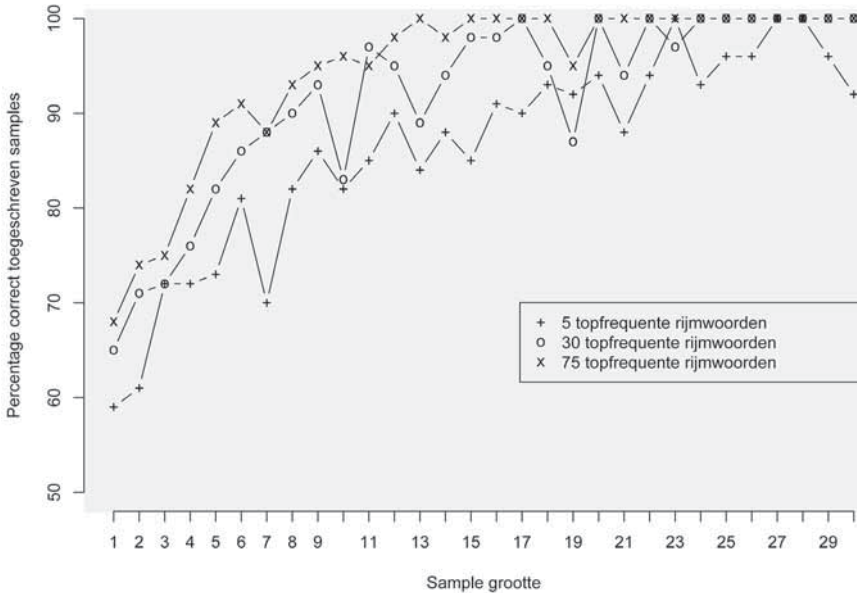
TABEL 5

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

Deze tabel geeft de resultaten weer van het voornaamste experiment in deze bijdrage. Opnieuw gaat het om een 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 (zie de twee kolommen uiterst links) en ten tweede, het effect van het aantal topfrequente rijmwoorden uit deze samples dat in de toeschrijving wordt betrokken (van 5 tot 75 topfrequente 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.

FIGUUR 2



Deze grafiek bevat een grafische voorstelling van de waarden uit tabel 5 voor de experimenten met respectievelijk 5, 30 en 75 topfrequente rijmwoorden. Het percentage correcte toeschrijvingen wordt weergegeven door de sample-grootte (in aantal kapittels) op de x-as uit te zetten tegen het aantal topfrequente rijmwoorden betrokken in het experiment (zie legende).

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. 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 samplegroottes de beste resultaten op.

Voor samples van minder dan twintig kapittels (ca. 1740 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 zou dat methodologisch moeilijk te verantwoorden zijn: uit de rankschikking van de hier gebruikte rijmwoorden (zie boven) 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 onverantwoord, aangezien de inhoud van de Tweede Partie (met veel passages over martelaren) veel vaker aanleiding geeft tot het gebruik van het woord en het geconstateerde verschil niet (enkel) aan de auteur kan verbonden worden. Eventueel kunnen dergelijke woorden manueel uit de lijst verwijderd worden, maar die selectie valt buiten de focus van deze bijdrage.

7 Slotbeschouwing

Deze bijdrage ging over auteursattributie in Middelnederlandse literatuur op basis van stijl en sluit aan bij de recente aandacht in de medioneerlandistiek voor niet-traditioneel, computerondersteund onderzoek hiernaar. Merkwaardig genoeg hebben onderzoekers de effectieve slaagkans van bepaalde toeschrijvings technieken tot op heden amper geverifieerd: de vraag is immers nooit gesteld of het auteurschap van Middelnederlandse teksten eigenlijk wel *kan* geverifieerd worden. Hierboven is casusgewijs een bijdrage geleverd aan het antwoord op deze vraag. Utenbroeke en Maerlant blijken in hun aandeel van de *Spiegel historiae* wel degelijk stilistisch te onderscheiden via een eenvoudig classificatiemodel. Zoals te verwachten was, leende het rijmvocabulaire van teksten zich in deze studie uitstekend tot auteursherkenning. Dit lijkt de claim te ondersteunen dat hoogfrequente rijmwoorden of stoplappen inderdaad gebruikt kunnen worden in niet-traditionele auteursattributie als een soort surrogaat voor moderne functiewoorden, waarvan eerder onderzoek uitwees dat die sterk door de overleving van teksten zijn beïnvloed. Toch moet de kracht van deze attributiemethodes voor middeleeuwse teksten momenteel niet overschat worden: het blijkt duidelijk dat een tekst voldoende groot moet zijn (ca. 1740 verzen) om een goed beeld te verschaffen van het rijmprofiel van de dichter. Ongetwijfeld kan de kwaliteit van de attributie gevoelig worden vergroot door het gebruik van meer geavanceerde leertechnieken, maar enige terughoudendheid en voorzichtigheid lijken vooralsnog geboden.

Het probleem van het ontwijken van de inhoud van teksten bij auteursherkenning blijft een moeilijkheid. Hoewel het gebruik van hoogfrequente, 'functione-

le' rijmwoorden of stoplappen in het middeleeuws auteursonderzoek goed is te verantwoorden, kan men over de mate van contextgebondenheid van een specifiek rijmwoord ongetwijfeld van mening verschillen. Een toepassing van de hier besproken methode op teksten uit verschillende genres is nu daarom hoognodig. Daarbij kan dan getraind worden op het oeuvre van auteurs in een genre (bv. ridderepiek) en getest worden op teksten van diezelfde auteurs in een ander genre (bv. historiografie). In verder onderzoek zullen wij deze kwestie nader behandelen. Het blootleggen van het styloom van middeleeuwse auteurs vormt het uiteindelijke doel van dit soort onderzoek en hoewel dit doel verre van bereikt is, zijn hopelijk toch stappen gezet in de goede richting.

Bibliografie

- Alpaydin 2004 – E. Alpaydin, *Introduction to Machine Learning*. Cambridge & Londen, 2004.
- Besamusca 1991 – B. Besamusca (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*. Assen, 1991 (Middelnederlandse Lancelotromans 5).
- Besamusca, Sleiderink & Warnar 2009 – B. Besamusca, R. Sleiderink & G. Warnar, 'Lodewijk van Velthem. Ter inleiding'. In: B. Besamusca, R. Sleiderink & G. Warnar (red.), *De boeken van Velthem. Auteur, oeuvre en overlevering*. Hilversum, 2009 (Middeleeuwse studies en bronnen 119), 8-30.
- Biemans 1997 – J. Biemans, '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. 2 delen*. Leuven, 1997 (Schrift en schriftdragers in de Nederlanden in de middeleeuwen 2).
- Burgers 1999 – J. Burgers, *De rijmkroniek van Holland en zijn auteurs. Historiografie in Holland door de Anonymus (1280-1282) en de grafelijke klerk Melis Stoke (begin veertiende eeuw)*. Hilversum, 1999 (Hollandse Studiën 35).
- Burrows 2002 – J. Burrows, "Delta". A Measure of Stylistic Difference and a Guide to Likely Authorship'. In: *Literary and Linguistic Computing* 17 (2002), 267-287.
- Burrows 2007 – J. Burrows, 'All the Way Through. Testing for Authorship in Different Frequency Strata'. In: *Literary and Linguistic Computing* 22 (2007), 27-47.
- Cd-rom Middelnederlands* 1998 – *Cd-rom Middelnederlands*. Antwerpen-Den Haag, 1998.
- Croenen 2005 – G. Croenen, 'Het dubbele auteurschap van de *Grimbergsche oorlog*'. In: R. Sleiderink, V. Uyttersprot & B. Besamusca (red.), *Maar er is meer. Avontuurlijk lezen in de epiek van de Lage Landen. Studies voor Jozef D. Janssens*. Leuven, 2005, 131-152.
- Daelemans & Van den Bosch 2005 – W. Daelemans & A. van den Bosch, *Memory-Based Language Processing*. Oxford, 2005.
- Geirnaert 2000 – D. Geirnaert, "'Membra disiecta': banden met het versneden verleden'. In: R. Jansen-Sieben, J. Janssens & F. Willaert (red.), *Medioneerlandistiek. Een inleiding tot de Middelnederlandse letterkunde*. Hilversum, 2000 (Middeleeuwse studies en bronnen 69), 85-101.
- Hinskens & Van Dalen-Oskam 2007 – F. Hinskens & K. van Dalen-Oskam, 'Kwantitatieve benaderingen in taal- en letterkundig onderzoek. Een ruwe schets'. In: *TNLT* 123 (2007), 1-21.
- Hogenbirk 2009 – M. Hogenbirk, 'Is hij het? Lodewijk van Velthem en de compiler'. In: B. Besamusca, R. Sleiderink & G. Warnar (red.), *De boeken van Velthem. Auteur, oeuvre en overlevering*. Hilversum, 2009 (Middeleeuwse studies en bronnen 119), 47-92.
- Holmes 1998 – D.I. Holmes, 'The Evolution of Stylometry in Humanities Scholarship'. In: *Literary and Linguistic Computing* 13 (1998), 111-117.
- Janssens 1988 – J. Janssens, *Dichter en publiek in creatief samenspel. Over interpretatie van Middelnederlandse ridderromans*. Leuven, 1988 (Leuvense studiën en tekstuitgaven (nieuwe reeks) 7).
- Kestemont, M., 'Een onderzoek met stijl' [recensie van: J. van Driel, *Prikkeling der zinnen. De stilistische diversiteit van de Middelnederlandse epische poëzie*. Zutphen, 2007]. In: *Queeste* 14 (2007), 174-181.

- Kestemont 2009 – M. Kestemont, 'Bloemlezen uit Velthem. Handschrift Wenen, ö.n.b., Cod. 13.708 in Herne ten tijde van het Westers Schisma'. In: B. Besamusca, R. Sleiderink & G. Warnar (red.), *De boeken van Velthem. Auteur, oeuvre en overlevering*. Hilversum, 2009 (Middeleeuwse studies en bronnen 119), 251-266.
- Kestemont & Van Dalen-Oskam 2009 – M. Kestemont & K. van Dalen-Oskam, 'Predicting the Past. Memory-based Copyist and Author Discrimination in Medieval Epics'. In: T. Calders, K. Tuyls & M. Pechenizkiy (eds.), *Proceedings of the Twenty-first Benelux Conference on Artificial Intelligence (BNAIC 2009)*. Eindhoven, 2009, 121-128.
- Kestemont, Daelemans & De Pauw 2010 – M. Kestemont, W. Daelemans & G. De Pauw, 'Weigh your words – Memory Based Lemmatization for Middle Dutch'. In: *Literary and Linguistic Computing* 25 (2010), 287-301.
- Kuiper 1989 – W. Kuiper, 'Die riddere metten witten scilde'. *Oorsprong, overlevering en auteurschap van de Middelnederlandse 'Ferguut', gevolgd door een diplomatische editie en een diplomatisch glossarium*. Amsterdam, 1989.
- Luyckx & Daelemans 2008 – K. Luyckx & W. Daelemans, 'Authorship Attribution and Verification with Many Authors and Limited Data'. In: D. Scott & H. Uszkoreit (eds.), *Proceedings of the 22nd Conference on Computational Linguistics (COLING 2008)*. Brighton, 513-520.
- Murk-Jansen 1988 – S. Murk-Jansen, 'Hadewijch's *Mengedichten*. An Experiment in Statistical Method'. In: *Dutch Crossing* 35 (1988), 26-39.
- Reynaert 2002 – J. Reynaert, 'Boendale of "Antwerpse school"? Over het auteurschap van "Melibeus" en "Dietsche Doctrinale"'. In: W. van Anrooij e.a., *Al t' Antwerpen in die stad. Jan van Boendale en de literaire cultuur van zijn tijd*. Amsterdam, 2002 (Nederlandse literatuur en cultuur in de middeleeuwen 24), 127-157 & 177-182.
- Sebastiani 2002 – F. Sebastiani, 'Machine Learning in Automated Text Categorization'. In: *Association for Computing Machinery Computing Surveys* 34 (2002), 1-47.
- Sonnemans 1995 – G. Sonnemans, *Functionele aspecten van Middelnederlandse versprologen*. Boxmeer, 1995.
- Stamatatos 2009 – E. Stamatatos, 'A Survey of Modern Authorship Attribution Methods'. In: *Journal of the American Society for Information Science and Technology* 60 (2009), 538-556.
- Stamatatos e.a. 2000 – E. Stamatatos, G. Kokkinakis & N. Fakotakis, 'Automatic Text Categorization in Terms of Genre and Author'. In: *Computational linguistics* 26 (2000), 471-495.
- Van den Berg 1983 – E. van den Berg, *Middelnederlandse versbouw en syntaxis. Ontwikkelingen in de versificatie van verhalende poëzie ca. 1200-1400*. Utrecht, 1983.
- Van den Berg 1986 – E. van den Berg, 'Over het lokaliseren van Middelnederlandse rijmteksten'. In: *Verslagen en Mededelingen van de Koninklijke Academie voor Nederlandse Taal- en Letterkunde* (1986), 305-322.
- Van den Berg 1992 – E. van den Berg, 'Nadrukformules in Middelnederlandse ridderepiek'. In: *De Nieuwe Taalgids* 85 (1992), 205-214.
- Van Daele 2005 – R. van Daele, 'De robotfoto van de Reynaertdichter. Bricoleren met de overgeleverde wrakstukken: "cisterci'nzers", "grafelijk hof" en "Reynaertmaterie"'. In: *Tiecelijn* 18 (2005), 179-205.
- Van Dalen-Oskam 2007 – K. van Dalen-Oskam, 'Kwantificeren van stijl'. In: *TNTL* 123 (2007), 37-54.
- Van Dalen-Oskam & Van Zundert 2007 – K. van Dalen-Oskam & J. van Zundert, 'Delta for Middle Dutch – Author and Copyist Distinction in *Walewein*'. In: *Literary and Linguistic Computing* 22 (2007), 345-362.
- Van Driel 2007 – J. van Driel, *Prikkeling der zinnen. De stilistische diversiteit van de Middelnederlandse epische poëzie*. Zutphen, 2007.
- Van Halteren e.a. 2005 – H. van Halteren e.a., 'New Machine Learning Methods Demonstrate the Existence of a Human Stylome'. In: *Journal of Quantitative Linguistics* 12 (2005), 65-77.
- Van Loey 1946 – A. van Loey, 'Rijmonderzoek en Middelnederlandse dialectologie'. In: *Handelingen van de Koninklijke Commissie voor Toponymie en Dialectologie* 20 (1946), 41-48.
- Van Oostrom 1992 – F. van Oostrom, 'Maerlant tussen Noord en Zuid. Contouren van een biografie'. In: F. van Oostrom, *Aanvaard dit werk. Over Middelnederlandse dichters en hun publiek*. Amsterdam, 1992 (Nederlandse literatuur en cultuur in de middeleeuwen 6), 185-216 & 299-306.
- Van Oostrom 1996 – F. van Oostrom, *Maerlants wereld*. Amsterdam, 1996.
- Van Oostrom 2006 – F. van Oostrom, *Stemmen op schrift. Geschiedenis van de Nederlandse literatuur van het begin tot 1300*. Amsterdam, 2006.

Westgeest 2001 – H. Westgeest, ‘De Leidse lapidariumfragmenten: delen van Maerlants “cortten lapydarys”?’ In: *Queeste* 8 (2001), 1-16.
Wollheim 1972 – R. Wollheim, *On Art and the Mind. Essays and Lectures*. Cambridge, 1972.

Adres van de auteur

Instituut voor de Studie van de Letterkunde in de Nederlanden (ISLN)
CLiPS Computational Linguistics Group
Universiteit Antwerpen, Stadscampus
Prinsstraat 13, kamer D.118
B-2000 Antwerpen
België
mike.kestemont@ua.ac.be