TNTL 123 (2007) 37-54

Karina van Dalen-Oskam

Kwantificeren van stijl*

Abstract — Quantitative methods to analyze style usually are applied in research which tries to attribute a text of uncertain authorship to a certain author. The uniqueness of authors (which is generally assumed) has been looked for in a.o. word length, sentence length, type-token ratio, lexical richness and lexical preferences. In recent years, researchers focused on textual and linguistic characteristics which are assumed to be the result of unconscious processes. In this context, measuring high frequency words became very important. The best results seem to be yielded by a measure called Burrows's Delta. In this paper the most important ways of measuring in the field of quantitative analysis of style and of non-traditional authorship attribution will be reviewed. The first results of a quantitative analysis of the lexicon of the Middle Dutch Arthurian romance *Walewein*, applying a form of Burrows's Delta procedure, are summarized to illustrate the possibilities of this type of research.

1 Inleiding

Kwantitatieve methoden voor het analyseren van stijl hebben tot op heden voornamelijk als doel gehad om auteurs van elkaar te helpen onderscheiden. Er wordt naar methoden gezocht die een tekst van onzekere herkomst met de grootst mogelijke betrouwbaarheid kunnen toeschrijven aan een auteur. Op basis van de resultaten wordt in een deel van de onderzoeken vervolgens gekeken welke verschillen in stijl op deze manier zijn gevonden en hoe deze zijn te beschrijven en verklaren vanuit de tekst die wordt onderzocht. Dit is te zien als de tekstanalytische onderbouwing. De fundamentele gedachte achter het onderzoek is 'individualiteit', het uitgangspunt dat elke auteur uniek is. Onderzoekers die een dergelijke individualiteit afwijzen, zullen geen waarde hechten aan dit type onderzoek, zoals Harold Love schrijft in zijn boek *Attributing authorship: an introduction* (Love 2002: 10). John Burrows, de belangrijkste vernieuwer in dit onderzoeksgebied uit de *humanities computing*, omschrijft het uitgangspunt in zijn meest recente publicatie (waarover meer aan het eind van paragraaf 2) als volgt:

> Evidence of authorship pervades whatever anybody writes. Provided appropriate procedures are employed in the analysis of an appropriate set of texts, it can almost always be elicited. It is inherent, however, not merely in statistical principle but in human behaviour at large, that such evidence cannot be absolute. The consistencies we observe are trends, not universals. Our many stabilities are offset by our capacity for change. (Burrows 2006: 2-3)

Die individualiteit wordt eveneens aangenomen en onderzocht in de cognitieve stilistiek, waarvan Ian Lancashire een belangrijke vertegenwoordiger is. Lancashire is op zoek naar het idiolect van auteurs en probeert door experimenten zicht te

* Mijn dank gaat uit naar Joris van Zundert (Huygens Instituut), die ervoor gezorgd heeft dat een aantal statistische aspecten nader toegelicht kon worden.

krijgen op hoe de menselijke hersenen zich gedragen in het proces van tekstproductie. Daarin zijn een onbewust en een bewust proces te onderscheiden: een creatieve, 'geïnspireerde' fase respectievelijk een analytisch proces waarin auteurs zinnen vormgeven en bewerken. Omdat de hersenen van ieder mens verschillen, zullen ook de kenmerken van door verschillende mensen geproduceerde teksten weer anders zijn, mede afhankelijk van sociale en culturele achtergronden en de tijd van schrijven. Die kenmerken 'are not unique, like fingerprints, but, taken together, they amount to sufficiently distinctive configurations to be useful in authorship attribution or text analysis' (Lancashire 2004: 397).

Kwantitatieve methoden zijn in de laatste vier decennia tot bloei gekomen, de nieuwe ontwikkelingen in de informatietechnologie zowel anticiperend als op de voet volgend. Kwalitatief onderzoek naar stijl- en auteursverschillen, bijvoorbeeld door het analyseren van thema's, motieven, stilistische voorkeuren zonder deze in getallen te willen vangen, vindt al veel langer plaats en is nog steeds waardevol, ook gecombineerd met moderne kwantitatieve methoden – het evalueren van kwantitatieve resultaten is voor een belangrijk deel een kwalitatief proces. In het al genoemde boek van Harold Love komt dit allemaal op voorbeeldige wijze aan de orde.

In het nu volgende zal de nadruk liggen op de aard en functie van de kwantitatieve methoden in het kader van letterkundig onderzoek. Een overzicht van de belangrijkste ontwikkelingen wordt gegeven in paragraaf 2; voor een uitvoerige beschrijving van de wiskundige en statistische kanten van de beschreven maten wordt verwezen naar de aangehaalde literatuur. De beschreven onderzoeksdiscipline heeft door het belang van de methodologie en statistiek een internationaal karakter, maar in het overzicht zal ook aandacht worden besteed aan opvallende publicaties over Nederlandse literaire teksten. Als toelichting op het algemene overzicht wordt in paragraaf 3 een *case study* meer in detail beschreven. Een blik op de toekomst wordt gepresenteerd in paragraaf 4, waarin wordt aangegeven welke sporen het veelbelovendst lijken voor vervolgonderzoek en waarin wordt geschetst wat er nodig is om nieuwe ontwikkelingen mogelijk te maken.

2 Ontwikkeling van het onderzoek

Stijl is een breed begrip. Het betreft niet alleen de manier waarop taal wordt gebruikt in een bepaalde context en met een bepaald doel (Leech & Short 1981: 10) maar ook inhoudelijke aspecten en de interactie tussen de (linguïstische) vorm en de inhoud (Van Eck & Streng 1997: 7). Een combinatie van eigenschappen bepaalt het idiolect van een auteur. Voor een beschrijving en een vergelijking van verschillende idiolecten kunnen alleen eigenschappen worden bestudeerd die objectief kwantificeerbaar zijn, dus eigenschappen die niet door elke onderzoeker weer anders omschreven (en dan wel of niet meegeteld) kunnen worden. Dit is over het algemeen gemakkelijker voor linguïstische dan voor inhoudelijke aspecten van teksten. Er zijn, zoals in de inleiding al is gezegd, ook andere manieren om stijl nader te analyseren zonder dat een dergelijk kwantitatief aspect nadrukkelijk aanwezig is, maar vaak zal er toch een zekere mate van kwantificering plaatsvinden, bijvoorbeeld als eerste fase in het onderzoek (vgl. Anbeek & Verhagen 2001). Hieronder

zullen de belangrijkste reeds onderzochte eigenschappen de revue passeren, met een beknopte evaluatie van de voor- en nadelen in het gebruik. Om geïnteresseerden verder op weg te helpen wordt de belangrijkste secundaire literatuur genoemd en wordt op enkele interessante toepassingsvoorbeelden iets nader ingegaan. Voor een korte bespreking van nog andere methoden wordt verwezen naar Holmes 1994 (vgl. Love 2002: 135). De volgorde waarin de verschillende maten aan de orde komen weerspiegelt ruwweg ook de chronologische volgorde waarin zij hun intrede hebben gedaan in de geschiedenis van de onderzoeksdiscipline.

Probeerde men aanvankelijk met een enkele maat het raadsel van auteurschap op te lossen, in de loop der jaren is men steeds meer verschillende maten naast elkaar gaan toepassen, om nadelen van het gebruik van een enkele maat te ondervangen. Ook werden nieuwe methoden ontwikkeld die steeds intensiever gebruik maakten van statistiek en nieuwe informatietechnologische ontwikkelingen. Hierbij werd het in de inleiding al genoemde verschil tussen bewust en onbewust tot stand gekomen tekstaspecten steeds belangrijker in de algehele argumentatie: tekstkenmerken die het resultaat zijn van bewuste processen kunnen gemakkelijk geïmiteerd worden voor misleidende of parodiërende doeleinden en zijn om die redenen minder betrouwbaar voor auteursherkenning dan uit onbewuste processen voortkomende tekstkenmerken (vgl. Love 2002: 179-193 en Burrows 2005). Ook een voortdurend punt van aandacht was welke verschillen nu eigenlijk werden ontdekt als een meting statistisch significante en dus in principe betekenisvolle, niet als toevallig te verklaren resultaten opleverde. Regelmatig leek een verschil in tijd van schrijven of van genre duidelijker opgemerkt te worden met de gebruikte maten dan een verschil in auteur, hetgeen consequenties heeft voor het selecteren van de te vergelijken teksten en auteurs en voor de breedte van de toepasbaarheid van de maten (vgl. Love 2002: 222). Dit resultaat is echter op zijn beurt interessant als men de computer zou willen inzetten voor het determineren van de tijdperiode waarin of de literaire stroming waarbinnen een tekst tot stand gekomen is, zoals Gillis Dorleijn in een zeer interessant artikel heeft voorgesteld (Dorleijn 1995). Kwantitatief onderzoek naar de veranderingen in stijl van een auteur door de jaren heen is nog niet vaak gedaan (Forsyth 1999).

2.1 Lengtematen

Verschillende 'lengtematen' zijn in het verleden verkend. Woordlengte is als onbetrouwbaar in het onderscheiden van auteurs afgedaan; het aantal syllaben per woord lijkt eerder talen dan auteurs van elkaar te onderscheiden, maar kan in sommige gevallen wellicht toch iets zeggen (Holmes 1994: 88). Zinslengte, gewoonlijk gemeten in het aantal woorden per zin, lijkt ook niet helemaal betrouwbaar, mede omdat deze tot stand kan komen 'under the conscious control of an author' of van een redacteur en dus imiteerbaar kan zijn (Holmes 1994: 89).

In 2001 publiceerde George Barr een artikel waarin een nieuwe uitwerking van de zinslengtemaat toch tot interessante observaties leidt. Kern van zijn aanpak is het relateren van tekstlengte aan de verzameling en frequentie van de gebruikte zinslengtes ('scale', iets wat grotendeels een onbewust compositieproces zou zijn), waarbij hij verder let op de relatie tussen het aantal langere en kortere zinnen ('contrast') en op significant groter gebruik van bijvoorbeeld extreem lange zin-

nen ('monumentality'). Met name dat laatste aspect kan wijzen op interessante tekstinterne en dus stijl- of auteursgerelateerde zaken. Maar voor de algemene uitkomsten moet Barr wel toegeven dat er ook erg veel overlap is in de gebruikte zinslengtes tussen verschillende teksten en tussen verschillende auteurs. Ook in zijn toepassing is de zinslengtemaat dus niet het wondermiddel om auteurs eenduidig van elkaar te onderscheiden, maar enkele aspecten van stijl of individualiteit kunnen er wel mee worden gesignaleerd. De door de onderzochte auteur gekozen syntactische structuren (van simpel tot complex) hebben immers invloed op de zinslengte, en teksten waarin bijvoorbeeld veel uitroepen voorkomen, bevatten daardoor relatief veel uiterst korte zinnen. Een dergelijk contrast kan weer iets zeggen over de 'mood' (zoals Barr dat noemt) van de tekst. Verder onderzoek naar Barrs aanpak als hulpmiddel voor het signaleren van tekst-, genre- en auteursgebonden stilistische verschillen zou zeker interessant zijn.

2.2 Woordenschat

De opbouw van de woordenschat van een tekst, oeuvre of corpus is in vele maten gebruikt als hulpmiddel voor het beschrijven van individualiteit. Een van de verkende maten is de 'type-token ratio', ofwel de verhouding tussen het aantal typen (woorden, dat wil zeggen lexicale items) en tokens (woordvormen, ofwel het aantal voorkomens in alle mogelijke grammaticale vormen van alle lexicale items). Dat houdt in dat voor het analyseren van de woordenschat elke woordvorm wordt gerekend tot een bepaald hoofdwoord waarvan het een (al dan niet) verbogen of vervoegde vorm is. Dus alle voorkomens van een bepaald werkwoord, in welke vervoeging dan ook – persoonsvormen in de eerste, tweede, derde persoon enkelvoud of meervoud, in de tegenwoordige of verleden tijd, als (on)voltooid deelwoord etc. – tellen bijvoorbeeld als instanties van de infinitief van dat werkwoord; de infinitiefvorm kan bijvoorbeeld worden gebruikt als het 'type'. De type-token ratio kan dan worden uitgedrukt als het gemiddeld aantal tokens per type. Deze simpele maat levert echter weinig zinvolle, scherp onderscheidende informatie op (Holmes 1994: 92). Interessanter is een volgende stap van woordenschatanalyse waarin geformuleerd wordt hoe divers, hoe 'rijk', het vocabulaire van een tekst of auteur is. De belangrijkste maat voor lexicale rijkdom (o.a. volgens Burrows 2002: 269) is 'Yule's Characteristic (*K*)', ontwikkeld door G. Udny Yule in 1944 en bijgesteld door G. Herdan in 1955 (vgl. Tweedie & Baayen 1998, Hoover 2003). Yule's *K* meet de mate van woordherhaling in een tekst, uitgaand van de basisgedachte dat het voorkomen van een bepaald woord random is (Holmes 1994: 92-93). Hierbij verstaat men onder woord gewoonlijk woordvorm, token. Hoe hoger *K*, hoe meer woordherhaling, en hoe lager *K*, hoe gevarieerder de woordenschat is. Wat *K* beter maakt dan type-token ratio is onder andere dat ook de lengte van de tekst in de berekening wordt betrokken.[1]

1 Dit is met een rekenvoorbeeld te demonstreren. In de tabel in deze noot vergelijken we vier teksten met elkaar. Twee teksten hebben tien woordvormen (N=10) en twee hebben er honderd (N=100). We berekenen dan de type-token ratio voor twee situaties: dat alle tokens tot hetzelfde type (lexicale item) behoren (V=1) ofwel tot tien verschillende typen (V=10). Dan blijkt dat de teksten met een even grote woordenschat (V) afhankelijk van de omvang een zeer verschillende ratio kunnen vertonen, en dat een tekst met een grotere woordenschat geen hogere ratio hoeft te vertonen dan een

Dit is een van de maten die zijn toegepast op de Middelnederlandse Arturroman *Walewein*, die geschreven is door twee auteurs, de verder onbekende 'Penninc' en 'Pieter Vostaert' (Van Dalen-Oskam & Van Zundert 2005: 214-215 en 224-225). De metingen zijn toegepast op de types in de tekst, en abstraheren dus van spelling, vervoeging en verbuiging zodat er daadwerkelijk voorkomens van lexicale items worden gemeten (alleen zijn homoniemen niet onderscheiden). Het laatste deel van de tekst bleek een aanzienlijk hogere *K* te hebben dan de delen ervoor en dan de gehele tekst; de woordherhaling in dit deel is dus substantieel groter. De eerdere delen van de tekst vertonen een duidelijk lagere *K* en zijn dus lexicaal rijker, hebben een gevarieerder vocabulaire. Na een hele reeks van metingen waarbij de lexicale rijkdom door de tekst heen gemeten werd, bleek het scharnierpunt in de grote verschillen in *K* zich rond vers 7880 te bevinden. Het is goed mogelijk dat dit ongeveer de plaats in de tekst is waar de tweede auteur het heeft overgenomen van de eerste. De resultaten van *K* voor een paar delen van de tekst illustreren dit:

Hele tekst (vers 1-11.202):	$K = 173{,}98$	
Vers 4581-7880	$K = 167{,}42$	-6,56 (-3,77%) t.o.v. hele tekst
Vers 7880-11.181	$K = 183{,}07$	+9,09 (+5,22%) t.o.v. hele tekst

Hierbij is de *K* vergeleken tussen de 3300 verzen voor en na vers 7880 (waarbij het grootste deel van de epiloog buiten beschouwing is gelaten en eventuele 'opstart'-verstoringen in het eerste deel de resultaten niet vervormen). Deze casus komt nader aan de orde in paragraaf 3. Dat Yule's *K* nog steeds in de belangstelling staat, blijkt ook uit een zeer recente publicatie van Miranda-García en Calle-Martín (2005).

Bij Yule's *K* wordt gekeken naar de rijkdom van een complete tekst; voor de volgende stap van het onderzoek van een literaire tekst is het nodig om te weten in welke woorden de verschillen tussen auteurs nu het duidelijkst zichtbaar worden. Door in te zoomen op de lexicale eenheden binnen een tekst kunnen de concrete verschillen tussen tekstdelen, teksten en auteurs in stilistisch opzicht zichtbaar gemaakt worden. Die stilistische verschillen kunnen vervolgens weer in verband worden gebracht met andere analyses, vanuit bijvoorbeeld onderzoek naar het doel van de auteur, het beoogd publiek, thematische lijnen, en zo verder. Voor het

tekst met een kleine woordenschat. Yule's *K* geeft als maat voor lexicale rijkdom een logischer beeld. In tekstnotatie: $-1/N + (\sum_{i=1..N}\{V_{(i,N)}\cdot(i/N)^2\})$ (We laten de vergrotingsfactor achterwege). De berekening van de type-roken ratio (TTR) staat voor het overzicht in dezelfde cel.

	N=10	N=100
V=1	K=-1/10 + { … + 0 + 0 + 1 * (10/10)^2 }	K=-1/100 + { … + 0 + 0 + 1 (100/100)^2 }
	= -1/10 + 1=9/10=0,90	= -1/10 + 1=9/10=0,90
	*TTR: 1/10*100=10*	*TTR: 1/100*100=1*
V=10	K=-1/10 + { 10 * (1/10)^2 + 0 + 0 + …}	K=-1/100 + { … + 0 + 0 + 10 * (10/100)^2 }
	= -1/10 + 1/10=0,00	= -1/100 + 10/100=9/100=0,09
	*TTR: 10/10*100=100*	*TTR: 10/100*100=10*

K geeft aan dat een tekst met een woordenschat bestaand uit één type (V=1) de kleinste lexicale rijkdom heeft, ongeachte de lengte van de tekst. Een tekst die meer typen kent (V=10) heeft een grotere lexicale rijkdom. En naarmate een tekst meer typen per tokens heeft, stijgt ook de lexicale rijkdom navenant. Tot maximaal oneindig als er precies evenveel typen zijn als tokens, zoals bij N=10 en V=10. Daar is 1/K=1/0,00 oftewel ∞. En dat is logisch: een grotere lexicale rijkdom is niet mogelijk.

beoordelen van de frequentie van woorden is veelvuldig gebruik gemaakt van de chi-kwadraat test.

Chi-kwadraat bepaalt de significantie van verschillen in de verdeling van een nominale variabele. Een nominale variabele verdeelt meetwaarden in categorieën die verder niet te ordenen zijn: een sekse valt in de categorie man of vrouw, de variabele religie categoriseert naar bijvoorbeeld christen, moslim, hindoe en boeddhist (de scores van numerieke variabelen daarentegen hebben een natuurlijke ordening: gewicht, lengte, aantal etc.). In een interessant artikel waarin de auteur zoekt naar de manier waarop kan worden vastgesteld welke woorden nu kenmerkend zijn voor een bepaalde tekst geeft Adam Kilgarriff aan waarom deze test in sommige gevallen problematisch is. De chi-kwadraat test gaat – net als Yule's *K* – er van uit dat een tekst een toevalsselectie van woorden bevat. Kilgarriff maakt aan de hand van corpusgebaseerde, dus empirische experimenten duidelijk dat die gedachte niet houdbaar is. De selectie van woorden in een tekst is volgens hem dus niet random, wat als consequentie heeft dat de meetresultaten van hoogfrequente en laagfrequente woorden verschillende interpretaties vereisen (Kilgarriff 1996: 2-3). Dat neemt niet weg dat de maat wel degelijk nuttige toepassingen heeft gehad. Voor Middelnederlandse teksten zijn belangwekkende resultaten verkregen door onder andere Evert van den Berg in zijn onderzoek naar verstechnische ontwikkelingen in Middelnederlandse epiek (Van den Berg 1983) en door Willem Kuiper voor het beschrijven van de verschillen in het eerste en het tweede gedeelte van de Middelnederlandse Arturroman *Ferguut* (Kuiper 1989).

2.3 Principal Components Analyse

Een andere maat die gebruikt kan worden om de stijl van een tekst nader te verkennen is 'Principal components analysis' ofwel factoranalyse. PCA heeft als doel 'to explain as much of the total variation in the data as possible with as few variables as possible' (Binongo & Smith 1999: 447; vgl. ook Holmes 1994: 99)). Die variatie kan worden bekeken in de woordenschat en/of in het gebruik van andere elementen, zoals interpunctie, zinslengte en dergelijke, maar ook kan bijvoorbeeld metafoorgebruik worden bekeken. Hiervoor moet de onderzoeker dan wel objectieve richtlijnen vastleggen aan de hand waarvan de tekst vervolgens van meetbare 'labels' wordt voorzien. Bij toepassing op de woordenschat van een tekst resulteert dat in een lijstje woorden of woordsoorten (de variabelen) die verdeeld zijn in enkele groepjes (ofwel componenten/factoren) die in aflopend belang de variatie in de tekst beschrijven. Verder inzoomen op die 'principal components' helpt vervolgens meer inzicht te verkrijgen in de manier waarop de onderzochte teksten van elkaar verschillen.

Een ruwe versie van PCA werd gebruikt door John Burrows, wiens dissertatie *Computation into criticism: a study of Jane Austen's novels and an experiment in method* (1987) vele latere onderzoekers heeft geïnspireerd. In zijn latere publicaties heeft Burrows belangrijke innovaties in de toepassing van PCA gepresenteerd. In zijn dissertatie pleitte hij voor het onderzoeken van de 'gewone' woorden in te analyseren teksten, en voor zijn eigen onderzoek definieert hij die als de dertig meest frequente woorden. In de romans van Austen heeft hij die dertig woorden onderzocht, onderscheid makend naar de dialoogtekst van de verschillende perso-

nages. Op deze manier wilde hij een beschrijving geven van de betreffende personages en van hun ontwikkeling (Burrows 1987: 1-12). In elk hoofdstuk past hij een andere maat toe om dit hoogstfrequente materiaal vanuit verschillende invalshoeken te bestuderen, en dat leidt tot belangrijke en inspirerende stilistische en tekstanalystische observaties. Een voorbeeld is Burrows' beschrijving van de veranderende verhouding tussen Emma en Mr. Knightley in *Emma* aan de hand van het gebruik van de twaalf frequentste woorden in hun dialoogtekst. Met name de veranderingen in het gebruik van het persoonlijk voornaamwoord *I* 'ik' blijken hierbij een rol te spelen (Burrows 1987: 200). Ook in zijn latere publicaties is de concentratie op de structuur en inhoud van het literaire werk zelf het meest innoverende aspect van Burrows' werk: hij blijft niet bij de cijfers als resultaten, maar ziet deze als stapsteen naar een nadere beschouwing van de stijl en inhoud van de onderzochte teksten. In zijn onderzoek verliest hij zijn uiteindelijke (letterkundige) doel, inzicht in de literaire werken, nooit uit het oog (vgl. ook Love 2002: 156). Burrows constateerde uiteindelijk dat 'the crucial point is that pca is not intrinsically a test of authorship but only of comparative resemblance'. De methode helpt dus wel in stijlonderzoek, maar niet als wichelroede voor auteursherkenning. Voor Burrows was dat de aanleiding om een andere test te ontwikkelen, die in paragraaf 2.4 uitvoerig beschreven zal worden. PCA blijft wel in zijn assortiment van maten zitten, maar pas in een later stadium van auteursonderscheidend onderzoek, als het aantal in aanmerking komende auteurs al zoveel mogelijk door andere metingen ingeperkt is (Burrows 2003: 8).

Alhoewel de wiskundige onderbouwing en de cijfermatige bewerkingen die ten grondslag liggen aan PCA lastig te doorgronden zijn (zie voor een uitleg van de mathematische kant Binongo & Smith 1999) wordt de methode regelmatig gebruikt. Jan Rybicki paste Burrows' methode toe om inzicht te krijgen in de verhouding tussen de stijl van een Poolse trilogie van Henryk Sienkiewicz en die van twee Engelse vertalingen ervan (Rybicki 2006), wat tot interessante observaties over stijlen van vertalen leidt. Een eigen stap verder doet Larry Stewart in zijn artikel 'Charles Brockden Brown: Quantitative Analysis and Literary Interpretation', dat beschouwd mag worden als een perfect voorbeeld van gebruik van kwantitatieve methoden (waaronder PCA) binnen een onderzoek dat is gestuurd door een concrete letterkundige vraagstelling (Stewart 2003). Object van onderzoek zijn twee romans van de Amerikaanse auteur Charles Brockden Brown, *Wieland or The Transformation* (1798) en *Memoirs of Carwin, The Biloquist* (1803-1805). De tweede roman wordt verteld vanuit hoofdpersonage Carwin, die ook voorkomt in de eerste roman. Het perspectief in *Wieland* ligt bij hoofdpersonage Clara, maar drie hoofdstukken worden gezien door de ogen van drie andere (mannelijke) personages en een van hen is Carwin. Stewart stelt zich de vraag

> whether quantitative analysis would indicate if Brown had created in Carwin a character and narrator with a distinctive voice, whether what we call the character Carwin is a distinct literary or linguistic entity. Is the voice of Carwin similar in the two texts and can it be differentiated from other narrative voices in *Wieland*? (Stewart 2003: 130)

Voor het onderzoeken van die 'narrative voices' stelde Stewart een lijst van 44 variabelen op: de dertig hoogstfrequente woorden, de frequentie van verschillende interpunctiesoorten (punt, komma, vraagteken, uitroepteken, puntkomma, dub-

bele punt en weglatingsteken). Verder telde hij zinslengte en alinealengte, de verhouding tussen korte en lange zinnen (waarbij kort vijf of minder woorden en lang 25 of meer woorden is), de verhouding tussen korte en lange alinea's (50 of minder respectievelijk 125 of meer woorden); gemiddeld aantal woorden per zin, gemiddeld aantal zinnen per alinea, en het percentage slechts eenmaal gebruikte woorden (hapax legomena). Stewart zette PCA in 'to reduce the forty-four variables to two dimensions' (ofwel componenten/factoren) en zette de resultaten voor deze twee belangrijkste factoren tegen elkaar uit in een grafiek. Dit leverde een duidelijk verschil op tussen de vier verschillende vertellers, waarbij de Carwins uit de beide romans inderdaad dicht bij elkaar in de buurt eindigden. De variabelen die de meeste invloed hadden op de geconstateerde verschillen waren het percentage korte alinea's, uitroeptekens, korte zinnen, woorden per alinea en komma's. Hierbij is Carwin steeds de focalisator (het personage 'door wiens ogen' de lezer het verhaal ervaart) met de 'beknoptere' stijl. Stewart keert terug naar de teksten en analyseert de kwantitatieve resultaten met de volgende observatie:

> Thus, while Clara's style frequently seems uncertain (as indicated by the question marks) and distracted and even wild (as indicated by the exclamation points and dashes), Carwin's effectiveness as a villain may be enhanced by his more understated and coolly logical style. (Stewart 2003: 132)

De 'narrative voice' van de twee andere mannelijke personages die ook elk een hoofdstuk van *Wieland* als verteller voor hun rekening nemen, lijkt erg veel op die van Carwin, en Stewart sluit zijn artikel af met een analyse van hun rol in de roman. De stilistische analyse leidt hier heel concreet tot een nieuwe interpretatie van de roman *Wieland* en tot een dieper inzicht in het werk (Stewart 2003: 134-138).[2]

2.4 Hoogfrequente woorden

In twee nauw met elkaar samenhangende artikelen introduceerde John Burrows in 2002 en 2003 een nieuwe maat voor auteursherkenning: Delta. Hij was op zoek naar een relatief simpele maat waarmee kan worden vastgesteld dat een tekst met grote waarschijnlijkheid kan worden toegeschreven aan de auteur van een andere tekst in een grote groep van teksten, of waarmee kan worden geconstateerd dat zich in dat grote corpus geen tekst bevindt van de gezochte auteur. Ook voor deze maat bepaalt Burrows zich tot woorden, omdat zij, zo beargumenteert hij, toegankelijk en betekenisvol zijn voor de onderzoeker. 'They help us, in particular, to form close and fruitful inferences about the outcome of an inquiry' (Burrows 2002: 268). Delta beschrijft in wezen het verschil in afwijking die twee (of meer) teksten vertonen in het gebruik van hoogfrequente woorden ten opzichte van een algemeen gemiddeld gebruik van die woorden. Het algemeen gemiddeld gebruik kan worden bepaald door voor een grote groep teksten de gemiddelde relatieve frequentie per type te berekenen. Dat gemiddelde berekenen we bijvoorbeeld

2 Niet elke methode biedt de mogelijkheid om aan de hand van de meetresultaten verder in te zoomen op de tekst. Om die reden wordt hier niet ingegaan op Monte-Carlo-Feature-Finding (meer hierover in Forsyth 1999) en op het gebruik van neurale netwerken (hierover is meer te vinden in Tweedie, Singh & Holmes 1996).

voor de honderdvijftig frequentste woorden uit de groep teksten. Vervolgens wordt voor twee teksten (bijvoorbeeld een tekst waarvan de auteur bekend is en een tekst waarvan de auteur onbekend is) de afwijking in het gebruik ten opzichte van dit algemene gemiddelde berekend. Dit doen we door per type te kijken wat het verschil is tussen de relatieve frequentie van de tekst en de gemiddelde relatieve frequentie in het corpus. Voor beide teksten tellen we al deze verschillen op en delen die door het aantal vergeleken typen (in dit geval 150). Dit geeft de gemiddelde afwijking van beide teksten ten opzichte van het algemene hoogfrequente woordgebruik in het corpus. Het verschil tussen deze twee gemiddelde afwijkingen noemt Burrows Delta.[3]

Burrows normaliseerde zijn materiaal wat de spelling betreft, gaf volledige in plaats van samengetrokken woordvormen (bijvoorbeeld *do not* in plaats van *don't*), en onderscheidde enkele homoniemen naar grammaticale functie. Zijn eerste onderzoek betrof een groep gedichten van de hand van vijfentwintig dichters uit de Engelse Restoration-periode, aangevuld met andere teksten waarvan eveneens in alle gevallen de auteur bekend was. Vervolgens keek Burrows voor steeds één tekst hoe deze zich verhield tot steeds één andere tekst uit de groep teksten, met de dertig frequentste woorden, daarna met de veertig, 60, 80, 100, 120 en 150 frequentste woorden. Omdat van alle gedichten de auteur als bekend werd verondersteld, kon vervolgens worden geëvalueerd bij welke invulling Delta de beste resultaten – zijnde de meeste correcte toeschrijvingen – opleverde. Hierbij betrok Burrows ook de lengte van de gedichten. Alleen bij heel korte gedichten (1-500 woorden) is het percentage correcte toeschrijvingen niet overtuigend. Maar voor de langere gedichten geldt dat de betrouwbaarheid van de toeschrijving aan een auteur uit de groep toeneemt hoe meer woorden er uit de top van de frequentielijst worden gebruikt. Zelfs voor gedichten van maar 100 woorden kreeg hij bij vergelijking met de 150 frequentste woorden uit de groep goede resultaten (Burrows 2002: 277).

Burrows' Delta is behalve door Burrows zelf ook uitvoerig getest door David Hoover. Hij paste Delta toe op prozateksten en liet zien dat de accuratesse van

3 Voor de volledigheid moet wel vermeld worden dat we feitelijk niet de verschillen in relatieve frequenties berekenen, maar de verschillen in z-scores voor die relatieve frequenties. De z-score wordt formeel gegeven door: $z_x=(x - \mu_x)/\sigma_x$. Dat wil zeggen: de z-score voor een meting met waarde *x* is gelijk aan *x* minus het gemiddelde van de meetreeks waar *x* onderdeel van uit maakt, gedeeld door de standaarddeviatie van diezelfde reeks. Feitelijk betekent dit dat de z-score aangeeft hoe groot de afwijking van een meting ten opzichte van het gemiddelde is, uitgedrukt in een aantal standaarddeviaties. De z-score geeft ons, kort gezegd, een maat voor de betekenisvolheid van een verschil in relatieve frequenties. Een voorbeeld kan dit duidelijk maken. Stel dat de relatieve frequentie van het type 'die' voor de individuele teksten in het corpus grote verschillen vertoont. In dat geval kunnen we misschien in een van beide teksten die we met het corpusgemiddelde vergelijken wel een heel groot verschil in frequentie vinden, maar we weten ook dat zo'n groot verschil niet erg uitzonderlijk is. Maar de z-score is in zo'n geval gering en deze afwijking telt daarom minder zwaar mee voor de totale Delta. Stel daarentegen dat het type 'ende' over alle teksten een heel consistente hoge frequentie laat zien, en stel dat we voor onze voorbeeldteksten juist vinden dat daarin de frequentie van 'ende' opvallend anders is (bijvoorbeeld veel lager), dan willen we dat verschil juist markeren als zwaarwegend. In dit geval is de z-score dan ook hoog en telt het verschil extra aan in de berekening voor de totale Delta. Op deze wijze benadrukt de z-score dus de opmerkelijke verschillen in hoogfrequent woordgebruik. Verder redeneren we dat twee teksten die ten opzichte van elkaar een heel kleine delta hebben (ofwel vrijwel dezelfde afwijking ten opzichte van een algemeen gemiddelde in het gebruik van hoogfrequente woorden) door dezelfde auteur geschreven kunnen zijn.

Delta hiervoor zelfs nog verder toenam als de lijst van 150 frequentste woorden werd uitgebreid tot de 800 frequentste woorden; dit geeft aan dat tekstlengte een belangrijke variabele voor de precieze invulling van de nieuwe maat kan zijn (Hoover 2004a). Door dieper in te gaan op de achterliggende statistiek wist Hoover de kracht van Delta nader te verklaren; hij stelde enkele modificaties voor die een nog beter resultaat opleverden (Hoover 2004b).

2.5 Minder frequente woorden

Burrows richtte zich vervolgens op de woorden die wat frequentie betreft meer in het 'middengebied' van de woordenschat van teksten en auteurs worden aangetroffen. Zijn eerste publicatie hierover betreft het auteurschap van *Shamela* (1741), een parodie op Samuel Richardsons *Pamela* (Burrows 2005). Het is een anonieme tekst die gewoonlijk wordt toegeschreven aan Henry Fielding en die wordt gekarakteriseerd als een briljante parodie. Toepassing van Delta levert wisselende resultaten op: soms sluit de tekst aan bij ander werk van Richardson, dan weer bij ander werk van Fielding. Burrows past zijn vraagstelling als volgt aan: welke woorden komen in *Shamela* opvallend vaker voor dan in andere teksten van Richardson en Fielding en bij welk van de auteurs sluit die tendens het meeste aan? Burrows zoekt naar een methode waarop deze, statistisch weinig relevante items, cumulatief bekeken kunnen worden en zo alsnog een aanvullende significante bijdrage kunnen leveren aan auteursonderscheidend en stilistisch onderzoek. De kern van zijn nieuwe maten in ontwikkeling, die hij in zijn meest recente publicatie (Burrows 2006) de Zeta test en de Iota test noemt, is het meten van het voorkomen van bij twee of meer auteurs voorkomende woorden in verschillende tekstblokken van (bijvoorbeeld) telkens 20.000 tokens. Vervolgens kijkt hij per woord in hoeveel tekstblokken het voorkomt, onafhankelijk van de frequentie, en vergelijkt hij dat per tekst met het voorkomen van die woorden in het overige materiaal dat als controlegroep dient. Voor *Shamela* wordt Fielding met deze maten eenduidig als de meest waarschijnlijke auteur aangemerkt.

De nieuwe maten van Burrows moeten, zoals hij zelf ook aangeeft, nog verder getest worden voordat ze breder toegepast kunnen worden naast Delta. Aangezien er, zoals eerder werd geconstateerd, verschillende statistische benaderingen nodig zijn voor hoogfrequente en middel- tot laagfrequente woorden in een corpus (vgl. Dunning 1993 en met name Kilgarriff 1996), is het van belang dat er nieuwe maten ontwikkeld worden.

3 Casus: *Walewein*

Een van de punten van kritiek op het onderzoek naar maten voor auteursherkenning is dat de onderzoekers een te simpel model van auteurschap hanteren. Auteurs kunnen samenwerken met anderen, maar ook als zij alleen werken zal het eindproduct invloeden van anderen vertonen, en bij het gereedmaken voor publicatie kan een redacteur een belangrijke rol spelen. Dit zijn allemaal zaken waarmee in het huidige kwantitatieve onderzoek naar stijl- en auteursverschillen weinig rekening wordt gehouden (Love 2002: 32-50). Misschien is dat een van de re-

denen waarom er nog relatief weinig kwantitatief onderzoek is gedaan naar middeleeuwse teksten in het kader van het verifiëren van auteurschap. Teksten werden handmatig gekopieerd en het is bekend dat kopiisten hun eigen invloed konden uitoefenen op de tekst die zij onder handen hadden, en dat zij alleen al door onbedoelde kopieerfouten in feite een nieuwe tekst creëerden. Bovendien konden kopiisten van legger wisselen, dus zich baseren op verschillende handschriften met 'dezelfde' (maar steeds unieke) tekst. Daarnaast zijn voor die periode maar weinig auteurs bij naam bekend en zijn veel teksten anoniem overgeleverd. Redenen genoeg om terughoudend te zijn. Hoe lastig deze situatie is, blijkt uit het onderzoek van Kari Anne Rand Schmidt (1993) naar het auteurschap van het veertiende-eeuwse Engelse *The Equatorie of the Planetis*, dat door eerdere onderzoekers wel aan Geoffrey Chaucer werd toegeschreven. Ondanks tegengeluiden werd de toeschrijving aan Chaucer steeds algemener. Rand Schmidt probeerde de casus met allerlei in de vorige paragraaf beschreven maten onbevooroordeeld te benaderen en moest concluderen dat de resultaten van de metingen niet significant genoeg waren voor een uiteindelijk beslissend antwoord, dus ook niet in het voordeel van Chaucer. De auteur keert daarom aan het slot van de studie terug tot een kwalitatieve evaluatie en houdt het erop dat het auteurschap van Chaucer onwaarschijnlijk is tot dat overtuigend is bewezen.

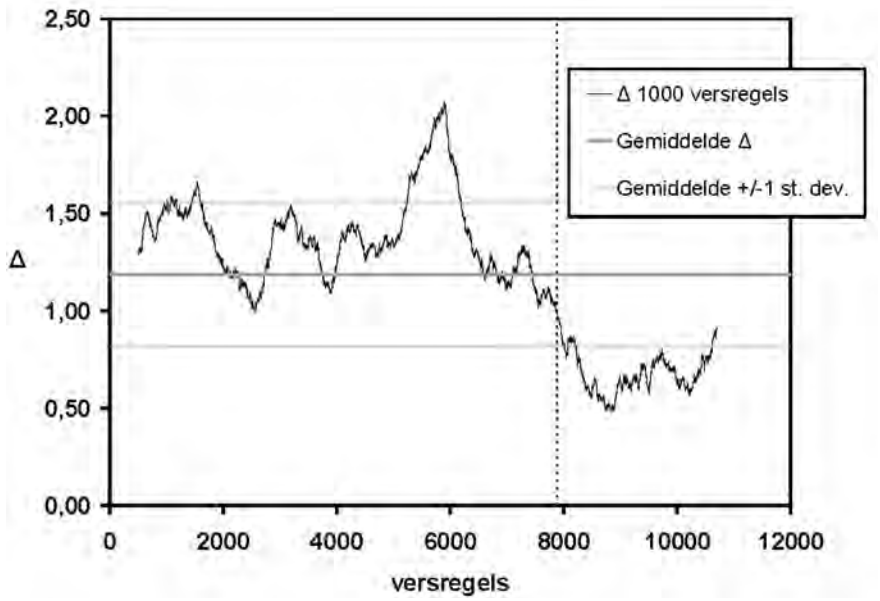
Dat neemt niet weg dat het onderzoeken van middeleeuwse teksten meer inzicht zou kunnen verschaffen in de verhouding tussen auteurs en kopiïsten en zodoende ook het onderzoek naar mogelijke verschillende invloeden in het productieproces van jongere teksten kan helpen verhelderen. Bij het ontbreken van voldoende teksten waarvan de auteur met zekerheid bekend is, is het dan zaak om te starten met een auteursprobleem dat relatief eenvoudig lijkt, en teksten betreft die tot hetzelfde genre gerekend kunnen worden en in dezelfde tijdperiode tot stand zijn gekomen. Pas als de methoden voor die teksten een aantoonbaar goed resultaat opleveren, zouden ze ook op ander middeleeuws materiaal uitgetest kunnen worden. Voor het Middelnederlands hebben we zo'n mogelijk vruchtbare casus in de Arturroman *Walewein.*

De in paragraaf 2.2 al genoemde *Walewein* is een paarsgewijs rijmende tekst van in totaal 11.202 verzen in het enige handschrift waarin het verhaal compleet is overgeleverd. Daarnaast zijn er twee fragmenten uit een ander handschrift over, met samen ongeveer 400 versregels. Het complete handschrift is geschreven door twee kopiisten, waarvan de eerste schreef tot en met regel 5783 en de tweede het manuscript uiteindelijk afsloot met de mededeling dat het in het jaar 1350 werd voltooid. In de tekst wordt expliciet vermeld dat het werk door een zekere Penninc werd bedacht, maar dat deze het helaas niet heeft afgemaakt. Pieter Vostaert vond dat jammer en heeft het werk daarom voltooid door er ongeveer 3300 verzen aan toe te voegen. Vostaerts werk volgt de structuur die Penninc heeft bedacht; hij beweegt zich dus in hetzelfde genre als Penninc had gedaan. In hoeverre een verschil in tijd invloed op de metingen kan uitoefenen is onduidelijk, omdat het niet zeker is wanneer Penninc en Vostaert hun afzonderlijke bijdragen aan de *Walewein* leverden.[4]

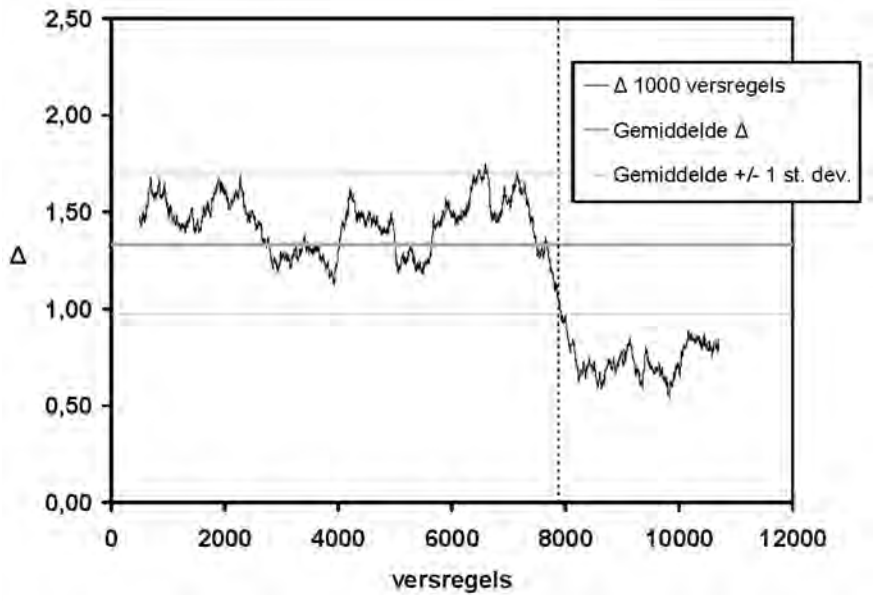
4 Het is theoretisch mogelijk dat Vostaert, de tweede auteur, geïdentificeerd moet worden met de tweede kopiist (die verantwoordelijk was voor vers 5784-11.202 van het enige complete handschrift). De specialisten houden dit over het algemeen echter voor onwaarschijnlijk, omdat zij vermoeden dat

Samen met Joris van Zundert heb ik onderzocht of moderne auteursherkenningsmethoden ons kunnen helpen de precieze plaats te vinden waar Pieter Vostaert start met zijn aanvulling op het verhaal van Penninc (zie Van Dalen-Oskam & Van Zundert 2005 en Te verschijnen). De resultaten van onze toepassing van Yule's *K*, voor het meten van lexicale rijkdom, zijn in paragraaf 2.2 van deze bijdrage al kort samengevat. Hiermee konden we geen precieze plaats aanwijzen, maar werd het wel duidelijk dat de lexicale rijkdom tot en met vers 7880 significant verschilt van die van vers 7880 tot aan het eind. Wij besloten om ook Burrows' Delta toe te passen op de tekst. Net als voor de meting van lexicale rijkdom lieten wij de meting 'door de tekst heenlopen' in de hoop dat ook bij het toepassen van deze maat een duidelijke breuk in de tekst zichtbaar zou worden die gerelateerd kon woorden aan het dubbele auteurschap. Zoals in de vorige paragraaf is uitgelegd, is Delta het resultaat van een vergelijking van steeds twee teksten of tekstverzamelingen. Omdat wij hier met maar één tekst te maken hebben, is dat als volgt opgelost. De gemiddelde z-scores van de 150 frequentste woorden uit elk tekstblok van 1000 versregels werden vergeleken met de gemiddelde z-scores van de 150 frequentste woorden uit een controletekst, die 3000 versregels bevatte ofwel uit het deel dat zeker van Penninc was ofwel uit het deel dat vrijwel zeker aan Vostaert is toe te schrijven. Het verschil tussen beide, Delta, werd afgebeeld in een grafiek. Door de 1000 versregels steeds een versregel in de tekst op te laten schuiven konden we een grafiek maken die de ontwikkelingen in de verhouding tussen een deel van de tekst in verhouding tot een controletekst afbeeldt. De Delta voor de vergelijking tussen versregel 1-1000 en een controletekst wordt afgebeeld op vers 500, die voor versregel 2-1001 op vers 501, etc. Een grafiek waarin de 150 frequentste woorden op deze manier werden vergeleken met het controledeel uit de 3000 versregels die vrijwel zeker van Vostaert zijn, leverde een grafiek op waarin het verschil tussen de twee auteurs het best zichtbaar is. Het is duidelijk dat het scharnierpunt tussen beide invloedssferen opnieuw rond vers 7880 ligt, maar een precies versnummer kan er ook hier niet uit worden afgeleid. De vraag waarmee wij begonnen kon dus niet beantwoord worden, maar er kwamen wel een aantal intrigerende nieuwe vragen naar voren uit onze verdere experimenten. De belangrijkste betroffen de verhouding tussen de twee auteurs van de *Walewein* en de twee kopiisten van het handschrift uit 1350. Deze ontstonden toen wij dezelfde grafieken produceerden voor steeds maar een gedeelte van die 150 frequentste woorden. De grafiek voor de woorden die in de rangorde van hoogstfrequent naar steeds minder frequent op plaats 1 tot en met 50 stonden (Fig. 1), bleek het auteursverschil wel aan te geven, maar een veel duidelijker breuk te vertonen op de plaats waar de tweede kopiist het overnam van de eerste kopiist (dus rond vers 5800 in plaats van rond vers 7900). In die hoogstfrequente woorden lijkt zich dus, voor deze tekst, een belangrijke lexicale bewegingsvrijheid van *kopiisten* te bevinden, en het unieke van *auteurs* iets minder nadrukkelijk aanwezig te zijn. In de grafiek die de woorden met de frequentierangorde 101-150 afbeeldde (Fig. 2) was het *auteurs*verschil het duidelijkst.

Vostaert niet lang na Penninc heeft gewerkt en zij Pennincs werk bijna een eeuw voor de datum van het handschrift plaatsen.

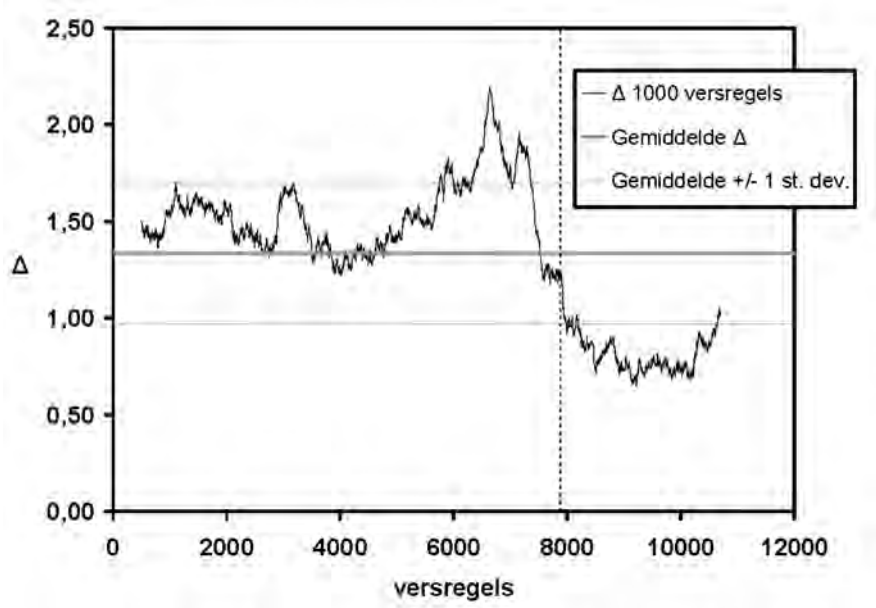


Figuur 1 De veranderingen in Delta door de tekst heen met een schuivend tekstdeel van 1000 versregels, vergeleken met vers 8000-11.000 van de Walewein. *Berekende waarden voor de woorden die in de rangorde van hoogstfrequent naar steeds minder frequent op plaats 1 tot en met 50 stonden.*



Figuur 2 De veranderingen in Delta door de tekst heen met een schuivend tekstdeel van 1000 versregels, vergeleken met vers 8000-11.000 van de Walewein. *Berekende waarden voor de woorden die in de rangorde van hoogstfrequent naar steeds minder frequent op plaats 101 tot en met 150 stonden.*

En in de tussenliggende groep, met frequentierangorde 51-100 (Fig. 3), zien we zowel het onderscheid tussen de kopiisten als tussen de auteurs naar voren komen. Bovendien lijkt er een tussengebiedje te zijn, weer rond vers 7880, van in totaal zo'n 400 verzen. Zou dat deel van de tekst, zo vragen wij ons af, een echte 'menging' van de taaleigenheden van Penninc en Vostaert vertonen, waar Vostaert het laatste stuk van het werk dat van Penninc restte bewerkte om het zo beter te laten aansluiten bij de verzen geheel van zijn eigen hand die daarna kwamen?



Figuur 3 De veranderingen in Delta door de tekst heen met een schuivend tekstdeel van 1000 versregels, vergeleken met vers 8000-11.000 van de Walewein*. Berekende waarden voor de woorden die in de rangorde van hoogstfrequent naar steeds minder frequent op plaats 51 tot en met 100 stonden.*

Deze hypotheses – want dat zijn het momenteel nog – zijn wij vanuit verschillende invalshoeken nader aan het onderzoeken en een definitief antwoord is nog niet in zicht. Wat ons allereerst interesseert is in welke delen van de woordenschat zich de belangrijkste vrijheden voor de twee *Walewein*kopiisten bevinden en in welk deel van de woordenschat de twee auteurs zich statistisch gezien het best lijken te onderscheiden. Wij hebben inmiddels nader verkend met welke woorden en woordsoorten we in die 150 frequentste woorden te maken hebben (Van Dalen-Oskam & Van Zundert 2006). Per groep van 50 hoogfrequente woorden, 1-50, 51-100, en 101-150 hebben we geïnventariseerd welk percentage verschillende woordsoorten daar van uitmaken.[5] Vervolgens hebben we berekend voor welke woordsoorten er statistisch significante verschillen tussen het voorkomen in de drie groepen te con-

5 Dit hebben we gedaan op basis van de typelijst, aangezien ons bronbestand nog niet is voorzien van woordsoortcodering op woordniveau. Wanneer die verrijking zal hebben plaatsgevonden, zullen de woordsoortfrequenties nog beter bestudeerd kunnen worden.

stateren zijn. Voor vier woordsoorten bleek er een significante ontwikkeling te zijn als we de groepen in ordening naar aflopende frequentie langslopen. De aanwezigheid van zelfstandig naamwoorden en werkwoorden bleek significant toe te nemen. In de groep frequenties waarbinnen het auteursverschil het duidelijkst optrad (101-150) is de kans dus groot dat deze twee woordsoorten daar voor een belangrijk deel verantwoordelijk voor zijn. De participatie van voornaamwoorden en voorzetsels nam echter significant af. In de 50 frequentste woorden, waarin de kopiisten zich het best van elkaar leken te onderscheiden, kunnen deze twee woordsoorten daarom een belangrijk deel van de kopiistenverschillen verklaren. Nader onderzoek in deze richting is zeer gewenst, want op deze manier kunnen we de concrete verschillen tussen de twee auteurs en de twee kopiisten van de *Walewein* formuleren in concrete verschillen in frequenties van (en dus in zekere zin bewegingsvrijheid in) of voorkeuren voor woorden, woordsoorten, concepten, en dergelijke. Pas als het onderzoek in die fasen is beland, verwachten we meer inzicht te verkrijgen in hoe auteurschap en editeur- of redacteurschap in de praktijk *van deze tekst* lijkt te werken. En dan is het tijd om de verkregen resultaten te testen op ander materiaal en zo de methode verder te ontwikkelen.

Een voorzichtige test van de toepassing van Delta op andere Middelnederlandse Arturromans hebben we ook al gedaan (Van Dalen-Oskam 2006). Uit de meetresultaten voor *Walewein ende Keye* bleek echter duidelijk dat de methode nog niet ver genoeg ontwikkeld is om op teksten met een complex ontstaans- en overleveringsproces losgelaten te worden. Met name de invloed van de lengte van de te analyseren tekst dient nader onderzocht te worden – het spreekt niet vanzelf dat de groepering in frequentie 1-50, 51-100 en 101-150 ook voor teksten korter dan de *Walewein* zinnig resultaat oplevert. Sterker nog: we weten dat ook niet zeker voor de *Walewein* zelf. Wat we op dit punt in ons onderzoek nodig hebben is de mogelijkheid om het doen van een grote hoeveelheid metingen te automatiseren, zodat de computer ons kan helpen om op basis van een combinatie van alle mogelijke parameters (grootte van het schuivende tekstblok, grootte van de basistekst waarmee wordt vergeleken, en de reeks uit de frequentierangorde) te bepalen bij welke combinatie de resultaten voor de *Walewein* het significantst zijn en wat er verandert als we de tekst kunstmatig verkleinen. We verwachten dat we hier pas mee aan de slag kunnen als we gebruik gaan maken van grid-computing (het uitbesteden van het rekenwerk aan een groot netwerk van aan elkaar gekoppelde computers die zo een enorme rekenkracht beschikbaar stellen), waardoor metingen die anders een paar dagen in beslag zouden nemen in aanzienlijk kortere tijd gedaan kunnen worden.

4 De toekomst

Het toepassen van kwantitatieve methoden om stijl te analyseren is nog niet heel gewoon. Hopelijk zijn de geschetste voorbeelden inspirerend voor onderzoekers die zich nog niet eerder op dit terrein hebben begeven. Ook in andere taalgebieden worden er nog weinig heel concrete literairhistorische vragen gesteld. Vermeldenswaard is een stilistische studie van Friedrich Dimpel naar Hendrik van Veldekes *Eneas*. Dimpel heeft onderzocht of de bewering van Veldeke dat hij een aantal jaren niet kon werken aan zijn roman omdat zijn handschrift hem was ont-

stolen ondersteund wordt door een stilistische vergelijking van het tekstdeel dat voor de roof tot stand kwam en het tekstdeel dat Veldeke na teruggave van het handschrift zou hebben toegevoegd (Dimpel 2006). Een kwantitatieve analyse van tien verschillende stilistische elementen levert geen significante verschillen in stijl op en zal tot een hernieuwd onderzoek naar de status van de betreffende mededeling in de tekst zelf moeten voeren (Dimpel 2006: 100). Andere interessante aanzetten tot een kwantitatieve benadering van Nederlandse literaire teksten zijn die van Boot & Stronks (2003), waarin het beoogde publiek van een bundel van Jacob Cats wordt onderzocht op basis van stijlkenmerken als aansprekingen in de tekst, en van Van Leuvensteijn & Wattel (2002), die het stijlkenmerk enjambement in toneelwerk van Vondel kwantitatief analyseren.

De beschreven studies leveren allemaal stappen in de richting van een beter begrip van de onderzochte literaire werken en hun ontstaan, opbouw of context. Definitieve antwoorden zijn zeldzaam, maar nieuwe vragen zijn er in grote hoeveelheden. Om die nieuwe vragen te helpen beantwoorden zijn er verschillende zaken nodig. Uiteraard zijn er digitale teksten nodig, maar dat is op zich bij lange na nog niet voldoende. Die teksten zouden bij voorkeur ook al gecodeerd moeten zijn met woordsoortaanduidingen en zijn gelemmatiseerd, zodat woordenschatonderzoek als beginfase van stijlonderzoek gemakkelijker kan worden uitgevoerd en sneller kan leiden tot diepgaandere vervolganalyses. Verder is het nodig dat de programma's die onderzoekers voor hun eigen werk (laten) schrijven beschikbaar komen voor andere onderzoekers, zodat metingen kunnen worden herhaald en gecontroleerd door latere onderzoekers. En voor het automatiseren van zeer omvangrijke metingen is behoefte aan grid computing.

Meer inhoudelijk is de volgende stap die moet worden genomen een nader onderzoek van de verschillende frequentiestrata in de woordenschat in het algemeen en in literaire teksten in het bijzonder. Onderzoekers die zich bezighouden met lexicografisch en lexicologisch werk vanuit een statistische achtergrond zullen hieraan een belangrijke bijdrage kunnen leveren. Waar het de analyse van stijl- en auteursverschillen betreft, heeft John Burrows zeer recent weer het initiatief genomen met zijn Zeta- en Iota-tests. Het wachten is op nadere uitwerking van zijn nieuwe maten aan de hand van uitvoerige tests op tekstmateriaal in verschillende talen. De doelstellingen van de kwantitatieve methoden zijn zeer divers: de wens om concrete stijlverschillen in teksten en bij auteurs, oeuvres of genres te krijgen, om meer inzicht te verkrijgen in relaties tussen teksten, genres, auteurs en zo verder. En zeker ook om een vraag te helpen beantwoorden die men zich tot op heden nog maar zelden heeft durven stellen: de wens om aanknopingspunten te verkrijgen voor een beter begrip van de glijdende schaal tussen auteur en kopiist/redacteur.

Bibliografie

Anbeek & Verhagen 2001 – Ton Anbeek & Arie Verhagen: 'Over stijl'. In: *Neerlandistiek.nl* 01.01, te vinden op www.neerlandistiek.nl/01.01/.

Barr 2001 – George K. Barr: 'Graphical analysis of the sentence length distribution curve and non-rational components'. In: *Literary and Linguistic Computing* 16 (2001), p. 375-388.

Van den Berg 1983 – E. van den Berg: *Middelnederlandse versbouw en syntaxis. Ontwikkelingen in de versifikatie van verhalende poëzie ca. 1200-ca. 1400.* Utrecht, 1983.

Binongo & Smith 1999 – José Nilo G. Binongo & M.W.A. Smith: 'The application of Principal Component Analysis to stylometry'. In: *Literary and Linguistic Computing* 14 (1999), p. 445-465.
Boot & Stronks 2003 – Peter Boot & Els Stronks: 'Ingrediënten van een sucesformule. Digitaal onderzoek naar Cats' Sinne- en minnebeelden'. In: *Nederlandse letterkunde* 8 (2003), p. 24-40.
Burrows 1987 – J.F. Burrows: *Computation into criticism. A study of Jane Austen's novels and an experiment in method*. Oxford, 1987.
Burrows 2002 – John Burrows: '"Delta": a measure of stylistic difference and a guide to likely authorship'. In: *Literary and Linguistic Computing* 17 (2002), p. 267-287.
Burrows 2003 – John Burrows: 'Questions of authorship: attribution and beyond'. In: *Computers and the Humanities* 37 (2003), p. 5-32.
Burrows 2005 – John Burrows: 'Who wrote *Shamela*? Verifying the authorship of a parodic text'. In: *Literary and Linguistic Computing* 20 (2005), p. 437-450.
Burrows 2006 – John Burrows: 'All the way through: testing for authorship in different frequency strata'. Advanced published in *Literary and Linguistic Computing* 2006, 21 pagina's.
Van Dalen-Oskam & Van Zundert 2005 – Karina van Dalen-Oskam m.m.v. Joris van Zundert: 'De list van het lexicon. Auteursonderscheiding met behulp van computer-ondersteunde woordenschatanalyse'. In: *Nederlandse letterkunde* 10 (2005), p. 212-233.
Van Dalen-Oskam 2006 – Karina van Dalen-Oskam: 'Textual relationships and author differentiation'. In: *Proceedings of the 78th meeting of the English Literary Society of Japan*, Nagoya, 2006, p. 164-166. Ook beschikbaar op www.huygensinstituut.knaw.nl >medewerkers >Dr K.H. van Dalen-Oskam >publicatielijst.
Van Dalen-Oskam & Van Zundert 2006 – Karina van Dalen-Oskam & Joris van Zundert: 'The quest for uniqueness. Author and copyist distinction in Middle Dutch Arthurian romances based on computer-assisted lexicon analysis'. Te verschijnen in: *Proceedings of the Third International Conference on Historical Lexicography and Lexicology (ICHLL 2006)*, 21-23 juni 2006, Leiden.
Van Dalen-Oskam & Van Zundert (Te verschijnen) – Karina van Dalen-Oskam & Joris van Zundert: 'Delta for Middle Dutch: Author and copyist distinction in *Walewein*'. Te verschijnen in: *Literary and Linguistic Computing*.
Dimpel 2006 – Friedrich Michael Dimpel: 'Der Verlust der "Eneas"-Handschrift als Fiktion – eine computergestützte, textstatistische Untersuchung'. In: *Amsterdamer Beiträge zur älteren Germanistik* 61 (2006), p. 87-102.
Dorleijn 1995 – G.J. Dorleijn: 'De periodiserende computer of stilistiek als instrument voor periodisering; een aanzet'. In: *De nieuwe taalgids* 88 (1995), p. 490-506.
Dunning 1993 – Ted Dunning: 'Accurate methods for the statistics of surprise and coincidence'. In: *Computational Linguistics* 19 (1993), p. 61-74.
Van Eck & Streng 1997 – Caroline van Eck & Toos Streng: 'Inleiding'. In: Caroline van Eck, Marijke Spies, Toos Streng (red.): *Een kwestie van stijl. Opvattingen over stijl in kunst en literatuur*. Historisch Seminarium van de Universiteit van Amsterdam, 1997. Amsterdamse historische reeks, kleine serie, deel 34, p. 7-18.
Forsyth 1999 – Richard S. Forsyth: 'Stylochronometry with substrings, or: a poet young and old'. In: *Literary and Linguistic Computing* 14 (1999), p. 467-477.
Holmes 1994 – David I. Holmes: 'Authorship attribution'. In: *Computers and the Humanities* 28 (1994), p. 87-106.
Hoover 2003 – David L. Hoover: 'Another perspective on vocabulary richness'. In: *Computers and the Humanities* 37 (2003), p. 151-178.
Hoover 2004a – David L. Hoover: 'Testing Burrows's Delta'. In: *Literary and Linguistic Computing* 19 (2004), p. 453-475.
Hoover 2004b – David L. Hoover: 'Delta Prime?'. In: *Literary and Linguistic Computing* 19 (2004), p. 477-495.
Kilgarriff 1996 – Adam Kilgarriff: 'Which words are particularly characteristic of a text? A survey of statistical methods'. University of Brighton, ITRI-96-08, 1996, www.kilgarriff.co.uk/Publications/1996-K-AISB.pdf, actief 28 augustus 2006. (Ook verschenen in *Proceedings AISB workshop on language engineering for document analysis and recognition*, Sussex university, April 1996, p. 33-40.)
Kuiper 1989 – W.Th.J.M. Kuiper: *Die riddere metten witten scilde. Oorsprong, overlevering en auteurschap van de Middelnederlandse Ferguut, gevolgd door een diplomatische editie en een diplomatisch glossarium*. Diss. Amsterdam, 1989.
Lancashire 2004 – Ian Lancashire: 'Cognitive stylistics and the literary imagination'. In: Susan

Schreibman, Ray Siemens and John Unsworth (ed.): *A companion to digital humanities*, Oxford, 2004, p. 397-414.
Leech & Short 1981 – Geoffrey N. Leech & Michael H. Short: *Style in fiction. A linguistic introduction to English fictional prose*. London / New York, 2003 (1981).
Van Leuvensteijn & Wattel 2002 – Arjan van Leuvensteijn & Evert Wattel: 'Een statistische methode voor stijlonderzoek. Vorm – inhoud correspondenties in Vondels *Jeptha*?' In: *Neerlandistiek.nl* 02.05, te vinden op www.neerlandistiek.nl/02.05/.
Love 2002 – Harold Love: *Attributing authorship: an introduction*. Cambridge, 2002.
Miranda-García & Calle-Martín 2005 – A. Miranda-García & J. Calle-Martín: 'Yule's Characterictis *K* revisited'. In: *Language resources and evaluation* 39 (2005), p. 287-294.
Rand Schmidt 1993 – Kari Anne Rand Schmidt: *The authorship of The Equatorie of the Planetis*. Cambridge, 1993.
Rybicki 2006 – Jan Rybicki: 'Burrowing into translation: character idiolects in Henryk Sienkiewicz's Trilogy and its two English translations'. In: *Literary and Linguistic Computing* 21 (2006), p. 91-103.
Stewart 2003 – Larry L. Stewart: 'Charles Brockden Brown: quantitative analysis and literary interpretation'. In: *Literary and Linguistic Computing* 18 (2003), p. 129-138.
Tweedie & Baayen 1998 – Fiona J. Tweedie & R. Harald Baayen: 'How variable may a constant be? Measures of lexical richness in perspective'. In: *Computers and the Humanities* 32 (1998), p. 323-352.
Tweedie, Singh & Holmes 1996 – F.J. Tweedie, S. Singh & D.I. Holmes: 'Neural network applications in stylometry: the *Federalist Papers*'. In: *Computers and the Humanities* 30 (1996), p. 1-10.

Adres van de auteur

Huygens Instituut, Postbus 90754, NL-2509 LT Den Haag,
karina.van.dalen@huygensinstituut.knaw.nl