PREPRINT artikel voor Madoc 2017, afl. 4, in de serie MAGNA DATA

Karina van Dalen-Oskam

Big data, magna data
Nieuwe mogelijkheden voor onderzoek naar teksten en handschriften

Big data is een term die je overal hoort en leest. Maar wat is (of zijn?) het eigenlijk en wat heb je er aan als mediëvist? 'In de Middeleeuwen zijn de bronnen zo schaars, daar zul je weinig big data-onderzoek zien', zei Frits van Oostrom onlangs in een interview naar aanleiding van zijn nieuwe boek *Nobel streven. Het onwaarschijnlijke maar waargebeurde verhaal van ridder Jan van Brederode* dat in oktober 2017 verscheen.[1] Jan van Brederode is de auteur van de Middelnederlandse *Des coninx summe*, de sublieme vertaling van de Franse *La somme le roi*, zo is definitief duidelijk geworden uit dit boek. Van Oostrom vervolgde in het interview met Bas Blokker voor *NRC*: 'Bij mij zeggen ze vaak "N=1". Dat is waar. Maar iemand zei me laatst: "Het meervoud van anekdote = data". Daar werd ik enigszins door getroost.' Met N=1 bedoelt Van Oostrom zoveel als: deze stelling is gebaseerd op een experiment waaraan maar één proefpersoon deelnam. En het is inderdaad aantrekkelijk om een hele stapel anekdotes samen te zien als beslist veel meer dan N=1. Maar de vraag is: hoe groot moet N zijn om 'big data' genoemd te worden? Aan wat voor soort mediëvistische data kunnen we denken? En zijn die inderdaad zo schaars als Van Oostrom beweert?

Big data zijn te omschrijven als gegevens die zo omvangrijk zijn dat ze heel lastig te onderzoeken zijn en logistiek moeilijk te benaderen, schrijft Christine L. Borgman in *Big data, little data, no data. Scholarship in the networked world*.[2] Hoe *magna* (groot) *big data* zijn, is dus relatief en houdt ook verband met de technologie die op dat moment beschikbaar is. Dat er digitaal steeds meer data tot onze beschikking komen is duidelijk, maar voor de mediëvist is er nog steeds veel meer niet dan wel in die vorm beschikbaar. Van Oostrom:

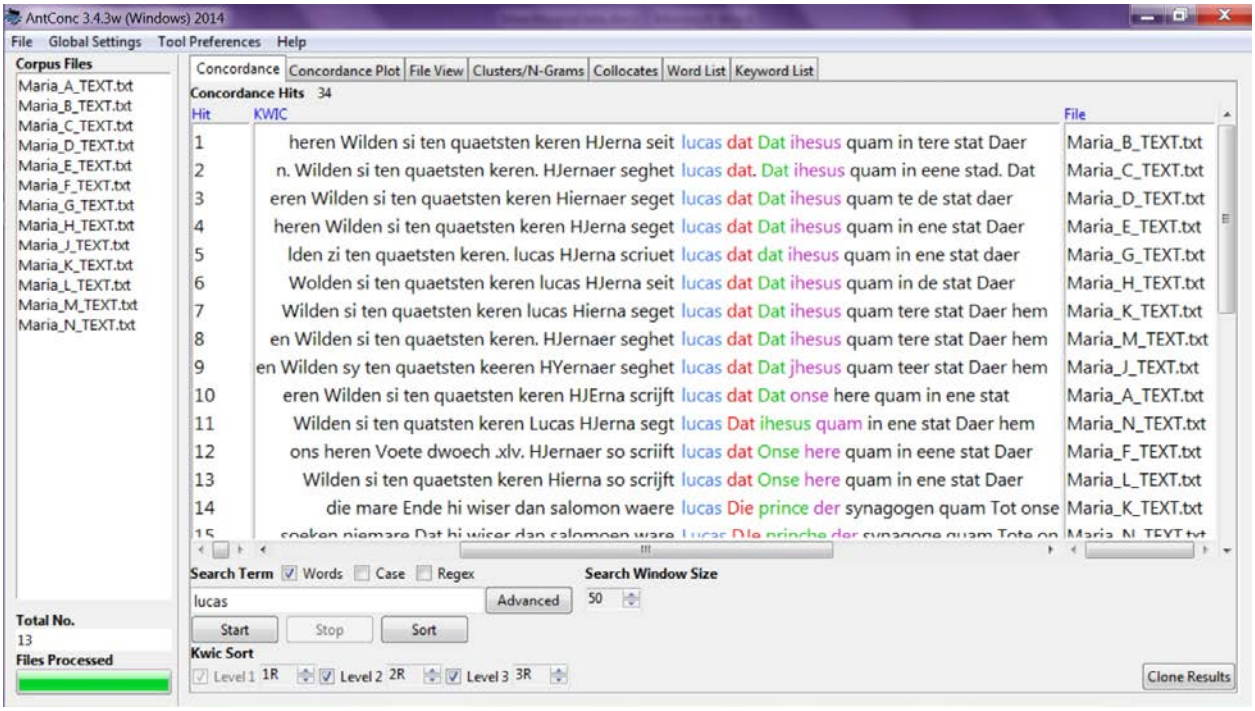
> Als je niet oppast (…) raak je verliefd op de apparatuur als nieuw altaar. Maar ik dacht aanvankelijk dat de computer ons vooral zou helpen bij de tekstverwerking en meer niet. Dat was ook een grote vergissing.[3]

Ik zal twee aspecten van mediëvistisch onderzoek bespreken in een *big data*-context, en wil daarmee een indruk geven van de huidige en toekomstige digitale mogelijkheden. Ze gaan over handschriften en schrijfhanden en over onderzoek naar verschillende versies van dezelfde tekst. Ik begin met het laatste.

Eerst wat *science fiction*

Als het over *big data* en teksten gaat, verwijs ik zelf graag naar een omschrijving van literatuurwetenschapper/statisticus Allen Riddell, die een tekstverzameling als *big data* beschouwt als die bestaat uit meer teksten dan een individuele onderzoeker normaal gesproken kan verwerken in een jaar van toegewijd lezen.[4] En tel maar na: daar zit je al gauw aan. Goed lezen kost tijd. Het is fantastisch dat digitaal beschikbare teksten met een zoekfunctie zijn uitgerust waardoor je veel sneller iets terug kunt vinden dan ooit tevoren en dat er programma's

zijn die het mogelijk maken om met enkele simpele handelingen alle vindplaatsen van hetzelfde woord in een tekst of tekstverzameling in hun context op je scherm te toveren (afb. 1) zonder dat je alles hoeft te herlezen, ook al is dat maar diagonaal. Want bij diagonaal lezen zie je immers ook snel iets over het hoofd.



Afb. 1 Concordantie van het woord ‘lucas’ op alfabetische volgorde van het volgende woord gesorteerd. Tekst: passage uit het Evangeliedeel van de *Rijmbijbel* van Jacob van Maerlant in alle beschikbare handschriften (exclusief fragmenten). Software: AntConc.

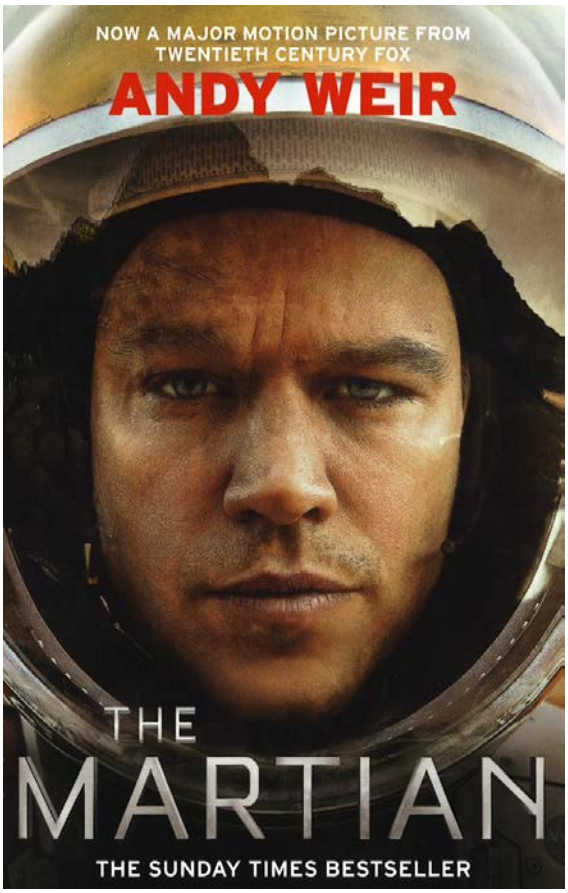
Bij het produceren van teksten zijn we tegenwoordig in zekere zin weer terug in de Middeleeuwen. Elke kopie kan net even anders zijn. We schrijven onze teksten in een tekstverwerker, verspreiden voorlopige versies onder proeflezers, en pas na heel veel schaven en schrappen komt er wellicht iets op papier en steeds vaker ook als e-book op de markt. Maar die eerdere versies zwerven dus nog ergens rond in ons digitale universum. Hoe vreemd dat ook klinkt: dat doet sterk denken aan de situatie in de Middeleeuwen. Teksten werden overgeschreven door verschillende personen voor verschillende personen of instellingen, en bij elk van die verveelvoudigingen kunnen net even andere doelstellingen en achtergronden een rol hebben gespeeld. En die leiden steevast tot unieke kopieën van wat we meestal nog wel kunnen beschouwen als ‘dezelfde’ tekst. De verschillen tussen al die versies kunnen ons heel veel leren over de tekst, de kopiisten, en de contexten waarin de tekst werd gekopieerd of moest gaan functioneren. Ook leren ze ons meer over de taal uit de betreffende tijdperiode of over lokale spellingconventies.

In moderne edities van middeleeuwse teksten worden regelmatig varianten vermeld, lezingen in andere handschriften. Het ligt dan aan het doel van de editeur welke varianten op die manier worden weergegeven, want het zijn er meestal veel te veel om allemaal op te nemen of zelfs maar te onderzoeken. Welke varianten kunnen weggelaten worden? Voor niet-taalkundigen zijn spellingvarianten vaak absoluut niet interessant, terwijl die voor de mediëvistiek als geheel wel degelijk van belang zijn. Indirect kunnen ze bijdragen aan meer kennis over de taal en dus bijdragen aan een verbetering van bestaande methoden voor lokalisering en datering van handschriften.[5] Maar de editeur van een tekst moet terughoudend zijn in het opnemen van varianten – het kost niet alleen te veel ruimte in een gedrukt boek, maar ook veel te veel tijd als de editeur niet jarenlang met het publiceren van het eindresultaat wil wachten. Nu tekstedities steeds vaker digitaal ter beschikking worden gesteld, worden langzamerhand andere keuzes gemaakt in het editieproces. Om een voorbeeld te geven: met afbeeldingen van pagina's uit de handschriften naast de transcriptie is het niet altijd meer nodig om bepaalde observaties ook in goed lopende zinnen te omschrijven. Een annotatie met een categorie-aanduiding ('Beschadiging') kan voldoende zijn, sterker nog: is zelfs beter voor de terugvindbaarheid in digitale context.

En we kunnen ook op een andere manier met varianten omgaan, in de geest van de Nieuwe Filologie.[6] Niet meer één basistekst kiezen en die naar eigen inzichten 'verbeteren' op grond van varianten, en niet meer kiezen welke varianten je in noten wilt opnemen omdat je nu eenmaal alleen 'belangrijke' dingen kunt noteren. Nee, als we alle versies volgens dezelfde strakke richtlijnen transcriberen kunnen we de computer inzetten om de verschillen tussen de tekstversies zichtbaar te maken, bijvoorbeeld met het speciaal hiervoor ontwikkelde programma CollateX .[7] Dan gebruiken we de computer dus als tool voor variantenanalyse. Dan ben je als onderzoeker niet meer afhankelijk van wat de editeur belangrijk genoeg vond om te annoteren – de computer maakt alle verschillen zichtbaar. Weg is het risico dat je conclusies trekt op basis van ontbrekende informatie die er best was, maar niet op waarde werd geschat binnen een ander onderzoeksperspectief, dat van de editeur.

Stel: we hebben een transcriptie van alle bekende handschriften en/of drukken van een bepaalde tekst, en we laten de computer een vergelijking maken. Welke versies lijken veel op andere versies, en welke wijken af? Wat kan dat zeggen over de ontstaansgeschiedenis van een handschrift, van een tekstversie, en over de belangstelling voor onderdelen van een tekst in verschillende omgevingen? Om de relevantie van die vragen toe te lichten maak ik een uitstapje naar fictie die tot stand is gekomen in de hedendaagse digitale wereld.

Het gaat om een roman uit het Science Fiction-genre: *The Martian*, geschreven door Andy Weir (in het Nederlands vertaald door Henk Moerdijk onder de titel *Mars*, maar in 2015 opnieuw uitgebracht onder de filmtitel *The Martian*). Astronaut en plantkundige Mark Watney is door onvoorziene omstandigheden alleen achtergebleven op Mars. We lezen in zijn logboek hoe hij zich weet te redden. Weir wilde het verhaal zo realistisch mogelijk maken, en dat zien we terug in de stijl van zijn roman. Die roman is in verschillende versies beschikbaar. De auteur had op basis van eerdere ervaringen namelijk weinig vertrouwen dat een uitgever zijn boek wilde publiceren en besloot in 2011 om *The Martian* gratis in afleveringen op zijn eigen website te plaatsen. Het verhaal werd zo populair dat er toch een uitgever geïnteresseerd raakte. Zo werd *The Martian* alsnog Weirs debuutroman, bij een gevestigde uitgeverij.



Afb. 2 Omslag van de roman *The Martian* van Andy Weir, in een druk uit 2015. De foto komt uit de film, waarin Matt Damon de rol van Mark Watney speelt.

Maar de uitgeverij had wel eisen. Ook al was de roman al online gepubliceerd, voor de boekdruk onderging de tekst een minutieuze redactieslag. De Digital Humanities-onderzoekers Erik Ketzan en Christof Schöch zagen een aantal interessante wijzigingen en wilden beter bekijken om wat voor veranderingen het nu ging.[8] Ze gebruikten verschillende computerprogramma's om de twee versies van Weirs roman met elkaar te vergelijken en vonden meer dan vijfduizend verschillen. De meeste waren op zich klein. Meer dan de helft, 2863 in totaal, waren tekstuele aanpassingen: correcties van spelfouten en van foutief gebruik van koppeltekens of hoofdletters, afkortingen die voluit werden geschreven. Veel getallen in arabische cijfers werden omgezet in woorden (8 werd *eight*, bijvoorbeeld). Veel scheldwoorden werden vervangen door mildere varianten. En aan het slot is het verhaal iets ingrijpender gewijzigd, waar ik niet verder op in zal gaan (dat voorkomt de noodzaak van een *spoiler alert*).

Ketzan en Schöch schetsen in hun verslag hoe deze op zich weinig relevant lijkende veranderingen (behalve die laatste dan) wel degelijk hebben gezorgd voor 'thematic and stylistic shifts' in de roman, wijzigingen in de thematiek en in de stijl. Het logboek van Watney is minder realistisch geworden door de correctie van zijn schrijffouten en door het omzetten van arabische cijfers in uitgeschreven getallen – je zou kunnen zeggen dat de *nerd* is weggeredigeerd. Het doel van de redactie, zo stellen Ketzan en Schöch, lijkt te zijn geweest om de hoofdpersoon

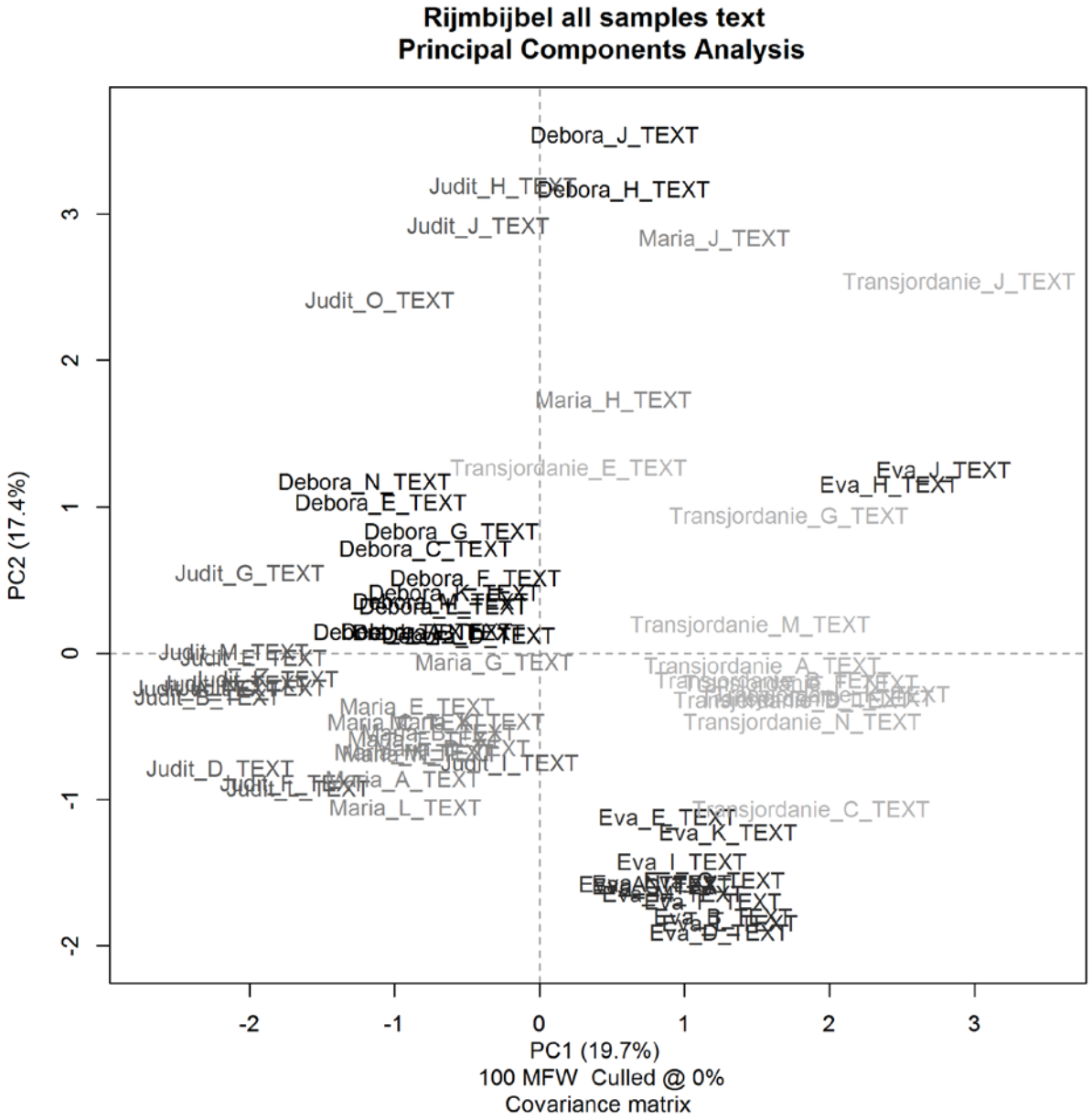
herkenbaarder en aardiger te maken, en op die manier de roman toegankelijker te maken voor het nog grotere publiek.

Van Mars naar Maerlant

In eerder onderzoek heb ik iets vergelijkbaars gedaan voor de Middelnederlandse *Rijmbijbel* (1271) van Jacob van Maerlant. Het is een tekst van zo'n 35.000 versregels, maar voorhanden in vijftien omvangrijke handschriften (die niet allemaal de complete tekst bevatten). De *Rijmbijbel* vertelt voor een breed publiek en in paarsgewijs rijmende verzen de meest aansprekende verhalen uit de Bijbel: narratieve delen uit het Oude Testament en de Apocriefen, het leven en de kruisdood van Jezus uit de Evangeliën (tot hier toe gebaseerd op de *Historia Scholastica* van Petrus Comestor), gevolgd door het beleg en de vernietiging van Jeruzalem door de Romeinen (sterk samengevat op basis van het werk van Flavius Josephus). Ik wilde weten hoeveel vrijheden de kopiisten van de *Rijmbijbel* zichzelf veroorloofden en wat dat kon zeggen over het functioneren van elk van die handschriften in hun eigen tijd en omgeving.

Ik koos vijf tekstpassages die alle over vrouwen gingen, drie uit het (grootste) Oudtestamentische deel van de *Rijmbijbel* (over Eva, Debora, en – apocrief – Judit), een uit de evangeliënharmonie (over de drie Maria's) en een uit het Josephus deel (over Maria uit Transjordanië). De passages van elk ongeveer tweehonderd versregels heb ik vervolgens diplomatisch getranscribeerd uit alle handschriften, gebruikmakend van microfiches en microfilms. Dat leverde in totaal zeventig korte digitale tekstbestanden op.

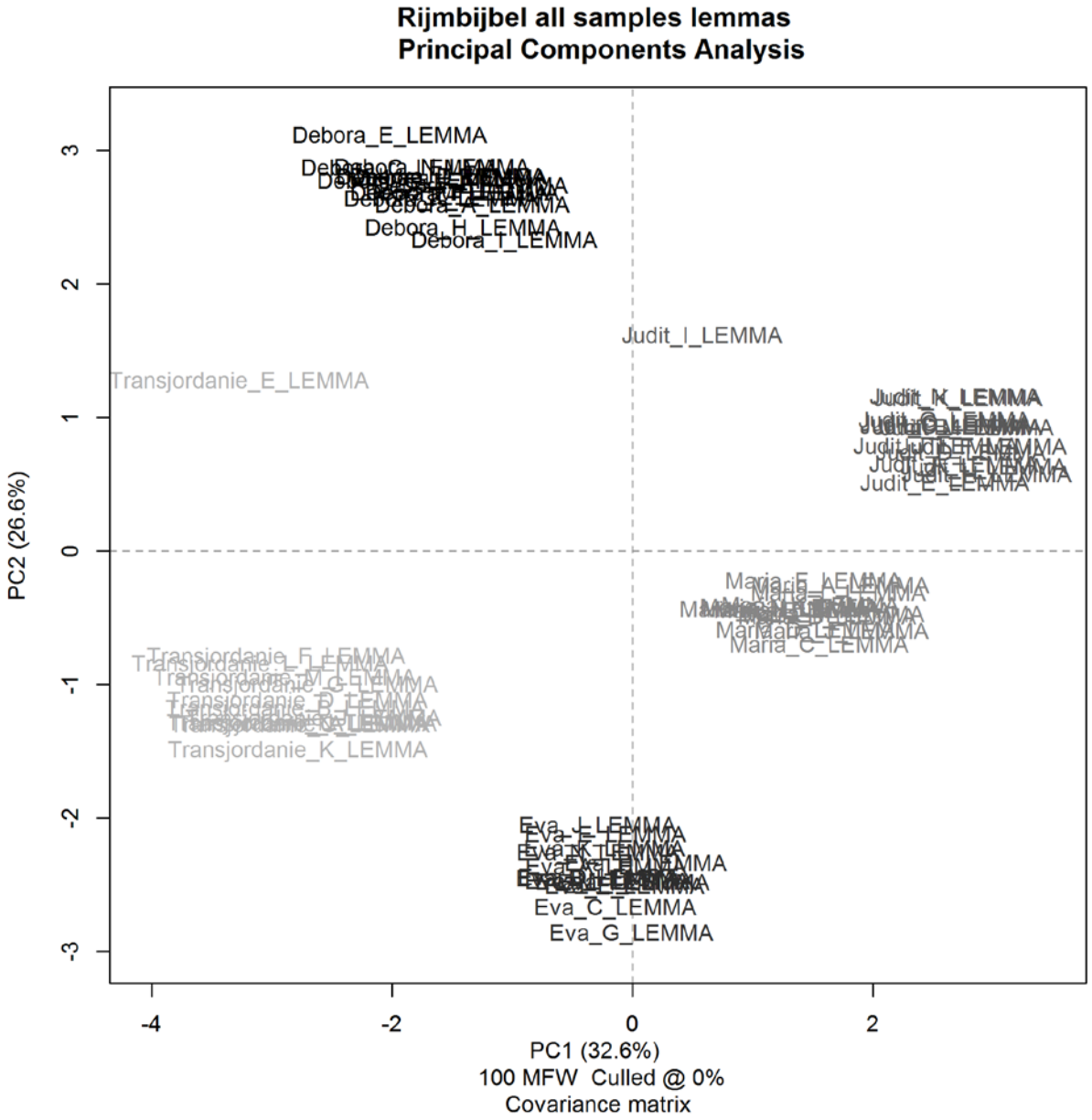
Met een softwareprogramma waarmee ik de woordenschat in teksten gemakkelijk kan doorrekenen, heb ik vervolgens al die *samples* met elkaar vergeleken - het *Stylo Package for R*. In afb. 3 staat een visualisatie van de verhouding tussen het vocabulaire van alle zeventig *samples* op basis van de spelling zoals die daadwerkelijk in de vijftien handschriften staat. Elk *sample* heeft een naam die de inhoud aanduidt, gevolgd door een letter voor elk handschrift (A tot en met O). Het werkt ongeveer zo: de software heeft alle woordvoorkomens in al die *samples* bij elkaar opgeteld en heeft berekend wat de relatieve frequentie in dit corpus is. Vervolgens wordt voor elk van de zeventig *samples* voor de woorden die in dat specifieke *sample* voorkomen gekeken wat de relatieve frequentie in dat *sample* is en die relatieve frequentie wordt vergeleken met die in het totale corpus van zeventig *samples*. De verschillen voor alle woorden worden bij elkaar opgeteld, en dat leidt voor elk *sample* tot een afstandsaanduiding ten opzichte van het gemiddelde. In afb. 3 is dat gevisualiseerd. Het kruispunt van de stippellijnen is te zien als het gemiddelde. We zien in de figuur dat een aantal versies van dezelfde tekstpassages bij elkaar clusteren, maar dat andere wat verder weg staan. Mogelijk zien we hier invloed van dialect op de plaats van met name de handschriften H en J bovenin de figuur.



Afb. 3 Alle *Rijmbijbel-samples* vergeleken op basis van de diplomatische editie. Software: Stylo Package for R. Principal components analyse (covariance) van de honderd meest frequente woorden.

Ik heb alle woorden in de teksten ook verrijkt met extra informatie, te weten een modern Nederlands lemma en een code die de woordsoort aangeeft. Twee voorbeelden: *die/des/der/den* hebben allemaal als lemma DIE en als code ‘aanwijzend voornaamwoord’, en de werkwoordsvormen *sijn/waren/es/is/ben* etc. allemaal het lemma ZIJN en een code die aangeeft dat het een werkwoord betreft. Op deze manier zijn spellingsverschillen dus verdwenen en kunnen we meer vergelijken op inhoudelijke aspecten. Dezelfde visualisatie op basis van alleen de lemma’s zien we in afb. 4. Het is duidelijk dat de spreiding nu een stuk minder wild is. Dat

geeft ons nog steeds geen heel duidelijke antwoorden op vragen naar het functioneren van elk manuscript in de eigen tijd en context, maar de visualisatie is een middel om uit te vinden waar zich mogelijk interessante zaken voordoen die het verdienen om als eerste onderzocht te worden. Deze figuur helpt ons zo om per tekstdeel te kijken waar meer inhoudelijke afwijkingen te vinden zijn. Voor het Judit-deel is dus handschrift I interessant om nader te onderzoeken, en voor het Transjordanië-deel handschrift E.



Afb. 4 Alle *Rijmbijbel-samples* vergeleken op basis van alleen de lemma's van de woorden. Software: Stylo Package for R. Principal components analyse (covariance) van de honderd meest frequente woorden.

Dergelijke visualisaties geven dus geen concreet antwoord op onderzoeksvragen, maar functioneren als tussenstap. Ze geven aanknopingspunten voor vervolgonderzoek. Hoe zit het met die 'outliers', wat is daar aan de hand? Ik heb daar inmiddels iets meer over gepubliceerd, maar nog lang niet alle vervolgmogelijkheden zijn uitgediept .[9] Op mijn verlanglijstje staan bijvoorbeeld nog vragen als: zien we een wijziging in de manier waarop versies van elkaar afwijken door de hele tekst heen? Ofwel: zit er minder variatie in het deel over Eva, nog in het allereerste deel van de *Rijmbijbel*, dan in de latere delen? Gaven kopiisten zichzelf meer ruimte als het apocriefe verhalen betrof, zoals ik voor het deel over Judit vermoed? En zien we een ander patroon in het Evangeliedeel, en zo ja, zou dat samen kunnen hangen met aparte circulatie van het Nieuwtestamentische deel van de *Rijmbijbel*? Hiervoor zou ik dus door de tekst heen willen wandelen en willen kunnen visualiseren hoe de verhoudingen veranderen. Daarvoor zijn nog veel meer diplomatische transcripties nodig, en verrijking van die transcripties. En ook de software moet nog worden uitgebreid. Op dit moment kunnen visualisaties zoals in afb. 3 en 4 nog niet door een tekst heen wandelen. Er zijn echter meer onderzoekers die iets vergelijkbaars willen, en daarom wordt er momenteel gewerkt aan een aanpassing door de makers ervan.

Zo werkt het over het algemeen in de Digital Humanities: meer data en nieuwe visualisaties van de data leiden tot nieuwe inzichten maar onmiddellijk ook tot nieuwe vragen en nieuwe software-wensen. Deze manier van samenwerken tussen mensen met overlappende en complementaire expertise is heel inspirerend. Met elkaar komen we steeds een stapje verder, en die tussenstappen die we maken zijn herhaalbaar en controleerbaar, waardoor ze soms net wat meer stevigheid aan het onderzoek kunnen geven dan als we uitsluitend op onze eigen lezende ogen zouden vertrouwen. Ik zie ernaar uit om mijn *Rijmbijbel*-dataset nog verder te verkennen zodra de nieuwe mogelijkheden beschikbaar komen. Die dataset is op het eerste gezicht misschien eerder *little data*, maar doordat het al veel meer gegevens zijn dan door een onderzoeker gemakkelijk overzien kan worden, zijn ze wel degelijk ook vanuit een *big/magna data*-perspectief te beschouwen. En er zijn veel meer teksten aan toe te voegen, op weg naar meer en meer data.

Terug naar *Des coninx summe*

Ik stel me voor dat een vergelijkbare benadering van de overlevering van Jan van Brederodes *Des coninx summe* interessant kan zijn. Van Oostrom vermeldt dat er twaalf handschriften zijn overgeleverd en zeker tien drukken. Het laatste deel van de tekst is van een andere auteur dan van Jan van Brederode. Van Oostrom signaleert dat het een duidelijk andere stijl heeft: vlakker en neutraler, zakelijker, het vuur ontbreekt een beetje. Wie dit deel vertaalde is onduidelijk. Het is alleen in de gedrukte versies aangetroffen, zodat het waarschijnlijk niet door een tijdgenoot is toegevoegd.[10]

Wat zou een vergelijking van delen van de versies van *Des coninx summe* kunnen opleveren? Ik zou zelf benieuwd zijn om vast te stellen of een analyse van het vocabulaire van het laatste deel inderdaad een ander patroon laat zien dan het deel van Jan van Brederode. Van Oostrom karakteriseert Brederode's tekst onder meer met de volgende metafoor (die alleen begrepen zal worden door gebruikers van oudere versies van een bepaald merk tekstverwerker): 'Het onderwaterscherm van *Des coninx summe* vertoont een negatieve visie op de wereld'.[11] Kunnen we dat met hulp van software zichtbaar maken? En kunnen we misschien zelfs op zoek gaan naar de mogelijke auteur van de toevoeging? Dat is waar de software die ik voor de tekstvergelijking heb gebruikt, het *Stylo Package for R*, in eerste instantie voor ontwikkeld is en

waar al veel ervaring mee is in de Digital Humanities, in het Nederlandstalige gebied ook voor middeleeuwse teksten.[12]

Maar er zal in de nabije toekomst nog heel veel meer kunnen. Zo kom ik op het tweede aspect dat ik in deze bijdrage (beknopt) wil bespreken. Het onderzoek dat Van Oostrom naar Jan van Brederode deed, leverde een nieuwe hypothese op over waar het bekende handschrift-Van Hulthem is neergeschreven. Dit in Brussel bewaarde handschrift met een schat aan Middelnederlandse literaire teksten blijkt namelijk volgens Van Oostrom dezelfde schrijfhand te bevatten als enkele archiefstukken die berusten in Utrecht en Den Haag. 'De bronnen liggen al sinds eeuwen voor het grijpen in de meest gerenommeerde instellingen op de twee vakgebieden. Maar alvorens neerlandici en historici daadwerkelijk over elkaars schutting kijken, is er heel wat nodig', schrijft Van Oostrom.[13] In een andere context formuleert hij het zo:

> Dit had overigens al een eeuw eerder kunnen gebeuren, want de enige vereiste was de koppeling van een beroemd letterkundig boek in Brussel aan archivalia te Den Haag en Utrecht. Maar kennelijk was daarvoor het water tussen beide disciplines te diep.[14]

Het koppelen van informatie is precies wat er in de digitale wereld steeds meer gebeurt. De praktische vereiste was het internet, waar informatie voor iedereen vanaf de eigen computer te vinden is. Hoe groter het internet wordt, hoe lastiger zaken terug te vinden zijn. Steeds vaker worden gegevens van verschillende websites op een handige manier met elkaar verbonden, zodat ze in één keer te doorzoeken zijn. Dat begon al met bibliotheekcatalogi, maar ook documenten zelf worden meer en meer compleet online beschikbaar gesteld, bijvoorbeeld in de Digitale Bibliotheek voor de Nederlandse Letteren, *dbnl*. Bibliotheken en archieven presenteren steeds meer afbeeldingen van manuscripten, van alle bladzijden met omschrijvende gegevens (metadata) erbij.

Er zijn nog veel meer nieuwe ontwikkelingen te verwachten. We kunnen nu al gemakkelijker dan ooit tevoren schrijfhanden vergelijken in documenten die zich bevinden in verschillende instellingen: de digitale afbeeldingen kunnen op het scherm naast elkaar geopend worden en zo worden bestudeerd. Maar dan moet de onderzoeker nog steeds zelf wel de documenten selecteren. Binnen enkele jaren gaat het echter nog een enorme stap verder. Onderzoekers van Huygens ING, Universiteit Leiden, en IISG zullen met geld uit het Onderzoeksfonds voor de KNAW instituten een *deep learning* systeem voor 'hand'-herkenning opzetten. *Deep learning* is het toepassen van zelflerende software, die op basis van trainingsdata patronen herkent die vervolgens worden vergeleken met andere data. Het resultaat zal naar verwachting zijn dat de onderzoeker van zo'n systeem suggesties kan krijgen welke documenten mogelijk door dezelfde persoon of in dezelfde plaats of tijd zijn geschreven, waar die documenten zich dan ook bevinden. De kloof die Van Oostrom tussen twee onderzoeksdisciplines zag, wordt op die manier definitief overbrugd. Tenminste, als de documenten op een voldoende gelijke manier digitaal via het internet worden aangeboden aan dergelijke algoritmen.

Al deze digitale koppelingen kunnen de op vele plaatsen voorhanden middeleeuwse documenten en data gezamenlijk beschikbaar en doorzoekbaar maken. En als het eenmaal zo ver is, dan zijn al die gegevens samen wel degelijk *big* data, *magna* data. Ik verwacht dat we dan aan het begin zullen staan van een gouden eeuw van nieuwe ontdekkingen.

Software

Laurence Anthony, *AntConc* (Version 3.4.3) Tokyo, 2014. Software en informatie beschikbaar via http://www.laurenceanthony.net/

Maciej Eder, Mike Kestemont, en Jan Rybicki, 'Stylometry with R: a suite of tools', in: *Digital Humanities 2013: Conference Abstracts*. Lincoln: University of Nebraska-Lincoln, 487-89. De software en meer informatie is te vinden via https://sites.google.com/site/computationalstylistics/

---

[1] Frits van Oostrom, *Nobel streven. Het onwaarschijnlijke maar waargebeurde verhaal van ridder Jan van Brederode (*Amsterdam 2017)

[2] Christine L. Borgman, *Big data, little data, no data. Scholarship in the networked world* (Cambridge Massachusetts-Londen 2015).

[3] Bas Blokker, 'Rechercheur van de Middeleeuwen. Interview Frits van Oostrom', in: *NRC*, 6 oktober 2017, https://www.nrc.nl/nieuws/2017/10/06/met-zevenmijlslaarzen-en-loep-dwars-door-de-middeleeuwen-13365045-a1576298.

[4] Allen Beye Riddell, 'How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models', in: Matt Erlin en Lynne Tatlock (red.), *Distant readings. Topologies of German culture in the long nineteenth century* (Rochester NY 2014) 91-113 (p. 92).

[5] Karina van Dalen-Oskam en Katrien Depuydt, 'Lexicography and Philology', in: K.H. van Dalen-Oskam e.a. (red.), *Dictionaries of Medieval Germanic Languages. A Survey of Current Lexicographical Projects* (Turnhout 1997) 189-197 (Selected Proceedings of the International Medieval Congress, University of Leeds, 4-7 July 1994).

[6] Karina van Dalen-Oskam, 'In praise of the variant analysis tool. A computational approach to medieval literature', in André Lardinois e.a. (red.), *Texts, Transmissions, Receptions: Modern Approaches to Narratives* (Leiden 2015) 35-54. Open access op http://booksandjournals.brillonline.com/content/books/9789004270848.

[7] Ronald Haentjens Dekker, Dirk van Hulle, Gregor Middell, Vincent Neyt, Joris van Zundert, 'Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project' in *LLC: The journal of digital scholarship in the Humanities* 30 (2015), 452-470; online 19 maart 2014: https://doi.org/10.1093/llc/fqu007.

[8] Erik Ketzan en Christof Schöch, 'What changed when Andy Weir's *The Martian* got edited?', in: Rhian Lewis e.a. (red.), *Digital Humanities 2017 Conference Abstracts. McGill University & Université de Montréal, Montréal, Canada, August 8-11*, 2017, 285-288. Online beschikbaar op https://dh2017.adho.org/program/abstracts/.

[9] Zie voor een overzicht en verwijzingen naar eerdere publicaties Matanja Hutter en Karina van Dalen-Oskam, 'Vrouwen en de *Rijmbijbel*'. In: Marjolein Hogenbirk & Lisa Kuitert (red.), *Schriftgeheimen. Opstellen over schrift en schriftcultuur* (Amsterdam 2017) 171-184.

[10] Van Oostrom 2017 (zie noot 1) 354-356.

[11] Van Oostrom 2017 (zie noot 1) 156.

[12] Karina van Dalen-Oskam en Joris van Zundert, 'Delta for Middle Dutch: Author and copyist distinction in "Walewein"', *Literary and Linguistic Computing* 22 (2007) 345-362. Online: 2 juni 2007: doi:10.1093/llc/fqm012 en het prachtige werk van Mike Kestemont, zie http://www.mike-kestemont.org/.

[13] Van Oostrom 2017 (zie noot 1) 184.

[14] Van Oostrom 2017 (zie noot 1) 313.