

Tijdens mijn studietijd in de jaren 70 deed de hoogleraar Russische literatuur Jan Meijer een onderzoek naar het woordgebruik van de Russische dichter Osip Mandelstam. Hij telde in het werk van deze dichter, die in 1938 in een Goelagkamp is gestorven, het aantal (rijm)woorden met een trieste betekenis of een sombere klank, zoals oe. Zijn conclusie was dat het aantal in de loop van de tijd toenam. Ik stond destijds vrij sceptisch tegenover dit type onderzoek. Dat werd niet beter door de rel die in 1990 uitbrak rond het anti-islamitische pamflet *De ondergang van Nederland* van Mohammed Rasol. De bekende tekstwetenschapper Teun van Dijk beweerde dat tekstanalyse ‘voor 95 procent zeker’ aantoonde dat Rasol een pseudoniem was van Gerrit Komrij, wat deze fel ontkende. Die conclusie baseerde Van Dijk onder andere op het voorkomen van een ‘ongebruikelijk’ woord als zwerfhond.

Genoemde onderzoekers keken naar bijzondere en opvallende naamwoorden en werkwoorden om de typerende stijl van een auteur vast te stellen. Die methode is intussen achterhaald. Tegenwoordig weten we dat juist onopvallende, hoogfrequente woorden als lidwoorden, voorzetsels, voegwoorden en voornaamwoorden geschikt zijn om de stijl van auteurs te karakteriseren en om anonieme werken aan een auteur toe te schrijven. Deze zogenaemde functiewoorden hebben geen eigen betekenis, maar alleen een grammaticale functie. Onderzoek heeft aangevend dat mensen niet letten op deze kleine woordjes, maar dat iedere schrijver ze wel op een eigen, unieke manier gebruikt. Met de computer kunnen verschillen in het gebruik van functiewoorden in grote tekstbestanden tegenwoordig eenvoudig worden geteld. Een mooi succes van de stylometrie, zoals dit

COLUMN

Stijl zit in kleine woordjes

NICOLINE VAN DER SIJS



soort kwantitatief stijlonderzoek wordt genoemd, werd deze zomer behaald. *The Sunday Times* maakte op 14 juli bekend dat een computeranalyse het vermoeden bevestigde dat de detective *The Cuckoo's Calling*, die op naam stond van debutant Robert Galbraith, in feite was geschreven door Harry Potter-schrijfster J.K. Rowling. Rowling gaf het meteen toe. Op de populaire taalkundige blog *Language Log* leggen taalkundigen Peter Millican en Patrick Julia uit hoe zij te werk zijn gegaan (<http://nrch.nl/32z7>). Met de computer analyseerden ze (‘in een halfuurtje’) onder meer de honderd frequentste woorden in *The Cuckoo's Calling*. Ze vergeleken die met de frequentste woorden in werk van andere auteurs. Rowling bleek de meest waarschijnlijke auteur te zijn. Ook onder Nederlandse en Belgische literatuuronderzoekers is stylometrie in opkomst. Zo presenteerde Karina van Dalen-Oskam in haar oratie eerder dit jaar een stilistische analyse van de briefromans die Elisabeth Wolff-Bekker en Agatha Deken eind 18de eeuw samen publiceerden. Aan de Universiteit van Antwerpen wordt momenteel innovatief onderzoek uitgevoerd door Mike Kestemont en collega's van het Centrum voor Computerlinguïstiek en Psycholinguïstiek. In zijn dissertatie uit 2012 onderzocht Kestemont welke delen van de 13de-eeuwse rijmtekst *Spiegel historiael* door Jacob van Maerlant waren geschreven en welke door Filip Utenbroeke. Voor het congres Digital Humanities dat op 18 juli in Lincoln, Nebraska, werd gehouden, maakte Kestemont een documentaire over tekstonderzoek naar de Latijnse teksten van Hildegard von Bingen. Deze 12de-eeuwse Duitse mystica dicteerde haar visioenen aan secretarissen, omdat haar kennis van het Latijn gebrekkig was. Uit het onderzoek blijkt dat haar laatste secretaris, Guibert van Gembloux, zo'n grote invloed had op

het taalgebruik dat hij als eigenlijke auteur van die bewuste teksten kan gelden. De informatieve en prachtig vormgegeven documentaire is te bekijken op <http://vimeo.com/70881172>. Antwerpenaren timmeren ook aan de weg: uit hun koker komt het computerprogramma Stylene, waar iedereen de stijl van een tekst kan laten analyseren (<http://nrch.nl/32z6>). Volgens Stylene lijkt de stijl van deze column het meest op die van Jeroen Brouwers - een conclusie waarmee ik goed kan leven. Stylometrie wordt niet alleen gebruikt voor auteurstoeschrijving. In 2011 voerden James M. Hughes e.a. een kwantitatieve stijlanalyse uit op een corpus literaire teksten vanaf 1550. Uit het artikel dat ze hierover publiceerden in *PNAS* bleek dat auteurs de meeste stilistische overeenkomsten hebben met tijdgenoten, en dat er zo iets als een literaire ‘tijdstijl’ bestaat. Politie en justitie gebruiken ‘forensische stylometrie’ om anonieme teksten (dreigbrieven, bewijsstukken) tot concrete personen te herleiden. En uiteraard wordt nu als antwoord gewerkt aan een computerprogramma dat de stijlkenmerken van een tekst automatisch anonimiseert. Stylometrisch onderzoek staat nog in de kinderschoenen en is zeker niet algemeen geaccepteerd - al was het maar omdat het kennis van statistiek vereist. De methode roept ook weerstand op: ‘Literature is not data’ zette de Canadese schrijver Stephen Marche boven een column. Hoe dit ook zij, de mogelijkheden van dit soort kwantitatief onderzoek zijn beperkt. De computer zal nooit willekeurige anonieme teksten onduidelijk tot één auteur kunnen herleiden: het wegen en duiden van de resultaten blijft - gelukkig - mensenwerk. Inmiddels ben ik wel benieuwd wie volgens de nieuwste stylometrische methode de auteur is van *De ondergang van Nederland*. Gerrit Komrij was het zeker niet.