Het ongezochte vinden

van der Sijs, N.

published in
NRC Handelsblad
2012

document version
Publisher's PDF, also known as Version of record

Het ongezochte vinden

Toen Bruno Becker, oprichter en eerste directeur van wat nu het Oost-Europa Instituut van de Amsterdamse universiteit is, ooit een bepaald citaat van Aristoteles nodig had, las hij het complete oeuvre van Aristoteles door om het te vinden. Tegenwoordig kunnen we zoveel geduld niet meer opbrengen: als we informatie zoeken, nemen we direct onze toevlucht tot digitale tekstbestanden die via bibliotheken of Google beschikbaar worden gesteld.

Geesteswetenschappers gebruiken digitale historische teksten nog vooral om woorden, namen, uitdrukkingen of citaten in op te zoeken. Ze willen achterhalen sinds wanneer een woord of een uitdrukking voorkomt, ze bekijken wat er in een bepaalde periode over een specifiek begrip – 'slavernij', 'democratie' – wordt gezegd, of hoe er op een bepaald moment werd gedacht over het werk van een auteur. Of ze bekijken op welk moment metaforen in zwang komen (drankzucht als 'kanker van de maatschappij'), omdat hieruit maatschappelijke veranderingen blijken. Voor dit soort onderzoek bieden digitale teksten geweldige mogelijkheden.

Dergelijk onderzoek – hoe interessant ook – is niet meer dan het automatiseren van het oude, handmatige leeswerk zoals door Becker en anderen vóór het digitale tijdperk werd ondernomen. Willen we echt nieuwe resultaten boeken in het geesteswetenschappelijk onderzoek, dan moeten we nieuwe onderzoeksmethoden aanwenden die de jonge digitale wereld ons aanreikt.

Een daarvan is dat we niet langer zoeken naar de bekende weg, maar dat we de computer serendipitair voor ons laten zoeken. Dat kan door grote hoeveelheden tekstbestanden automatisch te laten analyseren. Hiervoor bestaan allerlei programma's. Die programma's tellen alles in een tekst wat telbaar is, zoals de hoeveelheid zinnen, de gemiddelde lengte van een zin, de hoeveelheid woorden, de hoeveelheid lettergrepen per woord, de meest voorkomende woordcombinaties en de frequentie van woorden.

Dit komt u misschien bekend voor: dergelijke tellingen vormen de basis van de beroemde leesbaarheidsformule die de Amerikaan Rudolf Flesch in 1948 heeft ontwikkeld en die in de meeste versies van Microsoft Word is ingebouwd. De formule rekent op basis van de gemiddelde zinslengte en het gemiddelde aantal lettergrepen per woord uit hoe moeilijk een tekst is: een tekst met veel lange zinnen en woorden scoort laag in leesbaarheid.

Wellicht bent u sceptisch over het idee dat het domweg tellen van woorden kan leiden tot nieuwe inzichten, maar daarmee onderschat u de mogelijkheden van dit soort statistische tekstanalyse. Een tekstanalyseprogramma kan automatisch overeenkomsten en verschillen tussen twee tekstbestanden vaststellen, en aangeven welke woorden of woordcombinaties typerend zijn voor een bepaalde tekst of een bepaalde auteur. Uiteraard moet een onderzoeker de resultaten kritisch beoordelen.

Tekstanalyse kan bijvoorbeeld munitie leveren voor de discussie die wordt gevoerd over de vraag of de verschillen tussen het Standaardnederlands in Nederland en België toenemen of juist afnemen. Het feit dat Vlamingen steeds vaker kleed vervangen door jurk wijst op een afnemend verschil, maar hoogfrequente combinaties als 'zich interesseren aan' of 'Ik zou het zelf zo willen gezegd hebben' spreken dat weer tegen. Aan de universiteit van Leuven zijn onderzoekers al enkele jaren bezig de verschillen te meten tussen teksten uit Nederland en uit België. Hun voorlopige conclusie is dat het geschreven taalgebruik van de twee landen naar elkaar toe groeit. Uit ander onderzoek blijkt dat de uitspraakverschillen juist

Vlamingen vervangen steeds vaker kleed door jurk, maar interesseren zich nog wel 'aan' iets

groter worden.

Een spannend onderzoek waarvoor ik tekstanalyse graag zou inzetten, is het vergelijken van het taalgebruik van kranten uit Nederland en uit Nederlands-Indië in de eerste helft van de 20ste eeuw. Mijn hypothese is namelijk dat veel veranderingen in het Nederlands – nieuwe Nederlandse woorden of constructies – in Indië zijn ontstaan of geaccepteerd geraakt.

Ik kwam op het idee doordat het me opviel dat Indische kranten vaak de oudste bron zijn voor een Nederlands woord: haatzaaien bijvoorbeeld, maar ook 'gewone' woorden als reuzeleuk, piepklein, knoerthard en knotsgek. Misschien is dat gewoon toeval, maar misschien ook niet: de Indonesische maatschappij verschilde aanzienlijk van de Nederlandse, wat noopt tot neologismen. In de eerste helft van de 20ste eeuw waren er bovendien meer tweetalige Nederlanders en Indonesiërs dan ooit tevoren: uit onderzoek is bekend dat tweetaligheid leidt tot taalveranderingen en taalvernieuwingen.

Als tweetaligheid in Indië een motor van taalverandering is geweest, zou dat ook uit andere taalfeiten moeten blijken. Volwassenen die een tweede taal leren, hebben moeite onregelmatige vormen te leren. Lastig vinden zij bijvoorbeeld de verleden tijd van sterke werkwoorden (woei, joeg, ervoer, verschool), omdat de meeste Nederlandse werkwoorden zwak zijn. De verwachting is dan ook dat in Indische kranten vaker dan in Nederlandse sprake zal zijn van waaide, jaagde, ervaarde, verschuilde. En van 'een jonge meisje' in plaats van 'een jong meisje', omdat bijvoeglijke naamwoorden meestal op -e uitgaan.

Mijn handen jeuken om tekstanalytische programma's op kranten los te laten. Nu stuiten we echter op de praktische bezwaren. Niets dan lof voor de digitale krantenbestanden van de Koninklijke Bibliotheek, maar in hun huidige vorm zijn ze nog niet geschikt voor dit soort onderzoek. De computerprogramma's verslikken zich nog te vaak in de spellingvariatie van oude teksten. Maar daar wordt aan gewerkt. Zodra er vorderingen zijn, zal ik dat hier melden.