Culturomics: een nieuw vakgebied?

Begin december lanceerde Google de *Ngram Viewer*, een website waarop een selectie van ruim vijf miljoen boeken uit de Google Books-collectie doorzocht kan worden op reeksen van maximaal vijf woorden (n-grammen). De resultaten worden gepresenteerd in een grafiek die het relatieve voorkomen van de zoektermen of reeksen per jaar weergeeft. Tegelijk met de lancering publiceerde *Science* een artikel van de ontwikkelaars waarin zij uitgebreid ingaan op de mogelijkheden van de Ngram Viewer. Daarnaast presenteerden zij Culturomics, een nieuw vakgebied waarbinnen taal en cultuur bestudeerd worden aan de hand van grootschalige data-analyse.

De *Science*-publicatie zorgde voor opschudding onder met name geesteswetenschappers. Is deze *culturomics* een arrogante poging uit de bètahoek om de alfa's en de gamma's uit de brand te helpen, een brand die er helemaal niet is? Taal en cultuur worden toch al tijden onderzocht aan de hand van de analyse van data? Daarnaast is de Ngram Viewer nog helemaal niet zo bruikbaar voor onderzoek als de ontwikkelaars in hun artikel doen voorkomen.

Reden voor de Linguistic Society of America (LSA) om tijdens haar jaarlijkse bijeenkomst een extra sessie in te lassen met Jean-Baptiste Michel en Erez Lieberman Aiden, de hoofdauteurs van de Science-publicatie. Antal van den Bosch, hoogleraar Geheugen, taal en betekenis aan de Universiteit van Tilburg, was erbij.

'Het was erg duidelijk dat de ontwikkelaars zelf geen taalkundigen zijn; ze weten zelf amper wat ze doen', aldus Van de Bosch. 'Maar ze staan wel erg open voor suggesties om hun toepassing te verbeteren.' In de eerste week na de publicatie heeft deze houding al tot enkele essentiële verbeteringen geleid. Zo is er een dubbele indexering uitgevoerd, waardoor woorden als 'can't' niet meer alleen als een bigram ('can' en ''t'), maar ook als unigram ('can't') vindbaar zijn. Uit de kritieken uit de zaal tijdens de LSA-bijeenkomst sprak vooral de wens van een breder aanbod, zowel van cijfers als van metadata. Ook hier werken de ontwikkelaars graag aan mee – voor zover Google bereid is de benodigde informatie vrij te geven.

Momenteel geeft de Viewer als enige uitkomst het relatieve voorkomen van een woord per jaar ten opzichte van het totaal aantal woorden in dat jaar. Wetenschappers hopen binnenkort ook te beschikken over de absolute aantallen, zowel per token als per boek. Daarnaast zou een uitbreiding van de metadata erg waardevol zijn. In de huidige vorm is alleen een grove categorisering aangebracht tussen bijvoorbeeld fictie en nonfictie. 'Iemand die onderzoek doet naar poëzie uit Zuid-Frankrijk uit de tweede helft van de negentiende eeuw zal nu nog handmatig moeten zoeken', aldus Van den Bosch. 'Dat moet veel makkelijker kunnen.' Ten slotte zullen de publicaties van vóór 1900, die niet meer auteursrechtelijk beschermd zijn, binnen afzienbare tijd als *full text* gepubliceerd worden.

Dan blijven er nog twee grote struikelblokken over waar de ontwikkelaars weinig aan zullen kunnen veranderen. Ten eerste de kwaliteit van het toegepaste OCR-proces: die is nog altijd verre van perfect waardoor de resultaten niet altijd even betrouwbaar zijn. En ten tweede de eenzijdige bron van boeken. 'Boeken representeren de tijdsgeest maar ten dele,' legt Van den Bosch uit. 'Ze zijn representatief voor de mensen door wie en voor wie boeken werden geschreven. In de 19^{e} eeuw lag dat heel anders dan nu.' Maar de Ngram Viewer is nu eenmaal gebaseerd op het GoogleBooks-project.

En dan die term, culturomics, zal die beklijven? Van den Bosch verwacht van niet, maar wellicht kunnen we het over een paar jaar objectief bekijken met een verbeterde versie van de Ngram Viewer. Eentje waarin wellicht ook websites en tijdschriften zijn opgenomen, want zoals Van den Bosch het zegt, 'wat staat er tegenwoordig nog in boeken?'

Erica Renckens