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.

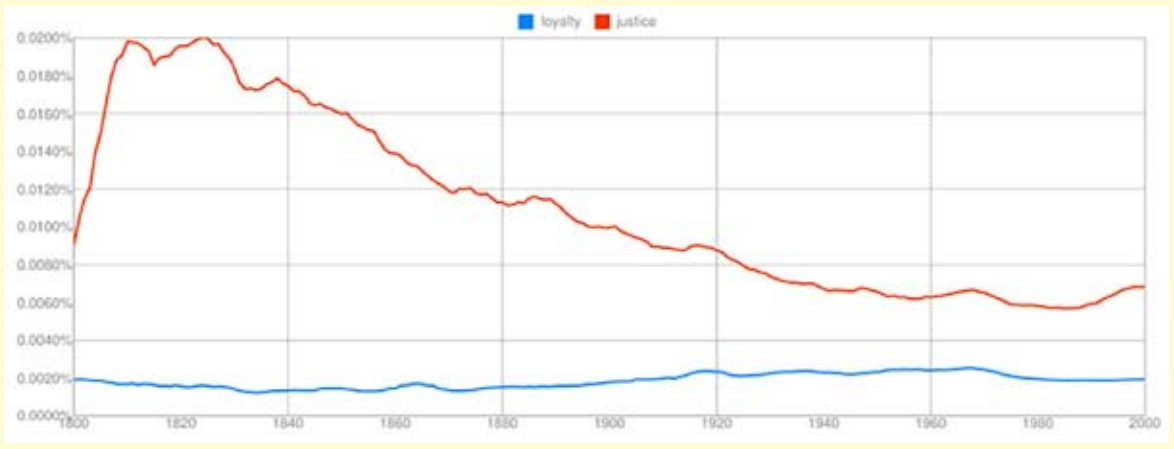
'Het was erg duidelijk dat de ontwikkelaars zelf geen taalkundigen zijn; ze weten zelf amper wat ze doen', aldus Van de Bosch. 'Maar ze staan wel erg open voor suggesties om hun toepassing te verbeteren.' In de eerste week na de publicatie heeft deze houding al tot enkele essentiële verbeteringen geleid. 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 19e 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?'

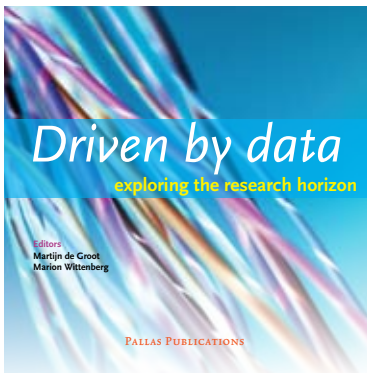
Erica Renckens



Uitkomst van de Ngram Viewer voor het voorkomen van de woorden 'loyalty' en 'justice' in de periode van 1800 tot 2000.

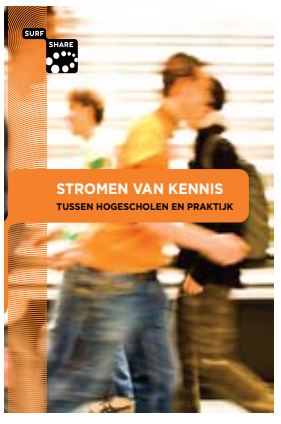
Gelezen

Heiko Tjalsma, Jeroen Rombouts and Annemiek van der Kuil (eds): *Selection of Research Data: Guidelines for appraising and selecting research data, a report by DANS and 3TU.Datacentrum*. Utrecht, Stichting SURF, 2010.
Er zijn algemene richtlijnen opgesteld voor het beoordelen en selecteren van onderzoeksdata om te bewaren. Die zijn te gebruiken door iedereen die iets met de beoordeling en selectie te maken heeft. Het rapport bevat de laatste stand van zaken op het gebied van het selecteren van onderzoeksdata, gebaseerd op literatuuronderzoek, een aantal interviews met belangrijke personen en de ervaringen van DANS en het 3TU Data centrum. Hoewel onderzoeksgegevens veelal bewaard moeten worden voor gebruik of hergebruik, of ter validatie van onderzoeksresultaten, geldt dit niet voor alle gegevens. Om te bepalen welke data waardevol zijn als (bron)materiaal voor onderzoek zijn er een aantal praktische richtlijnen opgesteld en samengebracht in een checklist.
ww.surffoundation.nl/nl/publicaties/



Martijn de Groot en Marion Wittenberg (eds): *Driven by data; Exploring the research horizon*. Amsterdam, Pallas Publications, Amsterdam University Press, 2010; ISBN: 978 90 8555 038 9
On 2 December 2009, the symposium 'Driven by data' took place in The Hague to celebrate the anniversaries of a number of archiving institutions that are now all part of DANS (Data Archiving and Networked Services). It was a lively meeting, in which attention was paid to various surprising and creative possibilities associated with the reuse of electronic data. DANS decided to publish this jubilee volume to give a larger public a taste of this versatility and creativity. The contributions cover topics as diverse as dendrochronology, politics, shipping traffic history, bird migration, philosophy of science and statistics. This publication can be downloaded free of charge from the DANS website: www.dans.knaw.nl

***Stromen van Kennis*. Utrecht, SURFshare, 2011**
In deze publicatie geven lectoren, docenten en studenten hun visie op het toegankelijk maken en delen van onderzoekresultaten van hogescholen. Dit is belangrijk omdat de HBO-raad in 2009 de 'Berlin Declaration on Open Access' heeft getekend. Daarmee geven de hogescholen aan zich te zullen inzetten voor het vrij beschikbaar stellen van hun onderzoeksresultaten. De behoefte aan vrije toegang tot kennis is groot. Zowel omdat praktijkgericht onderzoek een directe link heeft met de samenleving, alsook omdat, met de komst van lectoraten en kenniskringen, hogescholen steeds meer onderzoek doen. Het delen van de stroom kennis die momenteel vanuit het hbo los komt en waarmee hogescholen zich sterker willen profileren, vraagt om een gedegen kennisinfrastructuur. SURFfoundation ondersteunt dit proces, onder meer door het helpen realiseren van repositories, een onmisbaar onderdeel van die infrastructuur.
www.surffoundation.nl/nl/publicaties/



Column
Onno Crasborn

Onverwachte gasten

In 2008 publiceerden we het Corpus Nederlandse Gebarentaal (NGT). Zeventig uur video-opnames van dialogen tussen doven. Het leeuwendeel is openbaar, en dus door elke onderzoeker te gebruiken. 'Corpus' is een groot woord voor onze verzameling, leerden we steeds meer inzien: zonder uitgebreide annotaties is een video niet alleen voor gebarentaal-leken onbegrijpelijk, maar ook voor taalkundigen betekenisloos. In de twee jaar die de kleine investeringssubsidie van NWO ons eraan liet werken, was er nauwelijks tijd over om gebaren en zinnen te vertalen. Laat staan dat we een fonetische transcriptie hadden gemaakt van alle details. Vingerbewegingen, gezichtsuitdrukkingen, hoofdknikken, leunen naar links of naar achteren... je bent zo een hele dag bezig met een enkel minuutje video.

Gelukkig wisten we van tevoren dat de verzameling niet alleen interessant zou zijn voor kwantitatief onderzoek. Alleen al er naar kunnen kijken maakt de filmpjes bijzonder. Er is namelijk niet zoveel gebarentaal te zien geweest in het verleden. We hebben geen dovenomroep, er zijn geen dove speelfilms (met ondertiteling voor horenden), en er is geen Sesamstraat met een gebarende Pino. Af en toe is er een tolk op televisie. Wel zijn er twee planken met videobanden gemaakt sinds de jaren tachtig, en later nog een hele plank dvd's, maar je kunt er nog geen hoekje van een bibliotheek mee vullen. Het Corpus NGT bevat discussies die voor veel doven interessant zijn. Hebben dove ouders het recht om via IVF een dove embryo te selecteren? Is het positief of negatief dat er regionale variatie is in gebarentaal? Het vormt een aanzienlijke verrijking van het dove medialandschap, dachten wij zelf met gepaste trots.

En inderdaad, op de publiekssite www.corpusngt.nl kijken zowel doven als horende tolken en gebarentaaldocenten regelmatig rond. Maar het mailtje van collega's uit Hamburg hadden we niet verwacht. Of de publiekssite niet ook in het Duits vertaald kon worden, om de filmpjes te kunnen bekijken. Hè? Maar de Duitse Gebarentaal en de Nederlandse zijn toch twee verschillende talen? De dove oosterburen gaan de Nederlandse Gebarentaal toch niet begrijpen? Nee, dat was waar. Maar ze wilden het gewoon zien. Kunnen kijken. Er is geen vreemdetalenonderwijs voor gebarentaal, niet in Nederland en ook niet in Duitsland. Ontwikkeling, zo denkt men in het dovenonderwijs nog altijd, is gekoppeld aan gesproken taal. En daarom hebben de gebarentaalcorpora die nu overal gemaakt worden een bijzonder en onverwachte functie erbij gekregen: ze maken een blik op een andere taal mogelijk, voor doven uit allerlei landen.

Onno Crasborn is gebarentaaldeskundige en senior onderzoeker Taalwetenschap aan de Radboud Universiteit.