# Lofzang op de dataverzamelingen van het Meertens Instituut

*John Nerbonne*

Het verzamelen, organiseren en beschikbaar stellen van onderzoeksgegevens is een moeilijke taak maar ook een reusachtige dienst aan de wetenschappelijke gemeenschap (en soms daarbuiten), die te vaak wordt gezien als secretariële bezigheid, als niet-geïnspireerde 'boekhouding', of zelfs denigrerend als 'vlinders verzamelen'. Niets is minder waar, en de grote verzamelingen taalgegevens in het Meertens Instituut logenstraffen die opvatting.

De wetenschappelijke gegevensverzamelaars zijn een ondergewaardeerde groep, maar ze hebben de wetenschappelijke prestaties van vele collega's verhoogd, onder andere de mijne. Wie niet alleen gegevens verzamelt maar deze ook organiseert en voor de toekomst zeker stelt, verdient lof!

In 1925 begon de Belgische taalkundige Edgard Blancquaert de *Reeks Nederlandse Dialectatlassen* (RND), die onder meer vanwege de oorlog pas in 1982 door zijn student Willem Pée voltooid konden worden. Het is geweldig dat de Universiteit Gent een steekproef van de gegevens in de vorm van gescande pagina's beschikbaar heeft gemaakt (www.dialectzinnen.ugent.be/). Toen ik in 1995 tijdens de jaarlijkse conferentie van computationele taalkundigen in Europa een interessante voordracht over dialecten in Ierland hoorde, besloot ik het idee in een seminar in Groningen te behandelen en na te doen. Vervolgens zocht ik Nederland-

se data om de cursus voor de studenten in Groningen interessanter te maken, en daar bood de RND wat nodig was in de vorm van data, en dat ook op genereuze schaal. De verzamelaars hadden 159 zinnen in bijna 2000 plaatsen opgenomen, in fonetische transcriptie opgetekend, en goed en toegankelijk gearchiveerd. Zo was het mogelijk om de studenten in de ideeën en methoden in te voeren, zonder ook de tijdrovende dataverzameling voor onze rekening te nemen.

De studenten waren enthousiast over de mogelijkheid authentieke data te gebruiken en ze hebben uitmuntend gewerkt. De studie van dialectdata met de computer, de computationele dialectologie, is zelfs tot een van de voornaamste onderzoekslijnen van mij en anderen in Groningen uitgegroeid.

Het was wel nodig om de gedrukte transcripties te digitaliseren, want er bestond nog geen modern digitaal archief in het begin van de jaren 1980. Blancquaert en Pée hebben de data naar de toenmalige stand van het vak uitstekend opgeslagen en toegankelijk gemaakt, zodat toekomstige generaties belangrijke gegevens hadden waarmee ze oudere analysen konden repliceren en proberen te verbeteren, en waarmee de historische ontwikkeling van de Nederlandse streektalen in kaart gebracht kon worden.

Verder ben ik (zoals vele andere wetenschappers) ook schatplichtig aan het nieuwere Goeman-Taeldeman-Van Reenen-Project (GTRP), waar de dialectgegevens wel digitaal beschikbaar werden gemaakt, hetgeen de computationele analyses zeer vergemakkelijkt. In dat project werden vijftien jaar lang (1980-1995) gegevens verzameld, waarna veel werk werd besteed aan digitalisering en het organiseren van een database. Het GTRP bestaat uit digitale transcripties van 1876 elementen uit 613 plaatsen in Nederland en België en heeft tot een stortvloed van publicaties geleid.

We hebben in Groningen ook van de nieuwere collectie onze voordelen gehad. Wij richten daarom ook graag een dankwoord aan de oprichters en uitvoerders van GTRP. die hun gegevens vakkundig hebben georganiseerd en voor anderen beschikbaar hebben gemaakt.

Feitelijk waren Blancquaert en Pée, maar ook Goeman, Taeldeman en van Reenen, vooruitziend, en ze verdienen daarvoor onze dank! Ze bouwden tegelijk voort op een traditie in de dialectologie, want ook

Wenker, die vaak als stichter der dialectologie geldt, werkte tientallen jaren aan de organisatie en archivering van zijn hoofdwerk, *Der deutsche Sprachatlas*, dat pas lang na zijn dood in 1911 gepubliceerd kon worden. Ook de eersten dialectologen in Frankrijk, Jules Gilliéron en Edmond Edmont, hebben hun onderzoeksmateriaal aan het begin van de twintigste eeuw gepubliceerd.

Maar taalwetenschappers zijn – in het bijzonder in de laatste vijftig jaar – lang niet altijd goede rentmeesters van gegevens geweest, dus is onze waardering voor het werk van Blancquaert en Pée verdiend. Ik noemde de Meertens rentmeesters hierboven 'vooruitziend', want vandaag is het wetenschappelijke belang van goed gegevensbeheer gemeengoed.

Natuurlijk waarderen we vandaag de dag het goede beheer van wetenschappelijke gegevens. De wetenschap heeft een verplichting aan de waarheid! Iedere beoefenaar van de wetenschap zal deze uitspraak onmiddellijk beamen, misschien liever zonder uitroepteken, maar wel hartstochtelijk. Vroeger sprak men (bijv. Weber) zelfs van wetenschap als *roeping*, waarin ik meen een vleugje van 'roeping' in religieuze zin te horen doorklinken, waardoor dan ook een morele verantwoording ontstaat. De wetenschapper vertelt alles eerlijk en volledig, 'naar beste weten'.

Maar de associatie tussen eerlijkheid en wetenschap is een voorschrift, geen beschrijving van de werkelijkheid. Wij horen als wetenschappers eerlijk te zijn, want anders is de onderneming zinloos. Maar er zijn naast zorgvuldige, eerlijke collega's ook slordige en zelfs frauduleuze. Dit suggereert dat een aanpak die mikt op persoonlijke billijkheid tekortschiet. We bereiken nooit alle collega's op deze manier. Naast het (zeer correcte!) appel aan een *persoonlijk* beroepsethos hebben we *sociale* (of wel professionele) drempels en controles nodig om te garanderen dat wetenschappelijke bijdragen zo waarheidsgetrouw mogelijk zijn.

Het moderne antwoord – de openbaarheid van gegevens – neemt sinds een tiental jaren een hoge vlucht. De ouderwetse wetenschappelijke controle was de wetenschappelijke publicatie – in een tijdschrift, door vakgenoten gekeurd, en in principe voor iedereen toegankelijk. Maar deze vorm van controle bleek onvoldoende te zijn, zoals we te vaak hebben gezien. In plaats daarvan wordt men vandaag geacht ook de onderzoeks-

gegevens beschikbaar te maken waarop analysen en gevolgtrekkingen zijn gebaseerd. De reden voor deze ontwikkeling is veel breder dan alleen het voorkomen van fraude. Gegevens die goed gearchiveerd zijn, hergebruikt kunnen worden, hetgeen anderen veel tijd en energie bespaart, en verder ook de replicatie van analyses mogelijk maakt, hetgeen de betrouwbaarheid van conclusies verhoogt. Op het gebied van dialectwetenschap is er vaak ook een cultureel belang in de opslag van gegevens. Dus we zingen terecht de lof van de eerste verzamelaars en rentmeesters van gegevens voor ons vakgebied!

John Nerbonne is oud-professor alfa-informatica aan de Rijksuniversiteit Groningen en Honorarprofessor in Freiburg. Zijn onderzoek houdt zich bezig met computationele en kwantitatieve linguïstiek, vooral met vraagstukken op het gebied van taalvariatie en digitale geesteswetenschappen.

## Bronnen

De dataverzamelingen werden in de volgende stukken beschreven:

Blancquaert, Edgard, & Willem Pée. (1925-1980) *Reeks Nederlandse dialectatlassen*. Antwerpen: De Sikkel, 1930. (Scans) beschikbaar: https://www.dialectzinnen.ugent.be/

van Reenen, Pieter, Anton Goeman & Johan Taeldeman. 'Goeman-Taeldeman-Van Reenen-Project (GTRP), 1985-1995, Phonology & Morphology of Dutch & Frisian Dialects in 1.1 million transcriptions, version 2.2 (cd rom).' Beschikbaar: https://www.meertens.knaw.nl/mand/database/.

Taeldeman, Johan & Anton Goeman (1996) 'Fonologie en morfologie van de Nederlandse dialecten: Een nieuwe materiaalverzameling en twee nieuwe atlasprojecten' *Taal en Tongval*, *48*:38-59.

Het gebruik van de gegevens in Groningen werd in mijn afscheidsvoordracht naverteld:

Nerbonne, John (2017) *Humanities, exactly!/Letteren, exact!* Rijksuniversiteit Groningen: Letteren Faculteit.

Het belang van de archivering van gegevens wordt in veel beleidstukken benadrukt (https://dans.knaw.nl/nl/, https://researchdata.nl/, https://www.rd-al-

liance.org/, stukken van NWO), en Reardon laat zien hoe belangrijk goed gegevensbeheer was over het opsporen van fraude.

Reardon, Sarah (2021). Flawed ivermectin preprint highlights challenges of COVID drug studies. *Nature*, *596*(7871):173-174.