COLUMN NICOLINE VAN DER SIJS

Open dialectendata

Vroeger was er weinig te doen aan wetenschappelijke hamsterwoede

Rond 1900 groef de Engelse archeoloog Sir Arthur Evans bij Knossos op Kreta kleitabletten op met teksten in twee onbekende schriften, die hij Lineair A en B noemde. Toen het hem niet lukte Lineair B te ontcijferen, hield hij de kleitabletten moedwillig weg van andere onderzoekers. Pas tien jaar na zijn dood ontcijferde Michael Ventris het schrift, nadat hij had bewezen dat de teksten Grieks waren.

De Münchense onderzoekster Annelies Kammenhuber hamsterde decennialang kleitabletten met teksten in Hettitisch spijkerschrift, nadat zij rond 1965 was begonnen aan een uitgebreid woordenboek van het Hettitisch: een taal die van 1700 tot 1200 voor Christus werd gesproken in Turkije en die een verre verwant is van de Germaanse talen. Erg opschieten deed het woordenboek niet: bij haar dood in 1995 was zij gevorderd tot de G. Ondertussen was het materiaal al die tijd onbereikbaar voor haar Engelse collega's. Die schermden op hun beurt hun Hettitische teksten in hiërogliefenschrift voor de Duitse onderzoekers af.

Dit zijn slechts een paar van de vele verhalen over geesteswetenschappers die hun gegevens weghielden van anderen. Evans en Kammenhuber zijn nog positieve uitzonderingen omdat ze daadwerkelijk over hun onderwerp publiceerden. Er zijn ook verzamelaars die het stadium van verzamelen niet overstijgen.

Over dergelijke hamsterwoedes werd vroeger wel geroddeld, maar veel meer konden onderzoekers er niet aan doen.

Tegenwoordig ligt dat anders: de academische wereld stelt steeds meer regels op om te garanderen dat data toegankelijk zijn of komen voor andere onderzoekers. Zo pleit de Koninklijke Nederlandse Akademie van Wetenschappen in het vorig jaar verschenen advies *Zorgvuldig en integer omgaan met wetenschappelijke onderzoeksgegevens* voor vrije toegang tot wetenschappelijk materiaal, dus open data.

In de geesteswetenschappen blijven gegevens lange tijd waardevol. Dit in tegenstelling tot veel bèta-onderzoek: daar zijn resultaten vaak bouwstenen waarop onmiddellijk wordt voortgebouwd, waardoor onderzoekers niet meer teruggrijpen naar oorspronkelijke gegevens. In de geesteswetenschappen gaat het vaker over nieuwe interpretaties van oude gegevens. Geesteswetenschappers hoef je niet te vertellen dat ze dwergen zijn op de schouders van reuzen.

Die wijsheid uit de twaalfde eeuw is actueler dan ooit. In deze digitale tijd wordt het voortbouwen op eerder onderzoek steeds gemakkelijker. Grote hoeveelheden oude en nieuwe gegevens kunnen semi-automatisch aan elkaar worden gekoppeld. Juist deze aanpak zal leiden tot allerlei nieuwe vondsten en inzichten binnen de geesteswetenschappen. Een concreet voorbeeld hiervan, waaraan we momenteel hard werken, is de inrichting van een elektronische Woordenbank van de Nederlandse Dialecten.

In de loop van de tijd zijn er veel woordenboeken van Nederlandse dialecten gepubliceerd. Die dialectwoordenboeken zijn vervaardigd door professionele dialectlexicografen en amateurs, en ze beschrijven de woordenschat van een enkele plaats, zoals het Weerts, of van grotere gebieden, zoals het Limburgs, Brabants, Vlaams of Overijssels. De gegevens van deze woordenboeken zijn – laat ik het voorzichtig zeggen – niet optimaal toegankelijk. Als er digitale bestanden van bestaan, zwerven die meestal bij particulieren. De oudere gedrukte woordenboeken, daterend vanaf begin 19de eeuw, liggen in bibliotheken te verstoffen.

Onze kennis van de Nederlandse dialecten, en de veranderingen die deze sinds de 19de eeuw hebben doorgemaakt, zal sterk toenemen als we alle beschikbare gegevens uit de dialectwoordenboeken aan elkaar koppelen: op die manier ontstaat een compleet nieuw onderzoeksinstrumentarium voor dialectonderzoek. Om dit te verwezenlijken heb ik vorig jaar aan de Nederlandse dialectlexicografen en streektaalfunctionarissen gevraagd of ze hun digitale dialectwoordenboeken aan het Meertens Instituut willen overdragen. Vlaamse en Friese collega's verzamelen de woordenboekbestanden uit hun taalgebied.

Vrijwilligers zijn inmiddels begonnen met het corrigeren van gescande oudere dialectwoordenboeken en het overtikken van manuscripten. De Data Curation Service van de Radboud Universiteit Nijmegen (bekostigd door het nationale infrastructuurprogramma CLARIN) zal straks alle digitale bestanden omzetten naar een eenvormig computerformaat.

Om de verschillende dialectwoordenboeken aan elkaar te kunnen koppelen, worden ze verrijkt met extra gegevens: vrijwilligers en studenten voegen aan alle dialecttrefwoorden de Standaardnederlandse vormen toe. De Standaardnederlandse vorm kan dienen als input voor karteringssoftware. Die software tekent automatisch kaarten met de verbreiding van woorden, klanken, vervoegingen en verbuigingen over het Nederlandse taalgebied. Dat zal veel nieuwe inzichten opleveren. Veel dialectsprekers menen bijvoorbeeld dat in hun dialect unieke woorden of uitdrukkingen voorkomen – of dat waar is, zal nu aan het licht komen, en ik voorzie veel lange gezichten... Ook kan uit de kaarten blijken hoe de verbreiding van een dialectverschijnsel in de loop van de tijd is veranderd.

Het aan elkaar verbinden van zoveel mogelijk dialectwoorden biedt nog veel meer spannende mogelijkheden. Dat zal in de toekomst blijken, als de elektronische Woordenbank van de Nederlandse Dialecten is gelanceerd. Vooruitlopend daarop vraag ik lezers van deze krant die een digitaal dialectwoordenboek of ongepubliceerd manuscript op de plank hebben liggen, dit naar me op te sturen, zodat het opgenomen kan worden in het grote net van open dialectendata.