NRC Handelsblad - 12 juni 2012 dinsdag

Varen deed je met een wagen, maar wanneer was dat eigenlijk?

De instituten die erfgoedteksten beheren, zitten op goud. Maar bijna niemand kan bij de oude boeken, wetteksten en andere documenten, bij gebrek aan een vaste standaard bij het digitaliseren. Met Nederlab komt dat goed: via de website van deze superzoekmachine zal elke wetenschapper in de teksten vinden wat hij zoekt. En ook elke andere geïnteresseerde Nederlander.

Een superzoekmachine kan binnenkort alle gedigitaliseerde Nederlandse boeken, kranten en tijdschriften maar ook wetteksten, notulen en jaarverslagen van de afgelopen eeuwen doorzoeken. Op een website met de naam Nederlab zullen onderzoekers vanaf eind 2013 het gedigitaliseerde Nederlandstalige erfgoed van de achtste eeuw tot heden snel kunnen doorvlooien en analyseren. Het is een doorbraak. ,,In andere landen bestaat niets vergelijkbaars", zegt taalkundige Nicoline van der Sijs. Op het Meertens Instituut in Amsterdam is zij de coördinator van Nederlab.

Aan het portaal werken veel eigenaars van gedigitaliseerd drukwerk mee, zoals het Meertens Instituut, de universiteitsbibliotheken, de Koninklijke Bibliotheek en de wetenschappelijke instituten. De Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) geeft ruim 2 miljoen euro subsidie voor Nederlab. Het zal totaal 4 miljoen euro kosten.

,,Met Nederlab komt voor iedereen een gigantische bak gedigitaliseerde pagina's beschikbaar", zegt Hans Bennis, directeur van het Meertens Instituut. Alleen al deelnemer Digitale Bibliotheek voor de Nederlandse Letteren (DBNL) levert 3 miljoen digitale pagina's proza en poëzie van de Middeleeuwen tot nu.

Wat kun je ermee? Van der Sijs: ,,Het woord 'varen' betekende vroeger 'gaan' of 'lopen'. In oude teksten zie je het woord opduiken in combinatie met 'paarden' of 'wagens'. Pas later vind je 'varen' met bijvoorbeeld 'schuiten', dus in de huidige betekenis." Met Nederlab kun je straks de tekst vinden waarin varen voor het eerst met een boot gebeurde, en niet meer met een paard en wagen. En dat leert ons veel over de geschiedenis van de woordvorming in de Nederlandse taal, zegt Van der Sijs.

Zeker zo belangrijk is dat Nederlab een eind moet maken aan wat in deze krant 'het digitale drama' heette. De laatste 20 jaar zijn tientallen miljoenen pagina's in de computer gezet, van teksten in dialect bij het Meertens Instituut tot historische kranten bij de Koninklijke Bibliotheek. Dat kostte ruim 50 miljoen euro, maar taalkundigen en historici die de pagina's willen bestuderen, belanden in een doolhof waar gedigitaliseerde teksten, als ze al te vinden zijn, nauwelijks doorzocht kunnen worden.

Dat komt doordat erfgoedinstituten geen vaste standaard hanteren bij het digitaliseren. Pagina's moeten namelijk na het scannen worden omgezet van een plaatje in een voor de computer leesbare tekst. Deze omzetting verloopt niet foutloos, doordat oude teksten vaak lastig leesbaar zijn. Zo leest de computer het woord 'televisie' in een krant uit 1886, waar in werkelijkheid 'ter visie' staat. De teksten moeten dus worden gecorrigeerd, wat om financiële redenen nogal eens wordt nagelaten. Daar komt bij dat de instituten elk hun eigen digitalisering doen, met eigen software, die niet aansluit op die van andere instituten. Wie de ruim 100 Nederlandse kranten wil doorzoeken, zal daarom 50 websites moeten bezoeken. Het ideaal van één groot 'corpus' - dus één groot digitaal doorzoekbaar blok tekst van boeken, kranten en tijdschriften, is daarmee onmogelijk.

Geesteswetenschappers hebben hun zorgen daarover vaak geuit, en steeds tevergeefs. Hun belangrijkste klacht is dat ze geen gebruik kunnen maken van de ongelooflijke mogelijkheden van het digitale erfgoed. Er zijn incidentele vondsten in gedigitaliseerde teksten, zoals die van het woord 'cadeau' dat in 1798 voor het eerst bleek op te duiken in plaats van 'geschenk'. Maar dit is volgens Van der Sijs ,,een enkel goudklompje", terwijl het digitale erfgoed een onuitputtelijke goudmijn zou moeten zijn voor taalkundigen en historici.

Dat vonden ook de erfgoedinstituten, ze werken in Nederlab nu wèl samen. ,,Ik ben heel gelukkig dat er een einde is gekomen aan de versplintering", zegt Bennis.

Om een idee te krijgen wat Nederlab moest worden, heeft Van der Sijs 150 onderzoekers ondervraagd. ,,Historici hebben liefst zoveel mogelijk teksten tot hun beschikking, zo krijgen ze bijvoorbeeld een goed beeld van een maatschappelijke discussie vóór een nieuwe wet werd aangenomen", zegt Van der Sijs. ,,Er werd bijvoorbeeld al over slaven geschreven voordat in de negentiende eeuw de slavernij werd afgeschaft. Eerst op een neutrale manier, later steeds negatiever. Hoe en wanneer die omslag zich voltrok, is voor historici interessant."

En als Nederlab klaar is, is dat ook vrij snel te onderzoeken. De website biedt namelijk allerlei computerprogramma's voor verschillende manieren van onderzoek naar woorden en woordcombinaties - en dat kan ook op verschillende websites met uiteenlopende software. Daarnaast kunnen wetenschappers de teksten voorzien van voetnoten, die ook voor hun collega's te zien zijn.

Omdat veel bestanden nog niet goed zijn gedigitaliseerd, komen in eerste instantie de vrijwel foutloze teksten van DBNL in het portaal te hangen - en dan stap voor stap andere 'schone' tekstbestanden. Het corpus dat zo ontstaat biedt onderzoekers niet alleen veel nieuw onderzoeksmateriaal, maar verhoogt ook de kwaliteit van de geesteswetenschappen. ,,Het is nu haast ondoenlijk om een onderzoek van een historicus of taalwetenschapper over te doen. Maar het repliceren van onderzoek wordt straks heel makkelijk", zeg Van der Sijs. Je hoeft alleen maar dezelfde zoekcombinaties in te toetsen. ,,De alfawetenschappen krijgen zo meer een bètakarakter."

Iedere wetenschapper - of schrijver, journalist, student, hobbyist - kan een eigen virtuele werkruimte krijgen in het Nederlab. Alle data zullen worden opgeslagen in de 'cloud', de domeinen op internet.

Dat kost geld, net als bijvoorbeeld het up-to-date houden van de software en het onderhoud van het portaal. ,,Zeker wel een ton per jaar", zegt Bennis. ,,Dat is veel voor een klein instituut als het Meertens. Hoe meer het portaal gebruikt wordt, des te meer geld het kost. En ik ga er natuurlijk vanuit dat het veel gebruikt zal worden."

NWO trekt in totaal 15 miljoen euro aan subsidie uit voor apparatuur, dataverzamelingen en software. Behalve voor Nederlab wordt ook geld gereserveerd voor onder meer het verzamelen van hersenweefsel en de ontwikkeling van een neutronenmicroscoop.

Nicoline van der Sijs, pionier voor het digitaliseren van Nederlandse teksten

Nicoline van der Sijs (1955) is een taalkundige, die haar sporen onder meer heeft verdiend met haar onderzoek naar de herkomst van woorden in de Nederlandse taal. Toen Van der Van der Sijs vorig jaar werd benoemd tot Officier in de Orde van Oranje Nassau noemde staatssecretaris Zijlstra (Onderwijs, Cultuur en Wetenschap) haar ,,een van onze meest prominente taalwetenschappers die haar kennis over

etymologie en historische taalkunde heeft neergeschreven in een indrukwekkende serie boeken."
Haar boek De invloed van andere talen op het Nederlands (1996) geldt als het standaardwerk over Nederlandse leenwoorden. „Nicoline van der Sijs verdient een tuil met een gros gouden anjers voor haar dikke Leenwoordenboek", schreef taalexpert Battus daarover in de Volkskrant. Haar Etymologisch woordenboek (1997) en publicaties over onder meer de invloed van het Nederlands op andere talen en de chronologische ontwikkeling van het Nederlands werden eveneens goed ontvangen. Van der Sijs is ook een van de eersten die de mogelijkheden van de digitalisering van teksten zag. Zo tikten op haar initiatief in 2007 honderden vrijwilligers de Statenbijbel (1637) over. Sinds november 2011 werkt Van der Sijs bij het Meertens Instituut in Amsterdam, waar ongeveer 10.000 Nederlandstalige, zeventiende- en achttiende-eeuwse brieven van zeelieden worden gedigitaliseerd.