Nederland in ideeën

101 denkers over inzichten en innovaties
die ons land verander(d)en

Onder redactie van Mark Geels en Tim van Opijnen

MAVEN PUBLISHING

De digitalisering van het Nederlandstalige erfgoed

NICOLINE VAN DER SIJS

Hoogleraar historische taalkunde van het Nederlands in de digitale wereld, Radboud Universiteit Nijmegen; senior onderzoeker bij het Meertens Instituut; auteur van o.a. *Chronologisch woordenboek*, *Taal als mensenwerk: Het ontstaan van het ABN*

In de jaren negentig van de vorige eeuw is men begonnen met de grootschalige digitalisering van het Nederlandstalige erfgoed. Deze technische innovatie, die er uiteindelijk toe zal leiden dat alle Nederlandstalige teksten – van de achtste eeuw tot heden – door de computer kunnen worden doorzocht en geanalyseerd, vormt in mijn ogen een revolutie in de informatievoorziening die grote gevolgen zal hebben voor de kennis en interpretatie van ons talige en culturele verleden. Dit zal het leven van zowel leken als wetenschappers raken. Ook overheid, bedrijfsleven en onderzoeksinstellingen denken er zo over, want zij hebben Cultureel Erfgoed – waaronder ook het talige erfgoed – tot een onderdeel van de Topsector Creatieve Industrie uitgeroepen.

Tot eind vorige eeuw was kennis van de geschiedenis van de Nederlandse taal en cultuur een zorgvuldig bewaard geheim. Iedereen kón er kennis van nemen, maar moest daarvoor moeizaam zijn weg zoeken in bibliotheken en archieven. Door veel lezen kon men geleidelijk, na jarenlange studie, komen tot eruditie en specialisme. Zo puzzelde men globaal uit hoe de Nederlandse taal en cultuur in de loop van de eeuwen waren veranderd. Veel van de bestudeerde onderwerpen waren klein en gespecialiseerd. Het kostte veel tijd en er was literatuurkennis voor nodig om eenvoudige vragen te beantwoorden als: waar komt de uitdrukking 'ergens geen chocola van kunnen maken' vandaan; sinds wanneer

spreken we iemand aan met 'u' in plaats van 'gij'; sinds wanneer leest men keukenmeidenromans en waarom heten die zo?

Die situatie is totaal gewijzigd door de digitalisering van historische teksten. Doordat hele bibliotheken worden gedigitaliseerd, heeft iedereen die over een computer met internet beschikt, toegang tot het Nederlandstalige erfgoed, van elfde-eeuwse religieuze teksten, dertiende-eeuwse literatuur of zeventiende-eeuwse kranten tot verslagen van de Staten-Generaal vanaf 1814.

De digitalisering van oude teksten is een grote maatschappelijke doorbraak: terwijl het vinden van informatie in oude bronnen vroeger was voorbehouden aan een selecte groep, is het nu een privilege voor iedereen geworden. Theoretisch is er geen barrière meer om kennis te nemen van historische teksten en daarin allerlei vragen – van taalkundige, letterkundige en historische aard – op te zoeken. Zo kan iedereen leren van het verleden.

Iedereen? De praktijk blijkt toch weerbarstig. Terwijl vroeger het *vinden* van informatie het probleem was, is nu het vinden van de *juiste* informatie het probleem geworden. Vaak vindt men *te veel* informatie: wie in historische kranten de opvattingen over de letterkundige stroming 'romantiek' zoekt, krijgt bijna 82.000 resultaten. Ernstiger is dat men soms – zonder dat men zich daarvan bewust is – *te weinig* informatie vindt: iemand die wil achterhalen hoe oud het begrip 'geneeskunde' is, moet bijvoorbeeld weten dat hij naar de oudere vorm 'geneesconst' moet zoeken. Ook het wegen van de gevonden informatie vergt kennis van zaken. Hiermee zijn we weer terug bij af: de informatie kan alleen door wetenschappelijk gevormde experts worden verzameld en geïnterpreteerd.

Ook voor die specialisten heeft de digitalisering van het Nederlandstalige erfgoed onverwachte gevolgen. Zo bloeit het vakgebied historische taalkunde weer op. Dit vakgebied was sinds de jaren zestig naar de marge gedrongen door de intrede van de generatieve taalkunde van Noam Chomsky (die stelt dat het menselijk taalvermogen aangeboren is en dat met een eindige reeks regels

een oneindig aantal zinnen kan worden gegenereerd). Nu er zoveel digitale oude teksten beschikbaar komen, blijkt de kennis van historisch taalkundigen van onschatbare waarde voor het ontsluiten en interpreteren van de oude teksten.

Specialisten stellen andere vragen dan leken: ze wenden het gedigitaliseerde talige erfgoed aan voor theorievorming. De digitalisering heeft al geleid tot het ontstaan van enkele nieuwe vakgebieden. Zo worden binnen het vak 'culturomics' culturele trends in kaart gebracht. De Ngram Viewer die dit mogelijk maakt, is eind 2010 geïntroduceerd door de zoekmachine Google. Ook nieuw is de zogenoemde 'nomothetic approach': een benadering waarbij wetenschappers een statistische relatie zoeken tussen taalstructuur en sociale structuur. Uit deze benadering blijkt bijvoorbeeld dat talen met een klein aantal sprekers, een geringe geografische verbreiding en weinig contact met naburige talen een ingewikkelder systeem van verbuigingen en vervoegingen hebben dan talen met een groot aantal sprekers, een grote verbreiding en veel taalcontact. Dit inzicht biedt inspiratie voor nieuw onderzoek naar taalverandering en taalverbreiding.

In mijn ogen staan we nog maar aan het begin van de revolutie die door de digitalisering van het talige erfgoed veroorzaakt zal worden. Momenteel is slechts een klein deel van de gegevens gedigitaliseerd, en de mogelijkheden voor het analyseren en interpreteren staan nog in de kinderschoenen. Het is een kwestie van trial and error, en lang niet alle nieuwe analyses zijn geslaagd. Om me te beperken tot een enkel voorbeeld: bioloog Mark Pagel publiceerde in 2013 in het tijdschrift *Proceedings of the National Academy of Sciences of the United States of America* (*PNAS*) een onderzoek waarin hij op basis van een statistische analyse van veel taaldata concludeerde dat er 15.000 jaar geleden één Euraziatische oertaal was, waarvan hij drieëntwintig woorden reconstrueerde. Zijn conclusie is door taalkundigen neergesabeld: zowel de dataset als de analyse achtten zij onbetrouwbaar.

Desondanks ben ik ervan overtuigd dat de digitalisering van

het talige erfgoed in de toekomst tot veel nieuwe inzichten zal leiden. Wanneer het compleet overgeleverde tekstuele erfgoed is gedigitaliseerd, kunnen onderzoekers voor het eerst systematisch veranderingen in de taal en de cultuur over een langere periode in kaart brengen, en de achterliggende wetmatigheden en patronen van deze veranderingen zoeken. Aan de 'big data' kunnen grote vragen worden gesteld, bijvoorbeeld wat de interactie is tussen veranderingen in de cultuur, maatschappij, letteren en taal. Zo gaat men er binnen de taalkunde de laatste jaren van uit dat veel taalveranderingen het gevolg zijn van taalcontact als gevolg van migratie, dus van veranderingen in de maatschappij. Wordt dit bevestigd als gegevens over migratiecijfers en taalveranderingen in een bepaalde periode aan elkaar worden gekoppeld? Vinden er in tijden van oorlog of andere maatschappelijke onrust meer taalveranderingen plaats dan in tijden van vrede? En welke talige en culturele verschijnselen blijken in de loop van de geschiedenis constant te zijn?

De antwoorden op dit soort vragen zullen er in de loop van de eenentwintigste eeuw toe leiden dat ons inzicht in fundamentele kwesties wordt verdiept – bijvoorbeeld over de vraag welke menselijke eigenschappen zijn aangeboren en welke zijn aangeleerd (het nature-nurturedebat), hoe canons en nationale culturele identiteiten worden geconstrueerd en zich ontwikkelen, en hoe kennis, cultuur en taal worden verbreid. Het digitaliseren van het talige erfgoed zal bovendien leiden tot het stellen van *nieuwe* vragen. En nieuwe vragen leiden onvermijdelijk tot nieuwe antwoorden en nieuwe interpretaties. Welke dat zijn, valt niet te zeggen, want zoals bekend: voorspellen is moeilijk, vooral als het om de toekomst gaat.