Nicoline van der Sijs
Meertens Instituut, Amsterdam

Digitale vergezichten: Nederlab, een laboratorium voor nieuw onderzoek in oude teksten

Toen ik in 1997 voor het *Etymologisch woordenboek* van Van Dale van 30.000 trefwoorden onderzocht wanneer die voor het eerst werden gebruikt, moest ik me grotendeels behelpen met het bladeren in oude woordenboeken: zo leerde ik uit het *Kunstwoordenboek* van Petrus Weiland uit 1824 dat in die periode ondernemingen voor het eerst *riskant* werden genoemd en mensen elkaar verblijdden met *cadeaus*. Voordien waren ondernemingen *gevaarlijk* en gaf men elkaar *geschenken*. Aan digitale hulpmiddelen voor het dateren van Nederlandse woorden had ik op dat moment slechts de beschikking over een incomplete cd-rom van het *Woordenboek der Nederlandsche Taal*. Internet bestond al wel, maar was nog piepjong en bevatte nog weinig, voornamelijk jonge informatie.

Nu, vijftien jaar later, ziet de wereld er totaal anders uit. Bibliotheken, wetenschappelijke instituten en archieven zijn allemaal hard bezig het Nederlandstalige gedrukte erfgoed te digitaliseren en via internet beschikbaar te stellen. Om enkele voorbeelden te noemen: de Digitale Bibliotheek voor de Nederlandse Letteren bevat meer dan 3 miljoen pagina's Nederlandstalige literatuur van de middeleeuwen tot heden, zo bleek uit de beschrijving door René van Stipriaan in het septembernummer 2011 van Neerlandia/Nederlands van Nu. De universiteit van Gent laat haar hele bibliotheek door Google digitaliseren. De Koninklijke Bibliotheek in Den Haag presenteert een historische krantenbank, met een selectie kranten uit 1618 tot 1995, en de *Early Dutch Books Online* (EDBO), met 10.000 boeken uit 1781-1800. Instellingen als het Huygens ING, het Instituut voor Nederlandse Lexicologie en het Gentse Centrum voor Teksteditie en Bronnenstudie geven toegang tot digitale bestanden met teksten uit alle eeuwen. Via de website van het Meertens Instituut kun je informatie over Nederlandse dialecten in heden en verleden vinden.

Belangstellende lezers en onderzoekers weten massaal de weg naar de digitale bibliotheken te vinden, zo blijkt uit de jaarlijks toenemende bezoekersaantallen. Zij zoeken allerlei verschillende gegevens op, bijvoorbeeld: Sinds wanneer komen zogenaamde watermetaforen als *een stroom aan gegevens*, *een golf van geweld*, *een vloedgolf van protesten* en *een tsunami aan immigranten* voor? Was de samenstelling *beresterk* al bekend in de middeleeuwen, toen je in de Lage Landen nog wilde beren tegen het lijf kon lopen? Werd een begrip als *democratie* in de 19e eeuw positief of negatief beoordeeld? Wat kan ik vinden over de geschiedenis van mijn familie?

Overvloed en onbehagen

Aanvankelijk waren de bezoekers blij met ieder goudklompje dat ze uit de oude teksten wisten op te delven. Zoals de ontdekking dat *cadeau* blijkens *Early Dutch Books Online* al in 1798 voorkomt, wat aannemelijk maakt dat het woord is overgenomen in de Napoleontische tijd, toen Frankrijk de Lage Landen beheerste.

Maar naarmate de digitale bestanden vaker worden geraadpleegd, komen er steeds meer feilen aan het licht en ontpoppen de bezoekers zich steeds meer tot kritische consumenten. Na de aanvankelijke euforie zijn we beland in een fase van herbezinning. NRC Handelsblad wijdde op 10 september 2011 een alarmerend artikel aan *Het digitale drama*. Hierin wordt beschreven wat de gevolgen zijn van het ontbreken van standaardisering en samenwerking tussen de verschillende digitaliserende instellingen. Om te beginnen de slechte vindbaarheid en doorzoekbaarheid van de teksten: de historische bestanden worden op een groot aantal plaatsen op internet aangeboden, en niemand heeft een overzicht van wat er zoal is gedigitaliseerd.

Het meest opvallende probleem van al die verspreide gedigitaliseerde bestanden is het kwaliteitsverschil, dat veroorzaakt wordt door gebrekkige optische tekenherkenning. De meeste bibliotheken – de DBNL en wetenschappelijke onderzoeksinstellingen vormen een positieve uitzondering – digitaliseren hun drukwerk door boeken, kranten en tijdschriften onder een scanner te leggen en vervolgens te lezen met een programma voor optische tekenherkenning. Het tekenherkenningsprogramma maakt helaas, vooral in oudere teksten, veel fouten: letters en woorden worden door de computer niet goed herkend. Als je in de historische krantenbank van de KB Den Haag het woord *televisie* intypt en vervolgens een treffer krijgt voorgeschoteld in een tekst uit 1886, zul je waarschijnlijk snel het origineel gaan raadplegen, maar daar blijkt *ter visie* te staan. Als je voor *beresterk* een tekst uit 1921 vindt, ben je geneigd die datering voor waar aan te nemen. Als je de moeite neemt de door de computer verstrekte context na te gaan, dan lees je deze opmerkelijke zin: "De hoeren v.

BERESTERK en TEENSTRA (V. D.) zijn niet tevreden met het antwoord van den Minister." Bij raadpleging van het gedrukte origineel blijkt het te gaan om de *heren* (parlementariërs) Van *Beresteijn* en Teenstra. Voor wie het interesseert: de oudste datering voor *beresterk* is vooralsnog 1955. In middeleeuwse teksten is geen spoor van die samenstelling te vinden.

Ook de kwaliteit van de metadata bij de teksten laat te wensen over en varieert per bibliotheek en instelling. Metadata zijn extra gegevens over de brontekst, zoals de naam van de auteur, jaar en plaats van publicatie, en exacte titelgevens. Die metadata zijn noodzakelijk om zoekresultaten te sorteren, bijvoorbeeld chronologisch, op plaats of auteursnaam.

Nieuwe computertoepassingen, nieuwe eisen

De problemen blijken het duidelijkst uit een concreet voorbeeld. Stel: iemand wil nagaan hoe de populariteit van een bepaalde literaire schrijver – laten we zeggen Vondel – door de eeuwen heen heeft gefluctueerd, in fictie en in non-fictie. Dat wil hij doen door de computer te laten tellen hoe vaak het werk van Vondel in de verschillende eeuwen wordt geciteerd. Dat lijkt een fluitje van een cent, maar blijkt in de praktijk ondoenlijk. Het is namelijk niet mogelijk een totaalbeeld te krijgen. Je kunt slechts een groot aantal verschillende websites afzonderlijk raadplegen, die allemaal verschillende zoekmogelijkheden bieden, en dus ook verschillende, niet met elkaar vergelijkbare zoekresultaten! Lang niet alle websites geven de mogelijkheid de gegevens chronologisch te ordenen, en de meeste websites maken geen verschil in tekstsoort, zoals fictie en non-fictie. En dan laten we nu maar buiten beschouwing dat we alleen citaten van *Joost* van den Vondel willen tellen, en niet van zijn jongere broer *Willem* ...

De beschikbaarheid van omvangrijke digitale bestanden heeft geleid tot een nieuwe discipline binnen de computertechnologie, namelijk de ontwikkeling van *tools*, computerprogramma's die automatische analyses van teksten maken. Dergelijke programma's tellen bijvoorbeeld hoe vaak een bepaald woord of een bepaalde naam (zoals Vondel) in teksten voorkomt, en geven de resultaten weer in grafiekvorm of als een woordwolk (afbeelding 1). Ook kunnen die programma's automatisch tekstdelen met elkaar vergelijken. Automatische tekstvergelijking wordt bijvoorbeeld toegepast in plagiaatherkenners, waarmee ingeleverde werkstukken op scholen en universiteiten worden gecontroleerd. Ook de sluwe student die denkt weg te komen door bij het kopiëren enkele woorden in synoniemen te veranderen, komt van een koude kermis thuis, want inmiddels zijn de eerste parafraseherkenners op de markt.

Corpusanalysetools bepalen eigenschappen van een bepaald tekstbestand, bijvoorbeeld welke woorden en woordcombinaties typerend zijn en wat het aantal lettergrepen, de woordlengte en de zinslengte is. Ter illustratie is de tekst van dit stuk geanalyseerd door het programma Textalyser (afbeelding 2). Met behulp van een dergelijk analyseprogramma kun je verschillende tekstbestanden met elkaar vergelijken. Dat maakt spannend nieuw onderzoek mogelijk: zo kun je aan de hand van omvangrijke tekstcorpora bekijken welke verschillen er bestaan tussen het Belgisch-Nederlands en het Nederlands-Nederlands in de 21e eeuw en of die verschillen groter of kleiner zijn geworden op ijkpunten als 1950, 1850 en 1750. Zo ook kan de bewering dat de taal van de Statenvertaling representatief is voor het Nederlands zoals dat rond 1637 werd geschreven, worden getoetst aan de feiten door de tekst van de Statenvertaling te vergelijken met bijvoorbeeld een corpus literaire teksten uit dezelfde periode, of met een corpus geleerdenteksten of egodocumenten.

Afbeelding 1: een woordwolk van dit artikel

Textalyser Results — The complete results, incuding compexity factor, and other features

Total word count:	1475
Number of different words :	723
Complexity factor (Lexical Density) :	49%
Readability (Gunning-Fog Index) : (6-easy 20-hard)	13.4
Total number of characters :	12943
Number of characters without spaces :	10532
Average Syllables per Word :	1.98
Sentence count :	93
Average sentence length (words) :	19.84

Frequency and top words:			
Word	Occurrences	Frequency	Rank
van	66	4.5%	1
het	46	3.1%	2
een	43	2.9%	3
voor	21	1.4%	4
met	18	1.2%	5
als	18	1.2%	5
zijn	18	1.2%	5
uit	17	1.2%	5
teksten	16	1.1%	6
dat	16	1.1%	6

Afbeelding 2: een statistische analyse van dit artikel

Maar als we de analyse van teksten overlaten aan computerprogramma's, moeten we wel blind kunnen vertrouwen op de kwaliteit van de onderliggende gegevens. En die is, zoals gezegd, voor oude teksten nog onvoldoende.

Nederlab, een gebruikersvriendelijk laboratorium

Nu het aantal historische tekstbestanden explosief groeit en informatici steeds meer computerprogramma's ontwerpen waarmee die bestanden kunnen worden geanalyseerd, kunnen geesteswetenschappers allerlei interessante nieuwe onderzoeksvragen gaan stellen. Ook vragen die zich uitstrekken over een langere periode, liggen in het verschiet. Onderzochten we vroeger wat de invloed van een enkele auteur als Jacob van Maerlant op Noord-Nederlandse schrijvers in de middeleeuwen was, tegenwoordig willen we in breder verband achterhalen welke invloed Zuid-Nederlandse schrijvers hebben uitgeoefend op de Noord-Nederlandse literatuur van de middeleeuwen tot heden. Of wat de invloed van verschillende groepen immigranten op de Nederlandse taal en cultuur is geweest in de loop van de eeuwen.

Momenteel zijn dergelijke grote vragen nog niet te beantwoorden. Maar de mogelijkheden komen wel steeds dichterbij. Het Meertens Instituut heeft samen met een aantal andere onderzoeksinstellingen in 2011 het initiatief genomen om te bekijken welke voorzieningen noodzakelijk zijn. Daarvoor is een grote groep historisch taalkundigen, letterkundigen en historici uit Nederland en België geconsulteerd. De conclusie van die deskundigen was dat onderzoekers en studenten dringend behoefte hebben aan een centrale plaats – een portaal – van waaruit alle digitale bestanden met eenvoudige computerprogramma's kunnen worden doorzocht en geanalyseerd. De tekstbestanden moeten daarvoor op elkaar worden afgestemd, de metadata moeten worden geüniformeerd, en de tekstkwaliteit moet zeer hoog zijn: teksten met tekstherkenningsfouten moeten worden gecorrigeerd of kunnen worden uitgesloten uit het onderzoek.

Om die gewenste onderzoeksomgeving te realiseren is in november 2011 een aanvraag ingediend bij de Nederlandse Organisatie voor Wetenschappelijk Onderzoek NWO voor de oprichting van *Nederlab – Laboratorium voor onderzoek naar de veranderingspatronen in de Nederlandse taal en cultuur*. Het idee is dat Nederlab toegang geeft tot het complete gedigitaliseerde Nederlandstalige erfgoed, van de achtste eeuw tot heden, waar ook geschreven of gepubliceerd – dus Nederlandse teksten uit Nederland, Vlaanderen, maar ook uit bijvoorbeeld Suriname en Indonesië. Daarnaast biedt Nederlab een scala aan computertechnieken als hulpmiddel voor het doorzoeken en analyseren van de teksten.

De aanvragers zijn ervan overtuigd dat de door Nederlab geleverde infrastructuur baanbrekend onderzoek mogelijk maakt dat zal leiden tot een groot aantal nieuwe inzichten in onze

Op het STOKJE van
JOHAN VAN OLDENBARNEVELDT,
Vader des Vaderlands.

MYN wenſch behoed' u onverrot,
O Stok en ſtut: die geen Verrader,
Maar 's Vrijdoms Stut, en Hollands Vader
Geſtut hebt op dat Moordſcharvot,
Toen hij voor 't bloedig Zwaard moeſt knielen.
Vervordeeld als een Seneka,
Door Neroos haat en ongelijk,
Tot droeffenis der braafſte Zielen.
Gy zult noch Jaren achter een
Den uytgang van dien Held getuygen;
En hoe Geweld het Recht durfd buygen,
Tot ſmaad der onderdrukte Steen.
Hoe dikwijls ſtrekt gy, onder 't ſtappen
Naar 't Hof der Staten ſtadig aan,
Hem voor een derde voet in 't gaan,
En klimmen op de hooge trappen,
Als hy belaſt van ouderdom,
Papier en Schriften overleende,
En onder 't laſtig landſpak ſteende!
Wie ging, zoo krum gebukt, nooyt krom!
Gy ruſte van uw trouwe plichten,
Naar 't ruſten van dien ouden Stok,
Geknot door 't Bloedraads bitt'ren wrok:
Nu ſtut en ſtijft gij noch mijn Dichten.

J. v. VONDEL.

Oude teksten zijn moeilijk digitaal te doorzoeken doordat teksherkennings-programma's moeite hebben om letters juist te lezen. (*Het stockske van Joan van Oldenbarnevelt* (1657), door Joost van den Vondel) *[bron: http://nl.wikipedia.org/wiki/Bestand:Op_het_stokje..jpg]*

kennis van de geschiedenis van onze taal en cultuur. In de gedigitaliseerde teksten vinden we daarvan immers de neerslag. Taal en cultuur zijn voortdurend aan verandering onderhevig, maar dat gebeurt niet altijd in hetzelfde tempo. Onderzoekers willen erachter komen hoe dat komt. Wat is de relatie tussen veranderingen in de cultuur, maatschappij, letteren en taal? In hoeverre zijn taalveranderingen het gevolg van taalcontact als gevolg van migratie en immigratie, dus veranderingen in de maatschappij? Hoe wordt taal gebruikt als sturende en vormende factor in politieke en sociale veranderingen: welke metaforen (drankzucht als *kanker van de maatschappij*, slavernij als *zonde*) worden en werden gebruikt om de publieke opinie te sturen? In hoeverre vormen literaire werken een weerspiegeling van de contemporaine cultuur? De antwoorden op dat soort nieuwe en complexe vragen zullen helpen bij het begrijpen van de manier waarop canons en een nationale culturele identiteit ontstaan, en hoe kennis, cultuur en taal zich verbreiden.

Nederlab is in de eerste plaats, maar niet alleen, bedoeld voor wetenschappers. Studenten, scholieren, journalisten, schrijvers, iedereen die informatie zoekt over de geschiedenis van de Nederlandse taal en cultuur, zal met Nederlab aan de slag kunnen. Dankzij Nederlab zullen de reeds gedigitaliseerde bestanden beter en vaker benut worden dan momenteel het geval is. Als de subsidie wordt toegekend, zal waarschijnlijk eind 2013 een eerste versie van de website in de lucht zijn. Dat moment zal niet ongemerkt voorbijgaan …