

## Corpuslexicografie ligt binnen ieders handbereik<sup>1</sup>

Ewoud Sanders

Misschien zijn dit niet de goede woorden om een bijdrage in een academische huldebundel mee te beginnen, want ze klinken nogal bevlogen en wellicht zelfs een beetje hoogdravend, maar het is niet anders: ik heb een droom.

Ik heb een droom that one day, en het liefst een beetje snel, onderzoekers van de Nederlandse taal kunnen beschikken over een zo compleet mogelijke bibliotheek waarin je alle publicaties die tot nu toe over het Nederlands zijn verschenen, op een geavanceerde manier digitaal kunt doorzoeken. De droom heeft nog een staartje, namelijk dat er in die grote ‘digiTaalbibliotheek’ een forse zijkamer is ingericht waarin je alle Nederlandstalige encyclopedieën op een geavanceerde manier digitaal kunt doorzoeken.

Nou kun je verschillende dingen doen met een droom. Je kunt glimlachen om het onrealistische gehalte ervan – *alle* publicaties over het Nederlands, ha!, dat zijn er nogal wat – of je kunt besluiten op zijn minst een *poging* te wagen om je droom te verwezenlijken.

Ik heb besloten het laatste te doen, en daarom ben ik een kleine drie jaar geleden begonnen met het digitaliseren van mijn bibliotheek, die – dromen komen nooit helemaal uit het niets – twee grote afdelingen bevatte, namelijk een nagenoeg complete verzameling Nederlandstalige encyclopedieën van het begin van de 18de eeuw tot omstreeks 1970, en een zeer forse collectie boeken en tijdschriften over het Nederlands, waaronder honderden woordenboeken.

Over die collectie Nederlandstalige encyclopedieën kan ik kort zijn: die was groter dan ik kon behappen. Encyclopedieën bevatten vaak uitklapbare afbeeldingen en kaarten, en die kun je prima scannen, maar dat is extra veel werk. Dus heb ik besloten om dit deel van mijn collectie – tientallen meters boekenplank – in z’n geheel af te staan aan de Koninklijke Bibliotheek (KB), onder de voorwaarde dat zij al die encyclopedieën binnen één of twee jaar scannen en ontsluiten.

Dat deel van de droom zal dus zeker uitkomen, want de KB is een groot instituut dat veel geld steekt in (massa)digitalisering.<sup>2</sup>

Hoe zit het met het andere deel van die droom? Gaat het inderdaad lukken om dat te laten uitkomen? En wat kun je als lexicograaf of taalonderzoeker eigenlijk doen met een grote collectie gescande boeken en tijdschriften? Wat zijn, voor een instituut of voor een individuele onderzoeker, de voordelen van zo’n collectie boven internet, het grootste en meest diverse taalcorpus aller tijden?

### DigiTaalbibliotheek

Eerst iets over gedigitaliseerde Nederlandstalige boeken in het algemeen. Negen jaar geleden, begin 2000, is de Digitale Bibliotheek voor de Nederlandse Letteren (DBNL) begonnen met het digitaliseren van boeken en tijdschriften. De website bevat nu ruim 3.500 boeken en tijdschriften, samen goed voor ongeveer 1,2 miljoen pagina’s.

Een andere grote partij is Google, die op forse schaal boeken is gaan scannen uit 27 universiteitsbibliotheken. Voor ons taalgebied is de samenwerking van Google met de universiteitsbibliotheek van Gent het belangrijkste. Het is de bedoeling dat Google de komende

---

<sup>1</sup> Met dank aan Hans Bennis, Joep Kruijsen, Marc van Oostendorp en Piet van Sterkenburg voor hun commentaar op de eerste versie van dit artikel.

<sup>2</sup> De twee grootste digitaliseringsprojecten van de KB zijn momenteel: de Handelingen en Kamerstukken van de Staten-Generaal van 1814 tot 1995 met 2,3 miljoen pagina’s (zie [www.statengeneraaldigitaal.nl](http://www.statengeneraaldigitaal.nl)) en de Databank Digitale Dagbladen (8 miljoen krantenpagina’s vanaf 1618). Zie: <http://www.kb.nl/hrd/digi/ddd/>. Het eerste project moet in 2010 zijn afgerond, het tweede eind 2011.

vijf jaar 300.000 boeken uit de Gentse collectie gaat scannen – boeken die zijn verschenen van de 16de eeuw tot 1869. Die einddatum, die is vastgesteld door Google, moet problemen voorkomen met het auteursrecht, dat voorschrijft dat de rechten op een boek pas zeventig jaar na de dood van een auteur vervallen.<sup>3</sup>

Als het aanbod op internet al zo groot is en alleen maar groeit, waarom zou je dan veel tijd, geld en energie steken in de aanleg van een specialistische digitale taalbibliotheek?

Dat heeft te maken met enkele bezwaren die kleven aan de collecties van de DBNL en Google. Ik zal de voornaamste bezwaren hier kort opsommen.

### **Nadelen van de DBNL**

— De collectie van de DBNL mag groot lijken, in feite is ruim 3.500 titels en 1,2 miljoen pagina's in negen jaar niet veel, zeker niet als je bedenkt dat hier miljoenen euro's subsidie in zijn gestoken (de DBNL kreeg aanvankelijk 288.000 euro per jaar; voor een periode van vijf jaar krijgt ze nu 900.000 euro per jaar).

— Bij de DBNL ligt de nadruk sterk op de letterkunde. Taalkunde komt nauwelijks aan bod. De toestroom van taalkundige titels heeft zelfs een paar jaar helemaal stilgelegen omdat de subsidiegever – de Taalunie – dit toen verordonneerde.

— De DBNL laat boeken digitaliseren die door commissies zijn geselecteerd. Het doel van de site is om een soort canon van de Nederlandse literatuur te presenteren. Vanuit cultuurhistorisch perspectief is hier veel voor te zeggen, maar voor taalonderzoek wil je een zo breed mogelijk aanbod kunnen doorzoeken; niet alleen goede romans van bekende schrijvers, maar juist ook het werk van vergeten auteurs die bijvoorbeeld veel spreektaal in hun boeken vastlegden, van romans in dialect, van erotische werkjes, enzovoorts. Juist niet-canonieke teksten kunnen grote verrassingen opleveren – ook voor allerlei ander onderzoek trouwens.

— Bij het digitaliseren van boeken kiest men doorgaans uit twee opties. Men presenteert alleen de *tekst* uit een boek, of men presenteert *afbeeldingen* (in vakjargon *images* genoemd) van de oorspronkelijke bladzijden, met daaronder, op een tweede laag, de *tekst*. Een nadeel van de DBNL is dat van vrijwel alle boeken alleen de tekst wordt gepresenteerd, dus zonder images. Hiermee gaat informatie verloren die je uit de opmaak kunt halen. Door de afwezigheid van de images kun je bij twijfelgevallen – stond er in het origineel werkelijk *hoosd*, *Amfteldam* en *fchrijve*? – niet nakijken.

### **Nadelen van Google Books**

— Google Books bevat momenteel ruim zeven miljoen boeken en er komen dagelijks nieuwe titels bij. Ik vind dit een geweldig project, maar de collectie is extreem rommelig georganiseerd. Probeer bijvoorbeeld eens van een meerdelig werk de verschillende delen te vinden. Van het elfdelige *Nederduitsch taalkundig woordenboek* van Petrus Weiland zijn, met veel moeite, slechts vier delen te vinden, die niet naar elkaar verwijzen. Zo zijn er veel meer voorbeelden te geven. De boeken worden per plank gedigitaliseerd. Het is niet aannemelijk dat de bibliotheken die met Google samenwerken toevallig allemaal zeer incomplete reeksen hebben staan.

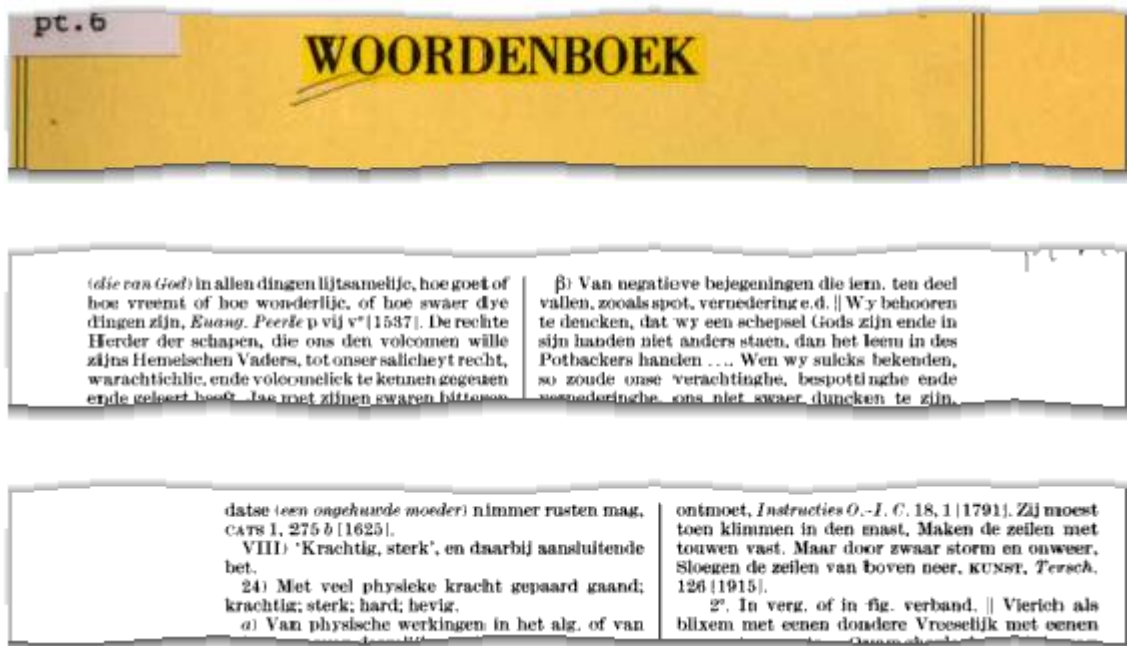
— Bij publicaties vanaf grofweg 1850 kun je met optische tekenherkenning (OCR) goede tot zeer goede resultaten bereiken. Bij oudere boeken holt die kwaliteit achteruit als je geen

---

<sup>3</sup> Voor meer informatie over de samenwerking tussen Gent en Google, zie Ewoud Sanders, 'Het ftinkdier flaaft', in: *NRC Handelsblad* (katern Wetenschap & Onderwijs) 23-2-2008, of: <<http://tinyurl.com/ftinkdier>>.

bewerkingsslagen toepast (in oude teksten lijkt de *s* bijvoorbeeld sterk op een *f*). Zoals gezegd stopt Google bij de Gentse collectie in 1869. Het is duidelijk dat Google geen verbeteringen aanbrengt in de OCR. Bij de oude boeken is die dan ook erg slecht.<sup>4</sup>

— Om problemen met het auteursrecht te voorkomen, krijg je bij boeken die grofweg na 1880 zijn gepubliceerd, slechts fragmenten te zien. Wie bijvoorbeeld *woordenboek* zoekt bij Google Books, krijgt momenteel als eerste vindplaats een (losse?) aflevering uit 1972 van het *Woordenboek der Nederlandsche Taal*. Het woord *woordenboek* komt volgens Google Books dertig keer in deze bron voor. Je krijgt drie fragmenten te zien. Hoe je andere 27 vindplaatsen moet benaderen, is volkomen onduidelijk.



Eerste resultaat van de zoekopdracht 'woordenboek' bij Google Books. Het gezochte woord staat alleen in het eerste fragment.

Over al deze bezwaren valt veel meer te vertellen, maar dat ga ik hier niet doen. Hier is het van belang om nog enkele beperkingen te noemen die gelden voor beide digitale boekencollecties.

### Nadelen van DBNL én Google Books

— De zoekmogelijkheden zijn zeer beperkt. Je kunt woorden bijvoorbeeld niet trunkeren. Zoekacties als *\*fiets, fiets\** of *\*fiets\** zijn onmogelijk. Ook is het niet mogelijk om thematisch of met jokertekens binnen een woord te zoeken.

— De zoekresultaten laten zich niet of slechts gebrekkig sorteren.

— Voor allerlei vormen van wetenschappelijk onderzoek is het prettig als je kunt zeggen dat je conclusies zijn gebaseerd op een statisch corpus. De DBNL en Google Books zijn dynamisch. Het is niet mogelijk uit deze collecties een bepaalde selectie vast te houden. Het

<sup>4</sup> Aan de verbetering van optische tekenherkenning bij oude teksten wordt gewerkt door IMPACT, een wetenschappelijk project dat wordt gefinancierd door de Europese Commissie. Voor meer informatie zie <<http://www.impact-project.eu>>.

gevolg is dat een bewering van vandaag ('we hebben dit woord hier niet gevonden'), morgen achterhaald kan zijn.

### **Opbouw**

Goed, tot zover de beperkingen van de grootste digitale boekencollecties die ons momenteel ter beschikking staan. Enkele van deze bezwaren – slechte kwaliteit van de OCR bij de oudere teksten, beperkte zoekmogelijkheden – gelden ook voor de historische krantenarchieven op internet, die ik hier verder buiten beschouwing laat.

Zoals gezegd heb ik een kleine drie jaar geleden besloten een poging te wagen om mijn droom – een zo compleet mogelijke digitale bibliotheek van publicaties óver het Nederlands – te verwezenlijken. Hoe heb ik dat aangepakt?

1. Ik ben zelf, met hulp van enkele scholieren, op grote schaal gaan scannen. Ik ben mijn eigen boeken gaan scannen en boeken van derden. Ik heb enkele taalauteurs benaderd met het verzoek mij een complete set van hun publicaties te geven, die ik in ruil daarvoor digitaliseerde en indexeerde. Ik heb enkele antiquariaten, waaronder De Slegte, benaderd met het verzoek om boeken die anders zouden worden weggegooid – opslagruimte is duur – aan mij te geven. Ik ben voor mij interessante partijen boeken gaan opkopen. Ik heb enkele vrienden geënthousiasmeerd om ook te gaan scannen.

2. Voor nieuwe publicaties over het Nederlands heb ik uitgevers en/of auteurs benaderd met het verzoek mij een pdf te sturen van de definitieve drukproef.

3. Ik ben boeken gaan downloaden die vrij beschikbaar zijn op internet, met de DBNL als grootste leverancier.<sup>5</sup> Daarnaast ben ik structureel artikelen over het Nederlands en woordenlijsten gaan *harvesten*, zoals dit in het internetjargon wordt genoemd. Je kunt wel lijstjes van hyperlinks gaan bijhouden, maar in de praktijk werken die vaak na een tijdje niet meer, bijvoorbeeld omdat de site is verdwenen. Tevens ben ik gedigitaliseerde bronnen die op dvd zijn uitgebracht, gaan verzamelen. Door dergelijke bronnen op je eigen pc te zetten, kun je er geavanceerde indexerings- en zoeksoftware op loslaten. Ik kom hier zo op terug.

### **Gestructureerd corpus**

Deze aanpak heeft in krap drie jaar tijd geleid tot een bibliotheek die momenteel 18.500 documenten telt, samen goed voor 3,2 miljoen bladzijden. Het gaat hier om een gestructureerd corpus van 2 miljard woorden, een corpus dat snel groeit, want wekelijks komen er tussen de honderd en honderdvijftig boeken bij.

Voordat ik verder ga met de beschrijving van de karakteristieken van deze bibliotheek, de indeling en de geavanceerde zoekmogelijkheden, eerst de belangrijkste vraag: waarom heb ik dit eigenlijk gedaan? Is dit niet een volledig uit de hand gelopen privéproject?

Ja, dat is het, maar ik heb het gedaan omdat ik merkte dat je jezelf als onderzoeker al snel beperkingen oplegt.

Een voorbeeld: in mijn bibliotheek had ik twee planken met spreekwoordenboeken. Op die planken stonden ruim 200 publicaties. Maar als ik iets wilde opzoeken over een

---

<sup>5</sup> Sinds kort zijn de meeste boeken bij de DBNL als pdf te downloaden (ook hiervoor geldt: alleen als tekst, zonder de images). Buitengewoon onpraktisch daarbij is dat je niet een compleet boek binnenhaalt, want het voorwerk (inleiding, inhoudsopgave) staat niet in de pdf, iets waar veel gebruikers geen erg in hebben. Als je bijvoorbeeld *Opperlandse taal- en letterkunde* van Battus downloadt, staat weliswaar in de downloadlink 'pagina 1 tot en met 100', maar na pagina 1 volgt pagina 5 en vervolgens pagina 11.

spreekwoord, keek ik doorgaans slechts in vijf boeken – boeken waar ik in het verleden het meest aan had gehad. Die andere 195 boeken waren in geen jaren van de plank geweest.

Zo had ik ook hele reeksen taaltijdschriften staan – ruim twee kasten vol – die ik vrijwel nooit raadpleegde, bijvoorbeeld omdat er geen register op was gemaakt. Ik had ze ooit wel een keer systematisch doorgebladerd, maar je zoekvragen veranderen voortdurend en het ontbrak me simpelweg aan de tijd om telkens weer die tientallen meters tijdschriften door te nemen.

De kwaliteit van taalonderzoek, is mijn stellige overtuiging, stijgt sterk als je conclusies zijn gebaseerd op zoveel mogelijk relevante data uit gestructureerd onderzoek. Gestructureerd onderzoek komt binnen handbereik als je kunt beschikken over een groot, flexibel corpus dat je – zo nodig per zoekvraag – kunt aanpassen.

Ziedaar het antwoord op de vraag waarom ik doorga met dit privéproject, hoewel ik als eerste zal toegeven dat het qua tijd en kosten volledig uit de hand is gelopen.

## **Afdelingen**

Goed, ik beschik momenteel dus over een digiTaalbibliotheek van 2 miljard woorden en ruim 18.500 documenten.

Alle gescande boeken worden opgeslagen in pdf, een formaat dat in 2008 door het internationaal normalisatie-instituut ISO is uitgeroepen tot officiële standaard (ISO 32000-1). Dit is van belang, want pdf is hiermee officieel een zogenoemde Open Standaard, wat wil zeggen dat het bestandsformaat onafhankelijk is van een bepaalde fabrikant. Ook de digitale duurzaamheid is gegarandeerd.

De pdf's bevatten twee lagen. Op de bovenste laag zie je de *image*, de afbeelding van de oorspronkelijke pagina. Op de tweede laag staat de tekst, die tot stand is gekomen door OCR. Door de image te laten zien blijft informatie over de oorspronkelijke opmaak – cursiveringen, spatiëringen enzovoorts – behouden. Bovendien: mocht er bij het OCR'en een herkenningsfout zijn gemaakt, en dat komt natuurlijk voor, dan kun je op de afbeelding van de oorspronkelijke pagina zien wat er had móéten staan.<sup>6</sup>

Alle titels van de documenten beginnen met een jaartal. Bijna altijd is dat het jaar van uitgave of het oorspronkelijke jaar van uitgave. Bij uitgaven van dagboeken of briefwisselingen kan dat de periode zijn waarin de oorspronkelijke bronnen zijn geschreven (een voorbeeld van zo'n titel: 1957-1968\_Kan, Wim\_Dagboeken (1988)). Dat alles is van belang omdat je de bronnen vervolgens chronologisch of omgekeerd chronologisch kunt doorzoeken.

De digiTaalbibliotheek telt twee grote afdelingen:

1. Primaire bronnen. Het gaat hier om ruim 5.000 romans en bundels met liedjes en poëzie van Nederlandstalige schrijvers. Ik heb ernaar gestreefd het werk van belangrijke auteurs min of meer compleet te krijgen, maar verder ben ik alles gaan scannen wat ik te pakken kon krijgen: streekromans, vergeten auteurs, derderangsschrijvers enzovoorts. Juist vanuit het idee dat een zo breed mogelijks corpus extra interessant is, hanteer ik nauwelijks of geen selectiecriteria.

2. Secundaire bronnen. Een collectie van duizenden publicaties over het Nederlands. Te denken valt aan vrijwel alle taaltijdschriften die sinds het begin van de 19de eeuw over het Nederlands zijn verschenen, een grote collectie naamkundige publicaties, honderden woordenboeken en woordenboekjes, enzovoorts.

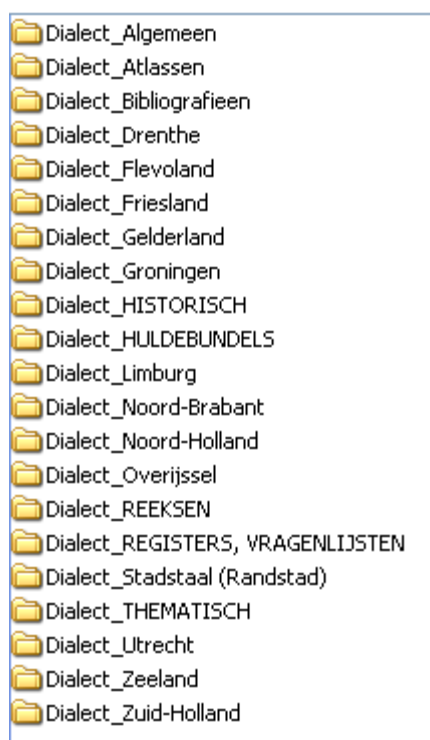
---

<sup>6</sup> Met hetzelfde gemak kun je gescande boeken uitlezen als platte tekst. Alle informatie over de opmaak gaat hiermee verloren, maar platte tekst is handig als je teksten wilt importeren in een database.

Die twee hoofdafdelingen zijn onderverdeeld in 145 thematische mappen. Het lijkt wellicht saai om hierover uit te weiden, maar hieronder zal ik het grote voordeel van zo'n mappenstructuur uiteenzetten. Hier enkele voorbeelden van thematische mappen:

— **Etymologie.** Deze map telt momenteel 200 gedigitaliseerde boeken, die zijn verdeeld over twee hoofdmappen: Etymologie\_Buitenlands en Etymologie\_Nederlands. In de map Buitenlands zitten submappen voor het Afrikaans, Frans, Duits, Engels, Spaans enzovoorts.

— **Dialecten en stadstalen.** De map 'Dialecten en stadstalen' telt momenteel ruim 450 titels, die zijn verdeeld per provincie (de Vlaamse dialecten staan in een aparte map).



Indeling van de map 'Dialecten en stadstalen'.

— **Primaire bronnen proza.** De map 'Primaire bronnen proza' telt momenteel ruim 4.800 titels, die zijn onderverdeeld per tijdvak. Zo zijn er mappen voor de tijdvakken 1501-1700, 1701-1800, 1801-1900, 1901-1950, en van 1951 tot nu.<sup>7</sup>

Hoe kun je deze bronnen nu geavanceerd doorzoeken? Door er indexeringssoftware op los te laten. De afgelopen jaren heb ik gezien dat taalinstututen bakken geld uitgeven aan de ontwikkeling, vaak in eigen huis, van slimme software om grote hoeveelheden tekst te kunnen doorzoeken. Natuurlijk heeft het nut als wetenschappelijke instituten onderzoeken hoe je dergelijke instrumenten bouwt. Het nadeel is echter dat je gegevens worden vastgeklonken in een softwareomgeving die goed moet worden onderhouden en snel kan verouderen.

---

<sup>7</sup> Auteursrechtelijk is het toegestaan een boek voor thuisgebruik – of voor gebruik binnen een instelling – te digitaliseren. Dit maakt het mogelijk om ook boeken uit de tweede helft van de 20ste eeuw te digitaliseren. Sterker nog: met ruim 3100 titels is dit bij mij, in de sectie 'primaire bronnen', de grootste afdeling en er komen wekelijks tientallen titels bij. De kwaliteit van de OCR bij dit soort moderne boeken is zeer goed: volgens de makers van OCR-software wordt zeker 99,8 procent van de tekens correct herkend.

Doorgaans zijn er slechts een paar mensen, meestal de oorspronkelijke ontwikkelaars, die helemaal doorgronden hoe een en ander in elkaar steekt, en als die er niet meer zijn, kun je de data niet meer benaderen.<sup>8</sup>

Ik heb ervoor gekozen mijn data – de gescande boeken en tijdschriften – zo vrij mogelijk te houden en slechts te benaderen met een indexeringsprogramma. Ik heb zeven indexeringsprogramma's getest, en als beste en meest gebruikersvriendelijke kwam Isys Desktop uit de bus, een programma dat wereldwijd door ruim 14.000 bedrijven en instellingen wordt gebruikt, waaronder veel grote justitiële instellingen.<sup>9</sup>

Ik vind Isys Desktop een prachtig programma, maar ik wil niet dat dit artikel uitloopt op een reclameboodschap. Wat ik hier wil laten zien is hoe nuttig zo'n grote collectie pdf's kan zijn voor met name taalonderzoek. Ik zal de belangrijkste mogelijkheden hier kort opsommen.

### Zoeken met Isys Desktop

Je kunt met Isys op drie manieren zoeken: Menu-Assisted, Command Based en Web Style. Handig voor taalonderzoek zijn onder meer de volgende functies:

— Je kunt niet alleen met de operatoren and, or, not zoeken, de 'gewone' booleaanse operatoren, maar ook met o.a. near, except, butnot en xor (dit is: het document moet de eerste of de tweede zoekterm bevatten, maar niet allebei).

— Je kunt zoeken op 'exacte formuleringen'. De zoekopdracht *fons is een* levert bijvoorbeeld de onderstaande zin op. Om u een indruk te geven van de snelheid: deze formulering werd binnen twee seconden gevonden in ruim 4.800 doorzochte romans, een deelcorpus van 320 miljoen woorden.<sup>10</sup>

**Fons was met z'n ouwe moeder en z'n broer op het dorp komen wonen, in het begin van de winter. Zij kwamen van hetzelfde gehucht als Merijntjes vader en zo waren ze met de Gijzens spoedig nader bevriend geworden. De simpele Fons was een merkwaardig en onbegrijpelijk wezen. Geestelijk was hij na zijn vierde of vijfde jaar niet meer gegroeid. Lichame-**

Bron: A.M. de Jong: *Merijntje Gijzen's jeugd*, 1931, p. 406.

— Je kunt woorden trunckeren. *\*fiets, fiets\**, *\*fiets\** leveren lange lijsten op met veel samenstellingen. Alle resultaten zijn alfabetisch gerangschikt, met achter het gevonden woord het aantal hits. Per woord zie je dus meteen de frequentie in het corpus dat je doorzoekt – een buitengewoon handig lexicografisch hulpmiddel.

---

<sup>8</sup> Er zijn voorbeelden bekend van uitgevers die hun data voor veel geld door een commercieel bedrijf in een softwareomgeving hadden laten vastzetten. Nadat het softwarebedrijf over de kop was gegaan, waren zij alle data kwijt. Iets dergelijks is onder andere gebeurd in de samenwerking tussen uitgeverij Van Dale en het softwarebedrijf AND.

<sup>9</sup> Voor meer informatie, zie: <<http://www.isys-search.com/technology/isysdesktop/>>. De belangrijkste feiten: met Isys kun je ruim 200 bestandsformaten doorzoeken, inclusief databases in sql-formaat (zoals FileMaker). Je kunt de data op je eigen pc zetten of op een server, waarna de indexen door oneindig veel gebruikers tegelijk kunnen worden doorzocht. Het programma kost ongeveer 400 euro, afhankelijk van de koers van het pond.

<sup>10</sup> Voor de oplettende lezer: dat de zoekvraag 'fons is een' als antwoord 'Fons was een' opleverde, komt doordat Isys is als stopwoord heeft opgeslagen in een zogenoemde stopwoordenlijst. Die lijst is met de hand aan te passen.

bakbromfiets	1
bakfiets	493
bakkersbakfiets	1
bakkersfiets	4
bedrijf fiets	1
bedrijfssnorfiets	2
bestelfiets	3
bezorgersfiets	1
bobterfiets	1
bongrofiets	1
boodschappenfiets	1
bromfiets	502
burgerfiets	1
busfiets	1
buurfiets	1
casionfiets	1
confectiefiets	1
crossfiets	8
damesbromfiets	1
damesfiets	112

Een selectie uit de resultaten van de zoekactie *\*fiets*.

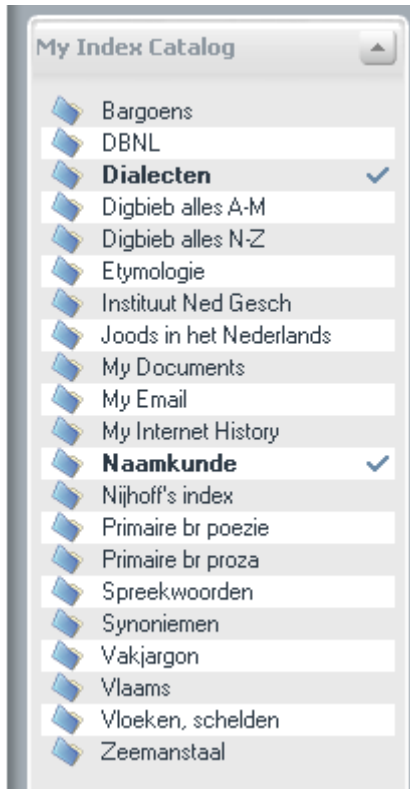
fietsafstand	7
fietsavon	1
fietsavonturen	3
fietsavontuur	4
fietsbaan	1
fietsbaand	1
fietsbakje	1
fietsban	3
fietsband	90
fietsbanden	97
fietsbandenfabriek	1
fietsbandsporen	1
fietsbanen	1
fietsbe	1
fietsbedevaart	1
fietsbel	125
fietsbeleid	2
fietsbellen	52
fietsbeltram	1
fietsbewegin	1
fietsbewegingen	4

Een selectie uit de resultaten van de zoekactie *fiets\**.

— Je kunt met jokertekens zoeken. Zo levert de zoekopdracht *m??rdijk\** bijvoorbeeld *moerdijker*, *moerdijkenaar*, *meerdijk* en *meerdijken* op. Zoek op *m\*dijk* en je vindt onder meer *maaidijk*, *maartendijk* en *monsterdijk*. Wil je weten in welke korte Nederlandse

woorden de lettercombinatie *oe* wordt gevolgd door *ij*, zoek: *\*oe\*ij\** of *?oe?ij?* (enkele resultaten: *doetijd*, *hoerije*, *koelijs*, *koedijk* en *poepijs*).

— Je kunt net zoveel indexen aanmaken als je wilt. Een index kan maximaal 2 miljard woorden bevatten. Je kunt 128 indexen tegelijk doorzoeken.



Zoeken door meerdere indexen.

- Alle resultaten worden gemarkeerd met een kleur (gehighlight). Je ziet de woorden in context. Je kijkt in de OCR-laag van de pdf, maar als je de oorspronkelijke bladzijde wilt zien, kun je die met één toetsaanslag openen.

V  
Rivier en sterren in een heldere droom:  
Waar de Moerdijk zich spant over de stroom  
Stond Marsman, en zijn norske kijken mat  
Het land van oeverzoom tot oeverzoom.

De OCR-laag.

V  
*Rivier en sterren in een heldere droom:  
Waar de Moerdijk zich spant over de stroom  
Stond Marsman, en zijn norske kijken mat  
Het land van oeverzoom tot oeverzoom.*

De oorspronkelijke bladzijde.

— Je kunt Isys zo instellen dat ook de interpunctie in de index wordt verwerkt. Wil je bijvoorbeeld weten hoe *desalniettemin* zich ‘gedraagt’ aan het begin van een zin, dan zoek je *.Desalniettemin*

— Je kunt in een zoekopdracht aangeven hoe ver woorden van elkaar verwijderd mogen zijn. Voorbeeld: bij de zoekopdracht *paard /10/ tillen* mag *tillen* niet verder dan 10 woorden van *paard* staan. Een uitgebreidere variant van deze zoekopdracht: *paard /10/ tillen OR paard /5/ getild*. Met een zoekopdracht als */-5, +10/* geef je het bereik aan: vanaf vijf posities vóór de gezochte woorden tot tien posities erna.








Er zijn nog veel meer zoekmogelijkheden, maar die moeten de lezers van dit stuk zelf maar eens bekijken: de testversie van Isys Desktop is gratis te downloaden.

## Mappenstructuur

Ik had nog beloofd om terug te komen op het nut van een digiTaalbibliotheek met een mappenstructuur. Waarom zou je niet, net als Google Books, alle boeken op één grote hoop gooien? Via een zoekmachine is uiteindelijk toch alles te vinden?

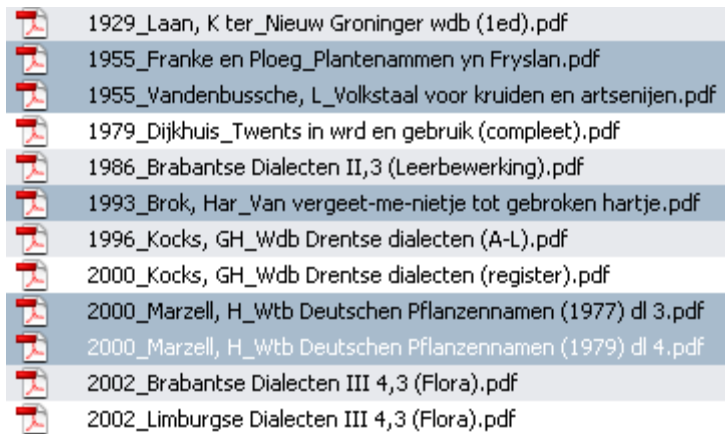
Zoals iedereen weet die vaak dingen op internet zoekt, is het probleem in toenemende mate dat je eerder te veel dan te weinig vindt. Zo’n mappenstructuur helpt om de zoekresultaten snel te ordenen. Hier drie voorbeelden van zoekacties.

1. Stel, je wilt weten in welke dialecten de plantennaam *jodenbaard* voorkomt. Je geeft dan bijvoorbeeld de volgende zoekopdracht: *jodenbaard\* OR j\*d\*b\*rd*. In de ruim 450 dialectbronnen die nu zijn gedigitaliseerd, vind je – in één seconde – vijftien hits waarin het woord in verschillende schrijfwijzen voorkomt: *jodenbaard*, *jeudenbaard*, *jeudnboard* en *joddenbaard*.

Words and Hits	
 jodenbaard	9
 jeudenbaard	3
 jeudnboard	1
 joddenbaard	1
 jodenbaard	9
 jöddenbaard	1
 Final Result:	15

Je kunt deze resultaten op verschillende manieren sorteren: op relevantie (de meeste hits eerst of juist laatst), op jaar (chronologisch of omgekeerd chronologisch), maar ook op de naam van de map, waardoor je in één oogopslag ziet dat *jodenbaard* in de huidige collectie voorkomt in de provincies Drenthe, Groningen, Limburg, Noord-Brabant en Overijssel.

Je hebt nu alleen gezocht in de afdeling dialecten, die is ontsloten met een aparte index, maar je kunt in meerdere indexen tegelijk zoeken, bijvoorbeeld in ‘dialecten’ en ‘naamkunde’, waarin ook een afdeling ‘plantennamen’ is opgenomen (ik heb de rubriek ‘naamkunde’ breed opgevat). Dit levert nog wat extra vindplaatsen op:



Een zoekactie door de complete bibliotheek, zoals gezegd 2 miljard woorden groot, levert in twee seconden de vroegste vindplaats op, namelijk *De Gids* van 1876, waar wij lezen: ‘Zij heeft verschillende namen: vroeger dien van *bedeguar*, *fungus cynobasti*, nu in Duitschland dien van *rozenkoning*, *rozenspons*, *slaapdoorn*, in de Veluwe van *wieperoos*, in de omstreek van Eibergen van *jodenbaard*.’<sup>11</sup>

Door naar de wetenschappelijke naam van deze plant te zoeken, *Saxifraga sarmentosa* of *Saxifraga stolonifera*, vind je al snel synoniemen voor *jodenbaard*.

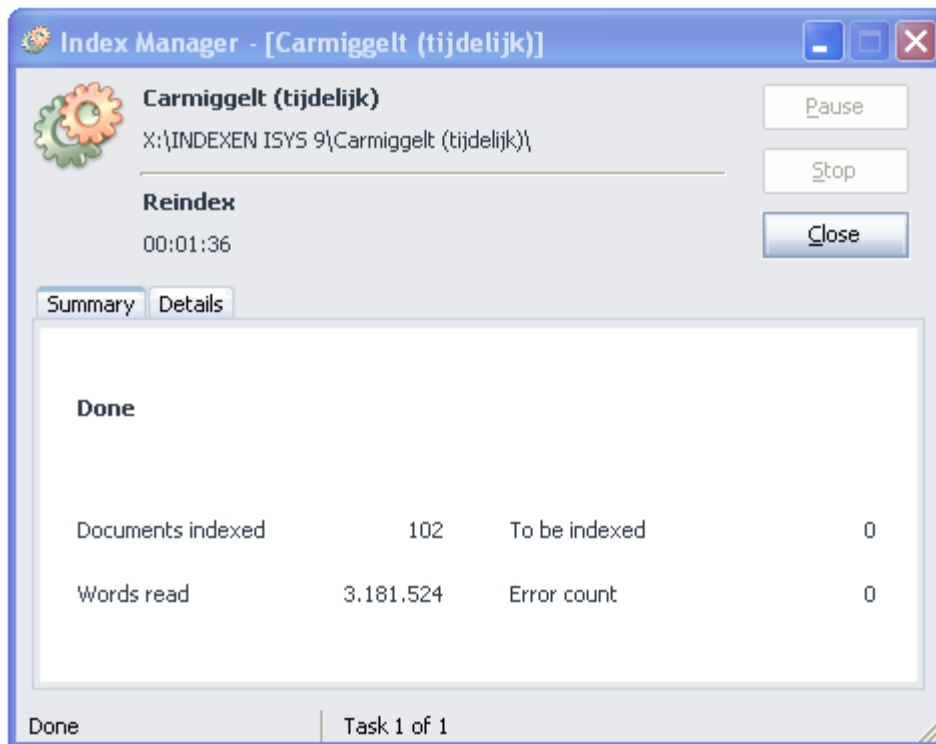
2. Stel, je wilt weten in welke periode het woord *boycot* is opgenomen in de ons omringende talen. Je zoekt *boycot*\* in de map etymologie, je sorteert de resultaten op de mapnaam en je ziet in één oogopslag dat het woord in de huidige collectie te vinden is in de etymologische woordenboeken voor het Duits, Engels, Frans, Italiaans, Spaans en Nederlands. Een snelle verkenning van die bronnen – een kwestie van scrollen – levert de dateringen op die in deze werken worden gegeven.

**boycotter** 1880, *le Parlement* ; angl. (to) *boycott*, du nom du capitaine en retraite *Boycott*, gérant de propriétés en Irlande, mis en quarantaine en 1880. || **boycott** 1888. || **boycottage** 1881, *le Figaro*. || **boycotteur** 1881.

*Boycot* in *Dictionnaire étymologique et historique du français* (1993)

3. Stel, je wilt het woordgebruik van een bepaalde schrijver onderzoeken (Gerard van het Reve, Simon Carmiggelt, Theo van Gogh). Hoe doe je dat? Door alle aanwezige titels van die auteur in een aparte map te zetten, om daar vervolgens een index op te maken. Om u een indruk te geven van de snelheid waarmee je zo’n index kunt maken: een index op 3,1 miljoen woorden in 102 boeken van Carmiggelt is voltooid in 1 minuut en 36 seconden.

<sup>11</sup> Het idee dat je voor dateringen van woorden wel kunt volstaan met de historische krantenarchieven, is onjuist. In het grootste historische krantenarchief dat momenteel voor het Nederlandse taalgebied beschikbaar is, het archief van de *Leeuwarder Courant*, komt *jodenbaard* vier keer voor, met als vroegste vindplaats 1927. Dit archief telt ruim 11 miljoen artikelen, van 1752 tot nu, op zo’n 800.000 pagina’s. Voor meer informatie, zie: <[www.ewoudsanders.nl/digitalisering/een-digitale-goudmijn/](http://www.ewoudsanders.nl/digitalisering/een-digitale-goudmijn/)>.



Een nieuwe index is razendsnel gemaakt

Op dezelfde manier kun je makkelijk een corpus aanleggen als basis voor bijvoorbeeld een periodewoordenboek. Wil je een woordenboek maken dat de periode 1975 tot 2009 bestrijkt? – het onderliggende corpus van honderden miljoenen woorden is binnen een paar uur geïndexeerd. Anders gezegd: uit het dynamische corpus, dat doorlopend groeit, kun je op ieder moment een relevant statisch corpus destilleren.

### **Droom**

Tot zover de zegeningen van Isys Desktop, die voor mijn taalonderzoek in ieder geval een grote sprong voorwaarts hebben betekend.

Terug naar een vraag die aan het begin van dit stuk werd gesteld, namelijk: gaat het lukken om een zo compleet mogelijke collectie aan te leggen van publicaties over het Nederlands?

De voortekenen zijn veelbelovend. Inmiddels begint het bij verschillende partijen te dagen dat de huidige collectie al bijzonder is. Bij mijn weten bestaat er ook voor het Engels, Duits of Frans niet zo'n gespecialiseerde digitale taalbibliotheek.

Om mijn droom te kunnen verwezenlijken is het nodig dat ook anderen gaan meewerken – door zelf te gaan scannen (volgens van tevoren afgesproken standaarden) of door boeken te leveren. Ook dit begint op gang te komen. Het Meertens Instituut leverde enkele scans en begint steeds meer boeken te leveren, een medewerker van het Instituut voor Nederlandse Lexicologie zorgde ervoor dat een complete reeks van het Corpus Gysseling (Middel nederlandse bronteksten) kon worden gedigitaliseerd, en de Zeeuwse Dialect Vereniging heeft opdracht gegeven om alle regiowoordenboeken voor het Zeeuws te digitaliseren, inclusief een belangrijk Zeeuws tijdschrift. Langzaam maar zeker begint een gespecialiseerde digiTaalbibliotheek werkelijk binnen handbereik te komen.

Echt compleet zal zo'n verzameling natuurlijk nooit worden, net zoals er geen echt complete woordenboeken bestaan. Maar ik hoop met deze bijdrage te hebben aangetoond dat er nieuwe wegen in de lexicografie kunnen worden ingeslagen. Niet zozeer in

wetenschappelijk opzicht, maar wel wat betreft een belangrijke basisvoorziening: een groot, laagdrempelig en flexibel corpus, zowel dynamisch als statisch, dat een brede collectie primaire en secundaire teksten bevat en dat je op een geavanceerde manier kunt doorzoeken. Het mooie vind ik dat deze nieuwe wegen niet alleen toegankelijk zijn voor grote instituten met veel geld en eigen systeemredacties, maar ook voor kleine vakgroepen met een kleine beurs, voor opleidingen Nederlands in het buitenland, voor dialectverenigingen en voor individuele taalonderzoekers. Uiteindelijk zal dit het onderzoek naar het Nederlands alleen maar ten goede kunnen komen, en daar is het mij in de eerste plaats om te doen.

**Bronverwijzing:**

Ewoud Sanders, 'Corpuslexicografie ligt binnen ieders handbereik', in: Egbert Beijk (red., e.a.) *Fons verborum. Feestbundel voor prof. dr. A.M.F.J. (Fons) Moerdijk, aangeboden door vrienden en collega's bij zijn afscheid van het Instituut voor Nederlandse Lexicologie* (Leiden: Instituut voor Nederlandse Lexicologie, 2009), pp. 223-236