De informatiedrager als informant: digitale tekstcorpora

door

Truus KRUYT

Abstract

This paper discusses the advantages of encoded digital text over printed text, from a researcher's perspective. The traditional notion of text corpus as a well-considered collection of texts is related to the huge amounts of digital texts that are currently available on the web. After examples of useful digitalization initiatives and available digital resources, information is given about the users and uses of the text corpora stored at the Institute for Dutch Lexicology. Attention is paid to some obstacles in building or using text collections. The conclusion is that up till now the digital medium primarily facilitates research rather than evokes new linguistic research questions.

INLEIDING

Dit artikel gaat in op de meerwaarde van een digitaal gecodeerd tekstbestand boven gedrukte tekst, vanuit het perspectief van de onderzoeker. Voorheen werd binnen een relatief beperkt aantal disciplines gewerkt met weloverwogen samengestelde tekstenverzamelingen ('tekstcorpora'). Door recente (inter)nationale ontwikkelingen is nu steeds meer digitale tekst beschikbaar, voor een veel breder scala aan onderzoek. Daarbij moet men niet alleen denken aan integrale teksten als basis voor onderzoek, maar ook aan allerlei tekstuele hulpbronnen (bibliografieën, secundaire literatuur e.d.). De technologie biedt de onderzoeker geavanceerde mogelijkheden om informatie op te vragen uit verschillende, al dan niet geografisch verspreide bronnen. De kwaliteit van een zoeksysteem is echter niet alleen een kwestie van technologie, maar hangt nauw samen met de mate waarin de oorspronkelijke tekst 'verrijkt' is met extra informatie in de vorm van coderingen. Dit zal worden geïllustreerd aan de hand van een aantal concrete zoeksystemen, met gegevens over gebruik en gebruikers. Als informant lijkt de digitale informatiedrager vooralsnog vooral faciliterend voor onderzoek te zijn: het maakt gegevens makkelijker beschikbaar en rela-

teerbaar. In hoeverre er daadwerkelijk nieuwe onderzoeksvragen en inzichten door worden uitgelokt, moet nog blijken.

WAT IS EEN TEKSTCORPUS?

Voordat er zoveel digitale tekst voorhanden was als nu, werd een tekstcorpus strikt onderscheiden van een tekstarchief of van een willekeurige verzameling teksten. Een tekstcorpus moest voldoen aan de eis dat de inhoudelijke samenstelling geschikt is voor een bepaald doel. Het gaat dus om een verzameling teksten die is opgebouwd volgens weloverwogen selectiecriteria en, indien van toepassing, volgens een weloverwogen proportionering van verschillende tekstypen. Tekstcorpora worden gebruikt voor onderzoek in diverse disciplines, bijv. corpuslinguïstiek, taalvergelijking, geschiedenis, theologie, kunstgeschiedenis, theaterwetenschap, archeologie. Andere domeinen waarin tekstcorpora gebruikt worden zijn taaltechnologie, computationele linguïstiek, (dialect)lexicografie, terminologie, vertaalkunde en het onderwijs. In principe vraagt elk doel om een specifieke corpussamenstelling (vgl. Kruyt 1998c). Toch wordt vaak gebruikgemaakt van een reeds bestaand corpus, omdat het aanleggen van een corpus erg arbeidsintensief is en het corpus 'slechts' een instrument is voor het eigenlijke doel. Maar er zijn grenzen aan dit hergebruik. Zo kan een corpus krantentaal heel geschikt zijn voor taalkundig onderzoek, maar het is niet geschikt voor onderzoek naar gesproken taal of kindertaal. In grote corpusprojecten wordt veelal gestreefd naar een samenstelling die geschikt is voor multifunctioneel gebruik. De toevoeging van classificerende metadata aan de teksten stelt de gebruiker in staat een subcorpus voor eigen doeleinden te selecteren. Een voorbeeld van een dergelijke opzet is het Corpus Gesproken Nederlands, een corpus van 10 miljoen woorden gesproken taal (http://lands.let.ru.nl/CGN/). Hetzelfde principe lag ten grondslag aan het project PAROLE van de Europese Commissie, waarin voor een groot aantal West-Europese talen een tekstcorpus van ca. 20 miljoen woorden is opgebouwd volgens dezelfde principes, bedoeld als basisvoorziening voor de ontwikkeling van (meertalige) taaltechnologische producten (Kruyt 1998a). Wegens de gigantische omvang ervan is ook het World Wide Web, waarvan de samenstelling en proportionering bepaald niet weloverwogen en evenmin stabiel genoemd kan worden, interessant geworden als corpus voor onderzoek (vgl. Van Oostendorp & Van der Wouden 1998).

Tekstcorpora bestonden vroeger vanzelfsprekend uit gedrukte of geschreven tekst. Sinds ca. 1970 werden, met name door grote lexico-

grafische instituten in onder meer Italië, Frankrijk, Zweden en Nederland, digitale tekstcorpora opgebouwd door digitalisering van gedrukte teksten. Pas sinds ca. 1990 is het mogelijk digitale teksten te betrekken van uitgevers, onderzoekers, tekstarchieven, digitale bibliotheken en het web. De hoeveelheid digitale tekst is sindsdien enorm toegenomen, onder meer door het gebruik van tekstverwerkers, door e-mail, door het web en door digitaliseringsprojecten. Instellingen als het Centrum voor Teksteditie en Bronnenstudie in Gent, de Digitale Bibliotheek der Nederlandse Letteren in Leiden en de Koninklijke Bibliotheek in Den Haag, beschikken over grote verzamelingen digitale teksten. Er zijn inmiddels ook veel andersoortige digitale taaldata, zoals digitale woordenboeken, computationele lexica (woordenboeken die door een computerprogramma kunnen worden aangeroepen), het Corpus Gesproken Nederlands en gebarentaalcorpora (www.let.kun.nl/sign-lang/). In tal van toepassingen is digitale tekst gecombineerd met beeld en/of geluid. Kortom, er zijn heel veel digitale talige gegevens, ze zijn divers van aard en ze zijn in principe alle herbruikbaar. De notie tekstcorpus in de oorspronkelijke opvatting is door deze ontwikkelingen enigszins verbleekt. Bij de huidige omvang en diversiteit van digitale taaldata wordt een goed instrumentarium voor de toegang tot selecties uit de totale dataverzameling steeds belangrijker (vgl. Offenga et al. 2006).

Functies van een digitaal en gecodeerd bestand

Ten opzichte van het gedrukte medium biedt digitale tekst het voordeel dat men er snel en flexibel (met behulp van Booleaanse operatoren en wildcards) in kan zoeken. Ook kunnen snel berekeningen worden gemaakt, bijvoorbeeld woordfrequenties en 'collocaties', paren van woorden die statistisch vaak in elkaars omgeving voorkomen. Dat zoeken en rekenen betreft de afzonderlijke woorden in een tekst, waarbij een 'woord' moet worden opgevat als een reeks van letters en tekens. Voor een computerprogramma zijn *corpus* (zonder vraagteken) en *corpus?* (met vraagteken) twee verschillende 'woorden'. Zoekmogelijkheden worden veel geavanceerder als de tekst verrijkt is met extra informatie in de vorm van coderingen. Dan kan het zoeken en rekenen niet alleen de tekst zelf betreffen maar ook de gecodeerde informatie. Afhankelijk van de aard van de gecodeerde informatie biedt een verrijkt tekstbestand diverse additionele mogelijkheden, zoals het zoeken binnen een bepaalde teksteenheid (bijv. een alinea, een hoofdstuk), het selecteren van een bepaald type tekst (bijv. editeurstekst in een editie,

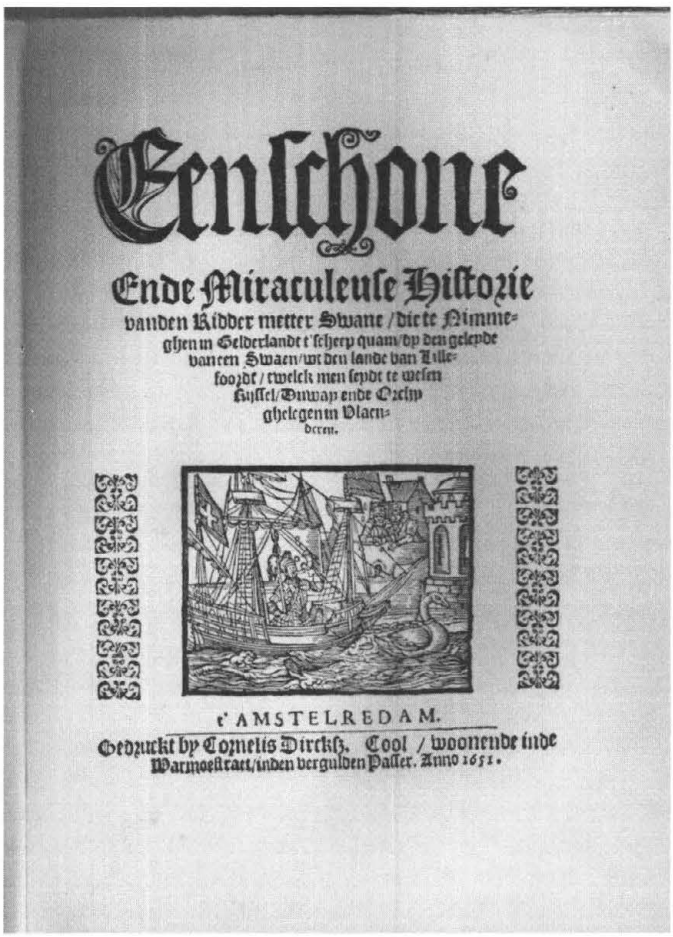
brieven in een prozatekst), het aanbrengen van links binnen een werk (bijv. van de voetnoot naar de voetnoottekst, van een verwijzing naar de desbetreffende plaats elders in het werk) en het linken en geïntegreerd opvraagbaar maken met andere data, die al dan niet elders voorhanden zijn. We komen hier later op terug.

Voor veel toepassingen is een minimaal taalkundig verrijkt bestand wenselijk. We illustreren dit met een voorbeeld. Als in een digitale tekst gezocht wordt naar woorden eindigend op *-ig*, dan bevat het resultaat onder meer *aardig*, *big*, *dienovereenkomstig*, *groenig*, *lig*, *nevelig*, *tuig*, *wig* enz. Is elk woord in de tekst verrijkt met woordsoortinformatie, dan is het mogelijk te zoeken naar uitsluitend de adjectieven eindigend op *-ig*. Het voorgaande resultaat wordt nu ingeperkt tot *aardig*, *groenig* en *nevelig*. Is aan ieder woord eveneens een trefwoord toegevoegd, dan levert het zoeken naar het adjectief *aardig* ook de daarbij behorende vormen op: *aardig*, *aardige*, *aardiger*, *aardigst*, *aardigste*. Als we zoeken naar een woord in een digitale tekst, bedoelen we vaak impliciet de trefwoordvorm met alle daarbij horende verbogen of vervoegde vormen. Voor een computerprogramma of een zoekmachine is er geen verband tussen al die vormen, tenzij ze elk afzonderlijk taalkundig verrijkt zijn met een expliciete trefwoordvorm. De toegevoegde trefwoordvorm is de parameter om alle vormen te vinden. Ook vindt zo'n zoekmachine bij het zoeken naar *vliegtuigonderdeel* niet zo maar *onderdeel van een vliegtuig* zonder extra taalkundige informatie. Kortom, verrijking maakt een computerprogramma 'intelligenter' (vgl. Kruyt 2007).

TEKSTCODERING: WAT EN HOE?

Naast taalkundige verrijking zijn tal van andere typen verrijking mogelijk. Zo bevat de titelpagina van Boekenoogen, G.J. (ed.) *Een schone ende miraculeuse historie vanden ridder metter swane (...) naar den Amsterdamschen druk van Cornelis dircksz. Cool uit het jaar 1651* (Leiden, 1931. Nederlandsche Volksboeken III; zie figuur 1) maar liefst zeventien potentieel te coderen soorten informatie met betrekking tot de inhoud en de lettertypen en -groottes. Inhoudelijke informatie betreft onder meer het titelblad als teksteenheid, de titel met daarbinnen twee subtitels, plaatsnamen in titel en subtitels, drukkersgegevens waarbinnen de drukker, het adres van de drukker, de plaats van drukken, het jaartal enz. Nu kun je die wel allemaal gaan coderen, maar dat is veel werk, te meer daar de coderingen lang niet altijd automatisch aan te brengen zijn. Het is dus zaak vooraf te over-

wegen wat het doel van de verrijking is. Dat kunnen de gewenste retrievalmogelijkheden zijn: wat niet gecodeerd is, kan niet gebruikt worden bij het zoeken, rekenen en linken. Een ander doel van de codering kan zijn de wijze waarop men een tekst wil representeren, bijv. op een computerscherm, waarbij men kan afwijken van de originele vormgeving (vgl. figuur 4). Ook kan men op basis van de codering nieuwe versies van teksten creëren, bijv. voor een bepaalde doelgroep. Al die doelen vragen om een andere codering.


Een schone
Ende Miraculeuse Hiſtorie
vanden Ridder metter Swane / die te Nimme-
ghen in Gelderlandt t'ſeheep quam/op den gelepde
vaneen Swaen/uit den lande van Lille-
foordt / tweleck men ſeydt te weſen
Kuſſel/Brunwap ende Oreſy
ghelegen in Vlaen-
deren.
t' AMSTELREDAM.
Gedruckt by Cornelis Dirckſz. Cool / woonende inde
Warmoeſtraet/inden vergulden Paſſer. Anno 1651.

Figuur 1. Titelpagina met veel potentieel te coderen soorten informatie

Is vastgesteld welke informatie voor codering in aanmerking komt, dan is de vraag in welke vorm die codering moet worden aangebracht. Vroeger deed iedereen dat op zijn eigen manier. Sinds 1988 is een internationale standaard in ontwikkeling: de TEI (Text Encoding Initiative). De TEI ontwikkelt richtlijnen voor een uniforme, taal- en platformonafhankelijke tekstcodering in XML (eXtensible Markup Language), ten behoeve van uitwisseling en hergebruik van tekstbestanden binnen de humaniora (www.tei-c.org). Die richtlijnen betreffen de tekst- en bestandsdocumentatie (bijv. bibliografische gegevens, metadata), de tekststructuur (bijv. proza: hoofdstuk, alinea enz.; poëzie: vers, versregel enz.), de typografie (font en lettergrootte), andere tekstuele elementen (bijv. noten, paginanummers, correcties) en, in dit verband heel relevant, de taalkundige en andersoortige verrijking. Figuur 2 laat een volgens de TEI gecodeerd fragment van een krantenbericht zien, met een vrij eenvoudige codering. Veel gedetailleerder is de codering in het tekstfragment in figuur 3.

```
- <TEI.2>
  + <TEIHEADER TYPE="text" STATUS="NEW">
  - <TEXT LANG="DU">
    - <BODY>
        <HEAD REND="EL-IT" TYPE="sub">NU GROTERE KANS OP EUROPESE SUBSIDIE</HEAD>
        <HEAD REND="XL-RO" TYPE="main">Nieuw leven in plan visserijmuseum Kuinre</HEAD>
      - <P>
          <NAME REND="CA" TYPE="location">KUINRE</NAME>
          -- Er wordt nieuw leven geblazen in het inititatief van anderhalf jaar geleden om
          te komen tot een historische scheepsbouwwerf in Kuinre, tevens een
          werkloosheidsproject. Was het mislopen van subsidie van het Europees Fonds
          anderhalf jaar geleden de spelbreker, de kans op een verstrekking van gelden in
          1996 is groter dan ooit, zegt Huub van Kesteren, één van de initiatiefnemers.
          Echter, om in aanmerking te komen voor deze gelden, moet het plan dit jaar
          gereed zijn en de aanvraag binnen zijn. Haast is geboden en de initiatiefnemers
          verzoeken de gemeente IJsselham om grond gratis beschikbaar te stellen. De
          lokatie blijft hetzelfde: achter de kerk in Kuinre-Noord.
        </P>
        <P>Huub van Kesteren is nu nog in dienst van Stichting Waterland. Over twee
          maanden loopt zijn contract af en gaat zich vanaf deze week vastbijten in het oude
          plan. Met vier andere personen, ondere andere Ivo van Veen en Gerard Viveen uit
          Kuinre, wil Van Kesteren binnenkort een stichting of een bv voor het toekomstige
```

Figuur 2. TEI-gecodeerd krantenbericht (fragment). Coderingen zijn omringd door punthaken. De codering moet strikt onderscheiden blijven van de eigenlijke tekst.

```
+ <front>
- <body>
    <pb n="1" />
  - <div0>
    - <head rend="align (C)">
      + <hi rend="align (C)">
        <hi rend="size (XL)">RABELAIS</hi>
        <lb rend="margin (top 1)" />
        <hi rend="align (C) size (L) slant (IT)">Genees-Heer.</hi>
      </head>
      <lb rend="margin (top 1)" />
    - <div1 type="part" n="1">
      + <head rend="align (C)">
        <lb rend="margin (top 1)" />
      - <div2 type="chapter" n="1">
        + <head rend="size (L) slant (IT)">
        - <p rend="margin (top 1)">
            <hi rend="size (XL)">H</hi>
            Eerlijke dingen en Helde daaden vang ik aan te beschrijven van den wijd-beroemden
            <hi rend="slant (IT)">Reuze Gargantua</hi>
            ; derhalven zal ik my nu niet ophouden met het verhaal van zijn afkomst end' Oudheid, die gy doch
            vervolgens zult vinden in 't groote Tijd-boek
          - <note n="1" anchored="yes" resp="aut">
              <hi rend="size (VS)">Chronique</hi>
            </note>
            van
            <hi rend="slant (IT)">Pantagruël</hi>
```

Figuur 3. Tekstfragment met gedetailleerde codering. Basis voor de min of meer originele weergave van dit fragment als is getoond in figuur 4.

De GEESTIGE WERKEN, Van *Mr. FRANCOIS*

RABELAIS

Genees-Heer.

GARGANTUA

EERSTE BOEK,

Eerste Hoofd-deel.

Van de Geslacht-reekeninge en Outheit van Gargantua.

HEerlijke dingen en Helde daaden vang ik aan te beschrijven van den wijd-beroemden *Reuze Gargantua*; derhalven zal ik my nu niet ophouden met het verhaal van zijn afkomst end' Oudheid, die gy doch vervolgens zult vinden in 't groote Tijd-boek[1] van *Pantagruël*. Uit de zelve zult gy in 't lang en <page 2>in 't breed gewaar konnen worden, hoe de *Reuzen* ter wereld quamen: en hoe van de zelve, door rechte en echte voorteeling, *Gargantua*, de Vader van *Pantagruel*, voort-quam: en neem dan niet qualijk, dat ik voor 't tegenwoordig my daar toe gedraag, en u beleefdelijk wijze: hoewel de zaak soodanig is, dat, hoe-mense meer vermeld, hoese uwe heerlijkheden meer behaagen zou: gelijk wy daar af de bevestiging hebben door *Plato* in *Philibo*, ook by *Gorgias* en *Flaccus*; zeggende, dat'er zommige saaken zijn (zonder twijfel zulke als dese) die zoo veel meerder vermaaken, alsmense meermaalen herhaalt.

Figuur 4. Min of meer originele tekstweergave op basis van het gecodeerde bestand in figuur 3. Het paginanummer is gecodeerd en de voetnoot bij het woord Tijd-boek is gelinkt aan de voetnoottekst, waardoor die op het scherm oproepbaar wordt (figuur 5).

De GEESTIGE WERKEN, Van *Mr. FRANCOIS*
RABELAIS
Genees-Heer.

GARGANTUA
EERSTE BOEK.

Eerste Hoofd-deel.
Van de Geslacht-reekeninge en Outheit van Gargantua.

HEerlijke dingen en Helde daaden vang ik aan te beschrijven van den wijd-beroemden *Reuze Gargantua*; derhalven zal ik my nu niet ophouder met het verhaal van zijn afkomst end' Oudheid, die gy doch vervolgens zult vinden in 't groote Tijd-boek– van *Pantagruel*. Uit de zelve zult gy in 't lang en <page 2>in 't breed gewaar konnen worden, hoe de *Reuzen* ter wereld quamen: en hoe van de zelve, door rechte en echte voorteeling, *Gargantua*, de Vader van *Pantagruel*, voort-quam: en neem dan niet qualijk, dat ik voor 't tegenwoordig my daar toe gedraag, en u beleefdelijk wijze: hoewel de zaak soodanig is, dat, hoe-mense meer vermeld, hoese uwe heerlijkheden meer behaagen zou: gelijk wy daar af le bevestiging hebben door *Plato* in *Philibo*, ook by *Gorgias* en *Flaccus*; zeggende, dat'er zommige saaken zijn (zonder twijfel zulke als dese) die zoo veel meerder vermaaken, alsmense meermaalen herhaalt.

Chronique

Figuur 5. De voetnoottekst "chronique" verschijnt in beeld op de plaats van de noot.

Nadat besloten is hoe gecodeerde tekstelementen moeten worden weergegeven (qua typografie, kleur enz.), kan die weergave met behulp van technologieën als XSL (eXtensible Stylesheet Language) en CSS (Cascading Style Sheets) worden gerealiseerd. Figuur 4 toont een min of meer originele weergave, met dit verschil dat de noot gelinkt is aan de voetnoottekst, die op het beeldscherm verschijnt door de noot aan te klikken (figuur 5).

TOEPASSINGEN

Er zijn tal van recente initiatieven tot het verlenen van toegang tot digitale teksten. We noemen er hier slechts enkele. De reeds genoemde Digitale Bibliotheek der Nederlandse Letteren (DBNL) breidt zich gestaag uit en heeft zijn belang meer dan bewezen in de universitaire wereld en het secundair onderwijs. Ook de Koninklijke Bibliotheek (KB) in Den Haag is erg actief op het gebied van (massa)digitalisering, onder meer blijkend uit een recent geïnitieerd project Databank Digitale Dagbladen (25 miljard woorden uit Nederlandse dagbladen vanaf de 17de eeuw). Op initiatief van het Nederlandse Ministerie van Onderwijs, Cultuur en Wetenschap wordt door de KB een Nationaal Programma Digitalisering voor de Geesteswetenschappen opgesteld.

Door de KNAW en NWO is het instituut DANS (Data Archiving and Networked Services) opgericht, dat zorgt voor opslag en blijvende toegankelijkheid van onderzoeksgegevens in de alfa- en gammawetenschappen. Op Europees niveau streeft het Digital Libraries Initiative naar een European Digital Library, die multilinguale toegang wil geven tot culturele collecties van alle lidstaten. Verder zijn via Google vele websites te vinden die toegang geven tot digitale teksten en/of hulpbronnen voor onderzoek (tijdschriften, secundaire literatuur, catalogi, documentie en informatie), van waaruit men al dan niet kan doorklikken naar integrale teksten. We noemen hier de elektronische tijdschriften www.neder-l.nl en www.neerlandistiek.nl, het bibliotheeksysteem van de Universiteit van Amsterdam www.uba.uva.nl, de informatiewebsites www.literatuurplein.nl en www.bibliotheek.nl, en de website van de vakgroep Nederlands van de Universiteit Leiden www.nederlands.leidenuniv.nl. Op de website van het Willem Frederik Hermans Instituut (www.willemfrederikhermans.nl) zijn ook interviews en door Hermans voorgelezen teksten te horen.

Laten we nu terugkeren naar de oude notie van tekstcorpus als een weloverwogen samengestelde tekstenverzameling en meer specifiek aandacht besteden aan de tekstcorpora die het Instituut voor Nederlandse Lexicologie via internet voor raadpleging ter beschikking stelt: het 5 Miljoen Woorden Corpus 1994 (Kruyt 1995), het 27 Miljoen Woorden Krantencorpus 1995 (Kruyt et al. 1996), het 38 Miljoen Woorden Corpus 1996 (Kruyt & Dutilh 1997) en het meer recente PAROLE-corpus 2004 (Van der Kamp & Kruyt 2004; Dutilh-Ruitenberg et al. 2005). Al deze corpora zijn bedoeld als onderzoeksinstrument voor intern en extern gebruik. De samenstelling is gevarieerd, met uitzondering van het 27 Miljoen Woorden Krantencorpus, dat afleveringen van NRC Handelsblad bevat. Alle corpusteksten zijn automatisch taalkundig verrijkt met woordsoort en trefwoord, waardoor een onderzoeker kan zoeken op de niveaus van woordvorm, woordsoort en trefwoord, en combinaties daarvan. Diverse functionaliteiten faciliteren het geavanceerd zoeken en het analyseren van de output. Het PAROLE-corpus is het enige corpus waarin de TEI-standaard is toegepast, zowel voor de taalkundige verrijking als voor de codering van de tekststructuur. Alle corpora zijn gratis raadpleegbaar met behulp van een retrievalsysteem waarmee tekstfragmenten (dus geen integrale teksten) kunnen worden opgevraagd. De enige voorwaarde voor gebruik is het tekenen van een gebruikersovereenkomst. Ter wille van de herbruikbaarheid voor verschillende doeleinden biedt het systeem diverse mogelijkheden voor het selecteren van een subcorpus waarop de zoekvraag wordt toegepast. Voor meer informatie verwijzen we naar

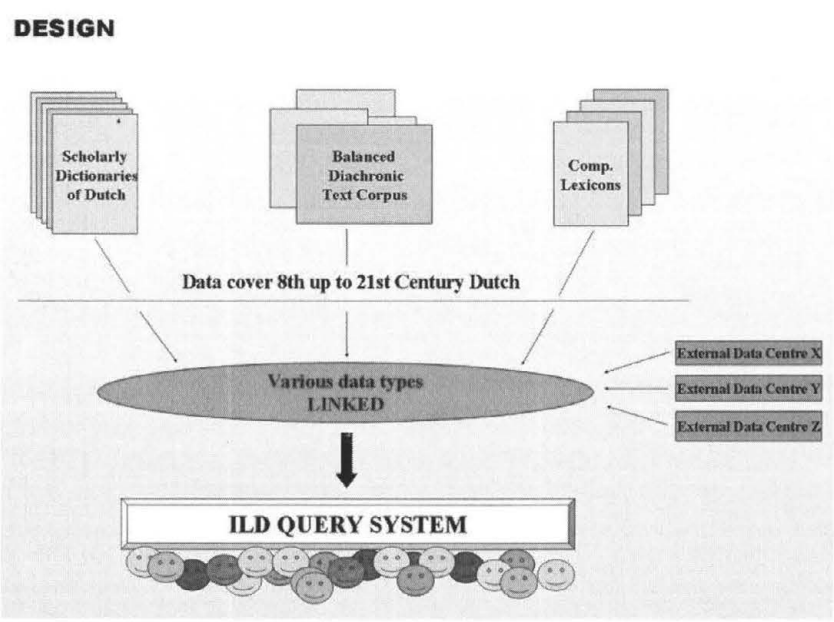
www.inl.nl onder producten, en de hier genoemde publicaties (alle van de website downloadbaar).

Deze corpora zijn en worden gebruikt voor diverse doeleinden (vgl. Kruyt 1998b): voor de bouw van internationale corpusgebaseerde lexica (o.a. Referentiebestand Nederlands, PAROLE-lexicon) en van meertalige woordenboeken met een component Nederlands, voor academisch onderwijs in de corpuslinguïstiek, psycholinguïstiek, lexicografie en de neerlandistiek, en voor de corpusgebaseerde studie van het Nederlands door onderzoekers en particuliere 'taalliefhebbers'. Al deze gebruikers, onder wie promovendi, gaven er dus de voorkeur aan een reeds bestaand corpus te gebruiken, in plaats van er zelf een op te bouwen. Op 1 oktober 2006 waren er 680 geregistreerde gebruikers van een of meer van de oude corpora: 275 van het 5 mln corpus, 325 van het 27 mln corpus en 524 van het 38 mln corpus. Het PAROLE-corpus, dat eind 2004 ter beschikking kwam had op 1 oktober 2006 198 geregistreerde gebruikers. 56 % van deze gebruikers komt uit Nederland, 26% uit België (met name uit Gent en Leuven) en 18% uit landen over de hele wereld. Het aantal gebruikers en het gebruik neemt nog steeds gestaag toe.

Behalve deze corpussystemen heeft het Instituut voor Nederlandse Lexicologie historische tekstcorpora ontwikkeld, alsmede elektronische (versies van de) wetenschappelijke historische woordenboeken van het Nederlands, computationele lexica van historisch en modern Nederlands, programmatuur voor automatische taalkundige verrijking en diverse retrievalsystemen als onderzoeksinstrumenten. Dit leidde tot het idee van een alomvattend onderzoeksinstrument voor het Nederlands: de Geïntegreerde Taalbank van de 8ste – de 21ste Eeuw (GTB) (Kruyt 2000, 2004). De GTB, die modulair wordt opgebouwd en in deelproducten ter beschikking wordt gesteld, zal uiteindelijk een flexibel instument zijn voor een breed scala aan synchroon en diachroon onderzoek naar de Nederlandse taal en cultuur door de eeuwen heen. Er zijn drie hoofdcomponenten onderscheiden: een component met de grote historische woordenboeken van het Nederlands, een component met een diachroon tekstcorpus met een weloverwogen samenstelling (Van Dalen-Oskam et al. 2002) en een component met synchrone en diachrone lexica. Woordenboeken en teksten zullen worden gecodeerd volgens de TEI in functie van de beoogde retrievalmogelijkheden (Depuydt et al. 2005). Er zullen links worden aangebracht tussen gegevens binnen afzonderlijke bronnen, tussen gegevens in bronnen binnen dezelfde component en tussen gegevens in bronnen in de verschillende componenten. Door die links kan een onderzoeker onder meer flexibel navigeren tussen corpusdata en woordenboeksgegevens.

Ook worden links voorzien met gerelateerde data van andere instituten, waardoor de onderzoeker toegang heeft tot een netwerk van geografisch verspreide data. Er zijn bijvoorbeeld links denkbaar tussen een Vlaams woord in het Woordenboek der Nederlandsche Taal en het Woordenboek van de Vlaamse Dialecten in Gent, tussen een woord in een GTB-corpustekst of -woordenboek met het Corpus Gesproken Nederlands en/of de DBNL, tussen een auteursnaam in de GTB en een bibliografische database enz. Figuur 6 toont een schematische voorstelling van het ontwerp van de GTB. Momenteel is een diachroon prototypecorpus gereed, waarbij de tekststructuur, de typografie en andere teksteigenschappen TEI-gecodeerd zijn (Kruyt 2004).

Begin 2007 is een retrievalsysteem voor de woordenboekencomponent van de GTB gereedgekomen, waarmee het Oudnederlands Woordenboek (ONW), het Vroegmiddelnederlands Woordenboek (VMNW), het Middelnederlands Woordenboek (MNW) en het Woordenboek der Nederlandsche Taal (WNT) on-line raadpleegbaar gesteld kunnen worden en al dan niet gecombineerd bevraagbaar zijn. Daarbij is rekening gehouden met de verschillen tussen de betreffende woordenboeken en is vastgesteld wat de gemeenschappelijke dan wel woordenboekspecifieke zoekmogelijkheden zijn. Het WNT is als eerste



Figuur 6. Schematische voorstelling van het GTB-ontwerp (ontleend aan een Engelstalige presentatie op het LREC-congres 2004; zie Kruyt 2004).

module van dit retrievalsysteem on-line beschikbaar gekomen in januari 2007, het VMNW als tweede module in mei 2007. Vanuit beide woordenboeken zijn er links naar andere data dan de oorspronkelijke woordenboekstekst, waaronder beeldmateriaal. Het ONW en het MNW volgen later: het ONW is nog in bewerking en het MNW vereist nog veel databewerking voor het als volwaardige module voor het publiek toegankelijk kan worden gemaakt.

Problemen

De ontwikkeling en het gebruik van tekstcorpora wordt gehinderd door een aantal factoren. Het auteursrecht behoort internationaal tot de belangrijkste belemmeringen. Het heeft betrekking op moderne teksten, op de editeurstekst in edities, op de taalkundige of andere verrijking, op foto's enz. Voor het ter beschikking stellen van een tekst aan anderen moet toestemming worden gevraagd van allen die een eigen creatief aandeel hebben gehad in de tot standkoming van de tekst in zijn huidige vorm, dus met eventuele verrijking, annotaties van een editeur, toegevoegde afbeeldingen enz. Bij de INL-corpora is daartoe een contract opgesteld tussen de corpusbouwer (het INL) en de auteursrechthebbenden (de tekstleveranciers) over de voorwaarden van gebruik. Er is eveneens een contract tussen de corpusinstelling (het INL) en de corpusgebruiker, waarin de gebruiker verklaart zich te houden aan de voorwaarden (de hierboven genoemde gebruikersovereenkomst). Dat de regels van het auteursrecht internationaal verschillend zijn, maakt het zorgvuldig ermee omgaan er niet gemakkelijker op.

Een tweede probleem is dat veel data niet aan anderen beschikbaar worden gesteld wegens commerciële belangen van bedrijven en uitgevers, of wegens de eigen belangen van onderzoekers of instellingen die veel geïnvesteerd hebben in de ontwikkeling van de data.

Van een geheel andere aard zijn de hindernissen die zich voordoen bij het gebruik van verschillende corpora naast elkaar, indien daarin verschillende systemen of modellen toegepast zijn. Zo verschillen corpora onderling in de theoretische achtergrond van de taalkundige verrijking. Inmiddels zijn daartoe richtlijnen ontwikkeld binnen het internationale standaardisatieproject EAGLES (www.ilc.cnr.it/EAGLES96/). Ook kan de toepassingsmethode van de taalkundige verrijking verschillen: wordt bijv. een adjectief in alle contexten als adjectief benoemd, of wordt het benoemd als bijwoord indien het als zodanig fungeert in de zin (vgl. Dutilh & Kruyt 2002). Hierboven is reeds aan de orde gesteld dat de TEI bedoeld is om meer uniformiteit

te bewerkstelligen in de codering van tekststructuur en verrijking. Ook voor de beschrijving van dataverzamelingen in termen van eigenschappen zijn verschillende metadatasystemen toegepast (bijv. Dublin Core, Open Language Archives Community). In het IMDI-project (www.mpi.nl/ IMDI/) is een nog gedetailleerder set van metadata ontworpen om specialisten de voor hen geschikte corpora en lexica te kunnen laten vinden in de enorme hoeveelheid dataverzamelingen die voorhanden zijn (Wittenburg et al. 2002). Tenslotte verschillen de corpussystemen ook puur technisch. Al die verschillen uniformeren voor een optimaal hergebruik van de data is niet haalbaar. Daarom zijn er nu internationale projecten gaande om zeer veel en zeer divers taalmateriaal dat geografisch verspreid is opgeslagen, toegankelijk te maken via één gebruikersinterface, waardoor het voor de onderzoeker lijkt alsof het gaat om toegang tot een enkele dataverzameling (DAM-LR, www.mpi.nl/dam-lr; CLARIN, www.mpi.nl/clarin).

Tenslotte is voor de ontwikkeling van corpora de enorme hoeveelheid benodigde databewerking een obstakel. Dit betreft niet alleen het aanbrengen van coderingen zoals hierboven besproken, maar ook kwesties als het opschonen van de verworven tekstbestanden, het uniformeren van bestaande coderingen daarin, een eventuele uniformering van afkortingen, persoons- en plaatsnamen ter wille van de retrieval enz. Er moet dus heel wat taalkundig georiënteerd werk gebeuren voordat een corpus geladen kan worden in een corpusretrievalsysteem. Het bouwen en toegankelijk maken van een corpus is bepaald geen sine cure.

Conclusie

Het thema van dit colloquium is: de informatiedrager als informant. In deze bijdrage is dit thema toegepast op digitale tekstcorpora, met name vanuit het perspectief van de meerwaarde van een gecodeerd digitaal tekstbestand boven het gedrukte medium. We hebben laten zien dat de digitale informatiedrager vooral faciliterend voor onderzoek is. Met dit medium kan de onderzoeker zoeken en rekenen met een snelheid en efficiëntie die met het gedrukte medium niet haalbaar is. Het hergebruik van en het voortbouwen op reeds ontwikkelde tekstdata is veel flexibeler dan met het gedrukte medium. Via internet heeft de onderzoeker wereldwijde toegang tot digitale tekstenverzamelingen. De mogelijkheden om flexibel toegang te krijgen tot geografisch gedistribueerde databestanden (inclusief hulpbronnen) worden steeds beter. Kortom, het digitale medium maakt gegevens gemakkelijker beschikbaar en relateerbaar.

Ook faciliterend voor onderzoek is de combineerbaarheid van digitale tekst met digitaal beeld en digitaal geluid. Bij het Corpus Gesproken Nederlands is de gesproken taal tegelijk opvraagbaar met de transcriptie. We noemden al de toepassing van tekst en geluid op de website van het Willem Frederik Hermans Instituut. De website van de Digitale Bibliotheek der Nederlandse Letteren (www.dbnl.nl) bevat naast teksten ook beeldmateriaal en geluid. Het digitale medium is ook faciliterend aan de productiekant. Het is nu mogelijk om gezamenlijk, maar op geografisch verschillende locaties via het web te werken aan bijvoorbeeld een editie (zogeheten 'collaboratories'). Ook de Wikipedia is hiervan een voorbeeld.

Het digitale medium is voor de taalkunde (in brede zin) dus onbetwistbaar faciliterend. Maar in hoeverre fungeert de informatiedrager daadwerkelijk als informant? Met andere woorden, levert gebruik van dit medium echt nieuwe gegevens, nieuwe onderzoeksvragen en nieuwe inzichten op? Hoewel dit ook voor dit wetenschapsdomein zeer wel mogelijk lijkt, laat een concreet antwoord op die vraag nog op zich wachten.

Referenties

Depuydt, K.A.C. et al. (2005), Basiscodering in TEI voor de Geïntegreerde Taalbank: Het tekstencorpus. *INL Working Papers* 2005-1. Leiden : INL. Op www.inl.nl.

Dutilh, T. & J.G. Kruyt (2002), Implementation and Evaluation of PAROLE PoS in a National Context. In: *Proceedings of the Third International Conference on Language Resources & Evaluation*, pp. 1615-1621. Ook op www.inl.nl.

Dutilh-Ruitenberg, M.W.F., J. de Does & J.G. Kruyt (2005), PAROLE: een nieuw tekstcorpus raadpleegbaar voor onderzoek. In: *Nederlandse Taalkunde* 10, pp. 326-334. Ook op www.inl.nl.

Kruyt, J.G. (1995), Nationale tekstcorpora in internationaal perspectief. In: *Forum der Letteren* 36 (1995), 47-58. Ook op www.inl.nl.

Kruyt, J.G. (1998a), Elektronische woordenboeken en tekstcorpora voor Europese taaltechnologie. In: *Trefwoord* 12, *Jaarboek Lexicografie 1997-1998*, pp. 28-42. Ook op www.inl.nl.

Kruyt, J.G. (1998b), Dutch Written Language Resources, their Users and Uses. In: *Proceedings of the First International Conference on Language Resources & Evaluation*, pp. 959-963. Ook op www.inl.nl.

Kruyt, J.G. (1998c), Valkuilen bij corpusonderzoek. In: *Nederlandse Taalkunde* 3, pp. 137-140. Ook op www.inl.nl.

Kruyt, J.G. (2000), Towards the Integrated Language Database of 8th-21st Century Dutch. In: *Revue française de linguistique appliquée* V-2 (Décembre 2000), pp. 33-44. Ook op www.inl.nl.

KRUYT, J.G. (2004), The Integrated Language Database of 8th-21st Century Dutch. In: *Proceedings of the Fourth International Conference on Language Resources & Evaluation*, pp. 1751-1754. Ook op www.inl.nl.

KRUYT J.G. (2007), Taaltechnologische vooruitgang: een reeks 'kleine onnozele ontdekkingjes'. In: F. Moerdijk, A. van Santen en R. Tempelaars (red.), *Leven met woorden. Opstellen aangeboden aan Piet van Sterkenburg bij zijn afscheid als directeur van het Instituut voor Nederlandse Lexicologie en als hoogleraar Lexicologie aan de Universiteit Leiden.* Leiden, Instituut voor Nederlandse Lexicologie/Uitgeverij Koninklijke Brill Leiden, pp.161-168.

KRUYT, J.G., S.A. Raaijmakers, P.H.J. van der Kamp & R. van Strien (1996), On-line Access to Linguistically Annotated Text Corpora of Dutch via Internet. In: H. Rettig (ed.), *Language Resources for Language technology, Proceedings of the first TELRI European Seminar in Tihany*, pp. 173-178. Ook op www.inl.nl.

KRUYT, J.G. & M.W.F. Dutilh (1997), A 38 Million Words Dutch Text Corpus and its Users. In: *Lexikos* 7 (Afrilex-reeks/series 7: 1997), pp. 229-244. Ook op www.inl.nl.

OFFENGA, F., D. Broeder, P. Wittenburg, J. Ducret & L. Romary (2006), Metadata Profile in the ISO Data Category Registry. In: *Proceedings of the Fifth International Conference on Language Resources & Evaluation*, pp. 1866-1869.

VAN DALEN-OSKAM, K.H., D.J.G. Geirnaert & J.G. Kruyt (2002), Text Typology and Selection Criteria for a Balanced Corpus: the Integrated Language Database of 8th - 21st-Century Dutch. In: *Proceedings of the Tenth EURALEX International Congress EURALEX 2002*, pp.401-406. Ook op www.inl.nl.

VAN DER KAMP, P.H.J. & J.G. Kruyt (2004), Putting the Dutch PAROLE Corpus to Work. In: *Proceedings of the Fourth International Conference on Language Resources & Evaluation*, pp. 1767-1770. Ook op www.inl.nl.

VAN OOSTENDORP, M. & T. van der Wouden (1998), Corpus Internet. In: *Nederlandse Taalkunde* 3, pp. 347-361.

WITTENBURG, P., W. Peters & D. Broeder (2002), Metadata Proposals for Corpora and Lexica. In: *Proceedings of the Third International Conference on Language Resources & Evaluation*, pp. 1321-1326.