Functionele aspecten van de gecomputeriseerde lexicografie

Joost Kist, *Lid van het dagelijks bestuur van de Stichting Instituut voor Nederlandse Lexicologie te Leiden, Nederland*

Abstract: Functional Aspects of Computerised Lexicography. The computerisation of lexicography has meant that traditional dictionaries are now supported, supplemented and — in some places — already supplanted by new electronic off- and online information carriers. The *Woordenboek der Nederlandsche Taal* (*WNT*) which was "completed" in 1998 and its electronic successors form a case study of these developments. In addition to a short description of the *WNT* project, this article also focuses on the functional aspects of computerised lexicography. Some of the more general aspects of information and knowledge technology are stressed, including the role of the user who needs to carry out his/her searches through the language banks of the future with the least possible effort.

Keywords: COMPUTERISED LEXICOGRAPHY, ELECTRONIC DICTIONARY, ELECTRONIC PUBLISHING, INFORMATION AND COMMUNICATION TECHNOLOGY, ON-LINE ACCESS VIA INTERNET, USER INTERFACE

Samenvatting: Computerisering van de lexicografie heeft er toe geleid dat het traditionele woordenboek thans wordt ondersteund, gesupplementeerd en — op sommige plaatsen — reeds verdrongen door nieuwe electronische off- en on-line informatiedragers. Het in 1998 "voltooide" *Woordenboek der Nederlandsche Taal* (*WNT*) en zijn electronische opvolgers vormen een case study met betrekking tot deze ontwikkeling. In dit artikel wordt — naast een korte beschrijving van het *WNT* project — aandacht besteed aan de functionele aspecten van de gecomputeriseerde lexicografie waarbij de nadruk ligt op enige meer algemene aspecten van de informatie- en kennistechnologie en de positie van de gebruiker die zijn/haar zoektochten door de taalbanken van de toekomst met zo min mogelijk moeite moet kunnen volvoeren.

Trefwoorden: COMPUTERONDERSTEUNDE LEXICOGRAFIE, ELECTRONISCH WOORDENBOEK, ELECTRONISCH UITGEVEN, INFORMATIE EN COMMUNICATIETECHNOLOGIE, ON-LINE TOEGANG TOT HET INTERNET, GEBRUIKERSINTERFACE

1. Inleiding

In dit artikel willen wij tegen de achtergrond van het gereedkomen van het *Woordenboek der Nederlandsche Taal* in 1998 en de ontwikkeling van nieuwe lexicografische producten en diensten van het Instituut voor Nederlandse Lexicologie (INL) een aantal functionele en meer algemene aspecten van het vervaar-

digen en gebruiken van — wat wij noemen — "informerende systemen" aan de orde stellen (Kist 1996) .

Telkens als er in de geschiedenis van de mensheid een nieuw informatiemedium wordt uitgevonden, zien wij dat de snelheid van informatieverstrekking wordt opgevoerd, terwijl gelijktijdig de hoeveelheid informatie gaat toenemen.

De eerste stap in dat technologisch proces was de uitvinding van het pictografisch schrift door de Sumeriërs en Egyptenaren, ruim 3 000 jaar voor Christus. Vervolgens werd dit "informerend systeem" via de uitvinding van syllabische en fonetische tekens versimpeld tot bruikbare alfabetten die voornamelijk toegepast werden bij de vervaardiging van "unicaten" van teksten zoals woordenlijsten, boekhouding, wetgeving en godsdienstige voorschriften. Stempel- en rolzegels boden overigens direct mogelijkheden tot duplicering "in druk" van persoonlijke gegevens. Wij kunnen dus in het verband van dit artikel al direct constateren dat woordenlijsten tot de allereerste cultuurproducten behoorden (Schaer 1996). De gelijktijdige uitvinding en/of toepassing van de losse loden letter, van de drukinkt en van de drukpers leidde omstreeks 1450 tot voorheen ongekende dupliceringsmogelijkheden van tekst en daardoor tot een universele geletterdheid van de mensheid (Kist 1988). Sinds de introductie en de implementatie van het informerend systeem van Gutenberg zijn er een viertal trends in de verspreiding van informatie waar te nemen:

— Een acceleratie van de productie van tekst in kwantitatieve en in kwalitatieve zin.
— Een segmentatie van de inhoud (gebruik van tekst voor meerdere doeleinden).
— Informatieverspreiding onder een groot publiek.
— En — gedurende de afgelopen decennia — een snel toenemende digitalisering van informatie die weer leidde tot een efficiente redactionele verwerking en tot goedkope opslag van grote hoeveelheden tekst, beeld en geluid.

Deze vierde trend versterkte de effecten van de eerste drie trends. Bij de invoering van de computer is bovendien de "connectiviteit" (de mogelijkheid om via een netwerk ieder stukje informatie van het ene naar het andere punt te transporteren) sterk toegenomen.

Inhoudelijke informatie wordt hierbij als het ware losgeweekt uit het informerend systeem (dus de gekozen informatiedrager) en gaat zich zelfstandig bewegen in de electronische omgeving. De klassieke informatieketen (bijv. auteur — uitgever — drukker — binder — boekhandel — bibliotheek — lezer) wordt nu verbroken en er gaan nieuwe diensten, producten en merken ontstaan die hun eigen, electronische route kunnen kiezen, waarbij ze de verschillende stations in de traditionele keten kunnen omzeilen. Er ontstaan zo nieuwe

informatiebronnen en informatiecombinaties, wat weer tot een grote verscheidenheid aan informerende systemen zal leiden. Een recent voorbeeld van deze ontwikkeling is het electronisch tijdschrift dat in een van zijn vele nieuwe verschijningsvormen rechtstreeks, zonder kwaliteitscontrole van een referee of tussenkomst van een uitgever, van auteur naar lezer wordt getransporteerd (Kist 1996).

Het is duidelijk dat de technologie van de gecomputeriseerde verwerking van tekst en het rechtstreeks in het verlengde daarvan liggende electronisch uitgeven van zeer grote betekenis is geworden voor de samenstelling en het gebruik van woordenboeken. In vorige afleveringen van *Lexikos* is hierover met betrekking tot het *WNT* al deskundig en uitvoerig gerapporteerd (Kruyt 1995 en Kruyt en Dutilh 1997), wat ons ontslaat van de plicht hierop opnieuw in te gaan bij het in kort bestek releveren van de ontstaansgeschiedenis van het *WNT*.

2. Het *WNT* en het Instituut voor Nederlandse Lexicologie

Het *WNT* is het grootste woordenboek ter wereld. Het kostte vijf generaties redacteuren bijkans honderdvijftig jaar om dit gigantische werk te voltooien. Het telt 40 banden, 45 805 bladzijden en het beschrijft tussen de 350 000 en 400 000 woorden uit de periode 1500 tot ongeveer 1921. Er zijn ongeveer 1,6 miljoen citaten uit bijna 10 000 bronnen verwerkt. Anderhalve eeuw is een lange tijd om aan een boek te werken en er is in die periode natuurlijk veel gebeurd. Uitgevers dreigden af te haken en er werd veel interne strijd geleverd. Nederlandse en Vlaamse overheden die het kostbare project financierden zijn enige malen van plan geweest het bijltje er bij neer te gooien. De beslissing, genomen in 1976, om geen bronnenmateriaal van na 1921 meer te verwerken gaf uiteindelijk de doorslag; zonder dit besluit was het *WNT* nooit meer afgekomen. Wel werd hierdoor het karakter van het *WNT* definitief bepaald: het is uitgegroeid tot een historisch-wetenschappelijk woordenboek dat op basis van miljoenen citaten de taal van 1500 tot 1921 analyseert. Ondanks bepaalde tekortkomingen is het *WNT* een onmisbaar wetenschappelijk instrument waarop andere Nederlandse woordenboeken zoals Van Dale, Koenen en Kramers zwaar steunen. Voor ons betoog is echter het meest interessant dat de bruikbaarheid van het *WNT* exponentieel is toegenomen door de vervaardiging van een cd-rom door AND Electronic Publishing in samenwerking met het INL. Deze cd-rom die te zijner tijd ook op het Internet geraadpleegd kan worden biedt de mogelijkheid, alle zoekopties electronisch bliksemsnel te exploreren en toont aldus de voordelen en toegevoegde waarde van het electronisch uitgeven onweerlegbaar aan.

Het nu voltooide *WNT* is niet het enige, maar vooralsnog wel het grootste project van de in 1967 opgerichte stichting Instituut voor Nederlandse Lexicolo-

gie (INL). De bijna vijftig Nederlandse en Vlaamse medewerkers leggen de woordenschat van heden en verleden vast in woordenboeken en woordenboekachtige producten, zoals electronische bestanden en cd-roms. Het INL mag men beschouwen als de schatbewaarder van de Nederlandse taal. Het *WNT* is nu voltooid maar een woordenboek is nooit af. Voor de medewerkers van het INL begint nu een nieuwe periode, met nieuwe en uitdagende projecten.

In 1999 zal aandacht besteed worden aan de voltooiing van het *Vroegmiddelnederlands Woordenboek*. Voorts zal de Taalbank van het INL, die miljoenen woorden aan (voornamelijk modern-)Nederlandse teksten bevat, geleidelijk worden uitgebouwd tot een Geïntegreerde Taalbank van het Nederlands van de 8ste tot de 21ste eeuw. In 1999 vangen eveneens de werkzaamheden aan voor een te produceren *Oudnederlands Woordenboek*, dat het alleroudste Nederlands uit de periode 750 tot 1150 zal beschrijven. Een ander nieuw en omvangrijk project betreft een *Woordenboek van het Eigentijds Nederlands*. Voorts blijft het INL verantwoordelijk voor het actualiseren van de *Woordenlijst Nederlandse Taal*, het bekende *Groene Boekje*. Tenslotte zal het voltooide *WNT* op enkele onderdelen nog kwantitatief en kwalitatief worden bijgewerkt, eerst in drie à vier delen "Aanvullingen" die op het supplementmateriaal zijn gebaseerd, vervolgens in een apart project "voortgezet WNT".

Op basis van de door het INL gevolgde strategie is het in dit stadium mogelijk een aantal potentiële producten en diensten te identificeren en doelgroepen te selecteren die voor verdere ontwikkeling en toepassing in aanmerking kunnen komen. Wij onderscheiden toepassingen ten behoeve van een groot aantal productcategorieën:

— In de categorie woordenboeken is een grote diversiteit van mogelijkheden potentieel aanwezig, zowel qua inhoud en doelgroep als in de technische vorm (folio, cd-rom, on-line enz.), hulpmiddelen op het gebied van semi-automatische vertaalsystemen enz.
— Uitgaven gebaseerd op uitwerking van specifiek taalkundig materiaal voor wetenschappelijke doeleinden (bepaalde grammaticalia, stilistische grammatica's respectievelijk woordenboeken en -lijsten van bepaalde taalkringen, lijsten per tijdsperiode, regio). Overigens kunnen in deze categorie ook producten voor meer algemene marktsegmenten worden geïdentificeerd, bijvoorbeeld puzzelwoordenboeken.
— Uitgaven gebaseerd op taalgebruik (taalgebruik van bepaalde auteurs of perioden, verzamelingen, uitdrukkingen, spreekwoorden, zegswijzen);
— Producten gericht op gebruik als spellingcheckers (voorzetselverbindingen, synoniemen of taxonomische informatie, ontledingen en moeilijke taalkundige woord- en spellingsvormen e.d.). Hiervoor zijn zowel algemene als specifieke doelgroepen of marktsegmenten aan te geven.
— Tevens zou een database ook specifieke, gerichte custom-made informatieproducten voor specifieke afnemers c.q. afnemersgroepen kunnen

opleveren terwijl de electronische ontsluiting en karakteristiek tevens output in verschillende vorm mogelijk zouden maken (on-line, cd-rom, Internet-downloading enz.).

2.2 Succesfactoren van de INL/*WNT* combinatie

Als succesfactoren die bij alle genoemde innovaties geleid hebben tot verbetering van de inhoud en van het proces van intellectuele creatie kunnen wij — het bovenstaande samenvattend — de volgende punten noemen:

— Toepassing van breed aanwezige interne redactionele expertise op diverse terreinen;
— Toegang tot betrouwbare informatiebronnen en documentatie informatiebronnen;
— Steun van betrokken ministers en overheden in Nederland en België;
— Waar mogelijk inschakeling van electronica (die gelukkig in de loop van de tijd steeds goedkoper werd hetgeen budgettair een meevaller was) en databanktechnologieën als hypertext en digital object identifiers; kennis van de informatie- en communicatietechnologie;
— Goede projectorganisatie en -leiding;
— Gestroomlijnde productie;
— Nieuwe producten en spin-off producten ("herverpakking" en verrijking van informatie) kunnen vrij gemakkelijk worden gegenereerd;
— Effectieve ontsluiting van informatie en herbenutten van meerwaarde van het bestand;
— Gedegen kennis van de gebruikerscategorieën;
— Internationale uitwisseling van expertise; en
— *Goed management.*

Mogelijke faalfactoren (die zich niet hebben voorgedaan) waren:

— Wegvallen van politieke steun voor de financiering;
— Afhankelijkheid van externe opslag en distributie; en
— Het niet kunnen aantrekken of behouden van deskundige medewerkers.

3. Functionele aspecten van de woordenboekenproductie

3.1 Aan welke functies heeft een gebruiker behoefte?

Als wij nu de voorgaande case study tot een meer algemene beschouwing over toekomst van de vervaardiging van woordenboeken trachten te verbreden dan is het noodzakelijk dat wij ons eerst verdiepen in de vraag, aan welke functies

de gebruiker van een informerend systeem en in het bijzonder van een woordenboek behoefte kan hebben. Wij komen dan tot de volgende opsomming:

— De gebruiker heeft allereerst natuurlijk behoefte aan een bepaalde inhoudelijke informatie (*kwaliteitsselectie*).
— Vervolgens wenst hij of zij een bepaalde diepgang of structuur (*niveauselectie*).
— Dan is er behoefte aan een bepaalde hoeveelheid (*kwantiteit*, niet te weinig maar zeker ook niet te veel).
— In het bijzonder bij de inzet van nieuwe media komt een behoefte aan een bepaalde *snelheid* naar voren (de levertijd die vroeger in uren of dagen werd gemeten mag nu niet langer dan seconden of nog minder duren).
— Ook zijn er specifieke wensen omtrent een bepaalde *technologische prestatie of verpakking*; men wil kunnen "browsen", "skimmen" en "scannen", men wil in "portals" rondneuzen of gegarandeerd veilige netwerken inschakelen. Veelal wil men de informatie toch weer op papier binnen bereik hebben (*output*).
— Belangrijk is ook het gebruiksgemak passend bij de gebruiksomgeving (*comfort*).
— Tenslotte is er natuurlijk de prijs (*prijselasticiteit*).

Deze opsomming is niet limitatief, maar de wensen van een gebruiker zijn altijd terug te brengen tot de klassieke "ijzeren driehoek": Prijs, Kwaliteit, Levertijd. De woordenboekenproducent zal ook in de 21ste eeuw met deze elementaire wensen rekening dienen te houden. Wij dienen ons dus af te vragen welke doelen wij bij de inzet van nieuwe media voor ogen moeten hebben.

3.2 Doelstellingen

Welke doelstelling willen wij door middel van digitalisering en computerisering bereiken? Wij doen het beste, deze doelstelling vanuit vier gezichtspunten te bezien:

— Met betrekking tot het vervaardigings- en vermenigvuldigingstraject;
— Met betrekking tot het informatieproces;
— Met betrekking tot de toevoeging van meerwaarde aan de inhoud; en
— Met betrekking tot verbetering van het rendement.

Verbetering van de productiviteit en de efficiency in het traditionele grafische vervaardigingsproces kan bereikt worden door digitalisering van de tekstver-

werking, door diverse innovaties in het vermenigvuldigings- en verspreidingstraject en waar nodig en mogelijk door introductie van nieuwe media, van cd-rom tot Internet. Deze productiviteits- en efficiencyverbetering zet nog steeds door en wij zagen dat allereerst de traditionele zetter in het grafisch bedrijf werd vervangen door tekstverwerking intern en in landen met lage lonen.

Na het geleidelijk uit het zicht verdwijnen van de klassieke drukkerij gaat men zich ook afvragen, welke rol de traditionele uitgever in de toekomst moet spelen, nu de regie van het redactionele apparaat geheel binnen de muren van een lexicografisch instituut kan worden uitgeoefend. Ook hier zullen drastische veranderingen in de informatieketen gaan optreden en mogelijk wisselende coalities in de plaats komen van de huidige lineaire verhoudingen.

Bij de verbetering van het informatieproces moeten wij vooral denken aan innovaties op het terrein van het redactionele werk met behulp van hardware en software. Controle op de uniformiteit van de teksten (wij gaven al eerder het voorbeeld van de cd-rom van het *WNT*, die wonderlijke inconsistenties en doublures opspoorde), het aanbrengen van hiërarchische relaties en de opbouw van relationele databases zijn hier enkele trefwoorden in een heel scala van te verwachten verbeteringen.

Heel belangrijk is ook de mogelijkheid van het toevoegen van meerwaarde aan en de tussentijdse verbetering van informatieve producten en diensten. Wij zien een voortdurende verbetering van gebruiksvriendelijkheid, betere navigatiemogelijkheden, uitgekiende helpfuncties, gelaagde presentaties en het aanbrengen van hyperlinks. Bestanden kunnen continu verrijkt worden en meer gegevens per trefwoord kunnen op de zoekplaats worden aangebracht. Hier staan wij nog maar aan het begin van vele nieuwe ontwikkelingen die lang niet altijd hun oorsprong vinden in de techniek maar vooral ook in de menselijke inventiviteit, de "intellectuele technologie".

De mogelijkheden tot verbetering van het financiële rendement zijn onder meer te vinden bij het uitgeven via het Internet. Bij een groeiend aantal gebruikers nemen de kosten van het (re)produceren — dit in tegenstelling tot het traditionele drukproces — nauwelijks toe. Bits zijn gemakkelijk te kopiëren en de communicatiekosten over het Internet zijn te verwaarlozen in vergelijking met de vaste kosten voor het creëren van informatiesystemen, de indexen en de inhoud. De inkomsten lopen derhalve snel op bij het toenemen van het aantal betalende (of gesubsidieerde) gebruikers terwijl de kosten nagenoeg constant blijven. Op het Internet bieden de zogenaamde portal sites informatie(-indexering), communicatie en andere functies aan. Het gemak voor de gebruiker bestaat o.m. hierin dat hij minder hoeft te klikken om zijn informatiebehoefte te bevredigen.

3.3 Conclusies met betrekking tot nieuwe functionaliteiten

Hoewel digitale tekstverwerkingsprocessen en electronic database publishing al sinds de jaren 70 bestaan, beginnen zij nu — na vele mislukkingen en kostbare vergissingen — door te breken, in het bijzonder voor documentaire netwerken en informerende systemen als woordenboeken en andere grote tekstcorpora. Wij zien een ontwikkeling van woordenboek naar woordenbank maar we zien terzelfdertijd dat verschillende media verschillende behoeften zullen bevredigen: "different media favour different content" (Kist 1996). Er is nog geen universele ideale informatiedrager zoals eens klei, papyrus of papier, geschikt voor alle doeleinden. Er zullen diverse dragers naast elkaar blijven bestaan, puttend uit een digitale woordenbank.

Het papier als medium voor opslag, presentatie en annotatie is overigens nog lang niet op alle fronten verslagen en het foliotijdperk loopt kennelijk nog niet ten einde. De informatiegebruiker wil tekst (bijv. van een website) veelal ook nog op papier zien of vasthouden, alleen al omdat veel electronische producten thans in wezen nog versies zijn van informatie die evengoed op papier had kunnen worden vastgelegd zonder de tussenkomst van computerprogramma's. Printers zijn een belangrijk onderdeel van elke computerconfiguratie. Ook de uitgevers constateren dat "puur" electronisch uitgeven nog lang niet universeel is doorgebroken, zoals zij nog kortgeleden triomfantelijk voorspelden. Het publiceren van documentaire informerende systemen — of dat nu commercieel of in universitair verband plaatsvindt — zal in het begin van de volgende eeuw in vele vormen en op vele manieren geschieden. Het gaat er vooral om, de eigen, "proprietary" informatiebestanden zo goed mogelijk te beheren, te bewaken, te verrijken en daarbij uit de vele mogelijkheden de beste verspreidings- en exploitatiemogelijkheden te kiezen: het kapitaal zit in de kwaliteit van de eigen redacteuren en medewerkers en in de breedte en diepte van de eigen taalbank.

Wij weten nog niet hoe de nieuwe informerende systemen van de volgende eeuw genoemd zullen worden als begrippen als "woordenboek" en "taalbank" verouderd zullen zijn maar ze zullen waarschijnlijk de volgende kenmerken vertonen:

— de informatie zal gesegmenteerd zijn in precies afgewogen en op de gebruiker toegesneden porties;
— de informatie zal gefiltreerd zijn (intelligent geordend en ontdaan van onnutte, ongecontroleerde gegevens en doublures);
— de informatie zal aangepast zijn bij het specifieke medium (mediumspecifiek);
— de informatie zal geïntegreerd zijn (tekst, bewegend beeld en zo nodig geluid, dus multimediaal);

— de informatie zal toegespitst zijn op het leveren van oplossingen verrijkt met bruikbaarheidscriteria; en
— de informatie zal interactie mogelijk maken (dialoog tussen bron en gebruiker).

In het nabije verleden, toen de databanken nog voornamelijk op naslag of research waren gericht, lag de nadruk op documentatie en volledigheid. De beschikbare apparatuur werd ingezet om alle informatie op een bepaald gebied te verwerken. De sterk toegenomen verwerkingscapaciteit en snelheid van de apparatuur hebben in vele gevallen geleid tot het beschikbaar komen van veel te veel gegevens, relevant, niet relevant, overbodig, doublerend, verouderd en hinderlijk. Parkinson zou een nieuwe wet hebben kunnen formuleren die luidt: "Information expands to fill up the capacity of the system." Het is duidelijk dat een van de belangrijkste toekomstige functies van de informatieverschaffer zal zijn het functioneren als sluiswachter (om een typisch Nederlandse metafoor te hanteren). Informatieverschaffers moeten betrouwbare, toegespitste informatie leveren, gericht op de specifieke behoefte van de individuele gebruiker, op mensen met een individueel wensenpatroon, niet te veel en niet te weinig. Nieuwe media en nieuwe diensten vereisen een nieuwe presentatiewijze en een aanpassing van de inhoud.

Bell (1973) noemde de informatievoorziening in de postindustriële maatschappij een "game between persons". Inderdaad begint en eindigt de kennisketen bij de mens. De intellectuele mens, de homo faber, de homo economicus of de homo ludens. Er is een nieuw soort van inspanning nodig om de intellectuele prestatie, de informatie en de kennis via de nieuwe informatieketens op de juiste wijze bij de intellectuele ontvanger, de lezer, de gebruiker te brengen. Daarvoor zijn nieuwe "intellectual tools" nodig. Wij hebben het gevoel dat de lexicografen en de lexicografische technologen (Kruyt 1995) en redacties van de nieuwe woordenboeken bij uitstek degenen zullen zijn die deze functie in de volgende eeuw zullen gaan vervullen.

Literatuur

Barquin, R. e.a. 1979. *Building, Using and Managing the Datawarehouse.* New York: Prentice Hall.

Bell, D. 1973. *The Coming of the Post-industria Society.* New York: Basic Books.

Boer, M. de. 1997. Organisationele innovatie door traditionele uitgeverijen: De transformatie van mono naar mixed media. *I&I, Informatie- en Informatiebeleid* 15(2): 92-100.

Kist, J. 1988. *Electronic Publishing: Looking for a Blueprint.* New York: Routledge.

Kist, J. 1995. Uitgevers tussen papier en electronica. *I&I, Informatie en Informatiebeleid* 13(1): 13-20.

Kist, J. 1996. *Bibliodynamica: Slaag- en faalkansen bij innovatie van informerende systemen, in het bijzonder in het uitgeversbedrijf.* Amsterdam: Cramwinckel.

Kruyt, J.G. 1995. Technologies in Computerized Lexicography. *Lexikos* 5: 117-137.

Kruyt, J.G. en M.W.F. Dutilh. 1997. A 38 Million Words Dutch Text Corpus and its Users. *Lexikos* 7: 229-244.

Schaer, R. (Red.). 1996. *Tous les savoirs du monde, encyclopédies et bibliothèques, de Sumer aux XXI siècle*. Parijs: Bibliothèque Nationale Française.

Verkuyl, Henk. 1998-1999. Een fusie tussen Van Dale en de Winkler Prins? *Trefwoord* 13: 135-151.