TAALTECHNOLOGIE VOOR HET NEDERLANDS: VORDERINGEN BIJ DE BOUW VAN EEN NEDERLANDSTALIG DIALOOG- EN AUTEURSYSTEEM

Gerard Kempen, Leo Konst & Koenraad De Smedt
Psychologisch Laboratorium K.U.
Nijmegen

De afgelopen tien jaar is flinke vooruitgang geboekt op het gebied van de automatische verwerking van natuurlijke taal door de computer. Al ligt het tempo van zijn taalverwerving aanzienlijk lager dan dat van kinderen, toch zijn de eerste nuttige toepassingen van zijn nog beperkte taalvaardigheid een feit. De computer kan nog lang niet zelfstandig optreden als gesprekspartner, tekstschrijver, taaldocent of vertaler, maar assistent-rollen zijn wel degelijk voor hem weggelegd, voornamelijk dank zij zijn onvermoeibaarheid, precisie en lage prijs.

Twee toepassingen van de taaitechnologie staan in dit artikel op de voorgrond: dialoogsystemen en auteur systemen. Deze benamingen zijn nogal pretentieus en slaan eigenlijk alleen op het einddoel waar de systeembouwers ooit nog een hopen aan te komen. Maar een kniesoor die daarop let: we protesteren ook niet wanneer van een kind dat net het aap-noot-mies te pakken heeft, gezegd wordt dat het kan lezen. Wel moet de koper van zulke producten zich realiseren dat hij geld neertelt voor systemen die nog in hun kinderschoenen staan. En de leverancier die een "dialoogsysteem" aan de man wil brengen mag niet de indruk wekken dat met dit systeem gezellige gesprekjes te voeren zijn terwijl het alleen maar in staat is om via een beperkt type vragen gegevens uit een bepaalde databank op te halen.

Dit artikel geeft een overzicht van het Taaltechnologie-project dat sinds eind 1982 aan de K.U.N, wordt uitgevoerd. Na een schets van de doelstellingen (par. 1) beschrijven we de stand van zaken per eind maart 1984 (par. 2). Paragraaf 3 is speciaal gewijd aan toekomstige vormen van

---

Het in dit artikel beschreven onderzoek wordt gesubsidieerd door het Directoraat-Generaal voor Wetenschapsbeleid van het Ministerie van Onderwijs en Wetenschappen, door het Ministerie van Economische Zaken en door het ESPRIT-programma van de Europese Gemeenschap.

tekstverwerking op basis van taalkennis die in de tekstverwerkende programmatuur is ingebouwd.

1. EEN TAALVAARDIG KANTOOR-WERKSTATION

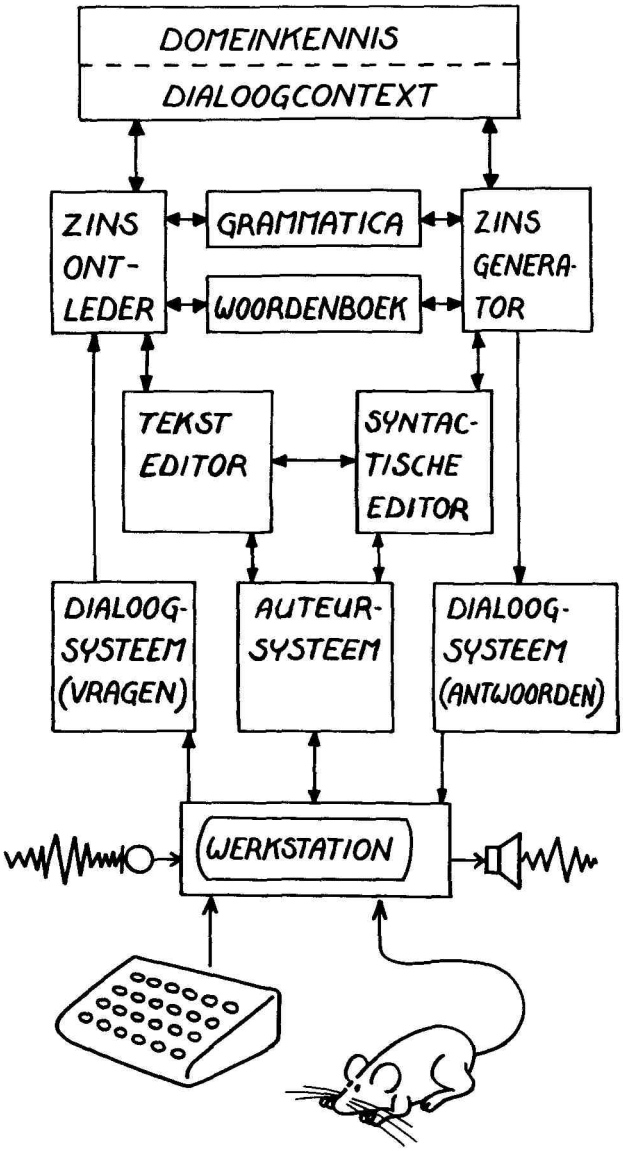
Onze inspanningen zijn gericht op het bouwen en evalueren van een taalcomponent die deel uitmaakt van een geavanceerd kantoorwerkstation (zie Figuur 1). De taaicomponent bestaat uit twee onderdelen: een dialoogsysteem, en een auteursysteem. Beide opereren in de context van een soort expertsysteem dat verstand heeft van het kantoor en de kantoororganisatie (domeinkennis). De vragen en antwoorden die het dialoogsysteem moet kunnen beantwoorden zullen dan ook betrekking hebben op personen die in de organisatie werken, op activiteiten die er plaats vinden, op documenten die hierbij een rol spelen, enz. We zullen nu niet verder ingaan op de vele linguïstische problemen waarvoor ontwerpers van dialoogsystemen zich geplaatst zien. Elders is daar uitvoerig over geschreven (zie bijvoorbeeld Kempen, 1983 en Bunt, 1984).

Het auteursysteem moet behulpzaam zijn bij het voorbereiden van documenten. Hedendaagse tekstverwerkers zijn geheel en al gespeend van linguïstische kennis [1]. Neem het volgende voorbeeld. Willen we op een gewone tekstverwerker een zelfstandig naamwoord dat vaak voorkomt in het meervoud zetten, dan kan dat tamelijk eenvoudig met behulp van een “globaal” commando. Echter, de verdere linguïstische veranderingen die deze wijziging met zich meebrengt, worden niet doorberekend. Bijvoorbeeld als het woord document wordt gepluraliseerd, moet ook het lidwoord het in de veranderen, moeten voornaamwoorden worden aangepast (dit naar deze, het naar de), moeten eventueel persoonsvormen in het meervoud worden gezet, enz. enz.

Bij huidige tekstverwerkers is dit een tijdrovend werk, vooral als geschreven wordt in een taal met een rijke morfologie, zoals Duits of Frans. Het auteursysteem dat we willen bouwen handelt deze voortplanting van wijzigingen zoveel mogelijk automatisch af. Op syntactisch niveau moet het auteursysteem eenvoudige commando’s bevatten om bijvoorbeeld hoofdzinnen om te zetten in bijzinnen met hun gewijzigde woordvolgorde, actieve zinnen in passieve, om nevengeschikte zinsdelen in te kunnen voegen, of om andere bewoordingen te suggereren. Ook zou het behulpzaam kunnen zijn bij het

---

[1] Een beperkte uitzondering maken we voor tekstverwerkingspakketten die regels bevatten voor woordafbreking en voor het signaleren van stijl- en spelfouten. Zie bijvoorbeeld Frase, 1983 en Heidorn et al., 1982.



Figuur 1. Dialoog- en auteursysteem.

signaleren van foutieve zinsbouw, van incorrect gespelde werkwoordsvormen, enz.

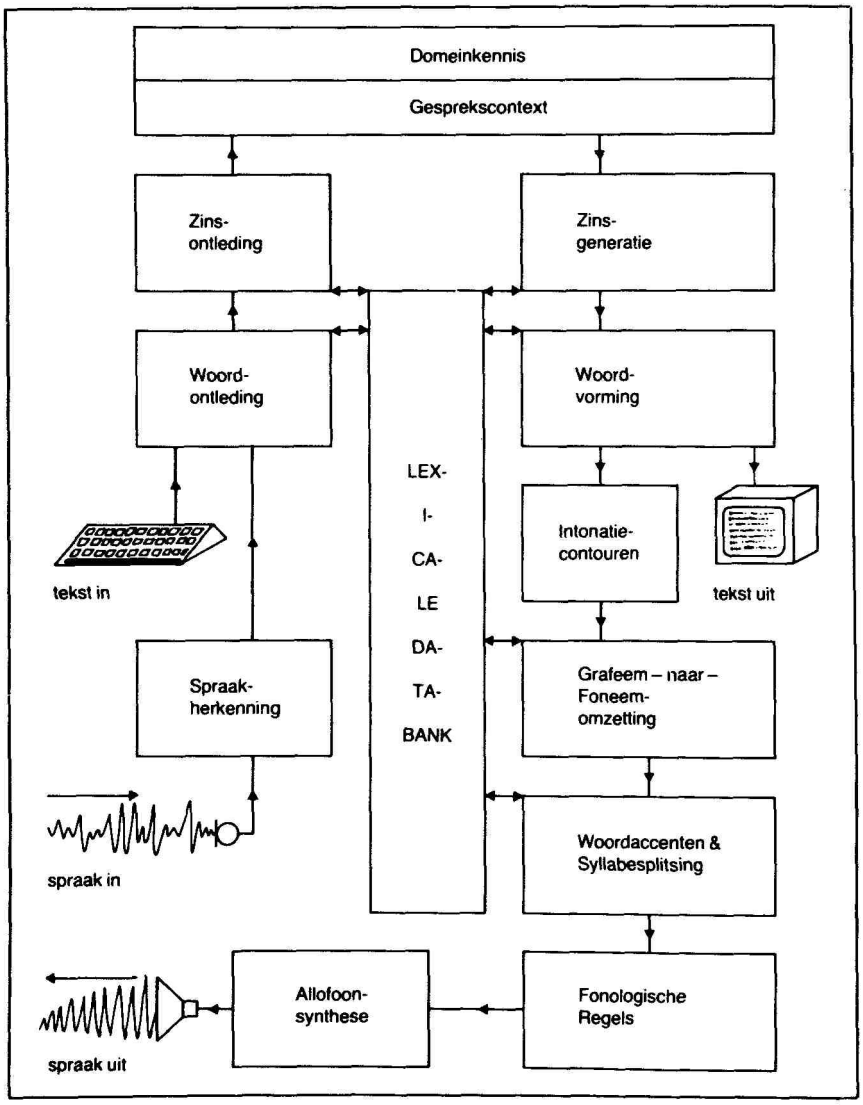
In de kantoorsituatie kan een auteursysteem zijn nut vooral bewijzen ten behoeve van het automatisch aanmaken van grote aantallen geïndividualiseerde standaarddocumenten (bijvoorbeeld zakenbrieven). De systemen die daarvoor momenteel in zwang zijn — meestal werkend op basis van "bouwstenen" of "menu's" — laten vrijwel geen linguïstische variatie toe en leiden daardoor tot houterige en saaie teksten. Het gevolg is dat ze slecht worden gelezen.

2. TAALMODULES VOOR HET NEDERLANDS

Het Nederlandstalige dialoog- en auteursysteem dat in Nijmegen wordt gebouwd is modulair van opzet. Figuur 2, die de samenhang van de diverse taalmodules in beeld brengt, kan gezien worden als een soort uitvergroting van Figuur 1. Bijvoorbeeld, van de zinsontleder in Figuur 1 wordt een afzonderlijke component voor Woordontleding afgesplitst. Bovendien houdt Figuur 2 rekening met de mogelijkheid van gesproken in- en uitvoer. Deze toevoeging vloeit voort uit onze samenwerking met het Spraaktechnologie-project van het Fonetisch Instituut van de K.U.N. Onder leiding van Dr. L. Boves gewerkt daar gewerkt aan spraaksynthese voor het Nederlands, onder meer voor de bouw van een hardopvoorleesmachine ("tekst-naar-spraak").

2.1 Woordvorming

We beschikken over een module die de verbogen of vervoegde vormen van vrijwel alle niet-samengestelde en van de meeste samengestelde Nederlandse woorden kan berekenen. Het programma, dat grotendeels door Drs. D. Bakker is vervaardigd, raadpleegt een indexed-sequential file die de fonologische en morfologische gegevens uit de "Dikke van Dale" bevat (zie par. 2.5). Een woord dat niet in dit woordenboek voorkomt, wordt volgens standaard morfologische regels verbogen of vervoegd nadat de gebruiker de woordsoort heeft aangegeven. Voor het berekenen van de vormen van samengestelde woorden is inzicht in hun morfologische structuur vereist. De van-Dale-file bevat echter niet altijd voldoende gegevens daaromtrent. Dit heeft onder meer tot gevolg dat trouwring (een samenstelling) op dezelfde manier wordt verbogen als haring (een enkelvoudig morfeem): de verkleinwoorden worden trouwrinkje en harinkje. Werkwoorden met scheidbaar prefix (op-bellen) zijn wèl als zodanig in de van-Dale-file gemarkeerd; deze krijgen dan ook de juiste vervoeging (op-ge-beld versus



Figuur 2. Taal- en spraakmodules. (Tekening overgenomen uit TNO Project, 1983, 11, pag. 105.)

ge-stofzuigd). Voor dit probleem is een oplossing in zicht. Vereist is een morfologische ontleder die samenstellingen als zodanig kan herkennen. Hierover handelt de volgende paragraaf.

2.2 Woordontleding

Drs. D. Bakker heeft samen met een van ons (De Smedt) een algoritme ontwikkeld voor morfologische ontleding van (al dan niet samengestelde) woorden. Deze bestaat uit twee delen. Eerst wordt het ingevoerde woord opgedeeld in letterreeksen die een woordstam zouden kunnen zijn. Reeksen die inderdaad in de Van-Dale-file blijken te staan, krijgen een woordsoort-label mee. De aldus herkende woorddelen komen vervolgens onder behandeling van een Parser (ontleder) die een honderdtal woordvormingsregels bevat. Alleen die combinaties van woorddelen die kloppen met deze regels worden geaccepteerd en voorzien van een structuurbeschrijving.

Tot nu toe wordt geparseerd met behulp van een zogenaamde Chart Parser (zie Winograd, 1983). Omdat deze nogal traag werkt zullen we binnenkort overstappen op een snellere methode. Een tweede snelheidswinst die we hopen te behalen is gebaseerd op fonotactische regels met behulp waarvan letterreeksen op een slimmere manier in woorddelen gesplitst kunnen worden. Hieraan wordt gewerkt samen met Drs. J. Wester van het Fonetisch Instituut. Een derde middel om het parseertempo op te voeren bestaat in het buiten beschouwing laten van ontledingen die een bepaald niveau van complexiteit overschrijden (bijvoorbeeld na ontleding van tabaksdozen tot tabaksdoos+en wordt niet verder gezocht naar tabak+s+doos+en). Voor veel toepassingen is dit niet nodig.

2.3 Zinsbouw

De zinsgenerator construeert Nederlandse zinnen op basis van de IPG (Incrementeel-Procedurele Grammatica van Kempen en Hoenkamp, 1982). Het programma krijgt betekenisrepresentaties in de vorm van “case-frames” als invoer. Makers van het programma zijn Ir. P. Desain en Drs. H. Schotel. De generator heeft heel wat syntactische constructies onder de knie maar is nog lang niet compleet. Zo zijn er nog geen voorzieningen voor pronominalisatie. Ook de volgorde van bijwoordelijke bepalingen, die op allerlei plaatsen in de zin kunnen verschijnen, is niet goed geregeld. In beide gevallen gaat het om erkende linguïstische problemen die nog slechts ten dele zijn opgelost. We moeten hier derhalve vol-

staan met onvolmaakte vuistregels. Het lexicon waarmee de zinsgenerator momenteel werkt omvat niet meer dan enkele tientallen woorden (die tot allerlei grammatische woordsoorten behoren), maar is wel gemakkelijk uitbreidbaar.

2.1 Zinsontleding

De zinsontleder die we gebouwd hebben — in hoofdzaak het werk van Drs. H. Schotel (1983) — is gebaseerd op de idee van "analyse door synthese". Een ingetypte zin wordt woord-voor-woord (van links naar rechts) afgehandeld. Voor elk binnenkomend woord wordt opgezocht welke betekenis het draagt en welke syntactische kenmerken het bezit. Deze twee typen informatie dienen als invoer tot twee processen: conceptuele ontleding en syntactische ontleding. Het ene berekent een betekenisrepresentatie voor de ingetypte zin, het andere een syntactische structuur. Berekening van de syntactische structuur komt in feite neer op het (re-)synthetiseren van de ingetypte zin: we gebruiken hiervoor de IPG-zinsgenerator. Het principe van analyse door synthese biedt het voordeel dat hetzelfde stelsel van syntactische regels en dezelfde syntactische processor gebruikt kunnen worden voor zinsproductie èn zinsontleding.

Momenteel Is de zinsontleder nog kwetsbaar. Dit hangt samen met de volgende omstandigheden. De programmatuur voor conceptuele ontleding is slechts rudimentair. Dit proces is in sterke mate afhankelijk van "kennis van de wereld". Het schrijven van programmatuur die in de praktijk bevredigend functioneert, kan daarom beter wachten tot bekend is op welk inhoudelijk domein de te ontleden zinnen betrekking zullen hebben. De tweede omstandigheid heeft te maken met ambiguïteit en onwelgevormdheid van de invoerzinnen. Een flinke proportie van de te ontleden zinnen blijkt in de praktijk meer dan èèn interpretatie toe te laten, of juist grammaticafouten te bevatten. Dit maakt het nodig om een soort boekhouding bij te houden van het ontstaan en verdwijnen van interpretatiemogelijkheden tijdens het doorlopen van de ingetypte zin. We werken aan een "filtermechanisme" dat deze boekhouding verzorgt. Aldus hopen we binnen afzienbare tijd over een meer flexibele en robuuste zinsontleder te beschikken.

2.5 Lexicale Databank

Drs. Schotel heeft een snelle en efficiënte lexicale databank opgezet uitgaande van de van-Dale-file die ongeveer 200.000 trefwoorden bevat. Deze van oorsprong sequentiële file is omgezet in een indexed-sequential file, zodat random access mogelijk werd. Vervolgens werd de structuur van de file ingrijpend gewijzigd. Alle trefwoorden — die zoals gebruikelijk in hun citatievorm vermeld stonden — werden omgezet tot stammen. En de onregelmatige vormen van alle trefwoorden werden als afzonderlijke trefwoorden (stammen) aan de file toegevoegd. Het aantal trefwoorden bedraagt nu ongeveer 225.000. Het hele bestand beslaat minder dan 10 Mbyte extern geheugen.

2.6 Zinsintonatie

Op basis van de intonatiegrammatica van het IPO te Eindhoven heeft Drs. C. van Wijk een programma geschreven dat Intonatiecontouren berekent voor mededelende en vragende zinnen. De invoer bestaat uit een door de zinsgenerator (IPG) geproduceerde "oppervlaktestructuur" waarin bepaalde woorden zijn gemarkeerd als drager van zinsaccent. (Deze markering wordt verkregen door bij de conceptuele invoer tot de zinsgenerator bepaalde betekenisfragmenten te voorzien van een +Accent-label.) De berekening van een contour geschiedt in twee stappen. Eerst wordt uit de oppervlaktestructuur de syntactische informatie geselecteerd die prosodisch relevant is. Aan de hand hiervan worden vervolgens de toonhoogtebewegingen bepaald die tezamen een welgevormde contour uitmaken (zie van Wijk en Kempen, 1984).

2.7 Domeinkennis en Gesprekscontext

Voor het representeren van inhoudelijke domeinkennis en van de dialoogcontext maken we gebruik van ORBIT (de Smedt, 1984a). ORBIT is een object-gericht kennisrepresentatiesysteem, vergelijkbaar met FLAVORS op de Symbolics Lisp Machine. Een belangrijk mechanisme in ORBIT is inheritance, waardoor objecten (kenniselementen, begrippen) op dynamische wijze eigenschappen van andere objecten kunnen overerven. Dit verschaft ons een handig uitgangspunt voor het opzetten van logische inferentieprocessen.

Dank zij zijn modulariteit en leesbaarheid is ORBIT een programmeeromgeving waarin niet alleen "kennis van de wereld" maar ook morfologische, syntactische, lexicale en andere

linguïstische kennis effectief kan worden behandeld (de Smedt, 1984b).

3. BLAUWDRUK VAN EEN NEDERLANDSTALIG AUTEURSYSTEEM

De meeste van de hierboven geschetste taalmodules zullen ook deel uitmaken van het auteursysteem. In feite streven we naar een maximale integratie van dialoog- en auteursystemen, zodat we met een minimum aan programmatuur een maximum aan functionaliteit behalen. De toegevoegde onderdelen zijn een Tekst-editor en een Syntactische Editor.

Aan het ontwerp van Figuur 1 ligt de eis ten grondslag dat teksten op twee manieren moeten kunnen worden behandeld: enerzijds als reeksen karakters die kunnen worden opgeslagen in een file en zichtbaar gemaakt op een beeldscherm, anderzijds als linguïstisch rijk gestructureerde objecten. Het auteursysteem dient beide typen tekstrepresentatie — zowel de orthografische als de linguïstische — te kunnen opbouwen en bewaren. Bovendien moeten ze in onderlinge samenhang kunnen worden gemanipuleerd. Aldus ontstaan allerlei nieuwe mogelijkheden voor het redigeren van teksten. Bijvoorbeeld, de gebruiker geeft aan het auteursysteem te kennen dat een bepaalde passieve zin in actieve vorm gezet moet worden. Hij/zij doet dit door een commando te geven aan de Syntactische Editor. Deze schakelt op zijn beurt de zinsgenerator in die de gewenste transformatie volgens grammaticaregels uitvoert. De woorden van de resulterende zin worden tenslotte in de tekstfile opgenomen en op het beeldscherm zichtbaar gemaakt. Een tweede voordeel betreft de mogelijkheid om linguïstisch gedefinieerde eenheden te kiezen als argument van editor-commando's. In huidige tekstverwerkers kunnen wel letters, woorden en tekstregels worden verplaatst, ingevoegd, weggehaald, aangewezen e.d., maar geen zinsdelen. De toekomstige gebruiker van het auteursysteem zal ook kunnen beschikken over commando's voor het verwisselen, weghalen e.d. van met name genoemde of aangewezen zinsdelen (bijvoorbeeld "verwissel de zinsdelen aan weerszijden van de cursor", "spring naar de persoonsvorm").

De gebruiker zal de gedaante van een zin bovendien kunnen beïnvloeden met behulp van de Tekst-editor, bijvoorbeeld door bij een zelfstandig naamwoord een meervoudsuitgang in te typen. Het systeem zal dan met behulp van de zinsontleder de achterliggende linguïstische representatie van de zin aanpassen, en de gevolgen daarvan voor andere woorden doorberekenen en zichtbaar maken.

De gebruiker van het auteursysteem beschikt aldus over twee methoden om een tekst te redigeren. Hij kan rechtstreeks de orthografische representatie van de tekst wijzigen door in een op het scherm afgebeelde tekstfile te typen. (Dit is de procedure die wordt gevolgd in gewone tekstverwerkers.) Of hij kan de Syntactische Editor aanroepen en vragen om bepaalde veranderingen in de linguïstische representatie aan te brengen. In het eerste geval zal de zinsontleder zorgdragen voor bijwerking van de linguïstische representatie. In het tweede geval zal de zinsgenerator de linguïstische structuur reconfigureren. Beide procedures leiden uiteindelijk tot aanpassing van de orthografische én linguïstische representaties van de tekst.

Een niet te onderschatten probleem bij de ontwikkeling van het toekomstige auteursysteem betreft het gebruikersinterface. De ingewikkeldheid ervan zal noodzaken tot het inschakelen van zeer geavanceerde grafische technieken, onder meer voor visuele weergave van de zinsbouw. Ook zal overvraging van de grammaticakennis bij de gebruiker moeten worden voorkomen.

* * *

We zijn ons ervan bewust dat het construeren van een dialoog- en auteursysteem dat in de praktijk voldoet nog heel wat voeten in de aarde zal hebben. Dit is evenwel geen excuus om het werk uit te stellen, temeer omdat in de ons omringende taalgebieden al forse onderzoeks- en ontwikkelingsprojecten lopen, al dan niet geïnspireerd op het Japanse Vijfde-Generatie-Computerplan. Zojuist is het eindrapport verschenen van de ZWO-Werkgroep Taal- en Spraaktechnologie die eind 1982 is ingesteld op initiatief van het Directoraat-Generaal voor Wetenschapsbeleid. De conclusies en aanbeveling van de Werkgroep wettigen de hoop dat het taal- en spraaktechnologisch onderzoek in Nederland spoedig met mèèr kracht dan tot nu toe ter hand zal worden genomen.

REFERENTIES

Bunt, H., Taal, kennis en computer. Openbare Rede, K.H. Tilburg, 1984.

Frase, L.T., The UNIX Writer’s Workbench Software: Philosophy. The Bell System Technical Journal, 1983, 62, 1883-1890.

Heidorn, G.E., K. Jensen, L.A. Miller, R.J. Byrd, & M.S. Chodorow, The EPISTLE text-critiquing system. IBM System Journal, 1982, 21., 305-326.

Kempen, G., Language facilities in informatlon systems: asset or liability? In: J. van Apeldoorn (ed.), Man and Information Technology: towards friendlier systems. Delft: Delft University Press.

Kempen, G. & E. Hoenkamp, An incremental procedural grammar for sentence formulation. Rapport, K.U. Nijmegen, 1982 (gaat verschijnen in Cognitive Science).

Schotel, H.$_t$ Analyse door synthese van Nederlandse zinnen. Rapport, K.U. Nijmegen, 1983.

Smedt, K. de, ORBIT: an object-oriented extension of LISP. Rapport, K.U. Nijmegen, 1984a.

Smedt, K. de, Using object-oriented knowledge-representation techniques in morphology and syntax programming. Proceedings of the Sixth European Conference on Artificial Intelligence. Pisa, 1984b.

Wijk, C. van & G. Kempen, From sentence structure to intonation contour: An algorithm for computing intonation contours on the basis of sentence accents and syntactic structure. In: B. Müller (ed.), Sprachsynthese. Hildesheim: Olms Verlag (Reihe Germanische Linguistik), 1984.

Winograd, T., Language as a cognitive process. Reading, Mass.: Addison-Wesley, 1983.