Hoofstuk 14

Mens en computer als taalgebruikers

1 Taal en rekenen

Een van de klassieke onderwerpen van de taalpsychologie staat bekend onder de naam 'linguïstische relativiteitstheorie'. De eerste en invloedrijkste formulering is afkomstig van Benjamin Whorf, een Amerikaanse chemicus en taalkundige die een tijdlang als verzekeringsexpert werkte. In die hoedanigheid was het hem opgevallen hoe vaak onbedachtzaam handelen uitgelokt wordt door een ongelukkig gebruik van woorden. Bijvoorbeeld het opsteken van een sigaret in de onmiddellijke nabijheid van een bord met het opschrift 'lege benzineblikken'. Het woordje 'leeg' wekte volgens Whorf de suggestie dat elk brandgevaar afwezig was, terwijl toch heel goed brandbare dampen in de blikken konden zijn achtergebleven. 's Mensen denken, doen en laten staan kennelijk sterk onder invloed van zijn taalgebruik.

Misschien is dit elementaire taalpsychologische verschijnsel mede debet aan het feit dat computers - 'rekenaars' - hoofdzakelijk worden geassocieerd met numerieke wiskunde. En dit terwijl een van de uitvindingen die tot de computer hebben geleid, afkomstig is van studenten die een elektrische schakeling bedachten om hun logicasommen automatisch te laten oplossen. Van meet af aan moet derhalve duidelijk zijn geweest dat computers veeleer 'symboolprocessoren' waren dan 'rekenmachines'. Niettemin is men bij het ontwerpen van zowel apparatuur als programmatuur vrijwel uitsluitend gericht geweest op de uitvoering van numerieke algoritmen. Pas na aanzienlijke prijsdalingen van de apparatuur is alom het inzicht doorgebroken dat computers heel goed inzetbaar zijn bij niet-numerieke aspecten van administratie, tekstverwerking, gegevensopslag enzovoorts. Heel aarzelend zien we momenteel de betekenis van het woord 'rekenen' zich verwijden tot 'het uitvoeren van een algoritme' of 'symboolmanipulatie'. *Rekenen op taal* - om een boektitel van Hugo Battus aan te halen - is daardoor niet langer beperkt tot woordentellerij en andere kwantitatieve behandelingen van taalmateriaal, maar kan nu slaan op alles wat nodig is voor het begrijpen en produceren van natuurlijke taal door computers.

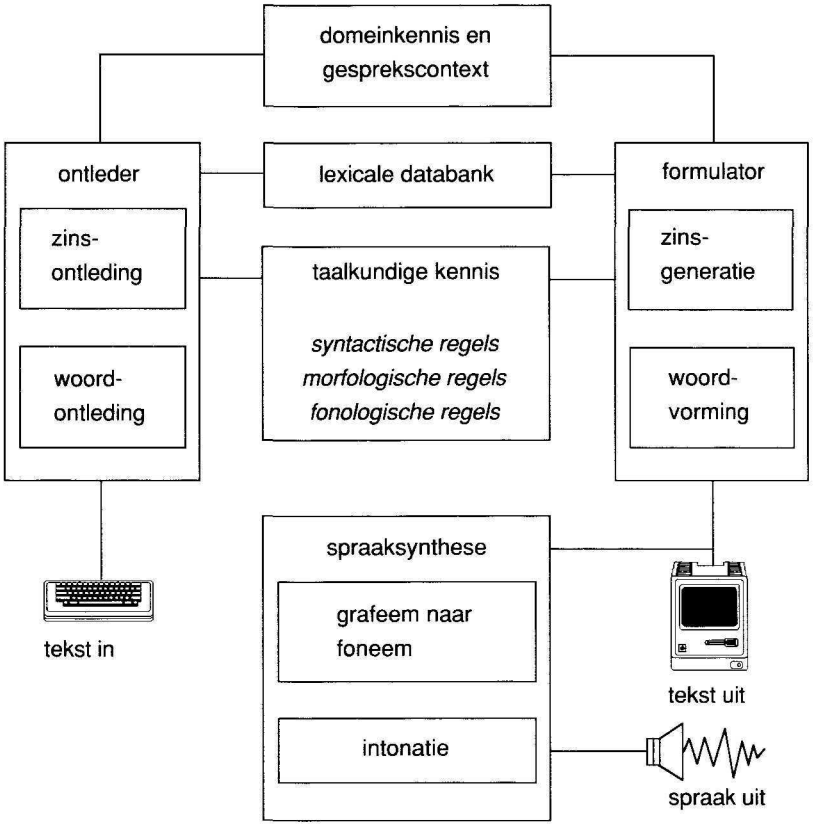
Deze inleidende beschouwing was niet geheel en al vrij van berekening. In vele sectoren der alfa- en gammawetenschappen neemt de graad van formalisering en algoritmisering snel toe. Tot die sectoren behoren ook de vakken waarin wij werkzaam zijn: de taalspsychologie - ook psycholinguïstiek genaamd - en de taalkunde (linguïstiek). Slechts een deel van deze toename mag op het blazoen van numerieke methoden worden geschreven. Dit betreft enerzijds waarschijnlijkheidsberekening en statistiek, anderzijds signaalverwerking en meet- en regeltechniek. Minstens even bepalend voor de stijgende graad van exactheid was de introductie van niet-numerieke methoden gebaseerd op formele logica, formele grammatica's, automatentheorie, zoekalgoritmen, produktiesystemen, schematheorie, unificatie en dergelijke. De belangrijkste stelling die we in dit hoofdstuk naar voren willen brengen, is nu de volgende:

De informatisering van de alfa- en gammawetenschappen blijft tot nu toe eenzijdig geconcentreerd op numerieke gegevensverzameling en -verwerking. De mogelijkheden en de noodzaak tot informatisering op niet-numeriek gebied worden buiten de eigen vakkring onvoldoende onderkend. Dit zet niet alleen een rem op de vooruitgang in de alfa- en gammawetenschappen, maar schuift ook een breed scala van maatschappelijk belangrijke informaticatoepassingen op de lange baan, met name die welke gebaseerd zijn op kennistechnologie.

Het zou weinig moeite kosten deze stelling te onderbouwen met cijfers en berekeningen omtrent het computergebruik door alfa- en gammawetenschappers en -studenten. Van meer gewicht vinden we evenwel dat beleidsmakers in informaticaland een indruk krijgen van wat de bedoelde niet-numerieke methoden en systemen concreet inhouden. We hebben daarom gekozen voor een illustratieve aanpak. Concrete voorbeelden, zo zegt ons vak, beklijven beter dan abstracte redeneringen. Tot zover het element van berekening in dit overigens geheel niet-numerieke verhaal.

2 Een mens-machinedialoog

Het voorbeeld dat we hier in enig detail willen uitwerken, heeft betrekking op zogenaamde *dialoogsystemen*. Deze dienen ervoor om in een natuurlijke taal zoals Nederlands of Engels te kunnen communiceren met computersystemen (zie ook hoofdstuk 11). Onder bepaalde omstandigheden en voor bepaalde typen van gebruikers is dit het aangewezen middel om de computer opdrachten te geven en antwoorden terug te krijgen. Het ontwikkelen van dergelijke systemen, die in enigszins bruikbare vorm voor nog slechts enkele talen beschikbaar zijn, vergt de gecombineerde inspanning van linguïsten, logici, psychologen en informatici. Het is een bijna ideaal proefveld voor het uittesten en integreren

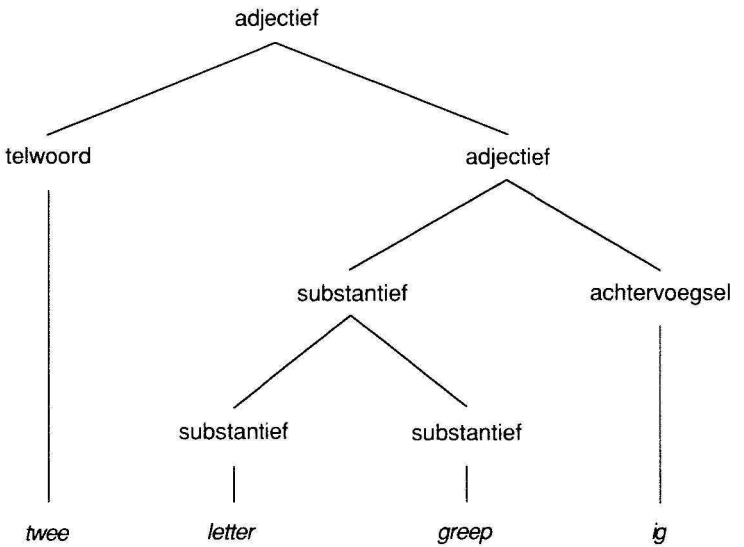


Figuur 1

van velerlei theorieën en methoden uit die wetenschappen. Een typisch stelsel van modulen waaruit een dialoogsysteem zou kunnen zijn opgebouwd, staat weergegeven in figuur 1. (Voor een ietwat verschillende opdeling zie figuur 2 van hoofdstuk 12.) Aan de linkerzijde treft u de modulen aan die betrokken zijn bij het interpreteren van een opdracht of vraag. Rechts staan de verschillende stappen die nodig zijn om een antwoord te formuleren in de gedaante van gesproken of geschreven tekst. In het midden en aan de bovenzijde van de figuur bevinden zich diverse kennisbestanden die de modules ten dienste staan. Deze kennis is ten dele linguïstisch van aard - met name de morfologische, lexicale en syntactische regels - maar heeft voor een ander deel betrekking op het inhoudelijke domein waar het de gebruiker in feite om te doen is. Het zou

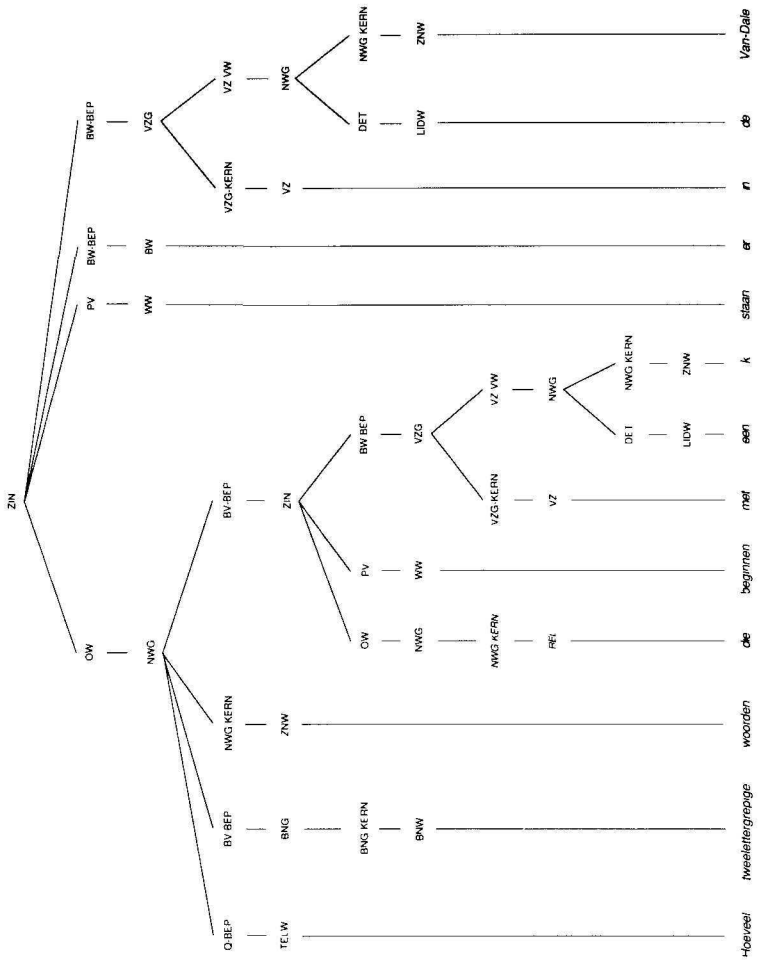
bijvoorbeeld een juridisch, medisch of technisch informatiebestand of expertsysteem kunnen zijn dat hij/zij wil raadplegen. In ons voorbeeld gaan we uit van een - nu nog denkbeeldige - taaldatabank die grote hoeveelheden informatie bevat over woorden en regels uit allerlei talen. Een vraag aan zo'n databank zou kunnen zijn: 'Hoeveel tweelettergrepige woorden die beginnen met een k staan er in de Van Dale?' Het - al dan niet correcte - antwoord zou bijvoorbeeld kunnen luiden: 'Van Dale bevat 3052 woorden van twee lettergrepen met k als beginletter.'

Laten we nu de symboolmanipulaties die de modules uitvoeren tijdens de behandeling van dit vraag-antwoordpaar, eens op de voet volgen. Een eerste taak is weggelegd voor de woordontleder die met behulp van morfologische regels de woorden ontdoet van voor- en achtervoegsels, en waar nodig samenstellingen herkent. Aan de hand van de aldus achterhaalde 'morfemen' stelt hij van alle individuele woorden de woordsoort en de betekenis vast via raadpleging van het lexicon. De structuur van het woord 'tweelettergrepig' is afgebeeld in figuur 2. Zulke gegevens zijn nodig als startpunt voor de volgende stap: zinsontleding.



Figuur 2

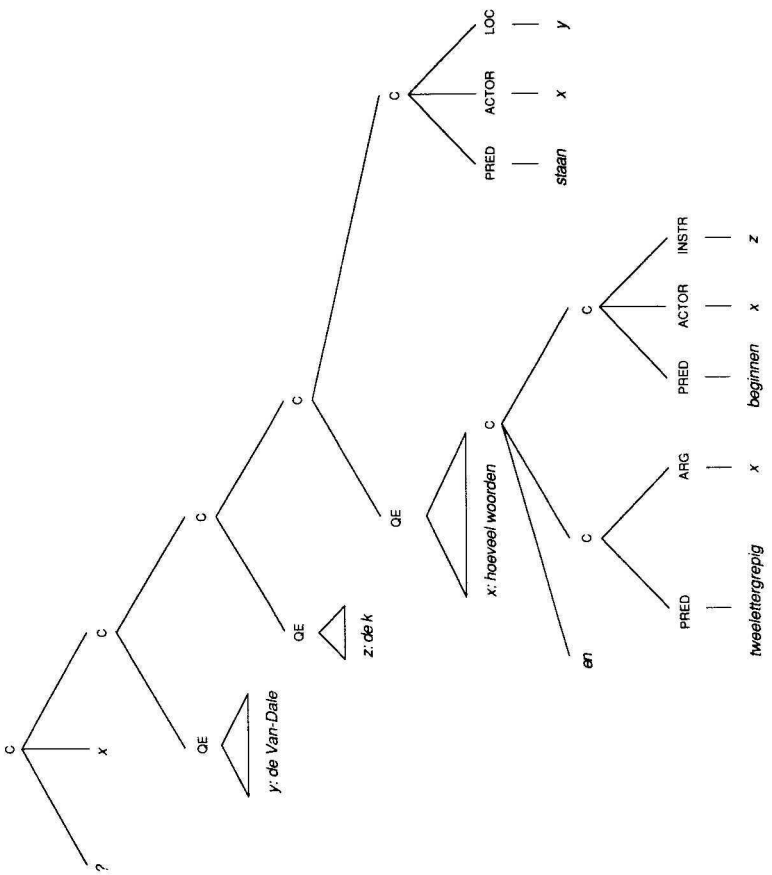
De zinsontleder spoort syntactische relaties tussen woorden op, herkent welke woorden samen een woordgroep vormen, en hoe deze groepen bij elkaar passen in een zinsverband. Het resultaat van deze analyses wordt



Figuur 3

gewoonlijk weergegeven in een boomdiagram (figuur 3) met op de knopen de namen van de opgespoorde woordsoorten, woordgroepen en syntactische relaties.

Tijdens de derde stap wordt een soort logische of conceptuele formule samengesteld die de betekenis van de zin nauwkeurig en ondubbelzinnig weergeeft (figuur 4). De symbolen C en QE staan voor 'conceptualisatie' en 'quantorexpressie'. De naamwoordgroepen uit de syntactische boom corresponderen met quantorexpressies in de logische formule. Ook de omzetting van syntactische boom naar conceptuele formule geschiedt in



Figuur 4

samenwerking met het lexicon; met name wordt intensief van betekenisinformatie gebruik gemaakt.

De conceptuele formule, die er voor menselijke beschouwers nogal afschrikwekkend uitziet, is in feite heel homogeen van opbouw en voor een computer gemakkelijk te interpreteren. De laatste stap die nog nodig is betreft de omzetting van de formule in een zoekopdracht die door de taaldatabank begrepen wordt. De formule in ons voorbeeld (figuur 5) behoort niet tot de zoektaal van een of ander gangbaar database management system, maar is een expressie uit een op de programmeertaal LISP gebaseerd kennisrepresentatiesysteem (ORBIT; De Smedt, 1984; zie hoofdstuk 12). De formule laat zich bijna als gewoon Nederlands lezen:

gevraagd wordt het getal dat de kardinaliteit aanduidt van de verzameling van alle woorden die voldoen aan de drie kenmerken.

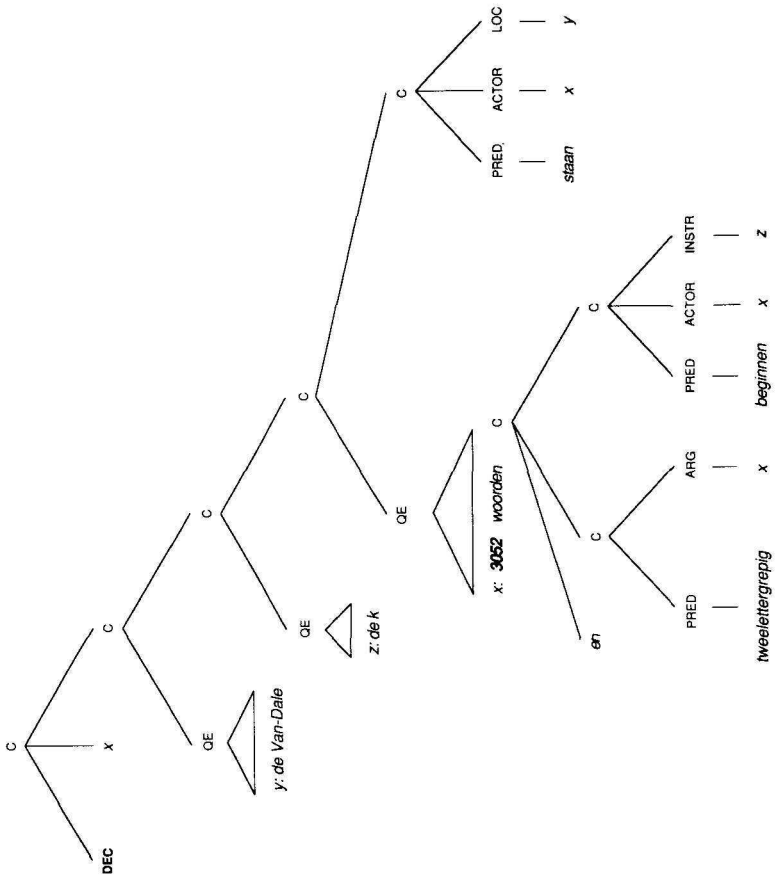
```
(een vraag
   (naar '(het getal
              (een kardinaliteit
                   (verzameling
                    (een verzameling
                         (intensie
                          '(alle 'woord
                              (met 'bestandsnaam 'Van-Dale)
                              (met 'aantal-lettergrepen 2)
                              (met 'beginletter 'k))))))))
```

Figuur 5

Zodra de vraag begrepen is en de zoekopdracht uitgevoerd, gaat het dialoogsysteem over tot de beantwoordingsfase. Het genereren van een antwoord in de vorm van een goede Nederlandse zin doorloopt dezelfde stappen, maar nu in omgekeerde volgorde. Uitgangspunt vormt de 'propositie' die het kennisrepresentatiesysteem heeft afgeleid in antwoord op de zoekvraag (figuur 6).

```
propositie-1
   + predikaat: (het getal
                     (een kardinaliteit
                          (verzameling
                            (een verzameling
                                 (intensie
                                  '(alle 'woord
                                      (met 'bestandsnaam 'Van-Dale)
                                      (met 'aantal-lettergrepen 2)
                                      (met 'beginletter 'k)))))))
   +argument:  3052
```

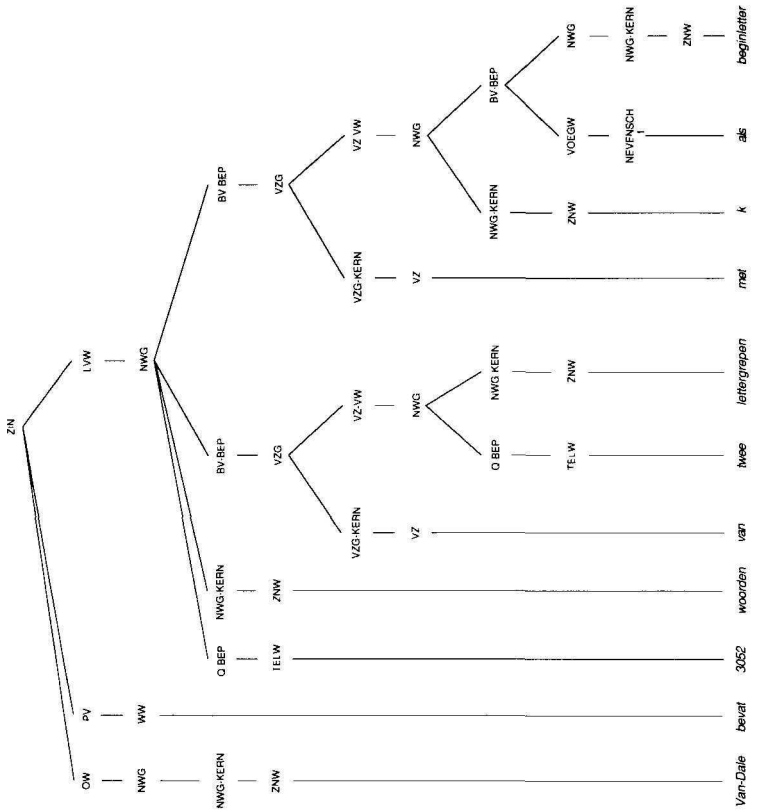
Figuur 6



Figuur 7

De meest directe manier om hieruit een zin samen te stellen, maakt gebruik van de logische formule van de bijbehorende vraagzin. Deze is bewaard gebleven, zo nemen we aan, als onderdeel van de gesprekscontext (vergelijk figuur 1). Het argument van figuur 6, namelijk het getal 3052, wordt ingevuld in de 'bevraagde' quantorexpressie, namelijk die welke hoort bij de index x in de top van die formule. Het vraagteken wordt vervangen door 'DEC(laratief)' ten teken dat het niet langer om een vragende maar om een bevestigende zin gaat. Deze formule (figuur 7) wordt nu ingevoerd in de zinsgenerator, die er een syntactisch boomdiagram uit construeert (figuur 8).

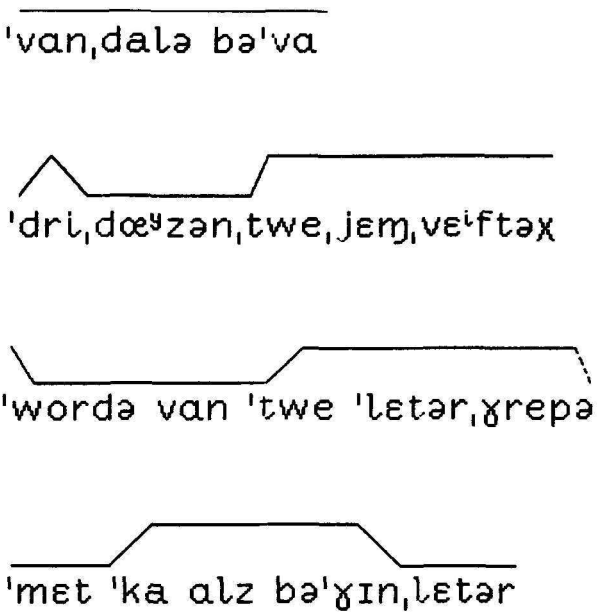
De eindknopen van deze boom komen vervolgens in behandeling bij de



Figuur 8

woordvormingsmodule. Deze berekent de precieze vorm van alle woorden volgens de regels van verbuiging en vervoeging. Het resultaat staat op een wat ongewone manier weergegeven in figuur 9. De lettertekens zijn afkomstig uit een fonetisch alfabet, en de lijnstukken boven de woorden markeren een ruwe intonatiecontour. De contour is berekend door een van de modules voor spraaksynthese. Soms is een antwoord in gesproken vorm handig of noodzakelijk, bijvoorbeeld in verband met de attentiewaarde, of wanneer het dialoogsysteem niet via een gewone terminal toegankelijk is.

Aan het eind van dit voorbeeld gekomen willen we nog twee opmerkingen toevoegen. Op de eerste plaats, de notaties en informatiestructuren die we ten tonele hebben gevoerd, zijn afkomstig uit een Nijmeegs project



Figuur 9

dat gericht is op de bouw van een Nederlandstalig dialoogsysteem (Kempen, Konst & De Smedt, 1984; zie hoofdstuk 12). Van alle benodigde taalmodules bestaan op dit moment prototypen. Ze vertonen echter nog allerlei lacunes en kinderziekten, en zijn bovendien nog niet allemaal op elkaar aangesloten. Ten tweede, we hebben overwogen om in plaats van een taal*technologisch* een taal*psychologisch* voorbeeld te behandelen. We zouden dan een plaatje getoond hebben van het cognitieve taalgebruiksysteem dat zich ergens in het hoofd van menselijke sprekers moet bevinden. Erg veel verschil zou die keus niet gemaakt hebben, want volgens gangbare theorieën is het taalgebruiksysteem in de mens ongeveer net zo gebouwd als het technische dialoogsysteem dat we beschreven hebben. Ook de informatiestructuren die in de cognitieve modules worden berekend, zijn in essentie dezelfde – alhoewel we hier flinke slagen om de arm moeten houden, want zó gemakkelijk geeft de menselijke geest zijn geheimen niet prijs.

De essentiële verschillen tussen de werking van een hedendaags artificieel dialoogsysteem en het cognitieve taalgebruiksysteem in de hersenen bevinden zich op dieper niveau, namelijk in de *architectuur* van de symboolprocessoren. Bij de talrijke pogingen die gedaan zijn om

menselijke cognitie en intelligentie op computers te simuleren, is men zoals bekend op vele obstakels gestoten. Maar het belangrijkste struikelblok is ongetwijfeld de beperkte mogelijkheid tot *parallelle* symboolverwerking. In recente psychologische theorievorming, maar misschien nog sterker in het nieuwere kunstmatige-intelligentieonderzoek, wordt uitgegaan van de idee van 'extreem parallellisme'. Langs vele wegen zoekt men naar handzame en beheersbare systeemarchitecturen waarin complexe taken worden verdeeld over grote aantallen – honderden, duizenden – eenvoudige processoren die in een netwerk zijn verbonden.

3 Informatisering van alfa- en gammawetenschappen

We komen nu toe aan enkele conclusies en suggesties. We zijn de stellige overtuiging toegedaan dat informatisering van alfa- en gammavakken zich in belangrijke mate dient af te spelen op het gebied van de *kennistechnologie* (vergelijk hoofdstuk 10). Taaltechnologie is daar slechts een onderdeel van. Toepassingen van kennistechnologie treffen we aan in de patroonherkenning, bij expertsystemen, bij nieuwe computerondersteunde onderwijsvormen en informatiediensten, en in de cognitieve ergonomie. Dit is overigens pas het begin. Op langere termijn zal bijvoorbeeld het *opinieonderzoek* een andere gedaante krijgen, zodra het mogelijk wordt om subjectieve opvattingen en meningen automatisch uit lopende tekst te extraheren en subjectieve redeneringen op de computer te simuleren.

Dergelijke onderzoeksthema's blijken zeer gespecialiseerde computervoorzieningen te vergen, nogal afwijkend van die welke ontworpen zijn voor numeriek rekenwerk, administratie, databankinrichting, procesbesturing, beeldverwerking, enzovoorts. Wie op dit moment een indruk wil krijgen van machinerie zoals een kennistechnoloog zich zou wensen, kan de zogenaamde LISP-machines in ogenschouw nemen die door enkele kleinere Amerikaanse firma's sinds een paar jaar worden uitgebracht. Dit soort machines vindt gretig aftrek en de verwachting is dat ook grotere computerfabrikanten zich erop zullen storten. Dit zal enerzijds leiden tot prijsdalingen, anderzijds tot grotere vraag. We voorspellen dat de bestuurs- en beheersorganen die in dit land te maken hebben met beoefening van alfa- en gammawetenschappen, geconfronteerd zullen worden met vele aanvragen voor aanschaf van dit nieuwe type apparatuur. Dit zal het kostenpeil van deze vakken flink opjagen. Aan de andere kant zal hun opbrengst eveneens aanzienlijk kunnen stijgen. Het is ons bijvoorbeeld niet ontgaan dat althans in Nijmegen fysici, informatici, chemici en medici opvallende interesse aan de dag leggen voor de apparatuur en programmatuur die eenvoudige alfa's en gamma's daar ten behoeve van cognitief en kennistechnologisch onderzoek hebben opgebouwd. Het niveau van specialisatie zal naar onze mening alleen nog maar toenemen

wanneer de nu opkomende behoefte aan extreem parallellisme en andere onorthodoxe architecturen verder doorzet. Dergelijke computerfaciliteiten zullen voor alfa- en gammawetenschappers even onmisbaar worden als CAD/CAM-systemen dat nu al zijn voor vliegtuigbouwers en chipontwerpers. Zonder zulke voorzieningen zullen informaticatoepassingen die berusten op kennistechnologie slechts moeizaam van de grond kunnen komen.

In het kielzog van deze ontwikkelingen zal sterke behoefte ontstaan aan on-line kennisbestanden van allerlei soort. Eén voorbeeld hebben we al gezien: voor taaltechnologisch werk zijn geautomatiseerde woordenboeken met bijbehorende programmatuur een vereiste. Er is goede hoop dat met steun van het Directoraat-Generaal voor Wetenschapsbeleid van het Ministerie van Onderwijs en Wetenschappen zo'n lexicale databank binnenkort van de grond gaat komen.[1] Ook andersoortige kennis zal in de vorm van programmamodules of -pakketten beschikbaar dienen te komen. Bijvoorbeeld, veel expertsystemen zullen zinnig moeten kunnen redeneren met alledaagse begrippen van ruimte en tijd. Wie een goede module voor dit stukje gezond-verstandkennis bouwt - geen sinecure overigens - zal deze graag ter beschikking stellen van vakgenoten, al dan niet met commercieel oogmerk. Aldus hoeft niet elke bouwer van een expertsysteem opnieuw het wiel uit te vinden. Vele soorten van algemene en specialistische kennis zullen verspreiding gaan vinden in een nieuwe gedaante: niet alleen in de vorm van het gedrukte woord maar ook als min of meer kant-en-klare programmabouwstenen.

Ten slotte. Voorzover uw denken over informatisering van wetenschapsbeoefening onderhevig was aan het linguïstische-relativiteitseffect waarmee we openden, hopen we dat enigszins gecorrigeerd te hebben. De soorteigen informatisering van alfa- en gammavakken zal zich, anders dan bij bêtavakken, niet zozeer op numeriek als wel op kennistechnologisch gebied gaan afspelen. Alfa's en gamma's zullen alfa's en gamma's blijven.