077715

182969064F

RGN·HSRC

SENTRUM VIR BIBLIOTEEK- EN INLIGTINGSDIENSTE

CENTRE FOR LIBRARY AND IMFORMATION SERVICES

VERVALDATUM/DATE DUE

24 JUL 1995

0 0 0 2 0 8 3 7 5 6

001.3072068 HSRC LEXI 7

208375

Rekenaartoepassings in die taalwetenskap

(TEXTNET-projek in die LEXINET-program)

Verslag LEXI-7

Rekenaartoepassings in die taalwetenskap

(TEXTNET-projek in die LEXINET-program)

V.N. Webb
T.J.D. Bothma
R. Morris

Pretoria
Raad vir Geesteswetenskaplike Navorsing
1988

RGN BIBLIOTEEK
1989 2. 08.
HSRC LIBRARY
STANDKODE
0013072068
HSRC LEXI 7
AANWINSNOMMER
077715

V.N. Webb
T.J.D. Bothma
R. Morris

Afdeling Leksikologie

Instituut vir Taal- en Kunstenavorsing
Uitvoerende Direkteur: Dr K P Prinsloo

ISBN 0 7969 0699 8

© Raad vir Geesteswetenskaplike Navorsing, 1988

Gedruk en uitgegee deur die RGN
Pretoriusstraat 134
Pretoria

VOORWOORD

Die Afdeling Leksikologie in hierdie Instituut het teen die einde van 1986 'n ondersoek geloods na die gebruik van die rekenaar in die verwerking van taaldata. In die buiteland vind daar snelle ontwikkelings op die gebied plaas terwyl relatief min aandag in Suid-Afrika daaraan bestee word. Die doel met die ondersoek was om die belangrikheidswaarde vir Suid-Afrika van die gebied te bepaal en om daarvolgens aanbevelings te doen oor moontlike ontwikkelingstappe.

Hierdie verslag is een van 'n reeks van sewe wat verskillende aspekte van gerekenariseerde taalverwerking dek, soos bespreek in hoofstuk een. Dan is daar ook 'n agste verslag as 'n samevatting van die hele reeks. Al die verslae het die gemeenskaplike kenmerk dat hulle verslae van 'n terreinverkenningsaard is. As sodanig is hulle belangrike inligtingsdokumente en vertrekstukke vir die verdere uitbouing van 'n betreklik nuwe vakgebied in Suid-Afrika, naamlik die rekenaarhantering van taal.

In die ondersoek waaroor dit in hierdie verslag gaan, het die volgende medewerkers bydraes verskaf:

Dr. P.J. Badenhorst, Departement Afrikaans en Nederlands UP
Prof. T.J.D. Bothma, Departement Semitistiek UNISA
Mnr. Q. Gee, Departement Rekenaarwetenskap WITS
Prof. R.M. Klopper, Departement Afrikaans UZ
Prof. E. Kotzé, Departement Afrikaans UZ
Dr. R. Morris, Instituut vir Taal- en Kunstenavorsing RGN
Prof. J.C. Roux, Departement Afrikatale US
Dr. J. Vorster, Instituut vir Taal- en Kunstenavorsing RGN
Prof. J-C. Lejosne, Universiteit van Metz, Frankryk
Prof. V.N. Webb, Departement Afrikaans en Nederlands UP.

Die projek is gelei deur prof. Webb en die verslag is deur hom, prof. Bothma en dr. Morris opgestel.

Opregte dank word betuig teenoor al die medewerkers aan hierdie ondersoek. Vir die uitbouing van die linguistiek in Suid-Afrika kan die inligting, besprekings en voorstelle in hierdie verslag van groot waarde wees.

Vir die wyse waarop prof. Webb, as projekleier en verslagskrywer, en prof. Bothma, as medeskrywer, hulle taak uitgevoer het, is daar groot waardering.

Ten slotte wil ek graag vir dr. R. Morris, die programleier van die LEXINET-ondersoek, hartlik bedank vir haar bydraes tot die TEXTNET-projek en vir haar bestuur en deurvoering van die program in sy geheel.

K.P. PRINSLOO
UITVOERENDE DIREKTEUR: INSTITUUT VIR TAAL- EN KUNSTENAVORSING

INHOUDSOPGAWE

HOOFSTUK 1: INLEIDING

1.1 RAAMWERK VAN DIE ONDERSOEK

Die ondersoek waaroor hierdie verslag handel (die TEXTNET-ondersoek) was een van die projekte in die navorsingsprogram LEXINET en die rekenarisering van taal.

Die LEXINET-program was gerig op navorsing wat met steun van die Wetenskaplike Adviesraad onderneem is ten einde rekenaartoepassings op natuurliketaalverwerking te ondersoek. Die oorsprong van die besluit om die navorsing te onderneem, was enersyds bevindinge uit vorige RGN-navorsing (wat daarop gedui het dat rekenarisering op taalgebiede in Suid-Afrika min benut word), en andersyds die waarneming dat rekenaartoepassings op die verwerking van taal in die buiteland 'n besonder hoë prioriteit geniet. In 'n aantal lande word daar grootskaals bestee aan die gebied, terwyl Engeland en Japan programme vir gerekenariseerde taalverwerking, en die toepassings ten opsigte van kunsmatige intelligensie wat daarmee saamhang, tot nasionale ondernemings verklaar het.

Die terrein van gerekenariseerde taalverwerking is wyd. Vir die doeleindes van die LEXINET-ondersoek is dit onderverdeel in sewe deelterreine, soos weerspieël in die titels van die sewe verslae wat nou verskyn (die verslae se noemname staan tussen hakies):

- Die rekenarisering van terminografiese prosesse (TERMNET)
- Die rekenarisering van leksikografiese prosesse (WORDNET)
- Rekenaargesteunde vertaling (TRANSNET)
- Rekenaartoepassings in die taalwetenskap (TEXTNET)
- Kunsmatige intelligensie en die prosessering van natuurlike taal (AILANG)
- Rekenaarfasiliteite vir die hantering van taal (PROLANG)
- Gerekenariseerde taaldokumentasie-databasisse (DOCNET)

Die sewe deelterreine hou op verskillende wyses met mekaar noue verband. Die samehang tussen hulle kan kortweg geïllustreer word aan die hand van toepassings op die gebied van mens-masjienraakvlakstelsels. 'n Stelsel wat aan gebruikers die moontlikheid bied om in gewone taal navrae aan 'n databasis te stel en om antwoorde in gewone taal te ontvang, moet onder andere komponente bevat wat taal kan ontleed, data kan sorteer en taal kan genereer. Hierdie funksionering veronderstel:

- gepaste apparatuur en programmatuur (die gebied wat deur die PROLANG-verslag gedek word);
- een of meer taaldatabasisse (die onderwerp van die DOCNET-verslag);

- een of meer woordeboeke (die TERMNET- en WORDNET-verslae bespreek onder meer ingeboude woordeboeke);
- 'n taalontleder *(parser)* en -genereerder (sodanige ontleders is kernkomponente in vertaalprogramme, soos onder andere bespreek in die TRANSNET-verslag;
- kunsmatige intelligensie en tegnieke op die terrein van taalverwerking (die gebied gedek deur die AILANG-verslag).

Hierdie illustrasie aan die hand van 'n bepaalde tipe stelsel is maar een voorbeeld van die samehang tussen die deelterreine. In die verslae word die saak verder bespreek, maar elke terrein word ook in eie reg behandel. So byvoorbeeld is daar in die verslag oor rekenaartoepassings in die taalwetenskap (die TEXTNET-verslag) enersyds sprake van die rekenaar in diens van taalnavorsing en andersyds bespreking van taalnavorsing in diens van gerekenariseerde taalverwerking.

Die aard van die LEXINET-ondersoek, naamlik 'n verkennings- en doenbaarheidstudie, bring mee dat die verslae as inligtingsdokumente, eerder as navorsingsverslae, beskou kan word. Ter ondersteuning van hierdie inligtingsfunksie word sommige van die verslae gekoppel aan 'n reeks meer tegniese bronstukke wat van die RGN se Afdeling Leksikologie bestel kan word.

Die inhoud van al die verslae word saamgevat in 'n hoofverslag getiteld LEXINET en die rekenarisering van taal. (Titel van die Engelse weergawe: LEXINET and the computer processing of language.)

1.2 TAALWETENSKAP EN DIE REKENAAR

Die TEXTNET-span se opdrag was om vas te stel hoe die taalwetenskap die rekenaar kan gebruik om sy hooftaak te verrig, naamlik om insig te verkry in die aard en funksionering van natuurlike menslike taal in die algemeen, en van spesifieke tale, soos Afrikaans en die hoof Afrikatale.

1.2.1 Die gebruik van die rekenaar in taalondersoek in die algemeen

Die waarde wat die rekenaar vir grammatikale ondersoek in die algemeen inhou, volg natuurlik direk op die funksies wat rekenaars kan verrig, soos herken/soek, sorteer/klassifiseer/orden, tel/bereken/uitvoer van statistiese toetse, simulering en die beheer en aandrywing van verskillende soorte apparaat. Hierdie vermoëns stel die linguis gevolglik daartoe in staat om take soos die volgende met behulp van die rekenaar te verrig:

- Om al die voorkomste van 'n bepaalde klank, morfeem, woord, sintaktiese konstruksie, ensovoorts in 'n bepaalde (gerekenariseerde) teks te identifiseer en in 'n lys/konkordansie uit te druk saam met hul kontekste en saam met hul presiese adresse in die teks (sodat elke voorbeeld vinnig opgespoor kan word). 'n Voorbeeld van die gebruik van patroonherkenning en patroonvergelyking/-

passing in sintaktiese analise kan gekry word in Boot se artikel oor die BOBRA-stelsel, Inligtingstuk 23.

- Om, afgesien van persentasies, gemiddeldes, standaardafwykings en die interrelasie tussen veranderlikes, die beduidendheid van die verskille tussen twee stelle data, ensovoorts te bepaal.

- Om, gegee die nodige taalkundige inligting en die nodige apparatuur, taalgedrag te simuleer, byvoorbeeld spraakklanke se artikulasieplek en/of hul klankkwaliteite sigbaar/hoorbaar voor te stel, en die verloop van gesprekke na te boots.

'n Tweede algemene nut van die rekenaar is dat dit binne 'n generatiewe benadering vir hipotesetoetsing gebruik kan word: in die geval van grammatikale beskrywings waarin oppervlaktevorme per reël afgelei word van onderliggende weergawes, kan die geldigheid van die geponeerde onderliggende vorme sowel as die geldigheid van die reëlordening per rekenaar vasgestel word.

Derdens: 'n besondere wins wat die rekenaar vir grammatikale ondersoek gebring het, is dat groot hoeveelhede data baie akkuraat ontleed kan word, asook dat uiters ingewikkelde interrelasies tussen verskillende verskynsels ondersoek kan word.

Elk van hierdie voordele sal hierna weer ter sprake kom.

1.2.2 Die gebruik van die rekenaar in taalondersoek en die gebruik van die taalwetenskap in die ontwikkeling van rekenaarstelsels

Soos reeds aangedui, gaan dit in hierdie verslag om die gebruik van die rekenaar as navorsingsinstrument in diens van taalwetenskaplike teorievorming en grammatikabeskrywing.

Hierdie gebruik van die rekenaar word vir die doeleindes van hierdie verslag rekenaartoepassings in taalondersoek genoem.

Naas rekenaartoepassings in taalondersoek staan die rekenaarlinguistiek. In die rekenaarlinguistiek gaan dit, vir die doeleindes van hierdie verslag, om die ontwikkeling van rekenaarstelsels waarmee tekste in natuurlike tale outomaties vertaal kan word (masjienvertaling: vgl. Hoofstuk 5) en waarmee daar met behulp van natuurlike tale dialoog tussen mens en masjien bewerkstellig kan word/inligting herwin kan word (kunsmatige intelligensie: vgl. Hoofstuk 2).

Die rekenaarlinguistiek is dus rekenaargerig, en die taalwetenskap staan daarin in diens van die rekenaar.

Hierdie verskil tussen rekenaartoepassings in taalondersoek en die rekenaarlinguistiek blyk netjies uit die verskil tussen VARBRUL en METAL.

VARBRUL is 'n pakket rekenaarprogramme waarmee die sisteem onderliggend aan variasieverskynsels ontdek kan word sodat die ondersoeker 'n reël vir die betrokke grammatikale proses kan skryf. VARBRUL is dus gewoon 'n analitiese instrument in diens van grammatikale ondersoek.

METAL is 'n rekenaarstelsel wat deur die Linguistic Research Center van die Universiteit van Texas (Austin) oor baie jare ontwikkel is vir die outomatiese vertaling van tegniese tekste uit Duits in Engels. Hoewel hierdie stelsel regstreeks op taalteoretiese en grammatikale werk steun, is die inligting aangepas by die vereistes wat die rekenaar stel met die gevolg dat dit prinsipieel verskil van die soort taalkundige inligting wat in grammatikas voorkom.

Hoewel dit vir die doeleindes van hierdie verslag nodig is om die onderskeid tussen rekenaartoepassings in taalondersoek en rekenaarlinguistiek deurentyd in gedagte te hou, moet die twee gebiede nie te skerp geskei word nie. Per slot van sake sluit hulle direk bymekaar aan:

- die werk wat aan rekenaargesteunde vertaling en kunsmatige intelligensie gedoen word, lewer insigte in die taalverskynsel wat van regstreekse waarde is vir die taalwetenskap (vgl. bv. De Stadler en Coetzer, Inligtingstuk 22);

- die werk wat aan rekenaargesteunde vertaling en kunsmatige intelligensie gedoen word, is direk afhanklik van taalkundige en grammatikale insigte.

As 'n illustrasie van die noue skakeling tussen rekenaartoepassings in taalondersoek en rekenaarlinguistiek kan verwys word na die volgende taalkundige werk wat by die Linguistic Research Center van die Universiteit van Texas (Austin), wat op rekenaargesteunde vertaling fokus, die afgelope tyd afgehandel is:

- beskrywings van die grammatikas van Duits en Engels, asook van Proto-Indo-Europees, Goties, Oud-Frans en Oud-Iers;
- 'n leksikologiese analise van die woordeskat wat opgeteken is in die Merriam-Webster-handwoordeboek; en
- 'n telekommunikasiewoordeboek vir Duits, die vierde masjienleesbare uitgawe van Feist se etimologiese Gotiese woordeboek en 'n Nahuat-woordeboek.

Daarby kan die grammatikakomponent van METAL onder andere die volgende verskynsels hanteer: woordplasing, inter- en intra-sinsanaforiek, morfologies gelede strukture, die bepalersisteem, bywoorde, prenominale modifieerders, ontkenning, vraagsinne en samegestelde sinne.

Rekenaartoepassings in taalondersoek en rekenaarlinguistiek oorvleuel dus aansienlik, en die verskille tussen hulle moet nie te sterk gestel word nie.

1.3 OOGMERKE VAN TEXTNET

TEXTNET wou sy opdrag uitvoer met 'n beskrywing van die volgende drie sake:

- 'n Beskrywing van rekenaartoepassings in taalondersoek in die algemeen
- 'n beskrywing van rekenaartoepassings in Suid-Afrika, en
- 'n reeks aanbevelings ten opsigte van rekenaartoepassings in Suid-Afrika oor onder meer die vestiging en bevordering daarvan, opleiding en navorsing daarin en die finansiering daarvan.

1.4 MEDEWERKERS

Die volgende persone het meegewerk aan die projek:

Dr. P.J. Badenhorst, UP
Prof. T.J.D. Bothma, UNISA
Mnr. Q. Gee, WITS
Prof. R.M. Klopper, UZ
Prof. E. Kotzé, UZ
Prof. J-C. Lejosne, Universiteit van Metz, Frankryk
Dr. R. Morris, RGN
Prof. J.C. ROUX, US
Dr. J. Vorster, RGN
Prof. V.N. Webb, UP (projekleier)

1.5 WERKSWYSE VAN DIE KOMITEE

'n Beplanningsvergadering, bygewoon deur alle medewerkers asook prof. W.T. Claassen van US, is gehou om die opdrag van die komitee in besonderhede te bespreek, op die werksverdeling te besluit, elke medewerker se presiese opdrag uit te stippel en 'n algemene werkswyse uit te werk.

Wat werkswyse betref, is daar op die volgende besluit:

- literatuurstudie;
- die uitstuur van 'n vraelys oor rekenaargesteunde taalondersoek in die RSA aan alle taaldepartemente en departemente vir rekenaarwetenskap aan Suid-Afrikaanse universiteite en technikons;
- gesprekke ter aanvulling van die literatuurstudie met plaaslike linguiste wat reeds ervaring van rekenaartoepassings in taalondersoek opgedoen het, met belangstellende plaaslike rekenaarwetenskaplikes, en met buitelandse betrokkenes indien moontlik.

Literatuurverwysings waarvan die sameroeper bewus was, is aan die medewerkers gestuur.

1.6 VERSLAGLEWERING

Die verslaglewering het twee fases gehad. In die eerste fase is elke komiteelid gevra om 'n geskrewe verslag in te dien oor die gebruik van die rekenaar as analitiese hulpmiddel op die taalkundige terrein wat aan hom/haar toegeken is. Hierdie verslae is vervolgens deur die projekleier tesame met bykomstige inligting saamgevat in een verslag. Hierdie samevattende verslag en die afsonderlike verslae van die spanlede is deur die RGN ingebind as 'n kantoorverslag, Rekenaargesteunde Taalondersoek (wat by die RGN se Afdeling Leksikologie ter insae is).

Vir publikasie-doeleindes is die stof anders ingedeel en geïntegreer. Intussen het prof. Jean-Claude Lejosne, 'n linguis van die Universiteit van Metz, Frankryk, heelwat gegewens oor die morfologiese en sintaktiese prosessering van taal verskaf en dit is ook in die verslag opgeneem.

Hierdie verslag is voorberei deur proff. V.N. Webb en T.J.D. Bothma, en dr. R. Morris. Die verskillende hoofstukke is soos volg op bydraes van die spanlede gebaseer:

Hoofstuk 1: Prof. V.N. Webb, dr. R. Morris en prof. T.J.D. Bothma
Hoofstuk 2: Volledig geskryf deur prof. J.C. Roux
Hoofstuk 3: Volledig geskryf deur prof. T.J.D. Bothma
Hoofstuk 4: Prof. T.J.D. Bothma, prof. J-C. Lejosne, prof. R.M. Klopper en dr. P.J. Badenhorst
Hoofstuk 5: Volledig geskryf deur dr. R. Morris
Hoofstuk 6: Volledig geskryf deur prof. V.N. Webb
Hoofstuk 7: Prof. V.N. Webb, prof. E. Kotzé en dr. J. Vorster
Hoofstuk 8: Prof. V.N. Webb en die meeste van die spanlede
Hoofstuk 9: Prof. V.N. Webb en prof. T.J.D. Bothma, gebaseer op aanbevelings in die afsonderlike verslae.

Daar is uit die staanspoor besef dat die ondersoek relatief onvolledig sou wees. 'n Vollediger ondersoek van rekenaartoepassings in taalondersoek was onmoontlik om die volgende redes:

- Suid-Afrikaanse linguiste het met enkele uitsonderings na baie min blootstelling aan rekenaartoepassings in taalondersoek gehad;
- rekenaarlinguistiek is 'n uiters komplekse en uitgebreide terrein, met ontwikkelinge wat deurlopend en besonder snel plaasvind;
- die relevante literatuur oor rekenaartoepassings in taalondersoek is omvattend, ingewikkeld en grootliks onbeskikbaar vir onmiddellike raadpleging, met die gevolg dat baie min daarvan werklik met groot vrug geraadpleeg kon word in die beskikbare tyd.

Hierdie beperkinge ten spyt, kan daar gestel word dat dit vir die Suid-Afrikaanse taalkunde belangrik is om ten minste toegang te verkry tot die inligting wat wel hier verskaf kan word.

HOOFSTUK 2: REKENAARTOEPASSINGS IN DIE FONETIEK EN FONOLOGIE

2.1 OPDRAG

Die opdrag was om ondersoek in te stel na rekenaartoepassings in die fonetiek en fonologie waaronder die volgende aspekte aangespreek moes word:

- Spraakanalise en -sintese
- Fonetiese transkripsie
- Die distribusie en frekwensie van spraakklanke
- Teks-na-spraaksisteme
- Spraakherkenning

2.1.1 Interpretasie van die opdrag

Gesien in die lig van

(a) die beperkte tyd vir die uitvoering van die opdrag, en

(b) die wesensaard van die opdrag, naamlik 'n doenbaarheidstudie, spreek hierdie verslag slegs breë tendense aan en bied dit nie noodwendig 'n 'stand van die wetenskap'-beskrywing aan nie. Die terrein is in ieder geval so wyd en ontwikkel teen so 'n tempo dat dit haas onmoontlik sal wees om op enige gegewe tydstip 'n volledige oorsig daaroor te gee.

2.1.2 Werkwyse

Hierdie verslag is die resultaat van:

(a) 'n basiese bronnestudie, en

(b) gesprekke voortspruitend uit 'n aantal studiebesoeke wat die afgelope agtien maande aan 'n aantal instansies binnelands sowel as buitelands gebring is (vgl. Bylae 1 aan die einde van hierdie hoofstuk).

2.2 KONTEMPORÊRE FONETIEK EN FONOLOGIE

2.2.1 Inleiding

In die hieropvolgende bespreking gaan dit enkel en alleen om rekenaartoepassings op die fisies-strukturele aspekte van menslike spraak. Die aandag word spesifiek gevestig op die rol wat die rekenaar tans speel, en ook in die toekoms behoort te speel op die vlak van spraakanalise en -sintese. Die fonologie en dit wat tradisioneel daaronder verstaan word,

word hier alleenlik betrek in die mate wat dit om fonotaktiese (oppervlakstruktuur-) verskynsels gaan.

Een belangrike punt behoort hier aangestip te word, naamlik dat die onderskeid tussen 'basiese' en 'toegepaste' navorsing op die gebied van fonetiek vandag al hoe sterker na vore tree, en daar word al selfs na die eiesoortige terrein van 'spraaktegnologie' verwys wanneer dit om toegepaste navorsing gaan.

In die geval van basiese navorsing gaan dit naamlik om die beskrywings (heel dikwels in isolasie) van artikulatoriese, akoestiese en perseptiewe eienskappe van menslike spraak ten einde insig te verkry in die aard en funksionering daarvan. Studies van hierdie aard kan en behoort 'n belangrike rol te speel in meer abstrakte fonologiese beskrywings, veral gedagtig aan die bekende uitspraak van Fromkin dat '..every phonologist should be a phonetician as well..' Hierdie tipe 'eksperimenteel-fonetiese' data is inderdaad uiters noodsaaklik nie alleen om impressionistiese beskrywings aan te vul nie, maar ook om geloofwaardigheid aan abstrakte fonologiese analises te verleen.

Die tweede tipe navorsing hou direk verband met fisiese aspekte van die menslike kommunikasieproses en die simulering daarvan deur rekenaars as klanklike invoer- en afvoersisteme. Hierdie tipe navorsing is nou verbind met ontwikkelinge op die gebied van die informasietegnologie, kunsmatige intelligensie, en die rekenaarwetenskap, en spreek 'n ander aspek van menslike spraak aan.

Om op 'n eenvoudige wyse 'n oorsig te kry van die aktiwiteite op hierdie twee verwante terreine, sou dit goed wees om te fokus op die programme van enkele van die mees verteenwoordigende internasionale konferensies op hierdie gebied, te wete dié van:

(a) The Tenth International Congress of Phonetic Sciences wat op 1-6 Augustus 1983 te Utrecht plaasgevind het (vgl. Inligtingstuk 1).

(b) The Eleventh International Congress of Phonetic Sciences wat op 1-7 Augustus 1987 in Tallinn, Estonia, USSR plaasgevind het (vgl. Inligtingstuk 2).

(c) Speech Tech '87 wat op 28-30 April 1987 in New York plaasgevind het (vgl. Inligtingstuk 3).

(d) European Conference on Speech Technology wat 2 tot 4 September 1987 in Edinburgh plaasgevind het (vgl. Inligtingstuk 4).

Uit 'n vergelyking van die aard en inhoud van die onderskeie voordragte is dit duidelik dat waar dit in (a) en (b) grotendeels gaan om 'basiese' aspekte rakende die klanklike kommunikasieproses en waar temas aangespreek

word soos 'Speech and Hearing', 'Relation between Speech Production and Speech Perception', 'Sociophonetics', 'Phonetics and Phonology' ensovoorts, gaan dit in (c) en (d) hoofsaaklik om 'Software systems for speech technology development and application', 'Aids for the handicapped' en 'Continuous Speech Systems', 'Text-to-Speech-Systems', 'Speech Modelling', ensovoorts.

Alhoewel daar uiteraard ook oorvleuelings ten opsigte van temas op verskillende konferensies mag wees (en inderdaad ook is), is dit baie duidelik dat die rol van die 'tradisionele fonetikus' vinnig besig is om te verander. In Inligtingstuk 5 word die hooftoespraak oor die onderwerp 'Phonetics and Speech Technology', gelewer tydens die 1983-kongres deur een van die bekendste persone in die veld, te wete prof. Gunnar Fant, weergegee. Die toespraak dui aan hoedanig ontwikkelinge op die vlak van rekenaartegnologie aanleiding gee tot 'n nuwe siening van die fonetiek as wetenskap.

2.2.2 Die rekenaar in fonetiekondersoek

Fant bied enkele gedagtes aan oor 'Computerized Phonetics' terwyl ander opmerkings oor die tema 'Fonetiek, Fonologie en die rekenaar' ook in Roux (1984, 1986) gevind kan word. In 'n sekere sin is dit onmoontlik om 'n oorsig oor apparatuur en programmatuur op hierdie terrein aan te bied, aangesien sisteme opgebou en programmatuur ontwikkel word om aan heel spesifieke behoeftes te voldoen. In die lig hiervan word daar vervolgens in breë trekke na verskillende ontwikkelinge verwys.

Dit is belangrik om daarop te let dat die rekenaar gebruik kan word

(a) as die aandrywer en/of beheersisteem van verskillende soorte apparaat waarmee ondersoek van 'n fonetiese aard onderneem word. Hier word spesifiek verwys na rekenaarbeheerde kinefluorografiese sisteme, ultrasoniese sisteme en verskillende sonarsisteme waarmee daar veral artikulatoriese data versamel kan word. Hierdie sisteme word veral in mediese navorsing aangetref en word as sodanig as toegewyde sisteme deur buitelandse maatskappye bemark;

(b) as 'n hulpmiddel by artikulatoriese, akoestiese en perseptiewe analises van menslike spraak.

In die hieropvolgende bespreking word daar grotendeels aandag gegee aan afdeling (b) hierbo aangesien dit vir die taalkundige as sodanig waarskynlik meer van belang is as die eerste afdeling.

Artikulatoriese analises

Vanuit die literatuur (vgl. veral Journal of Phonetics, Phonetica, Journal of Speech and Hearing Research, UCLA Working Papers in Phonetics) is dit duidelik dat daar talle tipes eksperimente op die artikulatoriese vlak uitgevoer kan word waar die versameling en verwerking van data deur rekenaars moontlik is. Hierdie navorsing dek ondersoeke na orale en nasale lugvloeiverskynsels, stembandvibrasie, tong-, lip- en kakebeenbewegings en tong-verhemeltekontak. Verder is dit moontlik om artikulatoriese posisies met behulp van 'n rekenaar te simuleer deur van akoestiese data as invoer gebruik te maak. Hierdie tipe aktiwiteit word gewoonlik met behulp van 'n hoofraamrekenaar verrig en wel met spesiaal ontwikkelde programme soos bv. dié van Haskins Laboratories en ook die bekende ILS (Interactive Laboratory System)-programmatuur.

Afgesien van die werk wat op hierdie vlak deur talle instansies wêreldwyd gedoen word, sou die volgende instansies wat alreeds bekendheid op hierdie vlak verwerf het, genoem kon word:

- Phonetics Laboratory, University of Reading, veral met betrekking tot gerekenariseerde elektropalatografie;
- Phonetisches Institut der Universität Koln, veral met betrekking tot gerekenariseerde artikulatoriese simulasie;
- Haskins Laboratories, Connecticut, New Haven, vir artikulatoriese navorsing *per se*.

In die praktyk blyk dit dat 'n span bestaande uit fisioloë en fonetici gewoonlik by die uitvoering van die meeste van hierdie tipe eksperimente betrokke is.

Akoestiese en perseptiewe analises

Navorsing op die gebied van die akoesties-perseptiewe analises van menslike spraak staan vandag sterk in die brandpunt. Die basiese uitgangspunt is dat die akoestiese spraaksinjaal as die draer van linguistiese intensie gesien word en dat 'n sorgvuldige analise en sintese daarvan die tersaaklike eienskappe van die klanklike kommunikasieproses kan blootlê. In die lig hiervan is dit dan so dat programme ontwikkel word om:

- die akoestiese sein te digitaliseer ten einde dit geskik te maak om as invoer te dien vir die rekenaar;
- die gedigitaliseerde sein op 'n interaktiewe wyse (of selfs ook outomaties, in sekere gevalle) te segmenteer en/of te manipuleer;

- die gemanipuleerde sein akoesties te ontleed ten einde inligting te bekom oor die eienskappe daarvan;

- nuwe seine te genereer op grond van 'nuwe' of 'ou' data ten einde data voor te berei vir persepsietoetse om die geslaagdheid van die analise te bepaal.

Afgesien van die talle programme wat indiwidueel deur navorsers ontwikkel is, sou die volgende spraakanalise en -sintese-programme waarskynlik as die beste beskikbare programme beskou kan word:

- ILS-programme vir hoofraam- sowel as persoonlike rekenaars, ontwikkel en versprei deur die firma STI, Goleta, California; pryse wissel tussen R2 500 en R20 000 (verdere inligting verkrygbaar by prof. J.C. Roux, US);

- AUDLAB : An interactive speech and signal analysis system, Center for Speech Technology Research, University of Edinburgh (Inligtingstuk 6);

- MIT SPEECHVAX-sisteem van D.H. Klatt wat tans wyd in die VSA en Europa gebruik word met as uitvloeisel die DECTALK-teks-na-spraak-sisteem.

Dit dien gemeld te word dat daar van tyd tot tyd kleinere sisteme in die handel verskyn waarmee spraak gedigitaliseer kan word en selfs ook gesintetiseer kan word (bv. 'n analoog-digitale omsetter vir die Apple MacIntosh). Hierdie sisteme het egter sekere beperkings en is nie sonder meer geskik vir ernstige navorsingsdoeleindes nie.

Fonotaktiese analises

Onderliggend aan die meeste toegepaste navorsingsprojekte wat gewoonlik op die vlak van spraaksintese en/of spraakherkenning lê, is daar noodwendig 'n verwysing na spesifieke struktuurverskynsels. Dit is so dat inligting oor die voorkoms en verspreiding van foneme en foneemkombinasies uiters noodsaaklik is in die ontwerp van sodanige sisteme en derhalwe word daar ook heelwat aandag gegee aan die ontwikkeling van programmatuur in hierdie verband.

In beginsel bestaan sodanige programmatuur uit minstens twee dele:

(a) 'n Ortografie-na-foneemomsettingsfase
(b) 'n Statistiese analisefase

In die geval van (a) moet bekende fonetiese transkripsiereëls gealgoritmiseer word sodat die mees optimale fonetiese transkripsie van 'n teks verkry kan word. Die resultate van die transkripsieprogram dien as invoer tot die analiseprogram waar enige voorstelbare inligting oor die fonotaktiese struktuur van die besondere taal bekom kan word.

Inligting oor foneemvoorkomsfrekwensies op 'n leksikaalkategoriese basis, asook as eenhede *per se* is van besondere belang vir enige werk op die vlakke van spraakherkenning, linguistiese verwagting, spraakoudiometrie en spraaksintetisering (vgl. veral Van den Broecke et al 1987, in hierdie verband).

Dit moet beklemtoon word dat sodanige inligting nie net vir toegepaste navorsing noodsaaklik is nie, maar ook vir basiese navorsing. Daar behoort op gelet te word dat verwysings na byvoorbeeld die 'mees frekwente' sillabestrukture in Afrikaans, wat so vry in handboeke verskyn, met omsigtigheid behandel behoort te word bloot op grond van die feit dat daar na die beste wete van die skrywer geen werklik verteenwoordigende databasisse bestaan waarop sodanige analises uitgevoer is nie.

2.3. TOEGEPASTE SPRAAKSISTEME

In sy hoofvoorlesing tydens die 1983-kongres in Utrecht (Inligtingstuk 5) beklemtoon Fant die steeds groeiende rol van die fonetiek in navorsing oor menslike funksies en wys hy daarop dat die fonetiek 'n tegniese profiel verkry het en dat spraaktegnologie direk op die fonetiek aangewys is om sy gevorderde doelstellings te bereik. Vanuit die amptelike program van Speech Tech '87 (Inligtingstuk 3) kan weer afgelei word juis hoeveel terreine in die menslike samelewing geraak kan word deur ontwikkelinge op die vlak van spraakinvoer- en -afvoersisteme. Hierdie sisteme is bruikbaar op die vlakke van:

- Elektroniese verbruikersgoedere
- Kantoor- en fabriekoutomatisasie
- Telekommunikasie
- Hulpmiddels vir gestremdes
- Sekuriteitsisteme (militêr en andersins)
- Opvoedkunde (onder andere as taalonderrigsisteme of as onderrigsisteme *per se* wat van natuurlike-taalinvoer en -afvoer gebruik maak)

Hierdie sisteme funksioneer grotendeels as

- spraaksintesesisteme (gewoonlik as teks-na-spraak-sisteme), en as
- spraakherkenningsisteme (met spraak en/of teks as invoer).

Twee belangrike punte behoort hier aangestip te word:

(a) Navorsing wat gedoen word in die daarstelling van sodanige toegepaste sisteme moet nie noodwendig gesien word as totaal losstaande en onverwant aan basiese fonetiekondersoek waar dit primêr gaan om insig in die aard en omvang van menslike spraak nie. Dit kan en behoort ook gesien te word in die lig daarvan dat die suksesvolle simulasie van spraakproduksie en spraakpersepsie juis kan lei tot beter begrip van die tersaaklike prosesse en van die menslike klanklike kommunikasieproses as sodanig. Dit is derhalwe dan ook nie verrassend dat daar ook op die 'tradisionele' fonetiekkongresse (vgl. Inligtingstukke 1 en 2) soveel aandag aan teks-na-spraak-sisteme en outomatiese spraakherkenning gegee word nie.

(b) Die ontwikkeling van bogenoemde sisteme is van belang vir die toekoms van die tale waarvoor dit ontwikkel word. Dit sal aan die betrokke tale 'n groot voorsprong gee as 'tegnologietale', terwyl tale waarvoor gerekenariseerde stelsels nie tot stand kom nie, uitgesluit sal word van die internasionale talenetwerk van die toekoms. Indien stelsels vir Afrikaans en sekere Afrikatale nie gebou word nie, kan daar met redelike sekerheid voorspel word dat Engels binne enkele dekades die enigste 'tegnologie-taal' in die RSA sal wees. Engels is reeds die taal waaraan die meeste ontwikkelingswerk op hierdie gebied gewy word. Tog is dit opmerklik hoedanig daar, veral in die Europese lande, klem gelê word op die ontwikkeling van sisteme in die inheemse tale. Dit gebeur heel dikwels dat die tegnologie as sodanig aangekoop word (of op 'n onderhandelingsbasis beskikbaar gestel word) en dat daaruit eietalige sisteme ontwikkel word. Die volgende twee gevalle dien as goeie voorbeelde hiervan:

DECTALK-verwerking aan die Fraunhofer-instituut in Stuttgart:

Die bekende DECTALK-teks-na-spraaksisteem van die maatskappy Digital Equipment word onder kontrak deur lede van hierdie navorsingsinstituut (deels verbonde aan die Universiteit van Stuttgart) as 'n Duitstalige sisteem omgebou, terwyl onderhandelinge tans gevoer word om die sisteem ook vir Sjinees aldaar aan te pas.

Q AND A-verwerking deur die firma TRANSMODUL in Saarbrücken:

Die firma TRANSMODUL het pas die Duitse omsetting van een van die mees suksesvolle databasisprogramme met 'n natuurliketaal-koppelvlak onderneem en bemark dit tans as die F(rage) + A(ntwort)-program. Die moedermaatskappy Symantec van die VSA het ook alreeds sisteme in Nederlands, Frans en Italiaans deur hierdie maatskappy laat ontwikkel. Weliswaar gaan dit in hierdie sisteem net om geskrewe teks, dog dit word in die vooruitsig gestel om 'n klankkomponent by te voeg. In die AILANG-verslag is daar volledíger oor hierdie program gerapporteer. Hierdie program wat op 'n persoonlike rekenaar werk, is alreeds op groot skaal deur die Belgiese regering aangekoop vir gebruik op administratiewe vlak.

Daar is met die bestuur van TRANSMODUL asook met verteenwoordigers van DEC in Engeland (as medewerkers in die Duitse projek) onderhandel oor die moontlike ontwikkeling van hierdie sisteme vir Afrikaans en sekere Afrikatale. Dit wil blyk asof samewerking wel moontlik is op voorwaarde dat daar aan sekere vereistes voldoen word en die nodige finansiële ondersteuning gevind sou kon word. (Spesifieke besonderhede is by prof. J.C. Roux beskikbaar.)

2.4 STAND VAN NAVORSING

2.4.1 Buitelands

Wanneer daar gekyk word na die aard en omvang van die 1983- en 1987-kongresse (Inligtingstukke 1 en 2), asook na die aantal deelnemers (651 referate) tydens die 1987-kongres, dan is dit opsigtelik juis hoeveel daar op hierdie terrein plaasvind. Indien daar verder na die adresse van die deelnemers, soos weergegee in Inligtingstuk 2 gekyk word, dan sal gemerk word juis hoeveel persone in die een of ander fonetieklaboratorium en/of -instituut werksaam is. Selfs derdewêreldlande soos Indië, Brasilië, Meksiko en Tunisië is relatief goed verteenwoordig, beide ten opsigte van die bestaan van sodanige inrigtings asook ten opsigte van kongresdeelname. Dit is opvallend dat die RSA hoegenaamd nêrens verteenwoordig is nie en geen rol op internasionale gebied speel nie.

2.4.2 Binnelands

Die feit dat die RSA geen rol in die buiteland speel nie hang direk saam met die aard van fonetiekonderrig en -navorsing alhier. Dit is bekend dat die tipe fonetiek wat aan universiteite bedryf word op die vlak van die prakties-impressionistiese lê. Een rede hiervoor is dat die taalkundige enige rekenaarmatige ondersoek, hetsy op die vlak van spraakproduksie of -persepsie as 'buite sy terrein' beskou het. Die gevolg hiervan was en is 'n proliferasie van wetenskaplik onhoudbare stellings (veral in handboeke) wat hoegenaamd nie rekening hou met ontwikkelinge op hierdie terrein nie.

Die bestaan van behoorlike en goedtoegeruste fonetieklaboratoria in die RSA is uiters beperk. Na die beste wete van die skrywer bestaan daar by die volgende instansies apparatuur vir fonetiekondersoek:

- Die Universiteit van Pretoria: Departement Spraakheelkunde
- UNISA: Interdepartementeel
- Die Universiteit van die Witwatersrand: Departemente Spraakterapie en Linguistiek
- Die Universiteit van Rhodes: Departemente Afrikatale en Engels
- Die Forensiese Laboratorium, Suid-Afrikaanse Polisie, Pretoria
- Die Universiteit van Stellenbosch: Departement Afrikatale

Slegs enkele instansies hierbo (nl. US, UP en nou moontlik ook WITS) beskik oor rekenaargebaseerde analise- en sintese-sisteme. Verder is dit so dat enkele ingenieursdepartemente aan Suid-Afrikaanse universiteite asook die WNNR oor gespesialiseerde apparatuur en programmatuur beskik vir spraakseinanalise en -sintese.

Aan die Universiteit van Stellenbosch bestaan 'n informele werkgroep vir seinverwerking (ingenieurs, kunsmatige-intelligensie-deskundiges, linguiste en fonetici) wat deurlopend betrokke is by 'n verskeidenheid projekte soos:

- Teks-na-spraaksisteme vir Afrikaans en Xhosa
- Die akoesties-perseptiewe eienskappe van Xhosa, Zulu, Noord-Sotho, Sesotho en Setswana
- Gerekenariseerde fonetiese databasisse vir die Nguni- en Sothotale

Die meeste van die aktiwiteite vind plaas in die elektronikalaboratorium asook in die fonetieklaboratorium van die Departement Afrikatale. In die proses is die volgende belangrikste apparatuur en programmatuur self ontwikkel en/of aangekoop:

Apparatuur:

- 1 VAX 785 hoofraamrekenaar met bykomstighede
- Netlan netwerkstelsel
- 'n aantal persoonlike rekenaars (ongeveer 8), Olivetti M24s en CW 16s (meestal toegerus met 20MG-starskywe en 8087-verwerkers)
- twee analoog-digitale-omsetters

Programmatuur:

- ILS: Vir verskillende tipes akoestiese analises
- ANALOG: Vir seinredigering en -analise

- KLATTALK: Vir spraaksintese
- MORFFON: Vir morfeemanalise in Afrikaans
- TEKSFON: Vir outomatiese fonetiese transkripsies in Xhosa, Zulu, Swazi, Noord-Sotho, Sesotho, Setswana, Venda, Tsonga
- FONSTAT: Vir statistiese data oor foneemvoorkoms
- STAP: Vir gevorderde statistiese verwerking met betrekking tot foneemverspreiding.

Bykomend hiertoe is 'n groot verskeidenheid nuts-programme, woordverwerkers, ensovoorts.

2.5 GEVOLGTREKKINGS

Die situasie soos hierbo geskets, dui daarop dat daar waarskynlik twee redes vir die gebrek aan rekenaargesteunde taalondersoek op klanklike vlak in die RSA bestaan:

(a) Nie-gerigte opleiding:

Dit is aangetoon dat opleidingsprogramme in die fonetiek, veral dié aan universiteite, nie daarop ingestel is om onderrig op hierdie vlak te gee nie. Taaldepartemente hou hulself besig met oppervlakkige impressionistiese beskrywings terwyl ingenieursdepartemente weer ewe oppervlakkig met fonetiese konsepte omgaan in hul akoestiese analise van die spraaksein.

In die buiteland bestaan daar al dekades lank kursusse in byvoorbeeld eksperimentele fonetiek, hetsy as onderdele van kursusse in die algemene of toegepaste linguistiek, of as volwaardige kursusse in 'n departement van fonetiek. Hierdie kursusse is normaalweg so gestruktureer dat dit toegang bied vir studente in sowel die geesteswetenskappe as die natuurwetenskappe en bied dan ook onderrig aan in rekenaartegnieke in spraakanalise en -sintese.

(b) Beperkte industriële navorsing:

Uit Inligtingstuk 3 sou afgelei kon word dat, ten einde sisteme wat op spraak kan reageer en/of self spraak kan genereer, te kan bemark, is 'n groot hoeveelheid navorsing en ontwikkelingswerk nodig. In baie gevalle word die ontwikkeling van sodanige sisteme deur groot rekenaarfirmas soos IBM, Siemens, Nixdorf en Texas Instruments onderneem in hul eie navorsingslaboratoria. 'n Belangrike verskynsel (waaroor daar ook volledig in die AILANG-verslag berig is) is egter dat daar in baie (ontwikkelde) lande op 'n werklik gestruktureerde wyse samewerkingsooreenkomste tussen die staat, handel en nywerheid en die universiteite tot stand kom, enersyds om te rasionaliseer en onnodige oorvleueling in navorsing te verhoed, en andersyds om die oordrag van kennis tussen universiteite en die privaatsektor te stimuleer.

In Wes-Duitsland bestaan daar die BMFT-ooreenkomste waar die Bundesministerium für Forschung und Technologie nie net gesamentlike projekte tussen universiteite en die privaatsektor finansier nie, maar ook self projekte op beperkte skaal inisieer. Dieselfde samewerking tussen universiteite en die privaatsektor word ook in Engeland aangetref, (vgl. die opmerkings van prof. John Laver in die inleiding tot die 1987 Edinburgh-kongres - Inligtingstuk 4) en in Japan.

Dit is te begrype dat binne die huidige situasie in die RSA industriële navorsing op hierdie vlak redelik beperk is en dat veral die groter maatskappye hul navorsing in die buiteland onderneem. Gevolglik was daar tot op hede waarskynlik relatief min aanvraag vir rekenaargeskoolde fonetici en derhalwe ook geen dringende noodsaak vir opleiding nie. Hierdie situasie is tans snel aan die verander. Veral in die lig van toenemende isolasie en boikotte sal eie sisteme eenvoudig ontwikkel moet word en sal daar op die hoogs moontlike vlak aandag aan hierdie saak gegee moet word.

Een punt is baie duidelik en daarvan behoort ook deeglik kennis geneem te word in die RSA:

Namate rekenaartegnologie ontwikkel en die 'vyfde generasie-rekenaarera betree word, waarin daar gestreef word na intelligente en gebruikersvriendelike rekenaars wat in natuurlike taal (ook op klanklike vlak) met gebruikers kommunikeer, sal die aanvraag vir kundigheid op die gebied van spraaktegnologie eskaleer. Indien enige sodanige spraakinvoer en -afvoersisteme vir tale soos Afrikaans en sekere Afrikatale ontwikkel sou wou word (wat op sigself uiters noodsaaklik is indien die tale nie totaal deur Engels as tegnologietaal verdring wil word nie), dan kan dit nie alleen aan tegnici en ingenieurs oorgelaat word nie, maar behoort die linguis-fonetikus sy bydrae daartoe te lewer. In hierdie verband sal daar dan spesifiek aandag gegee moet word aan die hersiening van leerplanne, interdissiplinêre samewerking en die stimulering van navorsing en opleiding op hierdie gebied.

BYLAE 1

LYS VAN BESOEKE IN VERBAND MET REKENAARTOEPASSINGS IN DIE FONETIEK EN FONOLOGIE

1986:

- Die Universiteit van die OVS:
  Departemente Algemene Taalwetenskap
  Afrikaans en Nederlands
  Bantoetale

- Die Akademie (Windhoek)
  Departement Afrikatale

- Die Universiteit van Pennsylvania in Philadelphia (VSA)
  Fonetiek-/Fonologielaboratorium

- Haskins Laboratories, Connecticut, New Haven, VSA.

- Linguistic Institute of the Linguistic Society of America, CUNY, New York:

  Opleidingsprogramme van ses weke in

  - Quantitative Analysis : Phonetics and Phonology (Labov)
  - Computer Implementation of Speech Analysis and Synthesis (Rubin)
  - Speech Perception (Raphael)
  - Articulatory Phonetics (Harris)

1987:

- Die Universiteit van Pretoria
  Departemente Afrikaans en Spraakwetenskap

- Die Fraunhofer Institut für Arbeitswirtschaft und Organisation, Stuttgart, Wes-Duitsland

- Die Institut für Kommunikationsforschung und Phonetik, Universiteit van Bonn, Wes-Duitsland

- Die Departement Informatik, Universiteit van Saarland, Saarbrücken, Wes-Duitsland

- Die maatskappy TRANSMODUL in Saarbrücken, Wes-Duitsland

HOOFSTUK 3: TAALTEORIEË BINNE DIE REKENAARGESTEUNDE MORFOLOGIE EN SINTAKSIS

3.1 INLEIDING

In hierdie hoofstuk word 'n aantal van die taalteorieë wat in die rekenaar-analise van taaldata aangewend word, bespreek. Aangesien geeneen van die teorieë tot dusver in Suid-Afrika in rekenaargesteunde analises toegepas is nie, en die rekenaarprogramme wat vir sodanige analises nodig is (nog) nie plaaslik beskikbaar is nie, is die bespreking noodwendig op die beskikbare oorsese literatuur gegrond.

Dit is 'n kenmerk van hoogs formele taalmodelle dat hulle uit stelle *constructs* bestaan. Aangesien laasgenoemdes geskepte begrippe *(objects of thought)* is, wat dikwels slegs binne die model betekenis het, kan dit wetenskaplik onverantwoord wees om in die geval van nuwe modelle termekwivalente in eie taal te skep voordat die plaaslike vakgemeenskap met die model-interne samehange deeglik vertroud geraak het. Hierdie hoofstuk bevat gevolglik vele Engelse terme en aanhalings uit die brontekste. Dit hou die nadeel in dat die hoofstuk moeiliker lees en dat taalsuiwerheidsnorme nie gehandhaaf kan word nie, maar in hierdie stadium lyk dit uit wetenskaplike hoek verkieslik om die oorspronklike terme te gebruik.

3.2 VERSKILLENDE TAALTEORIEË

3.2.1 Transformasioneel-generatiewe Grammatikas (TGG)

Hierdie teorie, wat tans die teoretiese linguistiek oorheers, het sedert die publikasie van Chomsky 1957 baie veranderinge ondergaan, as gevolg van Chomsky se eie navorsing (bv. Chomsky 1981 en 1982) en ook dié van ander taalkundiges. Rekenaartoepassings is gebaseer min of meer op Chomsky 1965; daarvolgens bestaan 'n TGG basies uit frasestruktuurreëls en transformasionele reëls. Frasestruktuurreëls het die vorm

A -> B C

S -> NP AUX VP

Die transformasionele reëls bestaan uit 'n strukturele indeks (SI) en 'n strukturele verandering (SC); as voorbeeld, die passieftransformasie:

SI	NP	AUX	V	NP
	1	2	3	4
SC	4	2	be+en 3	by 1

Volgens Grishman (1986:47) lê die voordeel van die transformasionele grammatika vir die rekenaarlinguistiek veral daarin dat *the paraphrastic re-*

lations captured by transformational grammar should simplify subsequent stages of language processing. Daar bestaan egter drie basiese probleme om die generatiewe proses om te draai:

- *assigning to a given sentence a set of parse trees which includes all the surface trees which would be assigned by the transformational grammar*
- *given a tree not in the base, determining which sequences of transformations might have applied to generate this tree*
- *having decided on a transformation whose result may be the present tree, undoing this transformation* (Grishman 1986:48).

Om die eerste probleem op te los, kan 'n konteksvrye grammatika ontwerp word *which will give all the surface trees assigned by the transformational grammar, and probably lots more; this context-free grammar is called a covering grammar* (Grishman 1986:48). So 'n *covering grammar* van die MITRE-groep het bestaan uit *550 productions to produce the surface trees and a set of 134 reverse transformational rules,* terwyl die kategoriale komponent bestaan het uit ongeveer 275 reëls en die transformasionele komponent uit 54 transformasies (Grishman 1986:48). Die stelsel het as volg gewerk: [...] *applying a series of reverse transformations, checking if the resulting tree can be generated by the base component, and then verifying the analysis by applying the forward transformations to the base tree* (Grishman 1986:49). Die sisteem wat deur Petrick ontwikkel is, het min of meer volgens dieselfde beginsels gewerk, maar dit is aansienlik gewysig sedert die eerste model (Grishman 1986:49-50). Vir meer besonderhede oor die MITRE-sisteem en dié van Petrick, vergelyk Grishman 1986:50-56 en Winograd 1983:383-7.

In die rekenaartoepassings wat in die laat sestigerjare en vroeë sewentigerjare ontwikkel is, kan die nuutste ontwikkelinge binne die TGG uiteraard nie gereflekteer word nie. (Vgl. Chomsky 1981:5 vir sy huidige siening oor die subkomponente van die reëlsisteem en die subsisteme van beginsels van die Government-Binding-teorie (GB). Vgl. verder Chomsky 1981, Chomsky 1982, Radford 1981 en Sells 1985).

GB word egter ook binne die rekenaarlinguistiek gebruik. In Wehrli 1984 word 'n *parser* vir Frans, gebaseer op 'n gewysigde weergawe van GB, beskryf. Die *parser* bestaan uit verskeie modules wat ooreenstem met sommige van die subsisteme van GB. Die *parser* is kragtig genoeg om al die grammatikale strukture van sinne van 'n redelik groot subdeel van Frans te ontleed, sonder om 'n veelvoud van alternatiewe analises voor te stel, selfs in die geval van redelik komplekse konstruksies.

Oor die bruikbaarheid van transformasionele grammatika binne die rekenaarlinguistiek, is daar egter nie eensgesindheid nie. Boot (1984:140) beweer byvoorbeeld *de transformationeel generatieve grammatica is geen model*

dat geschikt is voor een computer; de transformationeel generatieve grammatica hoort niet thuis in de computerlinguistiek. Volgens Berwick en Weinberg (1983:11) het rekenaarlinguistiek twee doelstellings, naamlik *It has aimed at computational explanations of distinctively human linguistic behaviour [...], it has accumulated a stock of engineering methods for building machines to deal with natural (and artificial) languages* en hulle beweer dat GB 'n bydrae kan lewer tot beide hierdie doelstellings. Winograd (1983:188-9) beweer: *Even if the overall transformational model does not serve as a psychological theory or as a basis for computer programming it is quite likely that some of the mechanisms that have been developed as parts of that model will prove useful* en hy beskou dit as noodsaaklik om, binne die rekenaarlinguistiek, bekend te wees met 'n transformasionele agtergrond.

3.2.2 Augmented Transition Networks (ATN)

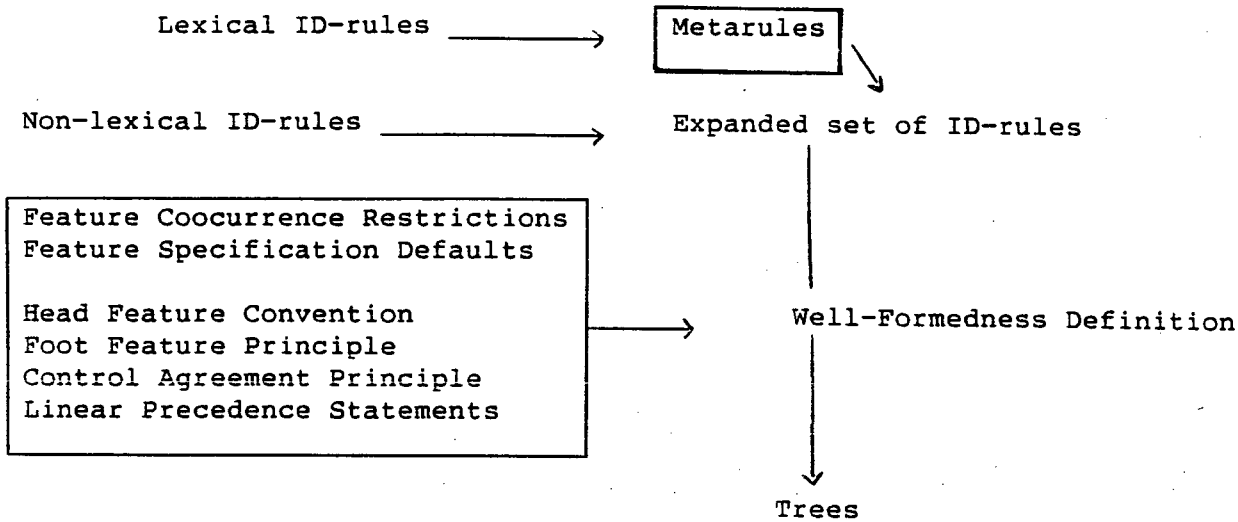
Die ATN-formalisme van Woods (1970) en verdere ontwikkelings het een van die mees algemene vorms geword om natuurlike taal in rekenaarsisteme te ontleed, veral vanweë *the flexibility of ATN parsing* (Winograd 1983:261). Grishman (1986:64) beskryf 'n oorgangsnetwerk as volg: *A transition network is a representation of a regular (or finite-state) grammar (...). The network is a directed graph whose arcs are labeled by terminal symbols (words or word categories). One node of the graph is designated as the start state; one or more nodes are marked as final states. A sentence is in the language defined by the network if there is a path from the start state to some final state such that the labels on the arcs of the path match the words of the sentence.* 'n Rekursiewe oorgangsnetwerk het die volgende uitbreiding: *arcs may be labeled with either node names or terminal symbols.* Voorts, *if we take a recursive transition network and add procedures (generally coded in LISP) for enforcing grammatical constraints and for generating a deep structure, we get an ATN* (Grishman 1986:65). (Vir meer besonderhede, vgl. Winograd 1983:195-267).

Verskillende ATN-sisteme is ontwikkel; vir meer besonderhede, vergelyk Grishman 1986:64-73, Winograd 1983:390-7, Stock, Castelfranchi en Parisi 1983, Lesmo en Torasso 1983. ATN's word ook gebruik vir die generering van sinne (vgl. Grishman 1986:164vv). In Shapiro 1982 word 'n veralgemening van die ATN-formalisme, op grond waarvan 'n ATN-grammatika geskryf kan word to *parse a semantic network and generate a surface string as its output,* bespreek; daar word ook 'n voorbeeld gegee van 'n *combined parsing-generating grammar that parses surface sentences, builds and queries a semantic network knowledge representation, and generates surface sentences in response* (Shapiro 1982:12).

3.2.3 Generalised Phrase Structure Grammar (GPSG)

GPSG het aan die einde van die sewentigerjare ontstaan uit die werk van Gazdar. Volgens Shieber (1986:38,46) is GPSG (en LFG, vgl. 3.2.4) 'n *unification-based formalism* wat as 'n linguistiese teorie (eerder as 'n *tool)* ontwikkel is. GPSG stel slegs een vlak van sintaktiese representasie, naamlik 'n oppervlakstruktuur, en slegs een tipe sintaktiese objek, naamlik

die frasestruktuurreël. Sells (1985:79) gee die volgende skematiese voorstelling van GPSG:



GPSG maak gebruik van 'n tweevlak X'-teorie wat verskil van dié van GB; byvoorbeeld, die balknotasie van 'n moeder en die hoofdogter is identies en S is 'n projeksie van V (Sells 1985:81). 'n Sintaktiese kategorie word beskou as 'n *set of feature value pairs* (Sells 1985:82), byvoorbeeld NP ($=N^2$) word beskou as 'n afkorting vir

{<N,+>,<V,->,<BAR,2>}

Die kategorie N het dieselfde kenmerke, met dié verskil dat die BALK-waarde 0 is. Frasestruktuurreëls verteenwoordig slegs onmiddellike dominansie en word ID-reëls genoem. Lineêre opeenvolging word deur 'n addisionele reël gespesifiseer. Reëls binne GPSG het dus die volgende vorm:

V --> V, NP,XP

waar die kommas aandui dat die elemente ongeorden is, tesame met die reël

V<NP<XP

wat die lineêre volgorde spesifiseer (Sells 1985:86, vereenvoudig).

Metareëls word uit ID-reëls afgelei; die passiewe metareël het byvoorbeeld die volgende vorm:

VP --> W, NP

⇓

VP[PAS] --> W, (PP[by])

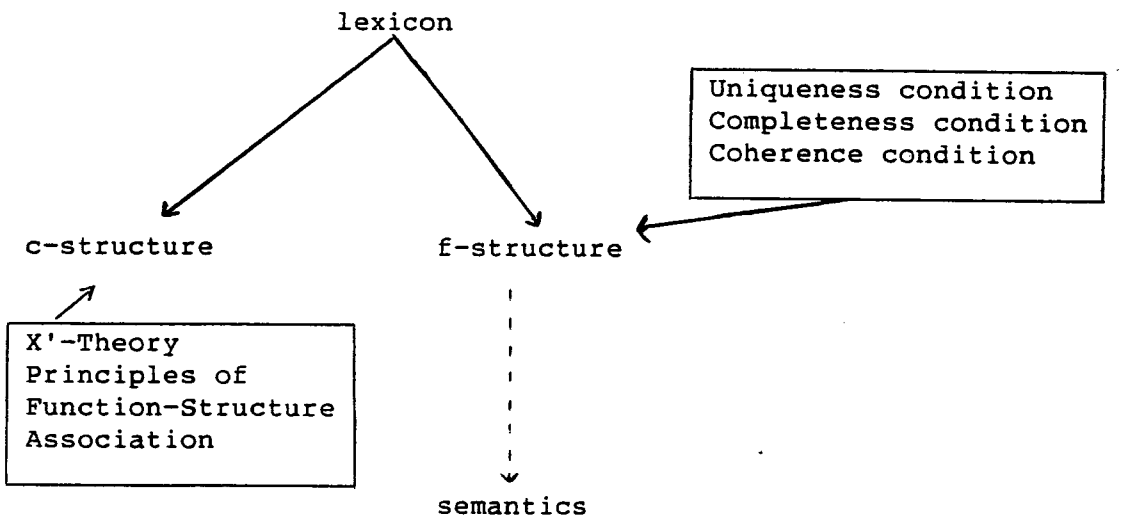
W is 'n veranderlike oor enige kategorie in 'n ID-reël (Sells 1985:91).

Vir meer besonderhede oor GPSG, vergelyk Sells 1986, Gazdar en Pullum 1982 en Gazdar, Klein, Pullum en Sag 1985. Onlangse variante van GPSG is voorgestel deur Pollard en sy kollegas te Hewlett-Packard (vgl. Shieber 1986:53-4 en veral Pollard 1987 en Sag 1987). Vir die ID/LP-formalisme binne taalanalise, vergelyk veral Shieber 1984a, asook Barton 1985, waarin die kompleksiteit van dié formalisme bespreek word. GPSG is gebruik as 'n basis vir 'n taalanalise-sisteem ontwikkel deur Hewlett-Packard; (Grishman 1986:84). Die aanvanklike sisteem is gebaseer op die pre-ID/LP-weergawe van GPSG (Uszkoreit 1983:111).

Insigte van die GPSG, veral die ID/LP-prinsipe, word in heelwat *unification-based* modelle aangetref.

3.2.4 Lexical-Functional Grammar (LFG)

LFG is in die laat sewentigerjare ontwikkel deur Bresnan en Kaplan (vgl. veral Kaplan en Bresnan 1982). LFG is, soos hierbo onder 3.2.3 gemeld, ook 'n *unification-based formalism* wat as 'n taalteorie ontwikkel is. Binne die teorie staan die teoretiese entiteite bekend as grammatiese funksies, en voorstellings wat hulle uitbeeld, staan bekend as funksionele strukture *(f-structures)* en bestaan uit funksies soos subjek ('SUBJ'), objek ('OBJ') en XCOMP, wat 'n oop komplement is (Sells 1985:135). Voorts, *LFG does not deny that there is a significant aspect of the representation of syntax that is characterised by phrase structure trees, and proposes a level of constituent structure (c-structure) that corresponds roughly to the level of PF [Phonetic form] in Government-Binding Theory and to surface structure in Generalised Phrase Structure Grammar* (Sells 1985:136). Sells gee die volgende skematiese voorstelling van LFG:



LFG veronderstel 'n redelike standaard stel frasestruktuurreëls en ook 'n X' teorie wat min of meer met dié van GB ooreenstem (met dié verskil dat S en S' in LFG nie projeksies van enige leksikale kategorie is nie). C-strukture dra inligting wat in f-struktuur voorgestel word en wat geannoteer word in *functional schemata*. Die S-reël in Engels word byvoorbeeld as volg geannoteer:

$$S \longrightarrow \underset{(\uparrow \text{SUBJ}) \downarrow =}{NP} \quad \underset{\uparrow = \downarrow}{VP}$$

Sells (1985:140) verklaar die annotasie as volg: *The up- and down-arrows [...] refer to the f-structure that corresponds to the c-structure node built by the rule. The 'up' refers to the f-structure of the mother node and the 'down' refers to the f-structure of the node itself. The annotations [...], which are read 'up's SUBJ is down' and 'up is down' indicate that (a), all the functional information carried by the NP (i.e., the NP's f-structure) goes into the subject part of the mother's f-structure (i.e., the S's f-structure), and (b), all the functional information carried by the VP (the VP's f-structure) is also direct information about the mother's f-structure. Grossly put, the annotations say of the NP that it is the subject, and of the VP that it is the functional head.*

Vir meer besonderhede, vergelyk Kaplan en Bresnan 1982, Bresnan (red.) 1982 en Sells 1985. In Berwick 1982 word die *computational complexity* van LFG bespreek. LFG word gebruik vir 'n natuurliketaal-prosesseringsisteem deur onder andere Frey en Reyle (1983).

3.2.5 Unifikasiegrammatikas

Die term *unifikasie* verwys na 'n begrip wat Shieber soos volg definieer:

> *This notion of combining the information from two feature structures to obtain a feature structure that includes all the information of both is central to unification-based formalisms, for it is the notion of 'unification' itself. (Shieber 1986:17).*

Shieber beskou die doeltreffendheid van unifikasie as onweerlegbaar: *the efficacy of unification as a tool for linguistic analysis and computation seems irrefutable.* (1986:67-8). Die sentrale rol wat unifikasie in hierdie teorieë speel, word ook kortliks behandel in Kay 1985.

Die teorieë wat onder hierdie hoof genoem word, is ontwikkel as instrumente in die rekenaarlinguistiek, eerder as linguistiese teorieë (vgl. Shieber 1986:38). Hieronder ressorteer onder andere FUG, DCG, TELEGRAM en PATR-II. FUG, DCG en TELEGRAM sal slegs kortliks genoem word, terwyl ietwat meer besonderhede oor PATR-II gegee sal word.

Functional Unification Grammar (FUG)

FUG het aanvanklik bekend gestaan as *unification grammar* (vgl. Kay 1983) en later as *functional grammar*. Dit is ontwikkel deur Kay *as a general linguistic tool using unification as its only operation* (Shieber 1986:38). Volgens Kay (1984:78) kan 'n masjienvertaalsisteem op FUG gebaseer word: *Any system implemented strictly within this framework will be reversible in the sense that, if it translates from language A to language B then, to the same extent, it translates from B to A. If the set 'S' is among the translations it delivers for 'a', then 'a' will be among the translations of each member of 'S'. I know of no system that comes close to these advantages.*

Definite-Clause Grammars (DCG)

DCG het ontstaan uit werk in Prolog deur Pereira en Warren. *DCG and its related formalisms (slot grammars, extraposition grammars, gapping grammars, modified structure grammars, etc.) all use a variety of unification based on term structures rather than feature structures. Term unification was originally developed for use in automatic theorem-proving* (Shieber 1986:44). Vergelyk verder Shieber 1986:44-6, Pereira en Warren 1980 en Perreira en Warren 1982.

TELEGRAM

TELEGRAM (Teleological grammar) het die doel om 'n unifikasiegrammatika te annoteer *with assertions about how grammatical choices are used to achieve various goals* en om die beplanner in staat te stel *to augment the functional description of an utterance as it is being unified* (Appelt 1983:74).

PATR-II

Volgens Shieber (1986:11, 37) is die PATR-II-formalisme verreweg die eenvoudigste van die verskillende *unification-based* formalismes.

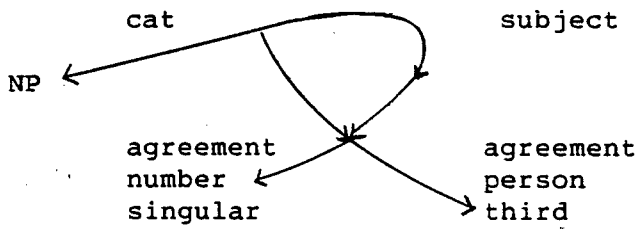
Net soos LFG maak PATR-II gebruik van f-strukture, ook genoem *dags* (as akroniem vir *directed acyclic graph)* (Shieber 1986:13). Die algemene notasie vir 'n f-struktuur is byvoorbeeld

number: *singular*
person: *third*

en so 'n struktuur word genoem D_{3sg}. Kenmerkwaardes kan self ook gestruktureerd wees, byvoorbeeld D_{3sg} kan die kongruensiekomponent van 'n enkelvoudige naamwoordfrase, D_{NP3sg} wees (Shieber 1986:13). F-strukture is kompleks (soos dié hierbo) of atomies (bv. die kenmerke *3* of *sg.*). 'n F-struktuur D_{NP} dra minder inligting of is meer algemeen of subsumeer D_{NP3sg}. Dus *a feature structure D subsumes a feature structure D' (notated* $D \sqsubseteq D'$*) if D contains a subset of the information in D'* (Shieber 1986:15). 'n F-struktuur D_{NPsg} en 'n f-struktuur D_{NP3} kan gesubsumeer word deur 'n f-

struktuur D_{NP3sg}. *This notion of combining the information from two feature structures to obtain a feature structure that includes all the information of both is central to unification-based formalisms, for it is the notion of 'unification' itself* (Shieber 1986:17). *In formal terms, we define the 'unification' of two feature structures D' and D'" as the most general feature structure D, such that D'* $\sqsubseteq$ *D and D'"* $\sqsubseteq$*D. We notate this D =D'*$\sqcup$*D'"* (Shieber 1986:17, 18). Indien f-strukture teenstrydige inligting bevat, kan unifikasie nie plaasvind nie.

Feature structures can be viewed as rooted, directed, acyclic graph structures (from which the term 'dag' is derived as an acronym) whose arcs are labelled with feature names. Each arc points to another such dag or an atomic symbol (Shieber 1986:20). Die f-struktuur $D_{NP3sgSubj}$ kan as volg in 'n *graph-structural notation* (Shieber 1986:20, 21) voorgestel word:



Die konkatenasie van stringe word aangedui met behulp van konteksvrye reëls. In die volgende voorbeeld word aangedui hoe die relevante f-strukture hiermee verbind word (Shieber 1986:25):

```
S      →      NP      VP
              <S head> = <VP head>
              <S head subject> = <NP head>
```

For these identities to hold, the 'head' value associated with the NP will have to be compatible with the VP's 'subject' *feature.* (Shieber 1986:22). Sien ook Shieber 1985:200-1. Vervolgens bied Shieber drie *sample grammars* aan, naamlik kongruensie van subjek en werkwoord vir persoon en getal, subkategorisasie van werkwoorde vir spesifieke postverbale komplemente en die semantiek van sinne, uitgedruk as enkoderings van logiese vorm (Shieber 1986:25-34; in *Appendix a,* pp. 69-87 gee hy die masjienleesbare weergawes van die grammatikas). In Shieber 1985:204 word 'n aantal addisionele voorbeelde van taalverskynsels wat binne die PATR-II-formalisme beskryf kan word, gegee:

* *Verbal subcategorisation for NPS, PPs, S's, VPs, including raising and equi phenomena, syntactic control, and auxiliary structure.*

* *Relation-changing rules including active/passive, 'there' insertion, and extraposition*

* *Unbounded dependencies including Wh-movement and relative clauses.*

* *Complex NPs and PPs.*

* *Adverbials of certain types.*

* *Semantics for these constructs, given as encodings of logical formulae in dag form.*

Vir meer besonderhede, sien Shieber 1986, asook Shieber, Uszkoreit, Pereira, Robinson en Tyson 1983, Karttunen 1984, Shieber 1984a, Shieber 1984b, Shieber, Karttunen en Pereira 1984 en Shieber 1985.

3.2.6 Extended Affix Grammars (EAG)

Oostdijk (1983:95) definieer 'n EAG as volg: *An EAG is a type of two-level grammar, based on a context-free grammar (CFG), i.e. the CFG constitutes the basic level and the affixes operate on the second level.* Enkele reëls binne die CFG is die volgende (Aarts en Van den Heuvel 1985:313-314):

```
SENTENCE     :   NP, VP.
NP           :   DET, HEAD.
DET          :   ;
                 'the'.
HEAD         :   NOUN STEM;
                 NOUN STEM, 's'.
NOUN STEM    :   'lion'; 'apple'.
VP           :   VERB STEM;
                 VERB STEM, 's'.
VERB STEM    :   'walk'; 'grow'; 'roar'
```

(Binne EAG word 'n dubbelpunt gebruik om 'herskryf as' aan te dui; alternatiewe word met behulp van 'n kommapunt aangedui, terwyl 'n komma gebruik word om die lede van 'n reël te verbind.)

Hierdie CFG kan nou aangevul word deur affikse. Byvoorbeeld om die kongruensie tussen subjek en werkwoord van 'n sin aan te dui, word die affikse getal en persoon bygevoeg. Die eerste reël hierbo word dus as volg uitgebrei (Oostdijk 1984:97):

```
SENTENCE:    NP(number, person),
             VP(number, person).
```

Metareëls, wat die omvang van die grammatika verminder en die leesbaarheid daarvan verhoog, spesifiseer die moontlike waardes wat 'n spesifieke affiks kan hê (Oostdijk 1983:99). Die volgende is 'n voorbeeld van 'n metareël

(Aarts en Van den Heuvel 1985:314; die notasie van metareëls verskil van dié van CF-reëls deur die gebruik van 'n dubbele dubbelpunt):

number::'SING'; 'PLU'.

Oostdijk (1983:99-121) bied 'n beskrywing van die NP in Engels in terme van EAG aan en in Aarts en Van den Heuvel 1985:313-318 word 'n beskrywing van 'n *small subset of English* aangebied. Binne die TOSCA-projek (Tools for Syntactic Corpora Analysis) is die Linguist's Workbench (LWB) ontwikkel, wat 'n interaktiewe analisesisteem is, asook die Linguistic Database (LDB), waarbinne die resultate van die analises gestoor kan word en daarna met behulp van 'n *query language* ondersoek kan word. Die LWB maak gebruik van 'n EAG-formalisme (maar die data van die LDB hoef nie noodwendig binne 'n EAG-formalisme ontleed te wees nie). Vir meer besonderhede, vergelyk Aarts en Van den Heuvel 1985.

3.3 STAND VAN NAVORSING IN SUID-AFRIKA

Volgens die vraelysondersoek *Standbeskrywing: Rekenaarondersteunde Taalnavorsing in Suid-Afrika* wat deur die RGN uitgevoer is, word geen navorsing tans in Suid-Afrika gedoen met die oog op die ontwikkeling van taalteoretiese modelle of -formalismes vir morfologiese/sintaktiese analise van taaldata nie. In die Departement Afrikaans en Nederlands aan die Universiteit van Stellenbosch word gewerk aan programme vir 'n morfologiese en sintaktiese analise vir Afrikaans. In die Departement Semitiese Tale aan die Universiteit van Stellenbosch is enkele programme ontwikkel vir morfologiese analise van taaldata en die programme word tans binne die Hebreeuse en Siriese projekte gebruik; hierdie programme word ook benut aan UNISA, waar tans ook alternatiewe analisesisteme ondersoek word, vir die Ugaritiese en Arabiese projekte; hierdie morfologies-geanaliseerde data sal later benut word vir sintaktiese analises. Met die oog hierop is UNISA besig met onderhandelinge om die LWB van die TOSCA-sisteem te bekom (die LDB is reeds ontvang); navorsing in dié rigting sal dus waarskynlik binne 'n (gewysigde) EAG-formalisme plaasvind.

3.4 SLOT

Om 'n goeie oorsig van die verskillende taalteorieë wat binne die rekenaarlinguistiek gebruik word, te bekom, kan veral Grishman 1986, Winograd 1983, Sells 1985 en Shieber 1986 geraadpleeg word, asook die bibliografieë wat in die werke aangebied word. Die genoemde werke is inleidend en vir meer tegniese besonderhede, behoort die verskillende navorsingsverslae bestudeer te word. Ontwikkelings in die veld van die rekenaarlinguistiek is besonder vinnig, veral wat betref toepassings. Die nuutste ontwikkelings binne die veld is nie in die verslag opgeneem nie, aangesien die materiaal nie in Suid-Afrika beskikbaar is nie. Kongresreferate van die Association for Computational Linguistics en die tydskrif Journal of Computational Linguistics is belangrike bronne.

Navorsing oor die struktuur van die tale van Suider-Afrika kan beslis bevorder word indien navorsers bewus sou wees van die voordele wat die rekenaaranalise van taaldata inhou (soos spoed en die noukeurigheid/toetsbaarheid van die formaliserings van taalreëls), en hierdie kundigheid produktief kan toepas. Verder, indien masjienvertaalsisteme ontwikkel kan word (wat een van die doelstellings van die rekenaarlinguistiek is) as gevolg van hierdie navorsing, sou dit geweldig belangrike implikasies vir Suider-Afrika met sy heterogeniteit van tale hê.

HOOFSTUK 4: REKENAARGESTEUNDE NAVORSING IN DIE MORFOLOGIE EN DIE SINTAKSIS

4.1 INLEIDING

Ooreenkomstig die doelwitte van die TEXTNET-ondersoek, is die doel met hierdie bespreking van die gebruik van die rekenaar in morfologiese en sintaktiese navorsing, eerstens om inligting te verskaf oor ontwikkelings in die buiteland, en tweedens om 'n beskrywing te gee van sodanige navorsing in Suid-Afrika. Die bespreking maak geensins aanspraak op volledigheid nie en daar word slegs gepoog om belangrike tendense en enkele belangriker programme kortliks te bespreek.

Dit sou kunsmatig wees om die verslag oor rekenaargesteunde navorsing in die morfologie en die sintaksis volledig te skei, aangesien beide aspekte nou verwant is en die morfologie en die sintaksis dikwels as twee komponente van dieselfde navorsingsprojek voorkom, of as modules van dieselfde program.

4.2 REKENAARGESTEUNDE NAVORSING IN DIE MORFOLOGIE

Gerekenariseerde morfologie is op sigself 'n konsep wat verskeie navorsings-aktiwiteite dek en is ontwerp om verskeie tipes voor- en na-redigerings- en teksprosesseringsprosedures te behels, onder andere:

- segmentering en outomatiese kaesura-programme;
- outomatiese lemmatisering;
- die uitgee van elektroniese woordeboeke;
- spelkontroleprosedures en die uitgee van spelkontroleprogramme;
- outomatiese herkenning en merking van rededele.

4.2.1 Rekenaargesteunde morfologie-navorsing in die buiteland

(1) Vroeë tegnieke: outomatiese lemmatisering

Die tipe lemmatiseringsprosedure wat ontwikkel word, hang uiteraard af van die uiteindelike aard van die lemma en die graad van inkorporering van linguistiese kennis. Lemmatiseringsprogramme verwys dus na 'n wye reeks programme, vanaf die mees eenvoudige programme waarin slegs 'n sekere aantal karakters aan die begin of einde van 'n woord afgekap word, tot die mees gesofistikeerde programme, waarin die woordvormingsteorie in verband gebring word met reëlmatige paradigmas, uitsonderings, lyste van produktiewe affikse ensovoorts.

Enkele voorbeelde van sodanige programme is:

(a) 'n Lemmatiseringsalgoritme vir 'n frekwensiewoordeboek van Sweeds, op die basis van die ontleding van woordpare; vergelyk Hellberg 1972.

(b) 'n Uitgangsparadigmabenadering toegepas op Deens; vergelyk Leunbach 1973.

(c) 'n Sisteem wat in Duitsland ontwikkel is vir die wiskundige afkapping van 'n voorafbepaalde aantal karakters, soos toegepas op Duitse tegniese tekste. In 'n tweede stadium is algoritmes ontwikkel vir die bepaling van stelle karakters wat inisiële, middel en terminale uitgange bepaal; vergelyk Hann 1973 en 1978.

(2) Morfologiese komponente van masjienvertaalsisteme

Masjienvertaling het ook 'n bydrae tot morfologiese analise gemaak. Ons verwys slegs na enkele tweede-generasie sisteme.

(a) Die Atef-sisteem deur Geta

Hierdie sisteem vorm die morfologiese analise-komponent van die Ariane-78 masjienvertaalsisteem. Dit werk met 'n woordeboek van affikse en is gebaseer op die kartering *(mapping)* van stringe karakters op benoemde bome, wat, via 'n *finite state transducer,* die invoer tot strukturele analise daarstel. Die proses behels die segmentering van woorde in morwe en die toekenning van waardes aan veranderlikes; die strategie is dus om alle moontlike segmenterings van 'n woord te doen deur in die woordeboek te soek vir alle patrone wat 'n volledige of gedeeltelike karakterstring genereer; die uitgange word met indekse geassosieer wanneer 'n wortel gevind word.

(b) Wobuso

Dit vorm die morfologiese analise van die Susy-sisteem en word vollediger in die TRANSNET-verslag bespreek.

(3) Onlangse ontwikkelings

Rekenaarmorfologie het in die afgelope aantal jare dramaties verander; dit het wegbeweeg van *ad hoc*-programmering tot die daarstelling van linguistiese teorieë oor woordvorming. Unifikasiegrammatikas het hierin 'n baie belangrike rol gespeel.

(a) Die Sewentigerjare

As voorbeeld hier kan genoem word die morfologiese komponent van die Mindsisteem; sien veral Cercone 1974 en 1978 en Kay 1970, 1973a, 1973b. Dit

bied 'n veeldoelige, gesofistikeerde, buigbare, taalonafhanklike, reëlgebaseerde raamwerk met behulp waarvan tekste geprosesseer kan word. Die sisteem bevat 'n gesofistikeerde morfologiese analiseprosedure wat woorde segmenteer en morfofonemiese reëls daarop toepas; verder identifiseer dit leksikale reekse wat in die woordeboek opgesoek moet word. Die sisteem is hoogs proseduraal en 'n tipiese analiseprosedure bestaan uit die volgende: morfografemiese herskryfreëls, 'n woordeboeknaslaanproses, 'n infleksiegenereerder, versoenbaarheidstoetsing en idioom- en samestelling-herkenning. Hierbenewens bevat die Mind-sisteem 'n effektiewe woordeboek-verwysingsisteem, met behulp waarvan inligting wat met 'n spesifieke string geassosieer word, teruggeroep kan word, asook 'n prosedure om die stel reëls wat op 'n woord toegepas word en die inligting wat uit die woordeboek oor dele daarvan verkry word, te beheer.

(b) Die tagtigerjare

Die volgende navorsing kan kortliks genoem word:

Beta (later Betatext) is ontwikkel om die mees waarskynlike morfotaktiese segmentering van 'n woord te bepaal, sonder die gebruik van 'n woordeboek; vergelyk Brodda en Karlsson 1980 en Brodda 1986.

Die Uppsala Chart Parser maak gebruik van 'n veeldoelige kaartgebaseerde *(chart based)* sisteem; vergelyk Sagvall-Hein 1978, 1983 en Carlsson 1982.

Karttunen (1981) het 'n sisteem ontwikkel op die basis van 'n inisiële stel gepostuleerde wortels wat ontdek is deur 'n foneem-vir-foneem links-na-regs *matching process.*

Die Morfin-sisteem (later Morfo) is ontwikkel aan die Helsinki University of Technology vir die ontleding van Finse infleksiemorfologie, gedeeltelik sonder die gebruikmaking van 'n woordeboek. Die sisteem is dus potensieel daartoe in staat om onbekende leksikale items te analiseer en die sisteem van enige taal uit te bou; vergelyk Jaeppinen 1983 en 1986.

Die Lexicrunch-sisteem is ontwikkel om reëls vir woordvorming outomaties te abstraheer op grond van 'n korpus van basisvorme en hulle geassosieerde geïnflekteerde vorme. Die idee is om 'n leksikon so kompak moontlik te representeer deur alle geïnflekteerde vorme uit te skakel. Die ontledingstrategie is daarop gebaseer dat daar geraai word watter transformasies plaasgevind het en om hierdie transformasies dan in omgekeerde volgorde te ontknoop; vergelyk verder Golding en Thompson 1985.

Jorge Hankamer se C-compiler-program vir VAX-rekenaars, in bedryf by die University of California, Santa Cruz, is aangepas uit 'n vroeëre weergawe wat by die Max Planck Instituut te Nijmegen vir UNIX-rekenaars ontwikkel is. Hankamer se program, waarvan daar ook 'n mikrorekenaarweergawe beskikbaar

is, sou volgens hom ewe goed vir tale soos Zulu of Afrikaans gebruik kon word indien fonetiese en leksikale spesifiseringsfilter-subprogramme vir hulle geskryf word.

Fought se morfo-sintaktiese ontledingsprogram wat aan die University of Pennsylvania bedryf word, benut die nie-numeriese programmeringstaal SNOBOL4 (vgl. Griswold, Poage and Polonsky 1971, Brillinger and Cohen 1972, Waite 1973 en Griswold and Griswold 1973), 'n veeldoelige koderingstaal wat woorde en afkortings (dus gewone taal) gebruik in plaas van syfer- en abstrakte simbole. Fought se prosedure begin met 'n outomatiese elektroniese inleser wat drukwerk in 'n aantal standaard-lettertipes (bv. *elite, pica, courier)* dubbelbladgewys in elektroniese kodes vertaal vir bewerking op middelslag UNIX-rekenaars deur middel van die morfo-sintaktiese programme wat in SNOBOL4 geskryf is.

Grimes se Paradigm Program vir die ontleding van affiksposisies en gelyktydige voorkoms *(co-occurrences)* van bepaalde morfeempare, vorm deel van 'n programpakket wat as die PTP Programs (the Programmable Text Processor) bekend staan, en inkorporeer benewens Grimes se program onder andere 'n glossarium-skryfprogram, twee woordeboekskryfprogramme, 'n inventaris- en frekwensieprogram asook 'n woordverwerkingsprogram. Sover vasgestel kon word, is slegs een van die bogenoemde programme op disket beskikbaar aan belangstellendes (vgl. Inligtingstuk 8).

Verdere programme wat ontwikkel is, is onder andere die Parspat- en Oracle-programme van die Universiteit van Amsterdam (Akkerman 1985), die Hamans-sisteem vir Duitse morfologie (Hoeppner 1983, 1984), Church se sisteem wat deel vorm van 'n teks-na-spraaksisteem (Church 1986), asook die werk van Byrd (1986) en Ralli en Galiotou (1987).

(4) Tweevlak-morfologie

'n Nuwe konsep en 'n radikale nuwe benadering, bekend as die tweevlak-model, se belangrikste kenmerke is die volgende:

- dit is bedoel vir sowel analise as sintese;
- die algoritme is taalonafhanklik;
- dit benut ooreenstemmings tussen oppervlak (tekstuele) en leksikale fenomene en skakel direk tussen leksikale en oppervlakvorme sonder om die tussenvorme te konstrueer;
- die reëls word geïmplimenteer as *finite state transducers* (FST);
- die reëls is deklaratief.

Hierdie nuwe benadering is voorgestel deur Koskeniemmi (1983), in navolging van vroeëre werk deur Kay. Latere werk is gedoen deur onder andere Gazdar en die ontwikkeling van GPSG (Generalized Phrase Structure Grammar). Die model is aanvanklik vir Fins en Engels opgestel, maar is reeds op baie tale toegepas, onder andere Japannees (Alam 1983), Arabies (Kay 1985, 1987), Oudkerkslawies (Lindstedt 1984) en Frans (Lun 1983).

4.2.2 Rekenaargesteunde morfologie-navorsing in Suid-Afrika

Navrae en die vraelys met betrekking tot rekenaargesteunde taalondersoek wat deur die RGN uitgestuur is, het min inligting oor morfologie-navorsing opgelewer. Blykbaar word daar slegs by twee universiteite rekenaargesteunde morfologie-navorsing gedoen, naamlik UPE en US.

Sover vasgestel kon word, bestaan daar drie Suid-Afrikaanse programme met morfeemontledingsfasiliteite. By al drie is morfeemontleding egter 'n middel tot 'n doel eerder as die doel self. Met die TEKSANA METAKEUSES-ontleedprogram van De Klerk (Inligtingstuk 9), kan enige gestipuleerde elemente, waaronder morfeme, van 'n masjienleesbare teks op verskeie maniere gemanipuleer word. Die program is volgens die inligting wat verskaf is onder andere in staat tot frekwensie-analise, voorbeeld-analise, frekwensie- en voorbeeld-analise, variant-analise, en variantfrekwensie-analise. De Klerk se pakket is geen egte rekenaar-morfeemontledingsprogram nie, aangesien dit oor geen morfeeminventaris beskik wat die korrekte morfeemkombinasies beheers nie, maar bloot enige gestipuleerde letterkombinasies vir verdere bewerking opspoor.

Dodds se Afrikaanse spelbeheerpakket wat tans met Combrink en Dodds se Retrograde-woordeboek van Afrikaans (1984) as vertrekpunt ontwikkel word, word volgens Combrink só ontwerp dat dit anders as in die geval van bestaande spelbeheerpakkette nie afsonderlike lekseme nie, maar bepaalde woordstamme met al hul affikskombinasies as grondslag neem vir spelbeheer. Op dié manier word meer as 'n honderd afsonderlike lekseeminskrywings met die stam *funk-* volgens Combrink ekonomies en kragtig gekombineer tot een inskrywing met verskillende affikskombineringsvoorwaardes. Hoewel die doelwit van Dodds se program spelbeheer is, berus dit op 'n uitgebreide morfeeminventaris wat sonder veel moeite as morfeemontledingsprogram aangepas sou kon word.

De Stadler en Coetzer (1988) se skrif-tot-spraak-omsetter neem morfeme met hulle allomorwe as basiese omsettingseenhede. Gevolglik inkorporeer een van die subroetines van die omsetter 'n morfeemwoordeboek met 'n groot aantal stammorfeme en affikse wat komplekse woorde in morfeme verdeel. Hierdie benadering hou groot moontlikhede in vir rekenaar-morfeemontleding, hoewel daar nog geen teikendatum vir die implementering van die omsetter is nie. Daarby is dit nie duidelik of die morfeemontledingsprogram, wat as deel van die skrif-tot-spraak-omsetter deur die Poswese gefinansier word, vir taalstudie beskikbaar sal wees nie.

4.3 REKENAARGESTEUNDE NAVORSING IN DIE SINTAKSIS

In rekenaargesteunde sintaktiese navorsing is daar 'n basiese onderskeid te tref tussen die benadering waar die rekenaar se gewone soek- en ander funksies benut word en die *parsing*-benadering waar ontleedprogramme op basis van gekodeerde korpusse data verwerk. Die bespreking wat volg, konsentreer

op laasgenoemde benadering, wat meer gesofistikeerd is, fyner analises oplewer en heuristies waardevol kan wees.

Die nuttigheid van die rekenaar gewoon as 'teksdeursoeker' word egter nie ontken nie. Met die soek- en sorteerfunksies kan teksmateriaal vinnig en op groot skaal deurgewerk word vir die uitlig van bepaalde verskynsels. Byvoorbeeld, in 'n ondersoek na die onderskeie funksies van *omdat* en *want*, kan data vinnig en akkuraat uit 'n korpus gelig en sorteer word. Frekwensie-, verspreidings- en ander statistiek kan dan ook met behulp van gewone (d.w.s. nie-taalondersoekspesifieke) programme bereken word.

Die literatuur, asook programme van internasionale kongresse, toon aan dat die hoofstroom in rekenaargesteunde sintaktiese navorsing in die konteks van rekenaargesteunde vertaling en kunsmatige intelligensie geleë is. Een van die doelwitte op laasgenoemde gebiede is die ontwerp van programme wat tekste volledig kan ontleed.

'n Oorsig van verskillende benaderings tot die ontwikkeling van sintaktiese ontleders *(parsers)* word vervolgens geskets.

4.3.1 Konteksvrye ontleders

Reeds sedert die vroeë vyftigerjare word navorsing gedoen oor ontleed-algoritmes *(parsing algorithms)* op die basis van konteksvrye grammatikas. 'n Konteksvrye grammatika beskik oor reëls met die vorm A -- X, waar A 'n nie-terminale simbool is en X 'n reeks van een of meer terminale en/of nie-terminale simbole is. Binne die raamwerk van 'n bepaalde grammatika geld die reëls ongeag die konteks.

'n Voorbeeld van 'n konteksvrye grammatika is die Linguistic String Theory van Harris (1962). 'n Volledige uiteensetting van hierdie grammatika word gegee deur Sager (1981). Hierdie teorie maak gebruik van sintaktiese kategorieë, 'n versameling elementêre reekse en kombinatoriese reëls wat elementêre reekse tot sinreekse saamvoeg. Die eenvoudigste sinne beskik slegs oor een elementêre reeks, genoem 'n *center string*. Hierdie reeks kan uitgebrei word deur adjunksie, konjunksie en vervanging. Elke woord in die taal word gekategoriseer op grond van sy grammatiese eienskappe. Hierdie kategorisering geld ten opsigte van die woord se gebruik in die taal as geheel en nie slegs sy gebruik in 'n bepaalde sin of teks nie. Dit is nodig dat inperkings op hierdie grammatikas geplaas word.

Ontleders word geklassifisieer as (a) neerwaartse of (b) opwaartse ontleders. 'n Neerwaartse ontleder ontleed vanaf die beginsimbool tot by die sin; 'n opwaartse ontleder ontleed vanaf die woorde of woordkategorieë deur middel van reduksie tot by die beginsimbool. Ontleders kan ook verdeel word in *backtracking*-algoritmes, wat een afleiding op 'n keer probeer en teruggaan wanneer vasgehaak word, en parallelle algoritmes wat alle

afleidingsreekse gelyktydig nagaan. Uitvoerige beskrywings en taksonomieë word deur Aho en Ullman (1972) gegee.

Die mees gebruikte ontleder is 'n neerwaartse *backtracking*-algoritme. Hierdie algoritme enumereer afleidings totdat dit 'n afleiding vind wat die invoersin genereer. Hierdie algoritme kan enige grammatika hanteer, behalwe een wat linksrekursief is. ('n Grammatika is linksrekursief indien daar vir 'n nie-terminale simbool A 'n reeks afleidings is wat begin met die simbool A.) Addisionele toetse moet vir linksrekursiewe grammatikas ingesluit word om herhalende lusse te vermy.

'n Ander algemeen gebruikte algoritme is die parallelle opwaartse ontleder. Hierdie ontleder bou gedeeltelike ontledings, waar elke gedeeltelike ontleding 'n analise van 'n subreeks van die sin verteenwoordig. 'n Komplikasie by die implementering van hierdie prosedure is om te voorkom dat dieselfde reduksie nie tweekeer plaasvind nie. Hierdie ontleder is bruikbaar vir enige grammatika wat nie oor nulelemente beskik nie. Hierdie probleem kan oorkom word deur die nulelemente uit te skakel of deur die nulelemente te merk.

Die spasievereistes vir die parallelle algoritmes is groter omdat alle gedeeltelike ontledings gestoor moet word. Die *backtracking*-algoritme daarenteen, vereis weinig meer spasie as wat nodig is vir 'n enkele ontledingsboom. 'n Ontleder se tydvereistes word bepaal deur die aantal produksies of reduksies wat uitgevoer word tydens die ontleding van 'n sin.

Die opwaartse ontleder sal in baie gevalle reduksies (gedeeltelike ontledings) bou waar die neerwaartse ontleder die ooreenkomstige produksies sou vermy het. 'n Voordeel van die opwaartse ontleder is dat 'n gedeeltelike ontleding slegs eenkeer gebou sal word. Die neerwaartse algoritme mag 'n gegewe simbool meermaal uitbrei deur by dieselfde woord te begin indien daardie simbool in verskillende kontekste voorkom.

Dit is moontlik om die voordele van die neerwaartse en opwaartse prosedure te kombineer deur een van hierdie algoritmes te neem en kenmerke van die ander te inkorporeer. Eksperimente om verskeie ontlederś te vergelyk, is gedoen deur Slocum (1981).

Die belangrikste van die vroeë konteksvrye ontlederś was die Harvard Predictive Analyzer (Kuno en Oettiger 1962). 'n Predictive Analyzer is 'n neerwaartse ontleder vir konteksvrye grammatikas en is geskryf in die Greibach normale vorm. (In die Greibach normale vorm, moet die eerste simbool aan die regterkant van elke produksie 'n terminale simbool wees.) Hierdie analiseerder het duidelik laat blyk dat 'n konteksvrye formulering van 'n grammatika vir 'n natuurlike taal nie suksesvol is nie. 'n Verskeidenheid inperkings is nodig om ongeldige analises uit te skakel.

'n Ander belangrike vroeë ontleder was die onmiddellike konstituentanaliseerder (die RAND-sisteem). Hierdie sisteem het 'n grammatika gebruik in die Chomsky standaardvorm en 'n ontleedalgoritme wat ontwerp is deur John Cocke, wat opwaarts geanaliseer het met 'n enkele links na regs ontleding van die sin (Hays 1967). Die RAND-sisteem was 'n vinnige algoritme wat alle ontledings gelyktydig gedoen het, maar het vir lang sinne baie spasie vereis. Hierdie sisteem was beperk tot sinne van ongeveer 30 woorde.

Die eerste *linguistic string grammar* (stringgrammatika) is ontwikkel aan die Universiteit van Pennsylvania (Harris 1962). Dit was 'n opwaartse ontleder en is spesiaal ontwikkel vir stringgrammatikas. Hierdie grammatika het gebruik gemaak van sintaktiese inperkings.

4.3.2 Transformasionele ontleders

Die eerste transformasionele ontleders is in die sestigerjare geskryf op die basis van Chomsky se TG, en veral op die insig dat baie sinne parafrases van mekaar is. Hierdie feit is vir die rekenaarkundige van besondere belang. Die rekenaarkundige se doel is nie slegs om 'n sin te ontleed nie, maar ook om vas te stel wat die sin beteken. Deur middel van transformasies kan die groot verskeidenheid sinne gereduseer word. Sinne wat parafrases van mekaar is, kan teruggevoer word na 'n enkele sin. Hierdie reduksie lei daartoe dat 'n kleiner reeks konstruksies op verskillende analise-stadiums gehanteer word.

Die generatiewe proses van die TG is baie meer ingewikkeld as vir 'n herskryfgrammatika. Dit bly veral problematies om 'n herkenningsprosedure te ontwerp. Grishman (1986:48) wys op verskeie probleme wat saamhang met die omkering van die generatiewe proses, waaronder:

- die toekenning van ontledingsbome wat alle oppervlakstrukture insluit wat deur die grammatika toegeken word;
- die vasstelling van die transformasiereekse wat gelei het tot die generering van 'n boomstruktuur wat nie in die basis is nie.

Die TG-analiseerder moet oppervlakbome kan saamstel en transformasies 'ongedaan' kan maak.

(1) Oppervlakstruktuurontleding

Binne 'n TG-benadering sal die oppervlakstruktuurbome baie strukture bevat wat nie deur die basiskomponent gegenereer word nie. Indien al die oppervlakstrukture deur 'n konteksvrye grammatika gegenereer moet word, sal dit nodig wees om die basis uit te brei deur 'n dekkingsgrammatika *(covering grammar)* op te stel. Omdat 'n transformasiereël 'n knoop kan vervang met twee knope, beteken dit dat 'n oneindige aantal reëls tot die basiskomponent toegevoeg moet word. Hierdie verrykte grammatika sal alle geldige oppervlakontledings lewer, maar sal ook baie ontledings lewer wat nie uit

die dieptestruktuur afgelei kan word nie. Aangesien elke 'onegte' oppervlakanalise 'n lang omgekeerde transformasionele proses moet ondergaan voordat dit as eg aanvaar word, bied dit vir hierdie benadering 'n ernstige probleem.

(2) Omgekeerde *(reverse)* transformasies

Hierdie komponent moet uit 'n oppervlakstruktuur alle dieptestrukture genereer waaruit die oppervlakstruktuur gegenereer kan word. Dit is nie altyd moontlik om so 'n komponent daar te stel nie. Indien 'n transformasie 'n deel van die ontledingsboom weglaat, sal dit onmoontlik wees om daardie gedeelte te rekonstrueer wanneer teruggewerk word vanaf die oppervlakstruktuur. Daar mag 'n oneindige aantal dieptestrukture wees wat een oppervlakstruktuur genereer. Dit is dan 'n geval van onherwinbare weglating.

Om hierdie probleme te oorkom, is twee sisteme in die middel sestigerjare ontwikkel wat beperkte sukses gehad het - die sisteem van Zwicky et al (die Mitre-groep) en dié van Petrick. Die Mitre-groep se grammatika het 'n basiskomponent gehad, bestaande uit ongeveer 275 reëls en 54 transformasies. Vir die herkenningsprosedure het hulle 'n konteksvrye dekkingsgrammatika opgestel met ongeveer 550 produksies en 14 omgekeerde transformasionele reëls. Die sisteem was baie stadig (ongeveer 36 minute om een 11-woord sin te genereer). Om die program vinniger te maak, het hulle gebruik gemaak van superbome (verskeie ontledingsbome in 'n enkele struktuur) en uitskakelingsreëls *(rejection rules)* waarmee sommige bome vroegtydig gedurende die ontleding uitgeskakel is. Die sisteem van Petrick (Petrick, 1965, 1966; Keyser en Petrick, 1967) het ook 'n reeks omgekeerde transformasies toegepas en dan nagegaan of die boom deur die basis gegenereer kan word. Daarna is die analise geverifieer deur transformasies op die basisboom toe te pas. Petrick se stelsel is heelwat hersien. In die huidige stelsel word die dekkingsgrammatika en die omgekeerde transformasies met die hand voorberei. Die transformasionele dekomposisie geld ten opsigte van 'n boomstruktuur en heelwat buigsaamheid is voorsien vir die stel van transformasies en hulle toepassingsvoorwaardes.

Corstius (1978:178) wys daarop dat die Transformasionele Grammatika meer gerig is op die spreker as op die hoorder. Die rekenaartaalkundige daarenteen toon 'n natuurlike voorkeur vir die hoorder. Transformasionele ontleders is baie meer bruikbaar vir generering as vir herkenning.

Sien ook hoofstuk 3,3.2.1, waar addisionele inligting oor die Mitre-sisteem gegee word, en ook inligting oor die gebruik van die Government-Binding-teorie binne 'n *parser* vir Frans (Wehrli 1984).

4.3.3 Verrykte konteksvrye ontleders *(Augmented context-free parsers)*

Uit die vroeë TG-analiseerders het dit duidelik geblyk dat daar veral aandag aan twee aspekte gegee moet word:

(a) die verfyning van die oppervlakanalise, sodat elke boomstruktuur minder struktuurbome lewer wat weer transformasioneel ontleed moet word, en

(b) die vasstelling van 'n basisstruktuur uit 'n oppervlakstruktuur op 'n betreklik direkte wyse.

Om te kom tot 'n meer suksesvolle oppervlakontleding moet addisionele inperkings gestel word. Vir die TG wat van 'n konteksvrye dekkingsgrammatika gebruik maak, blyk dit problematies te wees. Daar is kragtiger herskryfgrammatikas (kontekssensitiewe grammatikas en onbeperkte herskryfreëls), maar hulle blyk nie baie suksesvol te wees vir die stel van grammatiese inperkings op die oppervlakstruktuur nie.

Om hierdie probleme die hoof te bied, is gebruik gemaak van 'n konteksvrye grammatika en is klein prosedures by die individuele produksies van die grammatika gevoeg. Die konteksvrye grammatika spesifiseer die grammatika en die prosedures vervul twee funksies:

- hulle maak voorsiening vir die grammatikale inperkings wat nie in die konteksvrye grammatika opgeneem kan word nie, en

- hulle stel die bewerkings vas wat nodig is om die sin se diepte-struktuur te agterhaal.

Kombinasies van konteksvrye grammatikas en prosedures word verrykte konteksvrye grammatikas genoem. Twee belangrike verrykte konteksvrye grammatikas is reeds in gebruik, naamlik die Linguistic String Parser van 'n groep aan die Universiteit van New York onder leiding van Sager en die Augmented Transition Networks (ATN) van Woods.

(1) Die Linguistiese Stringontleder (Linguistic String Parser)

Hierdie ontleder maak gebruik van *restriction language* - 'n taal wat spesifiek ontwerp is vir die skryf van natuurliketaal-grammatikas.

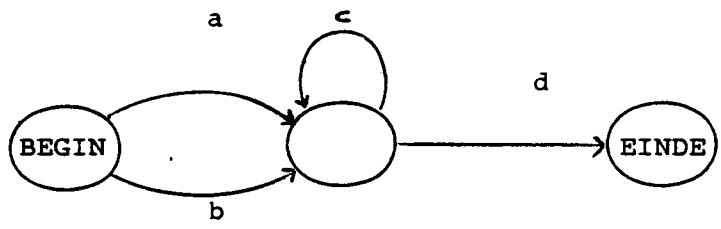
'n Restriction Language Grammar bestaan uit 'n konteksvrye komponent en 'n versameling inperkings. 'n Konteksvrye analise word slegs aanvaar indien 'n analise alle toepaslike inperkings bevredig. Sommige van hierdie inperkings is inperkings op die struktuur van bome, die meeste is egter toetse vir die attribute van bepaalde woorde. Die woordklasse vorm slegs die eerste orde kategorisasie van die grammatiese eienskappe van woorde. Om grammatiese inperkings, byvoorbeeld getal, subkategorisasie, seleksie, ensovoorts, te hanteer word 'n meer verfynde klassifikasie benodig. Dit word bewerkstellig deur attribute te voeg tot woorde in die leksikon. N kry byvoorbeeld die attribute ENK en MV; telbare naamwoorde NCOUNT, ensovoorts.

'n Belangrike kenmerk van die stringgrammatika is dat dit gerig is op die verhoudings tussen bepaalde woordpare in 'n sin. Twee verhoudings staan sentraal: die verhouding tussen twee woorde wat elemente van dieselfde string is en die verhouding tussen 'n kern *(host)* en sy adjunkte, byvoorbeeld bywoordelike sinsadjunkte of 'n adjunk van 'n woord in 'n ander string. Die meeste grammatikale inperkings geld ten opsigte van woorde wat verbind word deur hierdie stringverhoudings.

(2) Verrykte oorgangsnetwerke (Augmented Transition Networks)

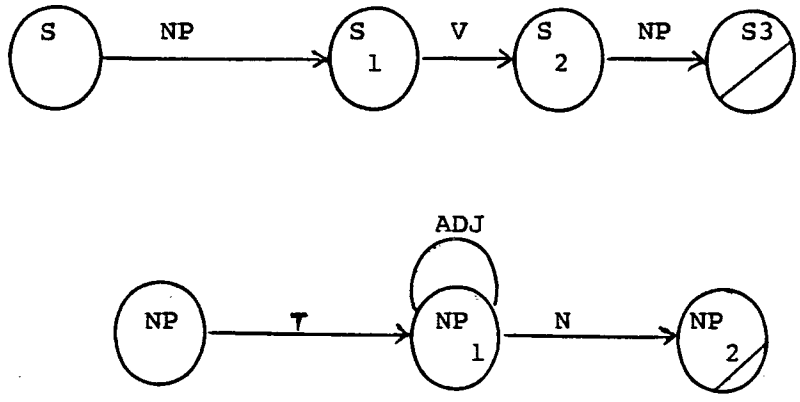
Hierdie formalisme, bekendgestel deur Woods (1980), is een van die bruikbaarste formalismes vir die skryf van natuurlike taalgrammatikas. 'n Oorgangsnetwerk verteenwoordig 'n eindstandgrammatika. 'n Oorgangsnetwerk is 'n netwerk waar die verbindings *(arcs)* benoem word deur terminale simbole (woorde of woordkategorieë). 'n Node van die netwerk word aangedui as die beginstand en een of meer nodes word gemerk as die eindstand. 'n Sin in die taal word deur die netwerk geïdentifiseer as 'n sin indien daar 'n pad is vanaf die beginstand tot die eindstand en die benaminge van die verbindings ooreenkom met die woorde van die sin. Die netwerk in Figuur 4.1 kom byvoorbeeld ooreen met sinne wat bestaan uit *a* of *b,* gevolg deur 'n sero of meer *c's,* gevolg deur 'n *d.*

Figuur 4.1



'n Rekursiewe verrykte netwerk is 'n netwerk met die volgende aanpassings: die nodes van die netwerk word benoem en die verbindings kan benoem word òf deur nodename òf terminale simbole. Hierdie nodename kom ooreen met die nie-terminale simbole van 'n konteksvrye grammatika. 'n Node wat benoem word met 'n nodenaam kan gekruis word indien daar 'n pad is (vanaf die benoemde node na die eindstand) wat ooreenkom met sommige dele van die sin of die res van die sin. Die rekursiewe netwerk in Figuur 4.2 (waar 'n deurkruiste sirkel 'n eindstand aandui) omskryf, byvoorbeeld 'n S as 'n reeks wat bestaan uit NS, WW, NS waar 'n NS 'n T is, wat gevolg word deur sero of een of meer ADJ'e en 'n N.

Figuur 4.2



'n Rekursiewe netwerk met bygevoegde prosedures (gewoonlik in LISP) om voorsiening te maak vir die stel van grammatikale inperkings en vir die generering van dieptestrukture, staan bekend as 'n verrykte oorgangsnetwerk.

Die ontleder vir 'n ATN moet 'n pad deur die netwerk, vanaf die beginnode tot by die laaste node, vind. Gewoonlik word 'n neerwaartse *backtracking*-algoritme gebruik. 'n Ordening kan vir die verbindings gespesifiseer word sodat sommige moontlikhede voor ander getoets kan word. Elke verbinding in die netwerk spesifiseer:

- die verbindingsbenaminge (woord, woordkategorie of nodenaam);
- enige voorwaardes vir die kruising van 'n verbinding;
- die prosedure nadat die verbinding gekruis is, en
- die nuwe stand wat bereik word na die kruising.

4.3.4 Onlangse ontwikkelings in die rekenaargesteunde sintaktiese navorsing

Die oorsig hieronder word verdeel volgens die doel van die sisteem, naamlik as deel van 'n 'nouer' sisteem of as deel van 'n masjienvertaalsisteem.

(1) As deel van 'n nouer sisteem

(a) Egte sintaktiese ontlederś

As voorbeeld hier kan die TOSCA-sisteem van die Universiteit van Nijmegen genoem word. Die primêre doel van hierdie projek is die ontwikkeling van navorsingshulpmiddels om 'n outomatiese maar interaktiewe sisteem vir die ontleding van 'n korpus tekste daar te stel. Dit word op twee maniere gedoen:

- deur die konstruksie van 'n sisteem onafhanklik van enige korpustaal om groot, ongeannoteerde korpora tekste te prosesseer met die doel om sintaktiese databasisse te produseer, en
- deur 'n formalisme te voorsien met behulp waarvan grammatikas geskryf kan word wat vir outomatiese sintaktiese korpusanalise gebruik kan word.

Die formele grammatika wat gebruik word, is die Extended affix grammar (EAG). Die sintaktiese komponent is die einde van 'n modulêre ketting wat *tokenization,* morfologiese ontleding, woordeboeknaslaanprosedures, lemmatisering en die ontleding van idiomatiese uitdrukkings insluit. Die sintaktiese ontleder opereer op die opeenvolging van woordklasse en idiomatiese uitdrukkings en dit lewer 'n sisteem van benoemde hakies, waarin kategorie- en funksie-informasie voorkom, op. Die sisteem is interaktief en vereis die ingrype van die linguis wanneer 'n veelheid van interpretasies voorkom (in die meerderheid van die gevalle), as gevolg van homografie, apposisie en onuitgedrukte beperkings. Huidige navorsing is gerig op die

uitbreiding van die Linguistic Database (LDB) en die ontwikkeling van GPSG-grammatikas, asook die daarstelling van korpora geanaliseerde Spaanse en Arabiese tekste. Vir meer besonderhede sien Aarts en Van den Heuvel 1985, asook hoofstuk 3 en die verwysings aldaar.

'n Verder voorbeeld van sodanige sintaktiese ontleders is die CLAWS-program, wat vir die outomatiese indeksering van die LOB-korpus (vgl. hoofstuk 7), met behulp van die TAGGIT-program, gebruik word. Vergelyk verder Johansson 1982, 1987 en Atwell 1982.

Verskeie unifikasie-gebaseerde sisteme is ook ontwikkel, waarvan die belangrikste die PATR-II-formalisme is (Shieber 1986); meer besonderhede oor enkele unifikasie-gebaseerde sisteme word in hoofstuk 3 gegee.

(b) Ander toepassings

'n Sekere mate van sintaktiese ontleding word ook vereis vir verskeie masjienvertaalsisteme wat vir spesifieke toepassings in geslote semantiek ontwikkel is; hiervan is die TAUM-meteosisteem 'n welbekende voorloper. Hierdie semantiese beperking kan, paradoksaal, lei tot baie spesifieke voorwaardes op die sintaktiese ontledingsgrammatika.

'n Verdere moderne toepassing wat tot 'n mate van sintaktiese ontleding gebruik maak, is sintaktiese en stylkontroleerders. Voorbeelde hiervan is die Right-program vir die toetsing van skryfvaardighede en die MR-program vir hulp met skryf. Hierdie tipes programme maak daarop aanspraak dat hulle kunsmatige-intelligensie-programme is. Hulle baan die weg vir nuwe navorsing op grond van aanvraag in die mark.

(2) As deel van 'n masjienvertaalsisteem

Die bestaande masjienvertaalsisteme *(computer-assisted or machine-aided translation systems)* is as interaktiewe sisteme ontwikkel, wat impliseer dat die sintaktiese ontleding slegs 'n module van die groter program vorm, met moontlik 'n addisionele module vir basiese sintaktiese generering.

Enkele sisteme word slegs kortliks genoem. Vir meer besonderhede vergelyk veral Hutchins 1986, Nirenburg 1987 en die TRANSNET-verslag.

(a) Die LOGOS-sisteem

Hierdie sisteem kan as 'n goeie voorbeeld van die belangrikste sisteme wat tans op die mark beskikbaar is, beskou word. Die sintaktiese ontleding is ontwerp as deel van 'n proses wat volg op 'n woordeboeknaslaanproses. In die woordeboeknaslaanproses word morfologiese kategorieë toegeken aan elke woord in die sin (na 'n segmentering van saamgestelde woorde en interaktiewe werk in die geval van woorde wat nie in die woordeboek voorkom nie).

Konseptueel bestaan die program uit 'n aantal herskryfreëls, wat elk weer uit 'n patroondeel en 'n aksiedeel bestaan. Die aksiedeel herskryf die patroon tipies na 'n meer abstrakte node in die ontledingsboom en annoteer dit ook. Deur van herskryfreël na herskryfreël te beweeg, werk die program deur die sin van onder na bo, van links na regs, totdat die ontleding voltooi is. Aan die genereringskant word sintaktiese konstruksie verkry deur die generering van drie basiese tipes: een vir werkwoordfrases, een vir naamwoordfrases en een vir elke tipe bysin. 'n *Double clause generator* hanteer gevalle van langafstandafhanklikheid. Die generasieproses word uiteraard sterk deur die annotasies en beperkinge wat van die ontledingsprogram geërf is, beïnvloed. Vergelyk verder die LOGOS-dokumentasie van die internasionale hoofkwartier in Dedham, MA, en die Europese hoofkwartier in Frankfurt.

(b) Systran

Benewens die sintaktiese komponent word in die sisteem baie aandag aan vaste en semi-vaste uitdrukkings gegee. Sogenaamde *uniterms* of stamme en semi-vaste uitdrukkings word verskillend behandel. Onder semi-vaste uitdrukkings word verstaan alle eenhede waarvoor daar geen *compositionality of meaning* is nie, of waarvoor verskeie sintaktiese lesings moontlik is.

Die navorsers het 'n uitgebreide beperkte semantiese woordeboek *(limited semantics dictionary)* met hierdie spesifieke probleme en die ooreenstemmende kontrole- en beheersisteem saamgestel.

Die sintaktiese komponent is ook saamgestel met die doel om spesiale sorg te dra vir:

- die probleem van NP-identifikasie in die geval van konkatenasie. Die probleem is groter in die Romaanse tale met hoogs polisemiese voorsetsels; dit vereis dus die toevoeging van 'n semantiese komponent;
- die probleem van diskontinue komponente, veral in die geval van die funksiedualiteit van relatiewe voornaamwoorde;
- homografiese strukture.

Sintaktiese beskrywing word met behulp van 'n grammatika wat volgens tradisionele strukture saamgestel is, gedoen. Met behulp van 'n stel reëls word wettige kombinasies van eenhede wat frases en sinne opmaak, beskryf. Dit is deel van 'n stapsgewyse prosedure wat na die naslaan in die beperkte semantiese woordeboek waarna hierbo verwys is, voorkom. Die proses sluit in:

- 'n sinsafbakeningsprogram;
- 'n identifikasieprogram wat van toepassing is op primêre sintaktiese verwantskappe *(governor/governed and modifier/modified relations)*

- 'n dergelike program wat op uitgebreide sintaktiese verbindings van toepassing is (essensieel koördinering en enumerasie).

Die sisteem is nie finaal nie en 'n permanente komitee is in bevel van optimalisering (met betrekking tot sowel woordeboeke as grammatikas). Vergelyk verder d'Hondt 1983 en verskeie artikels van Pigott.

(c) Werk in die VSSR en Japan

In die VSSR word verskeie masjienvertaalsisteme ontwikkel, onder andere AMPAR (vir Russies-Engels) en FRAP (vir Frans-Russies). Hierdie programme bied geen spesifieke oorspronklikheid in konsep en strukturering nie; dit bestaan uit 'n morfologiese komponent, 'n woordeboek (met insluiting van 'n idioomwoordeboek) en grammatiese ontleding met algoritmes vir die berekening van sintaktiese funksies ensovoorts. Vergelyk verder Marchuk 1985.

Dit lyk asof Fillmore se kasusgrammatika die beste geskik is vir die ontleding van Japannees. Die ontleding is dus op twee tipes inligting gebaseer: oppervlak- en diep-kasusraamwerke *(surface and deep case frames)*. Veral in die geval van ingebedde sinne, saamgestelde naamwoordfrases en *multiple parts of speech* kom baie probleme voor. Vergelyk verder Nagao 1986, Tomita 1987 en Matsumoto and Cacimura 1987.

(d) Eurotra, Susy en Metal

Hierdie sisteme word volledig bespreek in die verslag oor masjienvertaling, naamlik TRANSNET.

4.3.5 Rekenaargesteunde sintaktiese navorsing in Suid-Afrika

Die RGN-vraelysondersoek na rekenaargesteunde taalondersoek in Suid-Afrika het betreklik min inligting na vore gebring. Individue in drie of vier departemente eksperimenteer wel met ontlederś, onder andere die Departemente Afrikaans en Nederlands aan die Universiteit van Stellenbosch, die Departemente Linguistiek en Semitistiek aan UNISA en die RGN.

4.4 SLOT

Vanweë die rekenaar se vermoë om data akkuraat te sorteer en te manipuleer, en om groot hoeveelhede data te hanteer, hou morfologiese en sintaktiese navorsing met behulp van die rekenaar in beginsel groot moontlikhede in, veral op die gebiede van masjienvertaling, kunsmatige intelligensie en suiwer teoretiese navorsing in die taalkunde.

Die oorsig hierbo oor bestaande programme en projekte is baie kursories. Vir meer inligting kan veral Winograd 1983 geraadpleeg word, asook die kongresreferate van die Association for Computational Linguistics en die

tydskrif Journal of Computational Linguistics; enkele addisionele pakkette word ook in hoofstuk 3 vermeld.

Die gebrek aan aktiwiteite op die gebied van rekenaargesteunde morfologiese en sintaktiese analise in Suid-Afrika is waarskynlik toe te skryf aan die plaaslike inligtings- en kennisgaping wat veral op die gebied van die rekenaarlinguistiek blyk te bestaan. Die feit dat navorsing egter reeds by die RGN en sekere universiteite begin is, en in enkele gevalle reeds goed gevorder het, dui daarop dat hierdie baie belangrike navorsingsterrein in die toekoms in Suid-Afrika waarskynlik ook beter bestudeer en benut sal word.

HOOFSTUK 5: REKENAARTOEPASSINGS IN DIE SEMANTIEKNAVORSING

5.1 INLEIDING

5.1.1 Opdrag

Die opdrag was om ondersoek in te stel na rekenaartoepassings in die semantieknavorsing, waaronder die volgende aspekte ondersoek moes word:

- die verbande tussen semantieknavorsing en natuurliketaalprosessering;
- die stand van sake ten opsigte van rekenaartoepassings in die semantiek in die buiteland en hier te lande;
- die relevansie van die gebied vir Suid-Afrika.

Die ondersoek is 'n doenbaarheidstudie, dus is dit nie die bedoeling om 'n volledige beskrywing van die betrokke navorsingsterrein te probeer gee nie. Die terrein ontwikkel buitendien so snel dat 'n volledige oorsig op enige gegewe oomblik nie moontlik is nie, te meer omdat die inligtingsbronne oor die navorsing hoofsaaklik uit artikels en verslae bestaan, veel daarvan ongepubliseerd.

Aan die ander kant word daar in hierdie verslag rekening gehou met die moontlike behoeftes van lesers wat nie semantici is nie. Daarom word sekere basiese definisies en verduidelikings oor die vakgebied ook ingesluit. Omdat rekenaarondersteunde taalnavorsing in Suid-Afrika 'n relatief onbekende terrein is, is die verslag terselfdertyd ook bedoel om 'n inligtingstuk te wees, alhoewel op beperkte wyse.

5.1.2 Werkwyse

Hierdie verslag is hoofsaaklik die resultaat van 'n bronnestudie, aangevul deur vraelysdata en deur gegewens wat uit besprekings hier en in die buiteland verkry is.

5.2. DIE KONTEMPORÊRE SEMANTIEK AS NAVORSINGSTERREIN

5.2.1 Inleiding

Die semantiek is 'n subterrein van die taalwetenskap wat onder filosowe, logici en taalkundiges eeuelank reeds aandag geniet maar wat in die kontemporêre strewe na sistematiese teoriebou en formele formuleringstegnieke lank agterweë gebly het. Die kernredes hiervoor slaan enersyds op die

aard van die ondersoekterrein, en andersyds op die onlangse geskiedenis van die linguistiek as vakgebied.

Die aard van die terrein is omvangryk en verwikkeld. Die semantiekleer se basiese doelwit is om betekenis in taal op woord-, sins- en teks/diskoersvlak te ondersoek, sistematies te beskryf en te verduidelik.

Die ondersoekterrein sluit sowel 'interne' as 'eksterne' verbandlegging in, dit wil sê enersyds betekenisverbande binne die taal (die verband tussen die sintaktiese komponent en ander komponente, byvoorbeeld die prosodie - en die semantiek, die aard en funksionering van sinonimie, polisemie, en so meer), en andersyds die verbande tussen taal en denke, taal en die verwysingswêreld, en taal en kommunikasie, om slegs enkele van die hoofvraagstukke te noem.

In die onlangse geskiedenis van die taalkunde, en hier word spesifiek verwys na ontwikkelings sedert die sestigerjare toe die generatiewe taalkunde 'n oorheersende rol begin speel het, het die semantiekleer 'n tydlank in sekere opsigte 'n terugslag en verskraling ondergaan. Dit is hoofsaaklik te wyte aan die klemplasings van die generatiewe grammatika, veral die toespitsing op taalkennis *(competence)* eerder as op taal-in-gebruik *(performance)*, die atomisering van taalstruktuur (die sg. outonomie van die sintaktiese komponent en die sentrale rol wat dit speel), die aard van die data wat gebruik is (wat feitlik beperk is tot sinsvlakvoorbeelde wat uit die eie taalintuïsies van taalkundiges voortspruit) en die veronagsaming van betekenis, selfs sintaktiese betekenis.

Daarbenewens het die generatiewe taalkunde met sy klemplasing op vlakverskille *(linguistic levels)* 'n onderskeid probeer handhaaf tussen die semantiek en die pragmatiek, weer eens met nadelige gevolge vir die semantiekteorie.

Vervolgens word 'n kort oorsig gegee van onlangse semantiekteorieë en van die opkoms van die rekenaarlinguistiek, 'n rigting waarin die semantiek 'n meer sentrale rol speel as in sommige ander vertakkinge van die linguistiek.

5.2.2 Semantiekteorieë

Die semantieknavorsing van die afgelope dekades kan breedweg in drie hoofrigtings verdeel word: die generatiewe semantiek, die taalhandelingsteorie *(speech act theory)* en die tekslinguistiek. By laasgenoemde is die diskoers- en gespreksteorie (waarvan die rekenaarlinguistiek 'n verbandhoudende ontwikkeling is) ingesluit.

Binne die generatiewe skool is die bekendste navorsing gedoen deur Postal, Jackendoff, Lakoff en McCawley, terwyl Katz en Fodor se werk ook in sommige

opsigte by die generatiewe denkmodel inpas. Benewens die beperkings wat reeds hier bo genoem is, is die betrokke navorsing fragmentaries en abstrak. Partee (Sowa 1984:161) beweer in hierdie verband dat die betrokke teorieë *near vacuous* was, onder andere omdat *(they) give abstract syntax-like structures without simultaneously specifying a logic to operate on them.* Dowty (1981) is ewe krities, veral omdat die teorie nie aan die vereiste voldoen om te spesifiseer *how language connects with the world* nie.

Die Montague-teorie, waarvan Partee en Dowty aanhangers is, is die enigste teorie binne die raamwerk van die generatiewe skool en sy vertakkinge wat volledig geformaliseer is en daarom geskik sou kon wees vir gebruik in die rekenaarlinguistiek. Die hoofkenmerke van die teorie is dat dit model-teoreties is, gebaseer is op waarheidskondisies en, ter aanvulling van die leemtes van laasgenoemde, ook die *possible worlds*-benadering. Die formaliseringstaal is 'n uitgebreide predikaatlogika met 'n een-tot-een korrelasie tussen semantiese en sintaktiese reëls. Die teorie is ingebou in Gazdar se Generalized Phrase Structure Grammar en is op rekenaar geïmplementeer deur Hewlett Packard. Die program word onder andere by Stanford en Cambridge (VK) verder ontwikkel.

Die tweede hoofrigting wat die semantieknavorsing die afgelope dekades ingeslaan het, naamlik die taalhandelingsteorie, waarvan Austin (1962) die pionier was en Strawson en Searle van die bekendste uitbouers, toon 'n sterk nie-geformaliseerde benadering met klemplasing op taal-in-gebruik. In sekere opsigte behels hierdie benadering 'n samesnoering van die semantiek en die pragmatiek.

Vanweë die nie-geformaliseerde aard van die benadering is dit nie direk bruikbaar in rekenaarondersteunde taalnavorsing nie, maar insette daaruit word wel deeglik gebruik (vgl. Grishman 1986 en Sowa 1984). Die grondslag van Austin se bydrae is dat taaluitinge bepaalde kommunikatiewe oogmerke ten doel het op gespesifiseerde vlakke van kommunikasie (te wete inhoudsoordrag, intensie-oordrag en doelwitbereiking). Taaluitinge is doelgerigte aksies/gedrag. Die werk van Austin en sy opvolgers het binne die raamwerk van rekenaarverwante taalnavorsing meer spesifiek waarde vir navorsers wat in kunsmatige intelligensie mens-masjienraakvlakke ondersoek.

Die Situation Semantics-teorie van Barwise (1981) is ook 'n nie-geformaliseerde teorie wat nogtans in die rekenaarlinguistiek en in kunsmatige intelligensie belangstelling wek, want die grondslag daarvan is semantiese voorstellings wat op kennis van die wêreld *(real world knowledge)* berus.

In die derde van die semantieknavorsingsrigtings waaroor dit hier gaan, naamlik die tekslinguistiek (en daarby ingesluit ook diskoersontleding), is aansienlike bydraes gelewer tot die rekenaarverwerking van natuurlike taal op semantiese vlak. Die grondslae van die tekssemantiek is gelê deur bekendes soos Halliday en Hassan (1976), Van Dijk (1973), De Beaugrande en Dressler (1981).

Twee benaderings wat van groot belang vir die rekenaarlinguistiek is, word geïllustreer deur die werk van Schank en Minsky en hulle onderskeie medewerkers en opvolgers. Schank (1975) het vroeg al met sy Conceptual Dependency Theory die problematiek van begripsnetwerke en die toepassingsmoontlikhede daarvan in die rekenaarkonteks raakgesien, terwyl Minsky (1975) met sy Frame Theory, 'n ander benadering, wat meer pragmaties is, ontwikkel het. As gevolg van die invloed van hierdie twee benaderings word daar in Bylae 1 van hierdie hoofstuk 'n uiteensetting van die hoofaspekte van die betrokke navorsing gegee.

5.2.3 Die rekenaarlinguistiek (Computational Linguistics)

Die hoofdoel van die rekenaarlinguistiek is die ontwikkeling van programmatuur wat natuurliketaal-materiaal kan verwerk. Die doel is tweeledig: enersyds om ter wille van taalteoretiese insigte taalformaliseringsmoontlikhede te ondersoek (vgl. Berwick en Weinberg 1982), andersyds om praktiese toepassings te dien, soos rekenaargesteunde vertaling, inligtingherwinning en mens-masjiendialoog.

Rekenaargesteunde vertaling

In die vyftigerjare is daar met groot optimisme 'n aanvang geneem met die ontwikkeling van vertaalprogrammatuur. Die primêre doel was om programmatuur te ontwikkel wat rou vertalings sou voortbring vir die doeleindes van inligtingherwinning (in die konteks van militêre intelligensie). Die hoofdoelwit is in feite bereik, maar in die loop van die volgende dekades is daar tot die besef gekom dat verfynde rekenaargesteunde vertaling 'n onderneming is waarvan die ingewikkeldheid ernstig onderskat is.

By gebrek aan beskikbare rekenaarmatige taalteorieë, *(computational linguistic theories)* is die eerste programme op 'n a-teoretiese, soms uiters *ad hoc*-grondslag gebou. Soos dikwels die geval met *ad hoc*-teorieë van enige aard, het die vertaalprogramme al hoe minder hanteerbaar en effektief geraak namate uitbreidings en wysigings aangebring is. 'n Voorbeeld, hiervan is die SYSTRAN-stelsel by die EEG (die Europese Gemeenskapsmark) wat as gevolg van sy a-teoretiese basis en die toevoegings tot die stelsel, al hoe minder effektief funksioneer. Daarom word daar nou miljoene dollar belê in 'n vervangstelsel, naamlik EUROTRA.

Die EEG-geval is net een voorbeeld van vroeë terugslae wat gelei het tot groei in die akademiese navorsingsbelangstelling in taalteorieë wat geskoei is op natuurliketaal-vertolking binne die geformaliseerde konteks van die rekenaartegnologie.

RAAD VIR GEESTESWETENSKAPLIKE NAVORSING
HUMAN SCIENCES RESEARCH COUNCIL
077715

Inligtingherwinning

Stelsels vir inligtingherwinning uit gerekenariseerde databasisse, byvoorbeeld bibliografiese of terminologiese data, is reeds goed ontwikkel en word steeds verfyn. Die ontsluiting van inligting uit teksmateriaal (tydskrifte, verslae, boeke) is 'n groter uitdaging. Vir meer gevorderde herwinningstegnieke as net woordsoektogte word kennisvoorstellingstelsels *(knowledge representation systems)* benodig. Dit is die hoogste prioriteit in die vertakking van kunsmatige intelligensie wat die ontwikkeling van mens-masjien natuurliketaal-raakvlakke nastreef. Terselfdertyd is die algemene ontoereikendheid van bestaande kennisvoorstellingstelsels 'n hoofknelpunt in kunsmatige-intelligensie-navorsing. Vir semantici is dit dus 'n uitdagende navorsingsterrein. Alvorens taalkundig geldige voorstellingstelsels ontwerp word, sal doeltreffende programmatuur nie ontwikkel kan word nie.

Mens-masjienraakvlakke

Natuurlike taal sou die gerieflikste en maklikste kommunikasietaal wees vir die gebruiker van gerekenariseerde databasisse. Op vele vlakke, vanaf die beperkte dialoog wat 'n bankkliënt met die rekenaar van 'n bank op die sypaadjie voer, tot die gesofistikeerde wetenskaplike navraag, is daar 'n behoefte aan vinnige en direkte inligtingtoevoer.

Die interaktiewe aard van kommunikasie tussen mens en masjien is 'n fasiliteringsfaktor wat by die twee toepassings hierbo nie geld nie: by vraag-en-antwoordstelsels kan die gebruiker sy vraag herformuleer as dit blyk dat die woordeskat of konstruksie van die vraag buite die program se repertoire val. Gevolglik hoef vraag-en-antwoordstelsels nie so 'waterdig' te wees as byvoorbeeld outomatiese vertaalprogramme nie.

Daar word by tientalle navorsings- en nywerheidsinstansies vraag-en-antwoordstelsels ontwikkel. Vanselfsprekend is die mees geslaagde stelsels dié wat taakspesifiek is en slegs 'n klein afgebakende domein dek. 'n Voorbeeld hiervan is die GUS-stelsel waarvan 'n beskrywing gegee word in Bylae 2 van hierdie hoofstuk. Voorbeelde van ander stelsels wat in die literatuur beskryf word, is PLANES (Waltz, 1978), RENDEZVOUS (Codd, 1978) en LADDER (Hendrix *et al* 1978). PLANES hanteer inligting oor vlug- en instandhoudingsrekords van 'n stel vliegtuie; RENDEZVOUS gee toegang tot 'n databasis oor voorraad, onderdele en produkverskaffers; LADDER verskaf inligting in verband met vlootlogistiek.

Soos blyk uit die beskrywing van GUS in Bylae 2, kan sodanige programme betreklik eenvoudig wees omdat daar met 'n beperkte aantal sleutelwoorde en vraag-en-antwoordpatrone volstaan kan word. 'n Gereelde gebruiker van so 'n stelsel, byvoorbeeld 'n sakeman wat RENDEZVOUS gebruik vir voorraadopnames en -toevoeging, raak gou gewoond aan die program se woordeskat en grammatika en weet dus watter tipe vrae deur die stelsel hanteerbaar sal wees.

Vanselfsprekend word daar ook geëksperimenteer met meer gesofistikeerde programme. Die mediese vraag-en-antwoordstelsel LSP (Marsh en Sager 1982), wat 'n aansienlike spektrum navrae kan hanteer, kan onder andere fragment-data na volsinne uitbou en dit vertolk, byvoorbeeld *Nek styf en koors* word *Nek is styf en ly aan koors.*

'n Verdere voorbeeld van die gesofistikeerdheid van die hantering van inligting deur die rekenaar word deur Schank, Kolodner en De Jong (1980) beskryf. Hulle het 'n stelsel ontwerp om 'n belangrike openbare figuur (die voorbeeldgeval was die Amerikaanse politikus Cyrus Vance) se reise en uitsprake op spoedig toeganklike wyse op rekord te plaas. 'n Lees- en opsommingsprogram FRUMP gaan koerant- en ander berigte deur, berg die inligting onder rubrieke soos datums, ontmoetings, reisbestemmings, temas in toesprake, en so meer, en dan gee die vraag-program CYRUS aan gebruikers toegang tot die gebergde inligting.

Aangesien dit in die rekenaarlinguistiek 'n hoofdoelwit is om begrips-prosesse rekenaarmatig na te boots, neem die semantiek 'n ander posisie in in die rekenaarlinguistiek as in die teoretiese linguistiek. In die eerste plek is die semantiek 'n integrale komponent van 'n stelsel wat taal-in-gebruik verwerk. Daar word dus nie teoreties 'n onderskeid getref tussen die semantiek en die pragmatiek nie, maar ook nie tussen woord-, sins- en teksvlak nie.

In die tweede plek is die rekenaarlinguistiek prosesgerig, teenoor die neiging in die teoretiese taalkunde tot stellingsgerigtheid, dit wil sê die neiging om kennis as 'n versameling van feite, afleidings, ensovoorts na te vors, eerder as 'n middel tot 'n doel. Dit is 'n onderskeid wat oorspronklik deur Ryle (1949) gekenmerk is as *knowing that and knowing how* (die sg. *declarative-procedural*-onderskeid). Simon (1969) illustreer die onderskeid met die volgende voorbeelde:

Stellinggerig: *A circle is a locus of all points equidistant from a given point.* Prosesgerig: *To draw a circle rotate a compass with one arm fixed until the other arm has returned to its starting point.*

Prosesgerigtheid vereis 'n ander tipe instelling tot probleemoplossing as wat dikwels by tradisionele taalkundiges aangetref word.

'n Verbandhoudende kenmerk van die rekenaarlinguistiek is die eksperimentele, soms beslis a-teoretiese, aard van party van die navorsing. Sommige navorsers meen dat dit onnodig en selfs onwenslik is om vanuit 'n teoretiese agtergrond en met netjiese teorieë die ontwikkeling van programme aan te pak. Die teoretikus neig om 'n netjiese stel beginsels en 'n ideale, alomtoepasbare stelsel na te streef. Daarteenoor streef meer realistiese navorsers na praktiese stelsels wat meer *ad hoc* van aard en beperk is.

'n Besonder insiggewende debat oor die verhoudings tussen die teoretiese taalkunde en die rekenaarlinguistiek (met sy maat, kunsmatige intelligensie) kom in Modgil en Modgil (1987) voor.

5.3 REKENAARGESTEUNDE SEMANTIEKNAVORSING IN DIE BUITELAND: ENKELE PROGRAMME EN PROJEKTE

In die voorgaande bespreking is daarop gewys dat semantiese/pragmatiese navorsing by allerlei projekte van toegepaste aard, byvoorbeeld masjienvertaling en vraag-en-antwoordstelsels, geïntegreer word. Volgens die literatuur is sodanige toegepaste navorsing tans in die rekenaarlinguistiek oorheersend.

Daar word nogtans ook rekenaargesteunde navorsing gedoen wat op insig in semantiese en pragmatiese verskynsels as sodanig gerig is. Die vaktydskrifte Computers in the Humanities en ICAME Journal is van die meer taalkunde-gerigte publikasies op die gebied van die rekenaarlinguistiek.

Ter illustrasie van rekenaargesteunde navorsing wat meer op die taalkunde as op die rekenaar gerig is, word enkele projekte vervolgens beskryf.

Die BORIS-program (vgl. Dyer 1982) is 'n voorbeeld van rekenaargesteunde verhaalanalise. Die program spoor verhaalverloop op die vlak van oorsaak-en-gevolg op. Lehnert (1982) se program ontleed knoopeenhede *(plot units)* in verhale.

In Stenström (1987) word 'n voorbeeld gegee van navorsing oor gesprekanalise. Die doel met die navorsing was om vas te stel hoe die uitdrukkings *right, right'o, that's right, all right, that's all right,* funksioneer as interaksie-aanmoedigers *(carry-on-signals).* Die bron van die gesprekke wat as data gedien het, is die Survey of English Usage (University College London). Met behulp van die rekenaar is die voorkomsfrekwensie en die verspreiding van die betrokke uitdrukkings bepaal. Daarna is daar programmaties 'n ontleding gedoen van hoe uitdrukkings verskil in funksies soos gespreksfase-indeling en aan- of ontmoediging van beurtneming. In Aijmer (1986) word verslag gedoen van 'n reeks soortgelyke studies oor gespreksontleding.

'n Ander gebied waarop daar interessante semantiese navorsing met behulp van die rekenaar uitgevoer word, is die ondersoek van beeldspraak. Dit is natuurlik veral op die vlak van idiomatiese beeldspraak, byvoorbeeld in uitdrukkings soos *die tyd stap aan, koorsige aktiwiteite,* eerder as op die vlak van literêre beeldspraak wat rekenaaranalise slaag. Carbonell (1982) wys onder andere daarop dat 'n rekenaarprogram wat slegs met die letterlike betekenis van woorde en kollokasiepatrone voorsien is, fyner nie-letterlike gebruike opspoor as die mens. Dit is 'n ontledingstegniek wat byvoorbeeld

waardevolle toepassings sou kon vind in die ontleding van die fenomeen dat kindertaal in die eerste jare oënskynlik beeldspraakvry is.

In Somers (1982) word 'n beskrywing gegee van die PTOSYS-program wat op grond van betekenismerkers *(semantic features)* betekenisvoorstellings vir bepaalde werkwoorde op sinsvlak voortbring. Die kasusteorie wat die program rugsteun, is gebaseer op die werk van Fillmore, Chafe, Cook, Vendler en Dowty.

Devons (1986) skets 'n voorbeeld van die gebruik van groot korpusse vir 'n vergelykende studie. 'n Ontleding van *one's* ('n mens se') in die LOB-korpus van Britse Engels en die Brown-korpus van Amerikaanse Engels toon drie fyn semantiese verskille tussen die twee variante van Engels. Die bevinding is in 'n elisiteringseksperiment ingebou en die resultate daarvan het die korpusstudiebevindinge bevestig. Hierdie projek illustreer die feit dat die rekenaar besonder groot hoeveelhede data kan hanteer en dit presies kan sorteer.

'n Belangrike aspek van bogenoemde tipe navorsing is die beskikbaarheid, of die ontwikkeling, van masjienleesbare korpusse wat enersyds vir taalwetenskaplike navorsing geskik is, en andersyds bruikbaar geformateer en geënkodeer is. Die ontwikkeling van sodanige korpusse is die duurste aspek van rekenaargesteunde taalnavorsing (vgl. hoofstuk 7).

Benewens die finansiële aspekte van korpusbou, moet die mannekragimplikasies daarvan ook in ag geneem word. In beginsel kan 'n korpus meerdoelig aangewend word. Die voorwaarde vir meerdoelige aanwending is egter dat navorsers saamwerk en 'n ooreenkoms bereik oor dataverwerkings- en bergingskonvensies.

5.4 STANDBESKRYWING VIR SUID-AFRIKA

Die vraelysopname onder taalkundiges in Suid-Afrika oor rekenaargesteunde taalnavorsing wat onderneem word, het vir semantieknavorsing niks opgelewer nie.

5.5 SAMEVATTING

Heel kortliks gestel, lig die inhoud van hierdie hoofstuk die volgende hoofpunte uit:

- rekenaargesteunde semantiek- en semantiekverwante navorsing is in die buiteland 'n bedrywige gebied;
- semantiek-/pragmatieknavorsing figureer veral in die konteks van toegepaste projekte, soos die ontwikkeling van rekenaargesteunde

vertaling en van mens-masjien-raakvlakke, byvoorbeeld vraag-en-antwoordstelsels; in hierdie kunsmatige-intelligensie-toepassings van die verwerking van natuurlike taal is die semantiek/pragmatiek die hoof probleemarea; suksesse wat tot dusver behaal is, slaan op heel taakspesifieke, sub-taalspesifieke, en domeinspesifieke toepassings;

- meer basiese navorsing is ook nodig ter rugsteuning van bogenoemde tipes toepassings;

- in Suid-Afrika word die rekenaartegnologie onbeduidend min, dalk glad nie, by semantiek-/pragmatieknavorsing ingespan nie; die redes hiervoor is waarskynlik 'n gebrek aan inligting oor hierdie uitbreidende gebied en die gebrek aan opleiding in die rekenaarlinguistiek.

TEXTNET - SEMANTIEK

CONCEPTUAL DEPENDENCY THEORY, FRAME THEORY, SCRIPTS, PLANS
(uit Grishman 1986)

EXCERPTS FROM GRISHMAN 1986

SCHANK'S CONCEPTUAL DEPENDENCY THEORY

If our ultimate objective is the generation of a semantic representation of the sentence, why don't we do so directly, and use the syntactic constraints to guide this process?

The largest 'school' of computational linguists to adopt this approach was started by Roger Schank. In addition to the approach of analyzing text directly into semantic structures, this school is tied together by a common semantic representation, called Conceptual Dependency ('CD') networks.

Schank's semantic representation embodies an extreme view regarding predicate decomposition. Because his sentence analysis is primarily semantic rather than syntactic, he finds such decomposition (which makes explicit the roles of all the 'participants' in an action or predicate) particularly important. In particular, he wants a semantic representation such that any two synonymous sentences will have the same semantic representation. To achieve this, he has proposed a conceptual dependency network based on a small number of primitive actions (Schank 1975). The primitive actions of conceptual dependency are:

(1) ATRANS - the transfer of an abstract relationship such as possession, ownership, or control.

(2) PTRANS - the transfer of physical location of an object.

(3) PROPEL - the application of physical force to an object.

(4) MOVE - the movement of a bodypart of an animal by that animal.

(5) GRASP - the grasping of an object by an actor.

(6) INGEST - the taking in of an object by an animal.

(7) EXPEL - the expulsion from the body of an animal into the world.

(8) MTRANS - the transfer of mental information between animals or between the conscious processor, long-term memory, and sense organs of an animal.

(9) MBUILD - the construction by an animal of new information from old information.

(10) SPEAK - the action of producing sounds.

(11) ATTEND - the action of attending or focusing a sense organ towards a stimulus.

In addition, the graphical representation of conceptual dependency networks provides notation for causation, possession, containment, and a few other relationships (adjectives and nouns are not decomposed to primitives, so the objective of a canonical semantic representation is clearly not entirely achieved). As an example of his decomposition (without using his graphical representation),

> John hit Mary,

would become

> John's propelling a physical object from John to Mary caused a state of physical contact between that object and Mary.

Although the idea of decomposing into elementary actions is appealing, it does have several problems. First, it is not always evident what the proper decomposition is for a verb, whether the decomposition is unique, or whether the given 'primitive actions' are the appropriate ones for a semantic decomposition. Schank points out (Schank and Riesbeck 1981, p. 26) that the conceptual dependency representation was intended for stories about simple physical events and human interaction and may be inadequate for other domains. Both Schank and others have augmented the set of primitive acts in order to analyze richer texts. For complex technical vocabularies, very different decompositions may be needed.

The second objection is a strategic one. Even if such a reduction is possible, it is not clear that it should always be done. Decomposition should be done to the degree needed for inferencing; unnecessary decomposition produces unnecessarily large structures for subsequent processing. Just how this can be done will become clearer as the role and means of inferencing in natural language is better understood.

Using Conceptual Dependency Theory Schank and his school have developed various versions of conceptual analyzers. What these conceptual analyzers share is the use of skeletal semantic structures to guide the analysis. These skeletons are incomplete CD networks - networks which specify the primitive actions and the type of objects (people, foods, etc.) involved, and have places for filling in the specific objects involved in a particular instance. These skeletons are comparable to the selectional patterns we studied in the previous section (although one such skeleton may subsume several selectional patterns). Roughly speaking, then, we can characterize these analyzers as being guided by semantic (selectional) patterns and then applying (limited) syntactic checks, whereas most parsers are guided by syntactic patterns and then apply semantic checks. For this reason, we have included our presentation of conceptual analyzers here following the discussion of semantic constraints.

Consider for example the sentence 'John eats berries.'. The verb 'eat' corresponds to the primitive act INGEST. The skeleton for INGEST has two slots: one for an ACTOR (an animate being) and one for an OBJECT (an edible thing). A conceptual analyzer would find the word 'eat', create an INGEST skeleton, and then look for items in the sentence to fill the slots. It would find 'John', which satisfies the requirements for an ACTOR, and

'berries', which satisfies the requirements for an OBJECT of INGEST. Once these slots are filled in, the skeleton becomes a complete semantic representation.

Because it relies primarily on semantic factors, this approach has no trouble with syntactically ill-formed sentences, such as 'John eat berries.' or 'Berries John eats.'. However, the analyzer must sometimes make use of word order in order to produce the correct semantic analysis. For instance, 'give' is represented in CD by the primitive act ATRANS (transfer possession). ATRANS has four slots: the ACTOR (a person), the OBJECT (which is transferred), the person it is transferred FROM and the person it is transferred TO. The definition of 'gives' includes a skeletal A TRANS structure; it specifies that the same entity will fill the ACTOR and FROM slots; and it specifies that this entity normally precedes the verb, whereas the TO and OBJECT entities normally follow the verb. This word order information enables the analyzer to figure out who A TRANSed what to whom in a sentence such as 'Jack gave Jill some berries.'.

FRAME THEORY

Much of the recent work on organizing knowledge in order to process new data is based on the following observation: people generally assimilate new information by identifying it as an instance of a pattern with which they are already familiar. For example, when people see a room for the first time, they try to classify it as a bedroom, kitchen, dining room, etc. They may do this by a partial pattern match (for example, guessing that a room is a kitchen by peeking in and seeing an oven). The remainder of the pattern, and information associated with the pattern, can be used to generate inferences and expectations (for example, that the room probably also contains a sink and a refrigerator). Even if a situation or object doesn't match any pattern exactly, it will be remembered as an instance of a pattern with some exceptions (for example, a kitchen without a refrigerator). If we take this observation as guidance for the mechanical processing of discourse, we should make this process of classifying new information in terms of known patterns central to the analysis of texts. Accordingly, the groups of facts into which our knowledge is divided should be organized to facilitate and make use of such pattern matching.

The best known paper on this theme is Minsky's 'A framework for representing knowledge' (Minsky 1975). Minsky called these patterns frames, and systems using this approach have therefore become known as frame-based systems. Minsky described a general strategy for knowledge organization, and did not offer details of a knowledge representation formalism or how it might be applied to language analysis. However, since this paper was written, quite a few frame-based knowledge representation languages have been developed (such as FRL (Goldstein and Roberts 1977) and KRL (Bobrow and Winograd 1977), so it is possible for us now to describe a 'typical' frames formalism.

The basic 'chunk' of knowledge is the frame. There are prototype frames, which describe classes of objects or situations (for example, kitchens, buying groceries), and instance frames, which describe individual objects or situations (for example, Escoffier's kitchen, Sam Spade's trip to the supermarket last Friday). Each instance frame must be an instance (instantiation) of some prototype frame. In addition to this two-level class-instance hierarchy, there will usually be a multi-level 'generalization hierarchy' in which some prototypes (such as 'kitchen') are viewed as instances of a more general prototype (such as 'room').

Each frame contains a set of labeled slots. These specify the properties, constituents, and/or participants in the object or situation represented by the frame. The value of a slot may be an integer, a character string, or another frame (thus, a frame for an individual shopping trip may have a date slot, whose value is a character string, and a shopper slot, whose value is a frame representing the shopper - an instance of the person frame). The slots of a frame need not all be filled with values; this may be the case, for example, if a particular property is unknown or if an individual doesn't possess a particular constituent.

The frame slot structures form an alternative semantic representation; each prototype frame is in effect a predicate, with each of its slots an argument position. An input text may first be translated into logical forms based on more traditional predicates, closely corresponding to the vocabulary of the text. These logical forms can then be mapped into frames, with each argument of a traditional predicate typically filling one frame slot. The complexity of this mapping depends greatly on the richness of the set of frames, and hence on the semantic richness of the texts being processed. In the simple domains characterstic of today's systems, one can determine from each individual predicate, or possibly from the predicate with its immediate context within the sentence, which frame it fits into. In more complex domains, a predicate may correspond to slots in several different frames, so pattern matching over a larger context will be required to select the appropriate frame.

Texts may be classified both according to the subject domain involved and according to the type of discourse, such as narrative, argument, or instruction. As we have noted above, a language analyzer must have a model of the objects and relationships in the subject domain. In addition, the analyzer must understand the organizing principles for the particular type of discourse. These organizing principles enable us to understand, for example, why an instruction manual 'What to do when an earthquake strikes' is organized quite differently from a narrative 'How I survived the great 1989 earthquake'.

Narrative has been the most intensively studied form of discourse. Narrative is an appealing object of study because its primary organizing principle at least for simple stories - is straightforward: events are related in the order in which they occur. The task of understanding a narrative is therefore largely one of reconstructing a sequence of events and their

interrelationship from 'highlights' which appear explicitly in the narrative.

SCRIPTS

Schank and his co-workers at Yale have been at the forefront of the effort to understand narrative. Their studies have led them to define large semantic patterns similar to the frames we have been considering. Their first effort at defining such chunks of knowledge was the script (Schank and Abelson 1975, 1977). The script is intended to capture a person's knowledge about a stereotyped sequence of events.

Their 'classic' example is the restaurant script. From many visits to many kinds of restaurants, people build up detailed expectations - a script - of what will happen in a restaurant. The usual sequence of events includes (among other things) entering, being seated, ordering, receiving one's food, eating, receiving the check, paying, and leaving. Because people share such a detailed script, they can assemble coherent descriptions of such experiences from fragmentary information. In particular, the script contains actors and objects (the food, the check, etc.) which can be referred to in a text without having been previously introduced. The restaurant script would thus enable us to solve the reference problem for 'the check' in the paragraph at the beginning of our discussion of anaphora resolution. It would also allow us to figure out who 'he' is in

> The customer carefully gave his order to the waiter.
> Thirty minutes later he returned with the wrong entrée.

(since the script indicates that the waiter brings the food). Because the script supplies default information which is assumed in the absence of explicit information, we could also infer from a sentence. 'The food arrived cold,' that the food was probably brought by the waiter.

There are different kinds of restaurants, ranging from haute cuisine to fastfood, and our expectations of what will happen are different for different restaurants. In a fancy restaurant, we wait for the maître d' to seat us; in a fastfood place, we go up to a counter and order. This variety is reflected in different tracks in the script. Each track is a sequence of scenes, such as entering, ordering, eating, and leaving. Each scene is described in turn as a sequence of primitive actions, using conceptual dependency notation (section 3.1.6). The actions in the script involve a set of roles - the people in the script, such as the customer and waiter - and a set of props - the objects acted on, such as the menu and the food. We show in figure 4.1 the coffee shop track of Schank's restaurant script. Note that even within this single track there are several alternate paths, such as one for the situation where the food ordered by the customer is not available.

Scripts were used by a program called SAM (Script Applier Mechanism) for analyzing simple paragraphs of narrative text (Schank and Abelson 1975, 1977; Cullingford 1981). To determine when a script is applicable, various headers are associated with a script. For the restaurant script these include precondition headers ('the customer is hungry') and locale headers (something happened 'at a restaurant'). If a text matches one of the headers of a script and in addition mentions some action within the script, the script will be activated. SAM then fills in the script with information from the text. Using the script, it makes the inferences necessary to complete a causal chain connecting the events in the text.

PLANS

Scripts are only effective for analyzing texts which describe stereotyped situations. To analyze descriptions of novel sequences of events, Shank and Abelson proposed another type of knowledge, plans (Schank and Abelson 1977). To construct a causal chain for a novel sequence, a means - ends analysis must be performed.; that is, we must try to understand how later events in a text act to further previously stated goals. Often the connection will be indirect - an action will further an implicit subgoal which is a precondition of some previously mentioned goal. A plan consists of a goal, alternative sequence of actions for achieving that goal, and preconditions for applying the various sequences.

As an example of a direct connection involving a single plan, consider

> John wanted Bill's bicycle. He walked over to Bill and asked him if he would give it to him.

(from Wilensky 1978). The asking in the second sentence can be identified as part of an 'asking plan' corresponding to the goal of getting something. This goal can be identified with the goal of John wanting to get Bill's bicycle, as mentioned in the first sentence. A more complex chain is required to analyze

> Willa was hungry. She grabbed the Michelin guide. She got in her car,

(from Wilensky 1981, p.182). To understand the connections between these sentences, we must recognize that Willa can satisfy her hunger by dining at a restaurant (the 'eating-at-restaurant plan'), that to diner there she must find out where the restaurant is (which she can do using the Michelin guide) and must get to the restaurant (which she can do using her car).

TEXTNET - SEMANTIEK

GUS: A QUESTION AND ANSWER SYSTEM
(uit Grishman 1986)

EXAMPLE OF A QUESTION AND ANSWER SYSTEM: GUS

GUS (the 'Genial Understanding System') was developed at Xerox Palo Alto. GUS acted as a travel agent who was only able to arrange for plane trips to the west coast (Bobrow et al. 1977). A portion of a typical dialog handled by GUS follows:

GUS:	Hello. My name is GUS. I can help you plan a simple trip by air. Where do you want to go?
CLIENT:	I want to go to San Diego on May 28.
GUS:	What time do you want to leave?
CLIENT:	I must be in San Diego before 10 a.m.
GUS:	Would you like PSA flight 102 that arrives at 9.15 a.m.
CLIENT:	When is the next flight?
GUS:	Air California flight 310 that leaves at 8:30 a.m. and arrives at 10:00 a.m.
CLIENT:	I'll take the first one.

As this example shows, GUS is a mixed initiative system - either GUS or the 'client' may control the dialog. GUS will start by asking questions, but will allow the user to respond with a question of his own instead of a direct answer.

The information gathered in the dialog is stored in a set of frames. The frame structure is quite simple; the chief frame is a trip specification, which has the slots shown in table 1. Two of the slots are to be filled with instances of trip leg. another frame (table 2).

GUS is a frame-driven system: its 'goal' is to complete a trip specification frame. When GUS has the initiative, it looks for the first unfilled slot and executes the TO-FILL procedure associated with that slot. For many slots this procedure is Ask-client, which generates a question pertinent to that slot. GUS tries to parse the client's response as either a direct answer to the question or a full-sentence assertion or question (an ATN parser is used). If it is a direct answer or assertion, the output of the parse is mapped into an intermediate semantic representation and then into a set of operations for entering data into frame slots. If it is a question, GUS tries to answer the question. In the sample dialog given above, GUS begins by creating an instance of trip specification and then an instance of trip leg to fill the OUTWARDLEG slot. In trying to fill in trip leg, it encounters an Ask-client procedure attached to TOPLACE and generates the question 'Where do you want to go?' The response, 'I want to go to San Diego on May 28.', causes the TOPLACE and TRAVELDATE slots to be filled. Continuing in this way, GUS eventually fills all the slots. Although the frame structure here is obviously very simple, it does indicate how the frames serve to integrate the information from separate sentences and guide the dialog.

A simple mixed initiative system like GUS implicitly recognizes the existence of more than one goal in the conversation. GUS has its permanent goal of filling all the slots in the TRIP frame and its descendents. The user may express his own goal (for getting a particular piece of information) by asking a direct question. This goal takes precedence temporarily over GUS's permanent goal.

Table 1

Slot	Type of value	Procedures
HOMEPORT	city (default: Palo Alto)	
FOREIGNPORT	city	
OUTWARDLEG	trip leg	TO-FILL: create frame
AWAYSTAY	place stay	
INWARDLEG	trip leg	TO-FILL: create frame

Table 2

Slot	Type of value	Procedures
FROMPLACE	city	TO-FILL: find from HOMEPORT
TOPLACE	city	TO FILL: Ask-client
TRAVELDATE	date	TO-FILL: Ask-client
DEPARTURESPEC	timerange	WHEN-FILLED: propose flight by departure
ARRIVALSPEC	timerange	WHEN-filled: propose flight by arrival
PROPOSEDFLIGHTS	(set of flight)	
FLIGHTCHOSEN	flight	TO-FILL: Ask client
TRAVELER	person	TO-FILL: Ask client

HOOFSTUK 6: DIE GEBRUIK VAN DIE REKENAAR IN DIE VARIASIETAALKUNDE

6.1. INLEIDING

Die rekenaar word in die variasietaalkunde gebruik om die volgende take te verrig:

(a) Die konstruksie van datakorpusse en databasisse.

(b) Die analise van variasiedata met die doel om:

(i) die effek van die verskillende kondisies op die keuse tussen twee of meer variante vas te stel,

(ii) hipoteses oor die grammatika van linguistiese veranderlikes te toets,

(iii) die lede van 'n spraakgemeenskap volgens hul keuse van variante te groepeer en data-items in deelversamelings in te deel volgens die gedifferensieerde effek van die verskillende kondisionerende faktore, en

(iv) die teenwoordigheid van implikasionele skale in die data vas te stel.

(c) Die berekening van die waarskynlikheid dat die reëls van 'n grammatika in gegewe kontekste toegepas sal word (d.w.s. die konstruksie van probabilistiese grammatikas).

(d) Die ondersoek van verskynsels soos kodewisseling, akoesties-ontlede vokaalsisteme, denotasie en taalverwerwing.

6.2 DIE KONSTRUKSIE VAN DATAKORPUSSE EN DATABASISSE

Hoewel variasietaalkundiges hul databasisse dikwels regstreeks van bandopnames met die hand opstel, word die rekenaar ook daarvoor gebruik, soos in Poplack (ms.) bespreek (vgl. ook hoofstuk 7).

Poplack noem die volgende aspekte van die rekenarisering van data:

(a) Die datakorpus moet uiteraard voldoen aan die behoeftes van variasietaalkundige ondersoek ten opsigte van sosiale, etnografiese en stilistiese verteenwoordigheid, en omvang. (Die korpus waaroor sy rapporteer, verteenwoordig 270 uur se spraak, dit wil sê om en by drie-en-'n-halfmiljoen woorde.)

(b) Die korpus moet toeganklik wees vir die sistematiese elisitering van 'n databasis wat beantwoord aan die eise van toerekenbaarheid.

(c) Die transkripsie van die korpus vind plaas tydens die rekenarisering daarvan, en geskied in die tradisionele praktiese ortografie - dus nie in fonetiese skrif nie. Die redes hiervoor is dat laasgenoemde buitengewoon tydrowend is, dat verskillende navorsers noodwendig verskillende eise stel en dat rekenaarsoektogte gewoonlik geskied met verwysing na alfabetiese karakters. (Die insleuteling van die Poplack-korpusse het geskied teen 'n gemiddelde spoed van 'n halfuur se spraak per dag.)

(d) Na die insleuteling van die korpusse vind minstens vier stelle korreksies plaas. Die rede vir die intensiewe korrigering volg direk op die eis wat die variasietaalkunde ten opsigte van presiesheid stel. (Alle herhalings, gevalle waarin sprekers hulself in die rede val of 'n sin oorbegin, huiwerings, laggeluide en ander geluide word noukeurig in die transkripsie opgeneem.) Die korrigering van die datakorpusse neem gemiddeld tot 50 uur per korpuslêer.

(e) Nadat lêers ontfout is, word die korpusse met die hand geïndeksseer vir verskynsels wat gewoonlik vir variasietaalkundige ondersoeke van belang is, soos kodewisseling, leenwoorde en sprekeridentiteit (ouderdom, geslag, ens.).

(f) Geoutomatiseerde datahanteringsprogramme word hierna gebruik om inligting uit die datakorpus te onttrek (d.w.s. databasisse word opgebou). Die program wat die Poplack-groep gebruik het, is die Oxford Concordance Programme. Hoewel ons hierdie programpakket nie ter insae beskikbaar gehad het nie, werk dit na alle waarskynlikheid volgens die beginsel dat 'n woord of lys woorde as invoer verskaf word, dat die datakorpus daarna vir voorkomste van hierdie woord(e) deursoek word en dat die woorde daarna uitgedruk word binne die konteks van die sinne waarin hulle voorkom. Afgesien van hierdie soekfunksie kan die pakket ook woordfrekwensielyste saamstel, woordeskatstatistiek verskaf, konkordansies op verskillende maniere orden (alfabeties, toenemende/afnemende frekwensie, ens.) en kan dit ook die adresse van onttrekte konkordansie-items verskaf (vgl. ook hoofstuk 7).

6.3 DIE ONDERSOEK VAN LINGUISTIESE VERANDERLIKES

6.3.1 Inleiding

Die onmiddellike doel van die variasietaalkunde is om die aard en omvang van grammatiese diversiteit in 'n spraakgemeenskap te ondersoek en vas te stel hoe hierdie diversiteit grammatikale faktore asook die sosiale stratifikasie, die geografiese en situasionele differensiasie, taalverander-

ing, die kommunikatiewe funksie van bepaalde taalvorme en die dinamika van interpersoonlike interaksie reflekteer.

Die basiese analitiese en deskriptiewe eenheid van die variasietaalkunde is die *linguistiese veranderlike.*

Linguistiese veranderlikes behels die voorkoms van meer as een realisasievorm vir 'n abstrakte onderliggende element, byvoorbeeld die uitspraak van die Afrikaanse foneem /r/ as [r], [R] of 0 (/r/ is die veranderlike, en [r], [R] en 0 word 'variante' genoem) of die plasing van die onderwerp, die werkwoord en die voorwerp van 'n afhanklike stelsin in die volgorde SOV of SVO.

Die begrip *linguistiese veranderlike* verg 'n modelteoretiese opmerking. In die non-generatiewe strukturalisme word daar gewerk met die aanname dat die verhouding tussen die elemente in 'n sisteem van drie tipes is, naamlik opposisie, aanvullende spreiding en vrye variasie. Onlangs is egter aangetoon dat daar rekening gehou moet word met 'n vierde soort verhouding, naamlik *amperse komplementariteit (near* of *weak complementarity),* wat tussen aanvullende spreiding en vrye variasie lê in die sin dat die elemente wat daarby betrokke is nóg volledig aanvullend is ten opsigte van hul spreiding, nóg onderling volkome vryelik vervangbaar is. Die keuse tussen die elemente wat by hierdie verhouding betrokke is, word inderdaad gekondisioneer deur linguistiese (en nie-linguistiese) kontekste, maar dis 'n kwantitatiewe kondisionering. Die grammatikale ondersoek daarvan moet uiteraard dus kwantitatief van aard wees.

Die sentrale taak van die variasietaalkunde is om die sisteem/patroon onderliggend aan die voorkoms van die variante van 'n veranderlike te ondersoek. Anders gestel: die variasionis wil die heterogeniteit van variasiedata verklaar; hy soek invariansie, hy *raak ontslae* van die variasie.

Hierdie taak word verrig deur die effek van alle tersaaklike faktore (eienskappe van linguistiese en non-linguistiese kontekste) op die keuse tussen die variante van 'n veranderlike te bepaal. Indien daar vasgestel word dat sommige van hierdie faktore as kondisies op die voorkoms van die variante opereer, word aanvaar dat die reëlmaat, die sisteem onderliggend aan die variasie, ontdek is.

In die beginjare van die moderne variasietaalkunde (om en by 1960) is gemeen dat variasiedata netjies en patroonmatig ingedeel kan word. Die kwantitatiewe analise van data oor die inbedding van linguistiese veranderlikes/grammatikale reëls in die linguistiese en die sosiale orde is per hand onderneem, en variasioniste was tevrede met die vertoning van hul bevindinge in tabelle en grafieke.

Variasieondersoeke het egter gou aangetoon dat variasiedata tipieserwys baie oneweredig versprei is: die inhoud van selle (kontekste) mag wissel van baie groot tot baie klein terwyl baie selle selfs heeltemal leeg mag wees. Daarby is vasgestel dat die effek van kondisionerende faktore nie konstant is in alle kontekste nie. Gevolglik is dit onmoontlik om hul effekte deur gewone waarneming vas te stel. Groot hoeveelhede data, geënkodeer vir alle moontlike kondisionerende faktore, wat ontleed word met behulp van gesofistikeerde statistiese prosedures (soos die *maximum likelihood*-metodes en veelvuldige regressieanalise), is nodig vir 'n verantwoordbare ondersoek. Hiervoor is 'n rekenaar onontbeerlik.

'n Probabilistiese model word gebruik om die effek van hierdie faktore vas te stel. Hoewel daar sekerlik gerekenariseerde statistiese pakkette bestaan wat die nodige analises kan uitvoer, is 'n spesiale pakket programme daarvoor ontwikkel - hoofsaaklik omdat die data, soos hierbo genoem, in die geval van variasieondersoeke tipies oneweredig versprei is, 'n feit wat die effektiewe toepassing van die gewone statistiese pakkette bemoeilik.

Die betrokke pakket word VARBRUL genoem. Dit is ontwikkel deur variasioniste, statistici en programmeerders aan die Universiteite van Pennsylvania en Montréal. Twee weergawes daarvan bestaan, naamlik VARBRUL 2 en 3. VARBRUL 2 is aangepas vir gebruik op hoofraam, op 'n minirekenaar en op persoonlike rekenaar (genoem VARBRUL 2S). VARBRUL 3 is geskryf vir die hoofraam, maar word nog nie algemeen gebruik nie.

6.3.2 VARBRUL 2S

Die sentrale funksie van die VARBRUL-pakket is om op volkome outomatiese wyse op basis van data oor die frekwensie van 'n variant in die verskillende linguistiese en nie-linguistiese kontekste waarin die veranderlike voorkom, 'n waarde te bereken vir die effek van elke element in die linguistiese en nie-linguistiese omgewing op die keuse van die betrokke variant. (In reëlterme gestel: VARBRUL bereken die bydrae wat elke kondisie in die omgewingsgedeelte van 'n opsionele kontekssensitiewe reël op die waarskynlikheid van reëltoepassing het.) Vir 'n uiteensetting van die statistiese en wiskundige aspekte van die model waarop VARBRUL berus, vergelyk Sankoff 1978.

Dokumentasie oor VARBRUL 2S (met 'n uiteensetting van hoe die pakket werk, wat die formaat van die verskillende lêers moet wees, ens.) is op aparte disket saam met die programdisket gratis beskikbaar.

In breë terme geskied analises met behulp van VARBRUL 2S soos volg:

(a) 'n Datalêer word saamgestel. Hierdie datalêer bevat al die gevalle van die veranderlike wat bestudeer word, en wat in 'n datakorpus voorkom. Elke data-item in die lêer is beskryf ten opsigte van die linguistiese en nie-linguistiese faktore wat die ondersoeker meen 'n effek op die voorkoms van die veranderlike se variante mag hê. Laasgenoemde inligting is in gekodeerde vorm weergegee. VARBRUL 2S vereis dat die datalêer met 'n redigeerder geskep word (soos IBM se EDLIN, of PCWRITE) wat geen kontrolekarakters in die lêer inskryf nie. Die teks van die lêer moet 'n heel spesifieke formaat hê.

(b) Hierna word die datalêer gekontroleer vir foute en voorberei vir analise met die veranderlikereëlprogramme met behulp van 'n verskeidenheid van programme. Hierdie kontroleproses en voorbereidingsproses geskied met behulp van 'n aantal lêers wat met 'n gewone teksredigeerder geskep is.

(c) Laastens word die data ontleed met IVARB vir binomiale analise, TVARB vir veranderlikes met drie variante en MVARB vir multinomiale analise (vier of vyf variante).

Afgesien daarvan dat VARBRUL faktoreffekte bepaal, bepaal dit ook die beduidendheid van elke faktoreffek en van faktorgroepe as geheel en kan dit die geldigheid van 'n bepaalde model van die data vasstel. (Soos hieronder sal blyk, kan dit ook gebruik word om voorstelle in verband met reëlordening te evalueer.)

As deel van die veranderlikereëlanalise laat VARBRUL 2S mens natuurlik ook toe om datalêers vryelik te manipuleer (verdeel, kombineer, reduseer, uitbrei, ens.), en kan dit ook die volgende funksies verrig:

(i) Dit kan inligting verskaf oor die voorkomsfrekwensies van die variante in verskillende kontekste in tweedimensionele tabelle.

(ii) Dit kan data-items uit 'n datalêer haal en in 'n afsonderlike lêer plaas vir noukeuriger ondersoek.

(iii) Dit kan data-items sorteer (en in 'n aparte lêer plaas) volgens die voorkoms van die data-items in 'n spesifieke linguistiese konteks.

Die programme wat VARBRUL 2S bevat, is die volgende:

(a) CHECKTOK, wat die datalêer kontroleer vir foute.

(b) READTOK, wat 'n gekontroleerde datalêer oorskryf in 'n tweede datalêer met 'n formaat waarop verdere programme kan inwerk.

(c) MAKECELL, wat READTOK se afvoer weer herskep tot 'n lêer waarin die data in 'n formaat geberg word waarin dit verwerkbaar is deur 'n program wat die gespesifiseerde faktore se effekte op die veranderlikeproses wat ondersoek word, kan bereken.

(d) IVARB, TVARB en MVARB, wat die effek van die faktore wat die ondersoeker meen 'n uitwerking op die proses wat bestudeer word, mag hê, bepaal.

Afgesien van bogenoemde vier programme bevat VARBRUL ook etlike ander programme wat ander funksies verrig:

(e) COUNTUP, wat die voorkoms van bepaalde verskynsels bepaal en persentasies bereken.

(f) CROSSTAB, wat tweedimensionele tabelle van die data opstel.

(g) TSORT en TEXTSORT, wat sorteerfunksies verrig.

VARBRUL 2S word tans in die VSA, Kanada, Suid-Amerika, die Verenigde Koninkryk, Ierland, Iran, Duitsland, Nederland, Australië, Nieu-Seeland en Suid-Afrika vir die ondersoek van variasie gebruik.

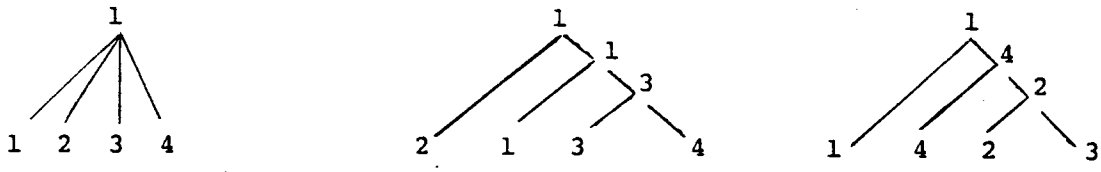
Vir voorbeelde van die gebruik van VARBRUL in die analise van linguistiese veranderlikes vergelyk Webb 1984 en 1987.

6.3.3 Hipotesetoetsing in die variasietaalkunde

Variasiebeskrywings wat binne die raamwerk van die generatiewe grammatika geskied, geskied uiteraard in die vorm van onderliggende weergawes en afgeleide weergawes met reëls wat laasgenoemde van eersgenoemde aflei. Vir sover hierdie afleidings volkome eksplisiet gestel kan word, is rekenaarprogramme ontwikkel wat die geldigheid van die voorgestelde modelle kan toets. 'n Voorbeeld van so 'n program is die RX-pakket van die Amerikaanse linguis W. Labov, wat die geldigheid van onderliggende weergawes, fonologiese reëls, die ordening van fonologiese reëls en die voorspellende krag van die model kan toets. Hierdie program is egter nog nie in die RSA beskikbaar nie.

VARBRUL is ook aangepas om hipoteses in sake reëlordening te toets. Hierdie pakket, wat vroeër slegs veranderlikes met twee variante kon hanteer (binomiale analise), is nou ook daartoe in staat om gevalle van veelvuldige variante te ontleed (die trinomiale of selfs multinomiale analise). In dergelike gevalle moet gevra word wat die derivasionele verhouding tussen die verskillende variante is: word almal afgelei van een onderliggende weergawe, of word een (of meer) variante afgelei van een onderliggende vorm en die res van die variante dan afgelei van die variant(e) wat vroeër afgelei is?

Die vraag oor die onderliggende vorm kan natuurlik nie statisties bepaal word nie. Dit is 'n suiwer linguistiese kwessie. Maar nadat daar besluit is wat die onderliggende vorm in 'n bepaalde geval is, kan die rekenaar gebruik word om die hiërargiese relasies tussen die variante te bepaal. Verskillende hipoteses kan naamlik opgestel word oor hul samehang. As voorbeeld:



Waar 1 die onderliggende weergawe asook een van die variante is en 2, 3 en 4 verdere variante verteenwoordig.

VARBRUL bereken p-waardes vir die faktore binne elk van hierdie konfigurasies, en stel dan vas watter van die samehange die data die beste pas. Die aangeduide samehang bepaal dan die reëlordening wat die data die beste beskryf.

Die reëlordeskemas wat by elk van bogenoemde bome pas, is die volgende:

	1	1 → 2	1 → 4
	2	1 → 3	4 → 2
1 →	3	3 → 4	2 → 3
	4		

6.3.4 <u>VARBRUL 3</u>

Kort na die ontwikkeling van VARBRUL 2 het Sankoff en Rousseau (vgl. Sankoff 1978) begin werk aan VARBRUL 3, wat etlike leemtes van VARBRUL 2 moes aanvul. Van die belangrikste eienskappe van VARBRUL 3 is die vermoë om die volgende funksies te verrig.

Die partisieprobleem

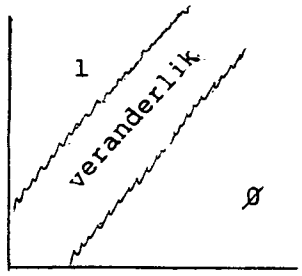
Dit kan gebeur dat VARBRUL 'n model voorstel wat die data nie goed pas nie, en dat die rede vir die swak passing geleë is in die feit dat die sprekers linguisties verskillend optree. Die oplossing vir hierdie probleem lê uiteraard in die ontdekking van watter sprekers linguisties eenders optree.

VARBRUL 3 het die vermoë om sprekers wat hul linguisties eenders gedra, op te spoor. (Dit kan kondisionerende faktore wat eenderse effekte op reëltoepassing het eweneens as deelversamelings identifiseer.)

Implikasionele skale

Linguiste soos die Amerikaner Bailey en die Australiër Bickerton het aangetoon dat linguistiese verskeidenheid soms slegs verstaan kan word in terme van *implikasionele skale*. Hiermee word bedoel dat sprekers soms georden kan word volgens hul gebruik van opsionele reëls en/of die variante van 'n veranderlike op so 'n wyse dat die gebruik van 'n bepaalde reël/variant deur 'n spreker impliseer dat hy alle reëls/variante links daarvan ook sal gebruik, of dat alle sprekers bokant hom in die ry ook die betrokke reël/variant sal gebruik. Met ander woorde daar kan 'n hiërargie in die struktuur van 'n sprekerskorps wees en die sprekers kan dan georden word tot 'n twee-dimensionele skikking waarin elke ry 'n spreker voorstel en elke kolom 'n ander reël/variant. Dergelike implikasionele skale kan ook opgestel word vir sprekers/veranderlikes en vir sprekers/variëteite, ensovoorts.

Voor VARBRUL 3 moes implikasionele skale per hand opgestel word, maar toe is besef dat implikasionele skale in beginsel die volgende struktuur mag hê:



Waar '1' = kategoriale reëltoepassing
en '∅' = kategoriale nie-toepassing van die reël

Die gevolg hiervan was dat VARBRUL daarvoor aangepas is om implikasionele skale outomaties op te stel. (Vgl. hiervoor Rousseau 1983 en Sankoff 1985.)

Algemene opmerkings oor VARBRUL 3

Afgesien van die eienskappe van VARBRUL 3 wat hierbo genoem is, het die jongste weergawe daarvan ook nog die volgende voordele bo die 1978 programme:

(a) Dit is beter gestruktureer, gebruik minder rekenaargeheue en minder rekenaartyd.

(b) Dit kan baie meer faktore en omgewings hanteer en kan dus groot datastelle ontleed.

(c) Dit kan faktore wat reëltoepassing kategoriaal vereis of verbied (die *knock-out factors)* hanteer. Dergelike faktore word geïdentifiseer en eenkant geplaas terwyl die program voortgaan om die oorblywende faktore se effek te bereken. Wanneer die verskillende faktore se p-waardes gedruk word, word hierdie faktore se waardes as 1 of as 0 aangegee.

6.3.5 Onopgeloste probleme

Op hierdie stadium is daar nog twee probleme waarvoor oplossings gesoek word, naamlik die hantering van kontinue veranderlikes (bv. ouderdom) en die hantering van rare veranderlikes.

6.4 PROBABILISTIESE GRAMMATIKAS

Soos bekend kan die basisreëls van 'n generatiewe grammatika 'n onbeperkte aantal sinne vir 'n bepaalde taal genereer. Dit beteken dat die grammatika sinne mag genereer wat selde of ooit in werklike linguistiese gedrag aangetref sal word. 'n Voorbeeld hiervan kry mens in veelvuldige sinsinbedding. 'n Generatiewe grammatika onderskei dus nie tussen moontlike sinne en onwaarskynlike sinne nie, en verskaf dus nie 'n *gedragsmodel* van 'n taal se grammatika nie.

Nou is dit moontlik om op basis van 'n ondersoek van datakorpusse inligting in te samel oor die voorkomsfrekwensie van die verskillende sinstipes/die toepassing van grammatikareëls, en om waarskynlikheidsgrammatikas op te stel. Sankoff 1976 en 1985 gee 'n uiteensetting van hierdie aspek van die variasietaalkunde, en wys daarop dat die konsep *probabilistiese grammatikas* reeds gebruik is in werk oor styl, moedertaalverwerwing, tweedetaalverwerwing, die bepaling van diskoerseffekte op NP-struktuur en kodewisseling.

6.5 VERDERE VARIASIETAALKUNDIGE ONDERSOEKE

Binne die variasietaalkunde is die rekenaar ook al gebruik vir die ondersoek van die volgende verskynsels:

(a) Kodewisseling: Die bekendste verslag hieroor is Sankoff en Poplack 1981, wat die linguistiese perke op die wisseling tussen die elemente/grammatikale kategorieë van twee tale ondersoek het.

(b) Die analise van akoesties gespesifiseerde vokaalsisteme: Die bekendste werk op hierdie gebied is Labov, Yaeger and Steiner 1972.

(c) Die semantiek van denotasie, waarin taalgebruikers se keuses tussen verskillende verwysingsterme vir dieselfde klas sake statisties ondersoek is. Labov (1973) lewer hieroor verslag.

(d) Die verwerwing van die taalsisteem: Labov en Labov (1976) lewer verslag oor 'n kind se geleidelike verwerwing van die sisteem onderliggend aan die inversiereël by vraagsinne in Engels.

6.6 NAVORSINGSONDERWERPE BINNE DIE VARIASIETAALKUNDE

Dit is natuurlik 'n onmoontlike taak om 'n lys te verskaf van al die onderwerpe waaroor daar reeds binne die variasietaalkunde met behulp van 'n rekenaar navorsing gedoen is. Die volgende lys wil dus hoogstens net 'n ruwe aanduiding van die terrein verskaf (die ondersoeke reeds vermeld word buite rekening gelaat):

segmentweglating in 'n groot verskeidenheid tale
nasaalverswakking in Sjinees
die reduksie en weglating van die kopula in Engels
kongruensie in die Romaanse tale
vokaalverskuiwing (stoot- en trekprosesse)
vokaalversmelting in Engels
tempusmarkering in Engels
die gebruik van *be* in swart Engels
die verlies van die ontkenningsmerker in Frans
die gedrag van die perifrastiese *do* in Engels
relatiefstrukture
komplementstrukture
tempus en aspek in Kanadese Frans
woordordeverskynsels in verskillende tale
spreidingspatrone in die gebruik van die onbepaalde voornaamwoord
die wisseling tussen hulpwerkwoorde in Frans.

6.7 DIE VARIASIETAALKUNDE IN DIE RSA

Hoewel die variasietaalkunde betreklik uitgebreid beoefen word in die RSA, is min daarvan kwantitatief van aard. Voorbeelde van kwantitatiewe analises is: Klopper 1976 en 1983, Kotzé 1983, Lanham 1979, Van der Zwan 1986 en Webb 1987. Van hierdie studies het slegs Webb 1987 gebruik gemaak van VARBRUL.

VARBRUL 2S is op die oomblik in die besit van verskeie Suid-Afrikaanse navorsers, by name by die universiteite van Port Elizabeth, Pretoria, Stellenbosch en Zululand. Die Departement Afrikaans aan die Universiteit van Pretoria beskik voorts ook nog oor 'n hoofraamweergawe van VARBRUL 2, wat deur lede van die Universiteit se Buro vir Rekenaardienste in PL/I ontwikkel is. Ten slotte beskik die Departement Afrikaans aan die Universiteit van Pretoria ook nog oor 'n kopie van VARBRUL 3, wat eersdaags hopelik op hoofraam operasioneel gemaak sal kan word.

Slegs enkele persone in die RSA is tans aktief betrokke by variasietaalkundige navorsing met behulp van VARBRUL. Afgesien van die skrywer van hierdie hoofstuk (V.N. Webb) is die ander: R.M. Klopper van die Universiteit van Zululand en I.A. Coetzee en A. Boshoff, albei studente in die Departement Afrikaans, UP. Etlike ander taalkundiges het egter reeds aangedui dat hulle daarin belang stel om opleiding in die gebruik daarvan te ontvang.

6.8 SLOT

Die variasietaalkunde is klaarblyklik 'n gebied waarop die rekenaar 'n belangrike rol kan speel. Soos blyk uit voorgaande bespreking is dit veral die rekenaar se vermoë om groot hoeveelhede data te hanteer, en om ingewikkelde berekenings uit te voer, wat dit so waardevol as navorsingshulpmiddel maak. Vanselfsprekend het die rekenaar dus toepassingspotensiaal op alle terreine waar groot datakorpusse en die berekening van menigvuldige gegewens ter sprake kom.

Verder is die variasietaalkunde 'n voorbeeld van 'n vakgebied waar buitelandse programme vir plaaslike aanwending geskik is.

HOOFSTUK 7: REKENAARGESTEUNDE ONDERSOEK OP ANDER TAALGEBIEDE

7.1 KORPUSLINGUISTIEK

7.1.1 Inleiding

Met 'korpuslinguistiek' word verwys na daardie terrein van die taalkunde wat hom besig hou met die versameling en voorbereiding van korpusse tekste in natuurlike taal met die oog daarop om hierdie tekste te rekenariseer en linguisties te indeksseer. Die doel met dergelike manipulasie is om datakorpusse saam te stel waaruit databanke vir die taalkundige analise van spesifieke verskynsels saamgestel kan word.

Die belangrikheid van die korpuslinguistiek is geleë in die toenemende waarde wat linguiste heg aan data uit werklike gevalle van taalgebruik.

Tot ongeveer 'n dekade gelede was die meerderheid linguiste tevrede met introspektiewe data, dit wil sê data wat hulle uit hul eie kennis van die taal wat hulle ondersoek het, onttrek het. Gaandeweg het al hoe meer linguiste egter besef dat dergelike data minstens twee nadele het, naamlik: (a) dat voorbeelde wat op hierdie wyse verkry is dikwels slegs 'n refleksie verskaf van hoe hulle self meen die taal *gepraat behoort te word* en dus nie noodwendig hoe die sprekers van die taal die taal werklik gebruik nie; en (b) dat dergelike data hoogstens 'n weerspieëling van net een soort taal kan wees, en dat verreweg die meerderheid variëteite in die taal nie in so 'n datastel verteenwoordig kan wees nie.

Verder is dit ook so dat intuïsiegebaseerde metodes die linguis slegs in staat stel om 'n beskrywing te gee van die tipes grammaties toelaatbare konstruksies wat in 'n taal kan voorkom, maar nie om ondersoek in te stel na die gebruikspatrone van 'n bepaalde taal nie.

Databanke/-stelle wat aan groot versamelings van natuurliketaal-tekste onttrek is, is dus van regstreekse belang vir die taalkundige, en kan op twee maniere 'n rol in sy werk speel, naamlik: dit kan as 'n kragtige middel dien tot die aanvulling en kontrole van linguiste se intuïsies omtrent die taalverskynsel wat hulle ondersoek, en tweedens, dit kan linguiste daartoe in staat stel om hul linguistiese hipoteses verder te toets.

Aanvanklik het linguiste hul databanke met die hand opgestel, dit wil sê hulle het hul tekste self deurgewerk en 'n lys voorbeelde daaruit saamgestel waarop hulle hul analises gebaseer het. 'n Voorbeeld van hierdie benadering tot data-insameling word gekry in die Laboviaanse variasietaalkunde waaroor hoofstuk 6 rapporteer.

Dit is sonder meer duidelik dat hierdie metode van data-insameling besonder tydrowend is (en dalk selfs nie eers 100% betroubaar is nie). Indien databanke met die rekenaar opgestel word, behoort daar 'n beduidende wins te wees ten opsigte van sowel tyd as betroubaarheid. So 'n wins sal egter slegs werklikheid word indien 'n effektiewe stelsel ontwikkel word vir die gebruik van die rekenaar in die korpuslinguistiek.

Die doel van hierdie deel van hierdie verslag is om 'n oorsig te verskaf van die rekenaarmatige korpuslinguistiek. Die inligting wat hierin verstrek word, is grootliks gebaseer op twee bronne, naamlik Johansson 1982 en Aarts en Meijs 1984.

7.1.2 Die doel van die korpuslinguistiek

Aarts en Van den Heuvel (Aarts en Meijs 1984:83) wys daarop dat die korpuslinguistiek twee hoofdoelwitte het, naamlik:

- Die konstruksie van 'n betroubare en doelmatige stelsel vir die prosessering van groot taalkorpusse en vir die beskikbaarstelling en berging van taalkundige inligting oor die korpus.
- Die ontwikkeling van 'n stelsel waarmee tekste taalkundig outomaties ontleed kan word sodat grammatikas geskryf kan word.

Wat die eerste van hierdie twee doelwitte betref, kan die volgende spesifieke take onderskei word:

- die opbou van 'n verteenwoordigende versameling korpusse;
- die omsetting van die tekste in masjienleesbare vorm;
- die voorbereiding van hierdie tekste sodat hulle linguisties op 'n effektiewe wyse geïndeksseer kan word;
- die prosessering van laasgenoemde tekste op so 'n manier dat ondersoekers enige inligting wat hulle uit die tekste wil onttrek, kan onttrek;
- die ontwikkeling van 'n stelsel waarmee belangstellende linguiste maklik toegang tot die korpusversameling kan kry.

In die res van hierdie afdeling word 'n oorsig van elk van hierdie fasette van die korpuslinguistiek gegee.

Die tweede doelwit hierbo word in 7.1.5 bespreek.

7.1.3 Die opbou van 'n versameling tekste

Die bruikbaarheid van 'n versameling van tekskorpusse is natuurlik regstreeks afhanklik van die mate waarin die versameling die linguis daartoe in staat stel om uitvoering te gee aan die ondersoek waarmee hy besig is. Daarom is dit essensieel dat die keuse van tekste op 'n beplande wyse geskied.

Die belangrikste vrae wat aan die begin van 'n projek in die korpuslinguistiek in hierdie verband gestel moet word, is:

- Wat is die universum waaruit die tekste getrek moet word?
- Hoe moet die teksmonster gestruktureer wees?
- Wat moet die grootte van die teksmonster wees?

Indien 'n linguis 'n bepaalde verskynsel wil ondersoek in 'n taal as geheel, dit wil sê soos dit voorkom in al die variëteite van die taal, sal hy sy data uiteraard uit 'n verteenwoordigende stel korpusse wil trek. 'n Toereikende versameling tekskorpusse sal dus minstens die volgende soorte tekste moet insluit: gesproke én geskrewe tekste, standaardtaaltekste én nie-standaardtaaltekste, volwassetaaltekste én kindertaaltekste, tekste uit skeppende werk, regsdokumente, koerante, tydskrifte, vakpublikasies en brosjures, ensovoorts. Alle moontlike genres sal gedek moet wees en tekste oor soveel onderwerpe as moontlik (soos godsdiens, politiek, sport en ontspanning, die beroepslewe, ens.) sal ingesluit moet wees.

Heelwat literatuur oor teksseleksie is beskikbaar, soos die werk van Johansson (red.) 1982 en Aarts en Meijs (reds.) 1984, die inligtingstuk van die Institut für deutsche Sprache (vgl. Inligtingstuk 25) en ICAME se nuusbriewe.

Etlike tekskorpusse is reeds oorsee beskikbaar. Voorbeelde is:

(a) Die Brown-korpus bevat 'n versameling van gepubliseerde Amerikaanse Engels van 1961 wat onder leiding van W. Nelson Francis van die Universiteit van Brown versamel is. Die versameling is nie verteenwoordigend van algemene Amerikaanse Engels nie en nie een van die konstituerende tekste is langer as 2000 woorde nie. Die versameling as geheel beslaan ongeveer een miljoen woorde en dek vyftien genres. Die bruikbaarheid daarvan is dus beperk. Tog dien dit nou al meer as 27 jaar as 'n baie nuttige navorsingsinstrument. Dit is tans op rekenaarband beskikbaar in twee formate. 'n Volledige konkordansie van al die woorde in die korpus, met statistiek oor die distribusie van die woorde in elk van die 15 teksgenres, is op mikrofiche beskikbaar.

(b) Die Lancaster-Oslo/Bergen-korpus bevat 'n versameling van Britse Standaard-Engels wat tussen 1970 en 1978 eers aan die Universiteit van

Lancaster en daarna aan die Universiteit van Oslo en die Bergen-sentrum byeengebring en versorg is. Dit bevat 500 uittreksels uit tekste waarvan elk ongeveer 2000 woorde lank is. (Totaal dus ongeveer een miljoen woorde.) Dit dek ook vyftien genres. Die teks is tans op rekenaarband beskikbaar in twee ongeïndeksseerde formate en twee geïndeksseerde formate. 'n Volledige konkordansie van al die woorde in die korpus, gesorteer per trefwoord én per woordsoortindeks, is ook op fiche beskikbaar.

(c) Die Londen-Lund-korpus bestaan uit tekste van opgevoede gesproke Britse Engels wat onder leiding van Randolph Quirk van die Universiteitskollege, Londen, en Jan Svartvik van die Universiteit van Lund, byeengebring is. Die versameling bestaan uit 87 tekste wat elk ongeveer 5000 woorde lank is. Vyf soorte gesprek is in die korpus gedek: persoonlike gesprekke tussen twee persone, 'n telefoongesprek, radio-uitsendings en radiodebatte, voorbereide, openbare voordragte en privaatgesprekke wat klandestien opgeneem is. Twee konkordansies van die korpus is ook beskikbaar.

(d) Die Birminghamse korpusversameling van geskrewe en gesproke Britse Engelse tekste, wat teen 1983 reeds 12 miljoen woorde beslaan het.

(e) Die Helsinkikorpus van diachroniese en dialektiese Engels. Eersgenoemde bestaan uit tekste wat verskyn het tussen die 8e en die 18e eeu. Die korpus sal uiteindelik ongeveer een en 'n half miljoen woorde bevat, en sal so verteenwoordigend as moontlik wees met verwysing na chronologie, dialek, tekstipe en styl. Die dialekkorpus sal ongeveer 'n halfmiljoen woorde beslaan, en sal bestaan uit bandopnames van onderhoude met bejaarde sprekers wat in 1970 opgeneem is.

(f) Die Melbourne-Surrey-versameling van Australiese Engels *(ICAME Journal* 11, pp. 39-43). Hierdie korpus bestaan uit 100,000 woorde en bevat die tekste van hoofartikels wat in die Australiese koerant *The Age* verskyn het oor 5 maande gedurende 1980/81.

(g) 'n Tweede korpus van Australiese Engels *(ICAME Journal* 11 pp. 27-38). Hierdie korpus is gestruktureer soos die Brown- en die LOB-korpusse, en bevat die volledige teks uit Australiese nasionale en metropolitaanse koerante van 1986.

(h) Die Kolhapur-korpus van Indiese Engels *(ICAME Journal* 12, pp. 15-26).

(i) Die Ottawa-Hull-versameling van Kanadese Frans, en

(j) Die Institut für deutsche Sprache se versameling van Duitse koeranttekste, wat sowat sewe miljoen woorde beslaan.

Die rekenaarmatige prosessering van kleinere korpusse (wat Sinclair in Johansson 1982:4 *sample corpora* noem) is reeds gevestig, en hulle word oor 'n wye front reeds beskou as 'n alledaagse navorsingsinstrument. Sisteme vir die hantering van groot korpusse (wat Sinclair *monitor corpora* noem) was teen 1982 nog aan die ontwikkel, en die gebruik daarvan as navorsingsinstrument sal vermoedelik eers later algemeen word.

Daar bestaan reeds etlike internasionale sentra vir die gerekenariseerde korpuslinguistiek, soos die ChiLDES-projek (vgl. later in hierdie hoofstuk), die Linguistische Datenverarbeitung van die Institut für deutsche Sprache (Inligtingstuk 25), en ICAME, die International Computer Archive of Modern English (vgl. Inligtingstuk 13). ICAME het in 1977 tot stand gekom met as doel die opbou van 'n argief van Engelse korpusse.

Die toename in die belangstelling in tekskorpusse blyk onder andere uit die feit dat ICAME in 1987 hul 8e internasionale kongres in Helsinki gehou het. Dit is bygewoon deur 74 deelnemers uit 12 lande.

Die bedrywigheid op hierdie gebied blyk ook uit die feit dat ICAME 'n rekenaarnetwerk opgestel het vir die uitruil van inligting en programmatuur vir gebruik in die korpuslinguistiek. Die netwerk wat gebruik word, is EARN/BITNET, maar dit kan bereik word deur enige netwerk wat toegang tot EARN/BITNET het, soos Uninet, Janet, ARPA en Csnet (vgl. *ICAME Journal* 11:65).

7.1.4 Die omsetting van tekste in masjienleesbare vorm

Dit is deesdae natuurlik moontlik om groot hoeveelhede teks wat reeds masjienleesbaar is, in die hande te kry. Die meeste koerante word byvoorbeeld rekenaarmatig geset, terwyl groot uitgewers hul publikasies ook op hierdie wyse set. (Rekenaargesette koerantteks is nie sonder meer bruikbaar in die korpuslinguistiek nie - soos Peters 1987:29 aantoon.)

Vir sover navorsers tekste in natuurlike taal self moet rekenariseer, moet die volgende twee stappe gevolg word:

Die eerste stap behels die omsetting daarvan in masjienleesbare formaat. Hierdie omsetting kan op die oomblik onder andere op die volgende maniere geskied:

- Die regstreekse insleutel van die teks per woordverwerker vanaf 'n ortografiese transkripsie daarvan (in die geval van gesproke tekste) of vanaf 'n gepubliseerde vorm daarvan;
- Die inlees daarvan per inleesapparaat vanaf 'n getikte of gedrukte kopie.

Wat die tweede metode betref, geld die volgende opmerkings. Tekste waarvan die letters proporsioneel gespasieer is, kan nie met enige inleesapparaat en enige inleesprogrammatuur effektief omgesit word in masjienleesbare vorm nie: die programmatuur wat gebruik moet word, moet proporsioneel gespasieerde letters kan herken. 'n Programpakket wat hierdie vermoë het, is 'n pakket met die naam S.P.O.T., wat slegs gebruik kan word op sekere inleesapparate. Die masjien wat gebruik is vir die inlees van die Brown- en die LOB-korpusse is 'n soort optiese aftaster ('n *scanner)* en staan bekend as die Kurzweil Data Entry Machine (die sg. KDEM - vgl. Johansson 1982:3 en Aarts en Meijs 1984:3 en 9-10). Vergelyk ook die WORDNET-verslag, afdeling 3.3.3.2.

'n Elektroniese vaslêmetode wat in die toekoms moontlik bruikbaar sal word, hou regstreeks verband met vordering op die gebied van outomatiese spraakherkenning. Indien navorsing op die gebied van mens-masjien-kommunikasie sy belofte vervul, is dit moontlik dat gesproke korpusse regstreeks in geskrewe, masjienleesbare vorm omgesit kan word. Die voorkoms in gesproke tekste van onderbrekings, die oorvleueling van sprekers se uitinge, huiweringsverskynsels, pouses, wysigings deur sprekers van hul eie uitinge, ensovoorts, maak die spoedige ontwikkeling van stelsels waarmee tekste op hierdie wyse vasgelê kan word, effens onwaarskynlik.

Wat hierdie eerste fase van die rekenarisering van tekskorpusse betref, moet daar nog verwys word na die minimum apparatuur wat daarvoor nodig is. Afgesien van die aftaster (met sy programmatuur) is 'n aantal mikrorekenaars en/of 'n hoofraamrekenaar nodig - die mikrorekenaars vir die hantering van tekste binne die groter korpus, en die hoofraam vir die berging van die teksversameling as 'n geheel. (Die 80386-mikrorekenaars wat tans beskikbaar word, sal vanweë hul groot geheuespasie die afhanklikheid van hoofraamrekenaars verminder.)

Die voorredigeringstadium

Die tweede stap in die rekenaarmatige manipulasie van tekste behels die voorbereiding van die teks vir die linguistiese indekssering daarvan. Hierdie stap staan bekend as die voorredigeringstadium.

Tekste, veral dié wat uit gepubliseerde bronne afkomstig is, bevat heel dikwels eienskappe waarmee stelsels vir die outomatiese linguistiese indekssering van die tekste probleme ondervind. Voorbeelde van hierdie eienskappe is: die voorkoms van numeriese karakters (soos datums), hoekhakies, aanhalingstekens, Romeinse karakters, bo- en onderskrifte, persentasietekens, tabelle en die voorkoms van hoofletters binne uitings (bv. vir eiename).

Die span persone wat verantwoordelik is vir die rekenarisering van die tekskorpusse moet dus 'n stelsel ontwerp waarmee problematiese eienskappe van tekste in die voorredigeerstadium gemerk kan word sodat die linguistiese indekssering van die teks effektief kan geskied.

Hierdie stadium van die hantering van tekskorpusse behels uiteraard ook die voorredigering van die tekste met behulp van 'n koderingstelsel.

Terwyl taalspesifieke eienaardighede altyd per hand gehanteer sal moet word, is navorsers tans besig met die ontwikkeling van 'n stelsel waarmee eienaardighede wat internasionaal van aard is, outomaties gemerk sal word. Hierdie internasionale standaard staan bekend as die Standard Generalized Markup Language (SGML), en in Europa as die Formex-program.

'n Eenvoudige, eenvormige, nasionaal-erkende transkripsiestelsel moet ontwikkel word vir die redigering van die tekste. In hierdie verband kan daar kennis geneem word van die transkripsiestelsel wat in die CHiLDES-projek gebruik word (vgl. Inligtingstukke 15-20).

Die berging van tekskorpusse

Hoewel die optiese skyf vir die toekoms moontlikhede inhou, is die gebruik van die magnetiese band vir die huidige waarskynlik die aangewese bergwyse vir omvattende tekskorpusse.

7.1.5 Die linguistiese prosessering van tekskorpusse

Die waarde van 'n gerekenariseerde tekskorpus is natuurlik regstreeks afhanklik van die mate waarin linguiste bruikbare inligting sonder moeite daaraan kan onttrek. Dit is dus van die grootste belang dat die gerekenariseerde tekskorpusse deursigtig gestruktureer sal wees en volledig genoeg geannoteer sal wees sodat linguiste alle inligting wat hulle nodig mag hê vir hul ondersoeke daaraan sal kan onttrek.

Daar is verskeie aspekte aan hierdie saak wat bespreek moet word. In die eerste instansie moet daar op gewys word dat linguisties ongeannoteerde masjienleesbare tekste nie vir die linguis waardeloos is nie. Nuttige voorbeeldmateriaal kan uit dergelike tekste verkry word deur eenvoudig gebruik te maak van die gewone patroonherkenningsfunksie van die rekenaar. Die rekenaar word opdrag gegee om bepaalde konfigurasies van karakters (soos bv. die kombinasie #w+a+t#, dit wil sê die merker van die relatiefsin of die vraagwoord) te soek en dan saam met, byvoorbeeld, die voorafgaande tien en die daaropvolgende tien woorde aan die korpus te onttrek en apart te stoor. Sodoende verkry die linguis 'n volledige lys voorbeelde van die verskynsel wat hy wil ondersoek (soos bv. die relatiefkonstruksies van 'n taal).

Hierdie metode is natuurlik nie waardeloos nie, maar is tog 'n 'onintelligente' wyse van doen aangesien materiaal aan die korpus onttrek mag word wat nie vir die linguis se ondersoek ter sake is nie soos gevalle van *wat*-vraagsinne terwyl die linguis moontlik slegs relatiefvorming wil ondersoek.

'n Vrugbaarder benadering behels die gebruik van linguisties geannoteerde tekskorpusse. Die grammatikale indeksering van 'n teks kan òf per hand òf (semi-)outomaties geskied, en kan verskillende vlakke van linguistiese gesofistikeerdheid bereik, naamlik slegs met woordsoortspesifikasie *(tagging)* of dáárby ook die spesifikasie van morfologiese struktuur en -funksie *(parsing)*, die spesifikasie van frasestruktuur en -funksie *(parsing)*, klousstruktuur en -funksie (eweneens *parsing)*, diskoersstruktuur en -funksie, ensovoorts.

As 'n illustrasie van 'n stelsel wat ontwikkel is vir die interaktiewe annotering van 'n teks kan IT genoem word. IT, die afkorting van Interlinear Text (System), is 'n produk van die Summer Institute of Linguistics in Dallas, Texas, VSA, en is die afgelope vyf jaar ontwikkel as 'n instrument vir die ontwikkeling van geanaliseerde tekskorpusse. Dit stel die navorser daartoe in staat om tekste interlineêr te annoteer ten opsigte van 'n verskeidenheid dimensies waarop die gebruiker self besluit, soos die fonetiese, die fonologiese, die allomorfiese, die morfologiese, die woordsoortlike, die struktuur en funksie van die frases en die klouse in die teks.

Die linguistiese annotasie van die teks geskied interaktief: terwyl tekste aan die hand van 'n voorafbepaalde annotasiemodel geïndekseer word, word 'n databasis van leksikale inligting opgebou wat dan weer tydens die annotasieproses as kontrole daarby dien.

Na afloop van die annotasieproses kan ander instrumente van die stelsel gebruik word vir die verdere manipulering van die teks en die leksikale databasis, en kan sintaktiese en leksikale inligting uit hierdie twee databronne onttrek word.

'n Ander stelsel van grammatiese indeksering is gebruik in die geval van die Londen/Lund-korpus. Johansson 1982:92-107 bespreek die koderingsbeginsels en die stelsel van woordsoortmarkering wat in dié geval gevolg is. En in Aarts en Meijs 1984:57-62 gee Eeg-Olofsson weer 'n uitvoerige uiteensetting van die sisteem wat in dieselfde korpus gebruik word vir die indekssering van frases en klouse. 'n Interessante aspek van hierdie sisteem is dat dit gebruik maak van herskryfreëls wat siklies toegepas word. Aangesien verskillende soorte frases en bysinne gekenmerk word deur die teenwoordigheid van verskillende woordsoortkombinasies kán hulle met behulp van sikliese herskryfreëls geïndeksseer word.

Stelsels vir hoërorde-indeksering (d.w.s. hoër as woordvlak) is in groot mate nog in 'n ontwikkelingstadium.

Hiermee is ons by 'n belangrike kwessie in die korpuslinguistiek, naamlik die ontwikkeling van stelsels waarmee tekskorpusse morfologies en sintakties outomaties ontleed kan word. Hierdie aspek word breedvoeriger bespreek in

hoofstukke 3 en 4 van hierdie verslag; hier word dus volstaan met enkele opmerkings.

Van die stelsels vir die (semi-)outomatiese grammatikale analise van korpusse wat reeds in hierdie verslag genoem is, is CLAWS, TOSCA, Grimes se PROGRAMMABLE TEXT PROCESSOR, Saarbrücken se IC-model, Fought (Universiteit van Pennsylvanië) se stelsel vir morfosintaktiese analise, PATR-II van Shieber en die ATN-model van Woods.

As 'n voorbeeld kan daar kortliks na TOSCA (TOols for Syntactic Corpus Analysis), wat ontwikkel is deur die Departement Engels aan die Universiteit van Nijmegen, Nederland, verwys word.

TOSCA is ontwikkel as 'n meganisme waarmee groot, linguisties onverwerkte taalkorpusse interaktief ontleed kan word met die doel om gedetailleerde sintaktiese inligting daaruit beskikbaar te stel. Daarbenewens was die TOSCA-projek ook daarop gerig om 'n formalisme te verskaf waarmee grammatikas geskryf kan word vir die outomatiese analise van 'n tekskorpus. Teen 1984 was daar met die hulp van TOSCA reeds grammatikas geskryf van die naamwoordstuk, die adjektiefstuk, die werkwoordstuk en die fleksiemorfologie van kontemporêre Engels.

Die stelsel neem 'n tekskorpus as toevoer, ontleed dit dan morfologies en sintakties en skep 'n leksikon uit die geanaliseerde korpus. Die morfologiese en sintaktiese analise geskied met behulp van outomatiese ontleders wat deur die stelsel se grammatikakomponent gegenereer word. Die grammatikamodel wat TOSCA gebruik, is die Extended Affix Grammar. Inligting hieroor is beskikbaar in die artikel van Aarts en Van den Heuvel en in 'n artikel deur Oostdijk, wat ook in Aarts en Meijs 1984 opgeneem is.

Daar is op die oomblik nie eenstemmigheid oor die mate van sukses waarmee outomatiese analiseerders ontwikkel is nie. Volgens Boot (Inligtingstuk 23) was daar byvoorbeeld in 1986 nog nie programme wat tekste/datakorpusse morfologies en leksikologies op bevredigende wyse kon annoteer nie. Uitgebreide linguistiese kodering van die tekste moet dus volgens hom steeds per hand gedoen word. Die Instituut vir toegepaste linguistiek en rekenaarlinguistiek, Universiteit van Utrecht, ontwikkel tans egter 'n stelsel met die naam BOBRA, wat masjienleesbare tekste morfologies outomaties sal kan annoteer.

Dit is belangrik dat die linguistiese indeksering van 'n teks sover as moontlik geoutomatiseer word. 'n Volledig outomatiese sintaktiese analise van 'n korpus is egter waarskynlik nie haalbaar nie vanweë die feit dat sintaktiese analise nie 'n uitsluitlik vormlike aktiwiteit is nie en 'n kennis van die alledaagse werklikheid soms nodig is daarvoor. Korpuslinguiste stem op die oomblik saam dat die doeltreffendste manier van linguistiese indekssering die interaktiewe benadering is. Een van die redes hiervoor is dat linguiste korpusse in elk geval waarskynlik self sal wil

annoteer onder andere omdat die annotasies van die korpusontwikkelaars nie volledig aan hul ondersoekvereistes mag voldoen nie.

'n Belangrike vereiste vir die modelle/raamwerke waarmee tekskorpusse grammatikaal geannoteer word, is dat hulle onkontroversieel sal wees, dit wil sê aanvaarbaar sal wees vir soveel linguiste as moontlik ten opsigte van byvoorbeeld hul grammatikale indekse vir woordsoorte, woordstruktuur, frasestruktuur, klousstruktuur, ensovoorts, sowel as ten opsigte van die strategieë waarmee prosesse soos neweskikking, onderskikking en ellips gehanteer word.

Ten slotte oor hierdie onderwerp: die WORDNET-verslag, afdeling 3.4, bevat ook nog inligting oor stelsels wat in die leksikografiebedryf en in masjienvertaling gebruik word vir die linguistiese analise van taalvorme in tekskorpusse. (o.a. die ASCOT-stelsel, die MARS-stelsel, die stelsels vir die morfologiese, sintaktiese en semantiese analise wat in die vertaalstelsel SUSY gebruik word en die komponente vir linguistiese analise wat in die LUTE-vertaalstelsel gebruik word).

Afgesien van grammatikale annotasie moet tekste natuurlik ook vir ander aspekte geïndeksseer word. Elke teks moet byvoorbeeld gemerk word vir die taalvariëteit waarin dit voorkom (bv. Standaard-Afrikaans, Griekwa-Afrikaans, Maleier-Afrikaans, ens.), die bron waaruit dit kom (boek, tydskrif, koerant, ens.), sy genretipe, sy medium (bv. gesproke of geskrewe), sprekeridentiteit (ouderdom, etnisiteit, geslag, ens.). Daarby moet elke reël van die teks ook genommer word.

Na afloop van die grammatikale en nie-grammatikale prosessering van tekste volg die laaste stadium, naamlik die stadium van naredigering. In hierdie stadium moet die linguisties versorgde teks deur die ondersoeker gekontroleer en ontfout word.

7.1.6 Die opstel van databanke

Nadat tekskorpusse grammaties (en pragmaties, ens.) geannoteer is, kan hulle as datakorpusse gebruik word waaruit databanke saamgestel kan word, met inligting oor byvoorbeeld die linguistiese omgewings waarin bepaalde taalvorme voorkom, die moontlike kombinasies waarin hulle kan optree, die woorde en/of woordsoorte wat tipieserwys in bepaalde sintaktiese konstruksies voorkom en die frekwensie van hul voorkoms.

In hierdie verband moet daar eers verwys word na konkordansieprogramme (vgl. ook die WORDNET-verslag. Renouf 1983 (Aarts en Meijs 11) noem in hierdie verband COCOA, die Oxford Concordance Programme en CLOC. Sy wys daarop dat die Oxford Concordance Programme op daardie tydstip nog in 'n eksperimentele stadium was en dat CLOC vir hulle doel onekonomies was ten opsigte van sy bergingsprosedures. In hulle projek is COCOA gevolglik gebruik. Inligting oor die Oxford Concordance Programme kan verkry word in Hockey en Marriot

1980, en oor CLOC in Reed 1978. 'n Ander konkordansieprogram wat in die literatuur genoem word, is WORD CRUNCHER, wat aan die Brigham Young-Universiteit ontwikkel is vir gebruik op die Brown-korpus. Hierdie program word bemark deur die Electronic Text Corporation (*ICAME Journal* 11, pp. 44-47).

Op 'n hoër vlak as konkordansieprogramme is die stelsels waarmee linguistiese databasisse saamgestel kan word. 'n Voorbeeld van so 'n stelsel is die Linguistic Databasestelsel (LDB) wat in TOSCA (vgl. ook hoofstukke 3 en 4) gebruik word. Die Linguistic Database onttrek data uit korpusse wat deur TOSCA voorberei is met behulp van 'n 'navraagtaal' ('n *query language),* en stoor die onttrekte data in 'n aparte lêer. Die data word geberg in die vorm van bome met geëtiketteerde node met inligting oor die items se kategoriale status, funksies en ander relasies tussen konstituente, woordvorme, struktuurvlakke, en so meer). 'n Voordeel van hierdie stelsel is dat die onttrekking van materiaal oor linguistiese patrone interaktief geskied, wat navorsers toelaat om die materiaal te ondersoek terwyl dit onttrek word, sodat nuwe hipoteses geformuleer kan word tydens die onttrekkingsprosedure.

7.1.7 Toegang tot gerekenariseerde tekskorpusse van buite

Indien tekskorpusse waarde moet hê vir linguiste van buite, moet weë tot stand gebring word waardeur hul formeel toegang tot die gerekenariseerde korpusse kan verkry. Die vernaamste maniere waarop linguiste toegang verkry tot die bestaande tekskorpusse is die volgende:

(a) Deur versoeke tot die beheerliggaam van die korpus per brief of faks te rig om spesifieke voorbeeldmateriaal/'n databank, wat dan per disket of per pos aangestuur word.

(b) Deur die voorbereiding van die hele korpus op diskette wat in mikrorekenaars gebruik word. Dit is byvoorbeeld die geval met die Brown-korpus.

(c) Deur die plasing van die korpus op kompakteskryf, en die beskikbaarstelling van 'n programpakket wat toegang tot die versameling verleen, soos in die geval van die New Oxford English Dictionary.

(d) Deur die ontwikkeling van 'n rekenaarnetwerk vir die hantering van navrae of die direkte toegang tot die korpus.

Hiermee is ons ook by 'n belangrike aspek van die korpuslinguistiek, naamlik die plasing van die versameling tekskorpusse. Daar is twee moontlikhede, naamlik die plasing daarvan op 'n sentrale plek, en die desentralisering daarvan. Albei moontlikhede het vanselfsprekende voordele, met die oorwig waarskynlik by eersgenoemde vanweë die groter bedryfskontrole wat dit toelaat.

7.1.8 Nut van die korpuslinguistiek

Die korpuslinguistiek kan belangrike bydraes lewer op verskillende terreine, soos byvoorbeeld taalondersoek, die rekenaarlinguistiek en taalonderrig. Die waarde van gerekenariseerde tekskorpusse vir linguistiese ondersoek spreek eintlik vanself, maar volledigheidshalwe moet daar tog aandag daaraan gegee word.

'n Eerste linguistiese nut daarvan is dat woordbanke daaruit saamgestel kan word, met inligting oor byvoorbeeld die frekwensie van die woorde en die linguistiese kontekste waarin elk kan verskyn. 'n Tweede voorbeeld is dat die verskillende variëteite van 'n taal (dialekte, style en registers) sowel as die gespreks- en geskrewe genres van die gemeenskap funksioneel en struktureel noukeurig daaruit ondersoek kan word. In hierdie verband was daar byvoorbeeld reeds studies oor verskynsels soos: die korrelasie tussen woordfrekwensie, genretipes en style; variasiepatrone in die keuse tussen variante soos *shall/will, maybe/perhaps, film/movie, shop/store,* in Australiese Engels (*ICAME Journal* 11); die verband tussen die strukturele kompleksiteit van die naamwoordstuk, sy funksies en linguistiese variëteite.

Dan kan empiries verantwoordbare studies van spesifieke fonologiese, morfologiese, sintaktiese, semantiese en leksikologiese verskynsels uiteraard ook onderneem word. Ter illustrasie hiervan word verwys na die artikel in Aarts en Meijs (1984) waarin die studie van nabepalingsklouse in die Engelse naamwoordstuk bespreek word. Die volgende aspekte van hierdie verskynsel is ondersoek:

- die interne struktuur daarvan, dit wil sê die sintaktiese patroon onderliggend daaraan sowel as die funksies van die voegwoorde, die tipe klouse, die werkwoordstukke wat daarin voorkom, ensovoorts;
- die perke op die voorkoms van hierdie klouse in naamwoordstukke;
- die plasing van naamwoordstukke wat hierdie klouse bevat in die sinne waarin hulle voorkom;
- die funksie van naamwoordstukke wat hierdie klouse bevat in die sinne waarin hulle voorkom; en
- die mate waarin hierdie konstruksie voorkom in die verskillende registers van die Engels wat ondersoek is.

Ander voorbeelde van sintaktiese ondersoeke is: die studie van ellips (in 'n sin soos: *You can do so if you want to),* ontkenningsinne, en die voorkoms en plasing van bywoorde.

'n Goeie voorbeeld van die waarde van tekskorpusse as navorsingsinstrumente is die grammatika van Quirk, R., Greenbaum, S., Leech, G.N. en J. Svartvik. 1972. A grammar of Contemporary English. London: Longman, wat gebaseer is op die Londen-Lund-korpus.

As 'n verdere illustrasie van die linguistiese waarde van tekskorpusse word verwys na die lys wat in Johansson 1982, pp. 148-156, verskyn, van studies wat gebruik gemaak het van die Brown-, die Londen-Oslo/Bergen- en die Londen-Lund-korpusse.

7.1.9 Die korpuslinguistiek in Suid-Afrika

Die korpuslinguistiek bestaan nie in Suid-Afrika as 'n georganiseerde werksterrein nie. Tog is daar reeds enkele korpusversamelings. Afgesien van elektronies vasgelegde boeke (o.a. die nuwe vertaling van die Bybel) en koerantteks, is daar ook die volgende teksversamelinge aan Suid-Afrikaanse universiteite:

Gerekenariseerde tekste:

Aan die Universiteit van Pretoria is verskeie bandopnames van sosiolinguisties gestruktureerde onderhoude met respondente in Danville, Villieria en Waterkloof in Pretoria reeds omgesit in masjienleesbare vorm. Dan is daar ook 'n omvangryke korpus gesproke Afrikaans by die Randse Afrikaanse Universiteit op rekenaar beskikbaar, en die kindertaalkorpus by die RGN.

Ongerekenariseerde tekste:

Ongerekenariseerde tekskorpusse van gesproke Afrikaans bestaan by die Potchefstroomse Universiteit vir CHO en die Universiteite van Pretoria, Stellenbosch en die Oranje Vrystaat.

Ten slotte moet daar genoem word dat die Departement Afrikaans aan die Universiteit van Pretoria tans 'n projek voorberei vir die opbou van 'n korpusversameling van moderne Afrikaans asook 'n historiese korpus. Wat laasgenoemde betref, word die dagboek van Louis Trigardt tans gerekenariseer.

7.1.10 Slotopmerking

Die ontwikkeling van tekskorpusse is nie 'n klein onderneming nie. In oorsese korpusontwikkelingsprojekte is daar in elke geval gebruik gemaak van 'n span navorsers, wat etlike jare lank aan die bepaalde projek gewerk het. Sinclair (Johanssen:1982.1) wys byvoorbeeld daarop dat dit tot tien jaar lank kan neem om 'n versameling tekste te versamel en voor te berei voordat bruikbare linguistiese inligting daaruit onttrek kan word.

7.2 PSIGOLINGUISTIESE NAVORSING

7.2.1 Inleiding

In Suid-Afrika is die psigolinguistiek 'n verwaarloosde studieveld en navorsing in die psigolinguistiek bestaan feitlik glad nie buite die RGN nie. Om dié redes is hierdie verslag nie omvattend nie. Die volgende sake word kortliks bespreek:

- Die gebruik van rekenaars in psigolinguistieknavorsing
- Korpusse vir ontwikkelingspsigolinguistiek
- Die stand van die psigolinguistiek in Suid-Afrika.

'n Groot hoeveelheid aanvullende inligting word verskaf in die inligtingstukke wat van die RGN se Afdeling Leksikologie bestel kan word.

7.2.2 Die gebruik van rekenaars in psigolinguistieknavorsing

Die nuwe psigolinguistiek wat deur die Chomsky-revolusie in die linguistiek aangespoor is, het op twee hoof navorsingsgebiede uitgeloop: eksperimentele en ontwikkelingspsigolinguistiek. Die doeleindes en metodes van eersgenoemde is nou verwant aan hoofstroom eksperimentele sielkunde en dit vermy normaalweg data wat naturalisties voortgebring is (vgl. Green, 1972, vir 'n oorsig van die eksperimentele psigolinguistiek). In teenstelling daarmee het die ontwikkelingspsigolinguistiek as sy hoof databronne, òf die uiting van 'n kind wat besig is om taal te verwerf (vgl. Brown, 1973, vir 'n beskrywing van die beroemde Harvard-projek), òf die uitinge van oppassers wat sulke kinders aanspreek (vgl. Vorster, 1975, vir 'n oorsig van die 'klassieke' tydperk in taalinvoernavorsing).

Voor die huidige beskikbaarheid van mikrorekenaars en programmatuur moes die ontwikkelingspsigolinguistiek tevrede wees met handgeskrewe en getikte transkripsies van klankopnames. Hul enigste moontlikheid was om met die hand die data deur te werk en kodes toe te ken. Daarna moes die resultate gekwantifiseer word voordat verdere prosessering kon geskied. Alhoewel die kwantifisering moeisaam was, het dit aansienlike penetrasie van die data meegebring. Dit is op hierdie wyse dat Roger Brown en sy kollegas die Harvard-projek uitgevoer het in die vroeë sestigerjare; so ook is die data wat die basis gevorm het vir die semanties-kognitiewe vertrek in kindertaalnavorsing in die vroeë sewentigerjare verwerk.

Beperkings van tradisionele teksanalisemetodes

Afgesien van die moeisaamheid daarvan het handverwerking van kindertaaltranskripsies ernstige beperkinge. Dit is uiteraard 'n geslote proses; terwyl dit in staat is om enige aspek van taalverwerwing te illumineer waarvoor die puntestel voorsiening maak (soos frekwensie van 'n nabootsing, vrae, bevele in 'n teks, die vermeerdering van longitudinale data van verskillende woordtipes en die verskyning van semantiese verhoudings), is

dit feitlik onmoontlik om hierdie metode te gebruik om komplekse interverhoudings tussen verskynsels in die data uit te lig en te bereken.

Laat ons hierdie punt illustreer. Een van die treffendste verskille tussen kinder- en volwassenes se spraak is kinders se weglating van sekere woorde wat in die omgangswêreld van moeder-en-kind min inligting oordra: kopulatiewe, hulpwerkwoorde, voornaamwoorde, subjek-naamwoordfrases, voorsetsels. Gevolglik is die studie van die sistematiese vermindering van weglatings in kinders se spraak hoogs leersaam. Met 'sistematiese vermindering' bedoel ons nie die *feit* dat sekere woorde met die tyd 'n weerstand opbou teen weglating nie, of die *tempo* waarteen dit geskied nie, maar die *kontekste* wat die weerstand blyk te voorspel - of omgekeerd, die kontekste wat weglatings blyk te veroorsaak. In die ondersoek na hulpwerkwoordskrapping in die RGN se Afrikaanse korpus, is dit gevind dat hierdie kontekste slegs bepaal kan word deur verwysing na die interverhoudings tussen die voorkoms van prenominale adjektiewe, bywoorde, voorsetselfrases en objek-naamwoordfrases (vgl. Vorster, 1983:163ff).

Sodanige diep ontledings van kindertaaldata kan onmoontlik met die hand uitgevoer word. Elke woord moet omvattend ontleed word vir kategorieë soos klasvorm en sintaktiese funksie, maar ook vir herkoms - met ander woorde, of die woord deur die kind gesê is, of voorsien is in die ondersoeker se parafrase van die kind se uiting.

Tipes van rekenaarvereistes vir psigolinguistiek-navorsing

Daar is essensieel vyf tipes van take waarvoor die rekenaar in die psigolinguistiek gebruik kan word:

(a) Die sorteer en tel van vooraf gekodeerde en ingevoerde data, en eenvoudige berekeninge van hierdie data soos persentasies, gemiddeldes en standaardafwykings. Die SAS-prosedures PROC SORT, PROC FREQ en PROC MEANS is ideaal vir hierdie take.

(b) Statistiese toetse om die waarde van die verskille tussen subdele van die data te bepaal, die interaksie tussen veranderlikes, item- en faktoranalises, ensovoorts. Die verskeie statistiese programme in die BMDP-pakket is ideaal vir hierdie doel.

(c) Rekenaarsimulasie van natuurlike taal en kognitiewe prosesse vir die ontwerp van kriteria vir leerbaarheid, probleemoplossingsvermoë, ensovoorts. Sover vasgestel kon word, is hierdie soort programme nog nie algemeen beskikbaar nie.

(d) Die outomatiese omskakeling van getranskribeerde data. Volle besonderhede van programmatuur wat tans gebruik word, word in Inligtingstuk 15 verskaf.

(e) Linguistiese analises van natuurlike taal. 'n Groot verskeidenheid van programme vir die herkenning van woorde, morfeme en sintaktiese stringe is reeds ontwikkel (vgl. vroeër in hierdie hoofstuk en die PROLANG-verslag).

Nie een van die bogenoemde take is uitsluitlik kenmerkend van die vereistes vir psigolinguistieknavorsing nie. Take onder (a) en (b) is gebruiklik in die eksperimentele sielkunde; take onder (c) val reg in die gebied van kunsmatige intelligensie, terwyl dié onder (d) en (e) ewe toepaslik is vir linguistiek in die algemeen.

7.2.3 Gerekenariseerde korpusse vir psigolinguistieknavorsing

In hierdie afdeling bespreek ons kortliks die internasionale gerekenariseerde data-uitruilsisteem vir kindertaal wat onlangs daargestel is, en die een rekenaar-toeganklike kindertaalkorpus in Suid-Afrika.

Die kindertaaldata-uitruilingsisteem (CHiLDES)

'n Klein aantal joernaalstudies van kindertaal, waarvan sommige meer as 'n honderd jaar oud is, was tot onlangs die enigste basis vir die studie van taalverwerwing. Daarbenewens was kindertaal vir ontwikkelingsielkunde en spraakterapie slegs van periferale belang. Tegnologiese vooruitgang in klankopnames het egter saamgeval met die koms van 'n hoogs invloedryke taalteorie wat in 'n belangrike sin ook 'n taalverwerwingsteorie is. Dit het gelei tot 'n groot uitbreiding van die veld die afgelope dertig jaar. In die sestigerjare is 'n paar klassieke kindertaalkorpusse daargestel; sedert die begin van die sewentigerjare is al hoe meer getranskribeerde data beskikbaar; en oor die afgelope paar jare het die tendens om hierdie data in rekenaarlêers te stoor eksponensieel gegroei.

Namate die gebied van kindertaalnavorsing uitgebrei het, het dit duidelik geword dat werkwyses lukraak, idiosinkraties en ongekoördineerd was. Die voordele van die rekenaartegnologie is weinig benut. Brian MacWhinney skets die situasie soos volg:

> *A glance at a fairly representative publication in child language - the annual Stanford Papers and Reports on Child Language Development - indicates the extent to which failure to use automatic schemes for searching is slowing progress in the field. In just the 20 papers in the 1983 volume from Stanford, there are three papers based on paper-and-pencil counts of numbers of occurrences in different word types in the corpus of data collected by Roger Brown. Moreover, there are three other papers based on analyses of corpora other than that of Brown that could have utilized automatic search and analysis techniques, researchers are not only wasting their time, but also increasing the likelihood of error and decreasing the explicitness of their analytic technique (MacWhinney, 1984:4).*

Ontevredenheid met die gebrek aan orde in die kindertaalhuishouding, gekoppel aan 'n aansienlike toekenning deur die John D. en Catherine T. MacArthur Stigting, het gelei tot die daarstelling van die Child Language Data Exchange System (CHiLDES) by die Carnegie-Mellon Universiteit in 1984 met Brian MacWhinney en Catherine Snow as trustees. Die sisteem bestaan tans uit 20 verskillende korpusse van normale middelklas Engels sowel as Afrikaans, Deens, Frans, Hebreeus, Hongaars, Italiaans, Spaans, Tamil en Turkse korpusse.

Uit die staanspoor was CHiLDES die leidende kindertaaldatabasis en dit sal na verwagting aanhou groei. In die ses uitgawes tot op datum van die CHiLDES nuusbrief *Transcript Analysis* word volledige beskrywings van die voorgestelde formaat en koderingskonvensies gegee, asook 'n groot hoeveelheid ander inligting. Kopieë van hierdie nuusbriewe kan van die RGN bestel word (Inligtingstukke 15-20).

Die RGN se Afrikaanse kindertaalkorpus

Die RGN se Afrikaanse kindertaalkorpus beslaan 100 halfuur klankopnames van moeder-kind-interaksie asook die getikte transkripsies daarvan. 'n Volledig gekodeerde substel van die korpus, bestaande uit 100 kinduitinge van elkeen van 40 voorbeelde en 'n totaal van 20 000 woorde is ingevoer as 'n SAS-datastel en is beskikbaar by die RGN.

Die koderingsisteem vir die korpus maak voorsiening vir die uitbreiding van die gerealiseerde uitinge na 'n voorstelling van hul bedoelde betekenisse. Voordat die data gekodeer is, is afwykende uitinge minimaal geparafraseer om welgevormdheid te verseker, en inligting met betrekking tot verskille tussen die parafrases - vaslegging van die kind se veronderstelde semantiese bedoeling - en die werklike realisasie van die uiting is in die kodering van elke woord geïnkorporeer.

Die volgende algemene inligting is vasgelê in die kodering van elke uiting: lengte, tipe (verklarend, vraend, ens.), funksie (verslag, kommentaar, ens.), en woordorde. Elke woord is gekodeer as behorende tot een van die 14 vormklasse. Per klas het die subkategoriespesifikasies van nege vir voornaamwoorde tot twee vir die tweede 'nie' gestrek. Die totale opsies beskikbaar vir die kodering van 'n voornaamwoord is 35, vir 'n naamwoord 29 en vir 'n werkwoord 28. (Besonderhede oor die samestelling van die korpus en die koderingskonvensies word in Inligtingstuk 21 gegee. 'n Beskrywing van die data-analises verskyn in Vorster, 1983.)

Die uiters delikate kodering van die data oorskry die vereistes van 'n tipiese kindertaalnavorsingsondersoek. As 'n metodologiese ondersoek was dit egter geregverdig. Die prosedure is maklik aanpasbaar vir enige taal, en sy aantrekkingskrag in vergelyking met handverwerking lê in sy onbeperktheid en sy potensiaal vir kumulatiewe kodering. Opeenvolgende navorsers kan slegs die gedeelte van die data wat vir hom/haar van belang is, kodeer en prosesseer. Mits dieselfde prosedure gevolg word, sal die

data dan meer toeganklik wees ná elke koderingsoperasie. Hierdie oorweging, naamlik dat die kumulatiewe insette van opeenvolgende ondersoekers die waarde van die data sou bly verhoog, was een van die hoofoorwegings by die ontwerp van die sisteem.

7.2.4 Die psigolinguistiek in Suid-Afrika

Die psigolinguistiek word net by enkele universiteite aangebied as 'n keusevak vir honneursstudie. Dit vorm ook 'n deel van die opleidingsprogram van spraakterapeute by die vier universiteite wat spraakpatologie en oudiologie aanbied.

Wat navorsing betref, vind daar ook nie veel plaas by plaaslike universiteite nie, ten spyte van die beskikbaarheid van die RGN se kindertaalkorpus.

7.2.5 Slot

Uit die voorgaande bespreking kan daar afgelei word dat

- die rekenaar noodsaaklik is vir die doeltreffende verwerking, berging en uitruil van kindertaaldata; en dat
- rekenaarvereistes vir psigolinguistieknavorsing nie noemenswaardig verskil van die vereistes vir taalondersoek in die algemeen nie.

Dit is klaarblyklik 'n voordeel dat Suid-Afrika reeds 'n bydraer tot die CHiLDES-stelsel is. Hierdie internasionale kontak sou met groot vrug vir psigolinguistieknavorsing in Suid-Afrika uitgebou kon word mits die universiteite meer nagraadse studente kon oorreed om die terrein te betree.

7.3. LEKSIKOLOGIESE ONDERSOEK

7.3.1 Algemeen

Die gebruik van die rekenaar in leksikologiese werk word volledig uiteengesit in die verslae van WORDNET en TERMNET. Die verslag wat hier volg, het dus uiteraard slegs 'n oorsigtelik funksie en word hier net ingesluit ter wille van lesers wie se primêre belangstelling taalondersoek in die algemeen eerder as leksikologiese ondersoek in die besonder is.

Afgesien van die glottochronologie van die diachroniese taalkunde (d.i. 'n tegniek waarmee die tydstip waarop twee histories verwante tale geskei het, via leksikale ondersoek bepaal kan word) en die leksikostatistiek ('n raamwerk waarmee die verwantskap van tale leksikaal bepaal kan word), wat albei lank reeds beoefen word, word die rekenaar in die leksikologie deesdae veral vir leksikografiese werk en vir die analise van die struktuur van natuurlike tale se woordeskatte gebruik.

Leksikografiese werk

Die rekenaarmatige opstel van verskillende soorte woordeboeke (standaard, sosiolinguisties, retrograde, ens.), asook tesourusse (d.i. versamelinge van woorde in semantiese of begripsvelde) is algemeen bekend: geïndeksseerde woorde in 'n rekenaar word onttrek en volgens bepaalde beginsels gesorteer.

Die gebruik van die rekenaar vir hierdie doel beteken nie noodwendig dat leksikografiese werk vakkundig beter gedoen sal word nie. Tog hou die gebruik daarvan sekere voordele in, soos o.a. die volgende:

(a) Die rekenaar kan groot hoeveelhede data hanteer. Die elektronifisering van die New Oxford English Dictionary het byvoorbeeld meegebring dat 60 miljoen woorde op 21,000 bladsye met 2 1/4 miljoen aanhalings en 425,000 kruisverwysings geberg kon word.

(b) Die bywerking en hersiening van inskrywings en die produksie van nuwe uitgawes kan ook makliker geskied.

'n Aspek van hierdie toepassing wat nog aandag ontvang, is die ontwikkeling van inligtingsherwinningstelsels waarmee tekste in natuurlike tale (boeke, tydskrifte en koerante) deurgewerk kan word en die leksikale items daarin outomaties geëkserpeer, geklassifiseer en gestoor kan word. Stelsels wat reeds hiervoor ontwikkel is, is dié van Grimes (U Texas, Arlington): die sogenaamde Programmable Text Processor wat lyste kan opstel van die voorkoms en verspreiding van die woorde in 'n teks, CELEX, die stelsels van die Institut für deutsche Sprache in Mannheim en die Oxford Concordance Program van die Universiteit van Oxford se buro vir rekenaardienste.

Leksikografiese databanke, dit wil sê versamelings waarin die frekwensie en verspreiding van leksikale eenhede gespesifiseer is, is reeds algemeen in die buiteland (vgl. 7.3.2).

Ook noemenswaardig is die werk wat by die Birminghamse Universiteit in die Verenigde Koninkryk gedoen word: hulle is besig met die leksikografiese analise van 'n 7,3 miljoen woordkorpus, met 80% van die werk reeds voltooi.

Leksikologiese werk

Die leksikografiese werk wat hierbo genoem is, se relevansie vir die taalkunde is vanselfsprekend. Om maar een voorbeeld te noem: lyste woorde wat in dieselfde linguistiese omgewing voorkom, of wat gemeenskaplike semantiese en pragmatiese kenmerke besit, kan sonder enige moeite uit leksikale datakorpusse onttrek word vir noukeuriger grammatikale ondersoek. Die rekenaar kan dus ook gebruik word vir die analise van tale se

woordeskatstruktuur. In paragraaf 7.3.2 word 'n uiteensetting gegee van 'n stelsel wat vir leksikologiese ondersoek ontwikkel is.

7.3.2 'n Nederlandse leksikologieprojek

Soos aangedui in die WORDNET-verslag is daar in die buiteland verskeie voorbeelde van stelsels vir die leksikologiese verwerking van taalmateriaal. In hierdie verslag volstaan ons dus met 'n kort uiteensetting van een dergelike stelsel, naamlik CELEX.

Die Nederlandse *Centre for Lexical Information,* of CELEX, het begin 1986 tot stand gekom. Dit was die resultaat van die gemeenskaplike inisiatief en samewerking van verskeie navorsingsinstitute, sowel as befondsing van regeringskant.

Die doel van hierdie projek is om gerekenariseerde, veeltalige, multifunksionele leksikale databasisse tot die beskikking te stel van belangstellende liggame deur middel van elektroniese skakelmetodes wat gebaseer is op die Nederlandse navorsingsnetwerk (SURFNET). Een van die motiverings wat verstrek word vir die noodsaaklikheid van die sentrum is die toenemende belangstelling in die verkryging van groter insig in die struktuur van leksikons. As 'n voorbeeld van 'n terrein waarop daar 'n dergelike belangstelling is, kan verwys word na spraakherkenning en -produksie. Goedontwerpte en omvangryke leksikons speel op hierdie gebied 'n sleutelrol.

Een van die eerste take van CELEX was 'n ondersoek na die moontlik logiese struktuur van 'n Nederlandse leksikale databasis. (Die skakel tussen die leksikon en korpus-hantering is hier duidelik op die voorgrond.) Inligting wat vir die tweetalige leksikale databasis (Engels en Nederlands) beoog word, sluit in ortografiese, fonologiese, morfologiese, sintaktiese en frekwensiegegewens. In die logiese struktuurbeskrywing word alle redundante inligting verwyder, asook alle inligting wat deur algoritmiese prosedures verkry kan word. Die gevolglike tabelformaat maak die databasis hanteerbaar, verminder die benodigde bergingskapasiteit en maak doeltreffende ontsluitingsprosedures moontlik.

Wat die mikrostruktuur van die afsonderlike inskrywings betref, is daar onder andere gepoog om morfologiese inligting vir alle woorde in die databasis te verstrek. Dit kon op een van ten minste twee maniere gedoen word: òf 'n reëlpakket òf -module, wat telkens direkte ontleding van enige woord moontlik maak, òf 'n volle spesifikasie van die morfologiese struktuur van elke woord. Voorkeur word gegee aan die tweede alternatief, om 'n verskeidenheid redes, hoofsaaklik die volgende:

- Die CELEX-databasis is bedoel om morfologiese teorieë te toets, nie om 'n teorie te verteenwoordig nie.
- 'n Volle spesifikasie van gegewens is die enigste manier om vinnige antwoorde op komplekse vrae moontlik te maak.

- Ten volle gespesifiseerde morfologiese ontledings verskaf meer inligting as 'n kombinasie van ontledingsreëls plus die totale woordeskat van onontlede items.

Die voordele is tweevoudig: Eerstens het 'n mens 'n maklik toeganklike lys van alle woorde in die databasis waarop sommige morfologiese prosesse opereer, en tweedens 'n ewe maklik toeganklike lys van al die uitsonderings op al die reëls in dieselfde basis.

Die morfologiese struktuur van die woorde word semi-outomaties verkry: Eers word die woorde ontleed met behulp van KASIMER, ('n kategoriale analiseerder *(parser)* wat in LISP geskryf is), en daarna word die ontledings per hand gekontroleer, ontfout en uitgebrei.

7.3.3 Gerekenariseerde leksikologieprojekte in die RSA

In Suid-Afrika bestaan daar nog nie leksikale databasisse van die CELEX-tipe nie, dit wil sê databasisse wat onder andere in leksikologienavorsing benut kan word nie. Die termbanke van die Nasionale Terminologiediens en die Weermag is nie multidoelig nie. Termverskaffing is hul uitsluitlike funksie. Volgens die RGN-vraelysondersoek ten opsigte van rekenaargesteunde taalnavorsing in Suid-Afrika word daar geen leksikologienavorsing met behulp van die rekenaar uitgevoer nie, en net drie universiteite gebruik die rekenaar vir die opstel van woordeboeke (UP, die Noorde en Zululand).

7.4 SLOTOPMERKINGS

In hierdie hoofstuk is aandag gegee aan korpuslinguistiek, die psigolinguistiek en die rekenaarmatige leksikologie. Daar is sonder twyfel nog etlike ander terreine van taalondersoek waarop die rekenaar reeds met groot vrug gebruik is, maar waaroor (vollediger) inligting nie beskikbaar was nie. Voorbeelde van sulke terreine is: kodewisseling, taalversteuring, ontlening, stilistiese ondersoek, taaltipologisering en die kwessie van probabilistiese grammatikas (vgl. hoofstuk 6).

HOOFSTUK 8: DIE GEBRUIK VAN DIE REKENAAR IN DIE SUID-AFRIKAANSE TAALKUNDE

8.1 INLEIDING

Inligting vir die TEXTNET-ondersoek is op twee maniere ingesamel, naamlik met behulp van 'n vraelys en uit die verslae van die lede van die ondersoekspan.

Om verskillende redes is dit nie moontlik om volledige gegewens in sodanige ondersoeke in te win nie. Alhoewel daar gepoog is om alle aktiewe taalnavorsers in die land te bereik, is die vraelys nie deur almal ontvang nie, terwyl talle ander navorsers weer nie gereageer het nie, òf omdat hulle die rekenaar nie gebruik nie, òf omdat hulle navorsing nie in 'n stadium is waarop hulle daaroor gegewens wil verskaf nie. Die vraelysondersoek is egter deur persoonlike navrae van die ondersoekspan gerugsteun, daarom lyk dit waarskynlik dat ten minste die projekte wat goed op dreef is wel in die opname ingesluit is. Intussen is daar sekerlik weer nuwe inisiatiewe geneem, maar as algemene aanduider van die stand van die gebruik van die rekenaar in die Suid-Afrikaanse taalkunde is die inligting wat ingewin is tog insiggewend.

8.2 DIE VRAELYSONDERSOEK

'n Kort vraelys, wat net basiese inligting oor bestaande rekenaargesteunde navorsing versoek het, is uitgestuur aan elke departement of afdeling aan die universiteite en technikons in Suid-Afrika wat moontlik met taalnavorsing besig is. Die responsie was teleurstellend beide ten opsigte van die aantal lyste wat teruggestuur is, en ten opsigte van die beskeie hoeveelheid rekenaargesteunde taalnavorsing wat onderneem word.

Die volgende inligting uit die vraelysantwoorde is vir hierdie verslag relevant:

(a) Slegs 24 responsies is ontvang, waarvan een sonder naam/adres en een betreklik negatief was. Die responsies was van die volgende universiteite/departemente:

Durban-Westville (Afrikaans)
Natal (D) (ATW en Kommunikasie, Rekenaarwetenskap)
Natal (P) (Zulu)
Noorde (Rekenaarwetenskap)
Port Elizabeth (Afrikaans, Engels, Nguni en Sotho)
Pretoria (Afrikaans, Afrikatale, Engels, Latyn)
Zululand (Afrikaans, Afrikatale)
PU vir CHO (Engels)
Rhodes (Afrikatale)

Stellenbosch (Afrikaans, Rekenaarwetenskap, Afrikatale)
UNISA (Semitistiek, Bedryfsleiding)
Vista (Afrikaans, Engels)
Wits (Afrikatale)

OPMERKING: Daar is verder vasgestel dat minstens die volgende departemente ook van die rekenaar in taalondersoek gebruik maak:

Pretoria: Semitiese Tale, Spraakheelkunde en Oudiologie;
Stellenbosch: Semitistiese Tale;
UNISA: Afrikaans en Linguistiek;
WITS: Fonetiek en Linguistiek.

(b) Negentien van die responsies was van taaldepartemente, drie van rekenaarwetenskapdepartemente en een van 'n departement vir bedryfsleiding.

(c) In sewe van die departemente wat gereageer het, word rekenaartoepassings in taalondersoek gedoen, en in vier word rekenaartoepassings in die toegepaste taalkunde gemaak (navorsing oor fouteanalise, diagnostiese toetse by die aanleer van Engels en vertaalwerk).

Die universiteite waar rekenaartoepassings in taalondersoek gedoen word, is die Noorde (dept. Rekenaarwetenskap), Pretoria (Afrikaans), Rhodes (Afrikatale), Stellenbosch (Afrikaans-Nederlands en Afrikatale), UNISA (Semitistiek) en Zululand (Afrikaans en Nederlands).

Drie universiteitsdepartemente beoog rekenaartoepassings in taalondersoek: UPE (Afrikaans-Nederlands) oor morfologie en Middelnederlands), UPE (Nguni en Sotho) oor linguistiek, en Vista (Afrikaans) oor leksikologie.

Drie universiteitsdepartemente beoog rekenaargesteunde navorsing in die toegepaste taalkunde: Natal (Durban - ATW) oor tweedetaalverwerwing, PU vir CHO (Engels) oor fouteanalise vir Engels, en Zululand (Afrikatale) - ongespesifiseer.

(d) Die taalkundegebiede waarop rekenaartoepassings in taalondersoek plaasvind, is: die fonetiek en fonologie (die Noorde, UP, US en blykbaar ook Rekenaarwetenskap by UPE), die morfologie (US, UPE), die leksikologie (die Noorde, UP, Zululand), taaltipologie (Rhodes), variasieondersoek (UP en Zululand) en morfologie en sintaksis (UNISA en Stellenbosch).

(e) Etlike navorsers beskik oor datakorpusse, maar min van hulle is gerekenariseer. UP beskik oor databasisse oor: /r/-weglating, *vir*-gebruik, *dat*-weglating, woordplasing in die afhanklike sin en nasalering in Afrikaans; US (Afrikatale) beskik oor

gedigitaliseerde spraakklankdata en foneties getranskribeerde teksversamelings vir Xhosa, Zulu, Sesotho, Setswana en Noord-Sotho, US (Afrikaans-Nederlands) beskik oor 'n geïndeksseerde bron oor taalprobleme in Afrikaans, en Zululand beskik oor 'n woordelys vir Afrikaans, Engels en Zulu.

(f) Afgesien van die gewone statistiese pakkette wat in die handel verkrygbaar is, beskik enkele navorsers oor gerigte programme, byvoorbeeld VARBRUL (UP, US en Zululand), terwyl die departement Afrikatale van US etlike programpakkette self ontwikkel het (STAP, TEKSFON en SUQU).

(g) Die enigste publikasies wat gerapporteer is oor rekenaartoepassings in taalondersoek is van Webb (UP) en Roux (US).

(h) Slegs drie universiteite (UP, US en Zululand) verskaf opleiding aan hul studente, en wel vanaf die derde studiejaar.

(i) Die RGN het onlangs begin met die ontwikkeling van woordeboeke- en vertaalprogrammatuur.

8.3 VERSLAE VAN DIE SPANLEDE

Volgens die verslae vind die meeste rekenaartoepassings in taalondersoek in die RSA plaas ten opsigte van fonetiese en fonologiese ondersoek, die morfologie, variasiestudies en psigolinguistiese ondersoek.

8.3.1 Fonetiese ondersoek

Hoewel rekenaargesteunde fonetiese ondersoek ook voorkom by die Universiteit van Pretoria, Witwatersrand en die WNNR is dit veral die Departemente Afrikaans-Nederlands en Afrikatale van Stellenbosch wat op hierdie gebied bedrywig is. 'n Werksgroep vir seinverwerking by die universiteit (bestaande uit linguiste, ingenieurs en rekenaardeskundiges) werk aan teks-na-spraaksisteme vir Afrikaans en Xhosa, die analise van spraakklanke se akoesties-perseptiewe eienskappe in Xhosa, Zulu, Noord-Sotho, Sesotho en Setswana en uitgebreide fonetiese databasisse vir die Nguni- en Sothotale.

Etlike programme is reeds deur die groep ontwikkel, insluitende:

TEKSFON, wat gebruik word vir outomatiese fonetiese transkripsie;

FONSTAT, wat gebruik word om statistiese data oor foneemverspreiding te verkry;

STAP, wat gebruik word vir gevorderde statistiese verwerking met betrekking tot foneemverspreiding; en

MORFFON, wat gebruik word vir morfeemanalise.

Afgesien van hierdie werk is progamme ook al ontwikkel vir die transkripsie van Hebreeuse karakters en vir die hantering van Siriese teks op die rekenaarskerm sowel as die drukker.

8.3.2 Morfologiese ondersoek

Morfologiese ondersoek vind plaas in die Departemente Afrikaans en Nederlands en Rekenaarwetenskap aan die Universiteit van Stellenbosch, waar 'n eie pakket programme ontwikkel word vir die outomatiese morfologiese analise van Afrikaanse woorde as deel van 'n projek vir die ontwikkeling van 'n skrif-na-spraak-sisteem. Hierdie pakket bevat reeds die volgende: woordvormingsreëls, 'n morfologiese woordeboek met 10 000 inskrywings en 'verbuigingsreëls' (vgl. De Stadler en Coetzer, Inligtingstuk 21).

Morfologiese analise word ook gedoen in die Departemente Semitiese Tale van US en Semitistiek van UNISA. In hierdie twee departemente word daar gewerk aan Hebreeuse, Siriese, Ugaritiese en Arabiese tekste.

8.3.3 Leksikale ondersoek

Op hierdie terrein is die volgende werk gedoen:

(a) Die ontwikkeling van die woordeboek The Source Dictionary deur R.P.J. Gerber, Dept. Klassieke Tale, Universiteit Zululand, wat gebruik kan word saam met enige woordverwerkingspakket vir naslaandoeleindes.

(b) Die ontwikkeling van 'n Afrikaanse tesourus met 200 000 woorde en 67 000 kruisverwysings (bykans gereed) asook 'n woordeboek vir Grieks/Afrikaans/Engels in die Griekse en Romeinse alfabette, eweneens deur Gerber.

(c) Die ontwikkeling van 'n program met die naam TEKSANA deur De Klerk, Pretoria, wat 'n teks kan deursoek en inligting daaruit kan onttrek oor die frekwensie, variasie en verspreiding van vorme daarin (vgl. Inligtingstuk 8).

(d) Verskeie gerekenariseerde itembanke is beskikbaar:
  (i) frekwensietellings (o.a. by die SAW, UPE, PU vir CHO, RAU en US),
  (ii) termbanke/terminologiese databasisse (o.a. by die SAW, die WNNR, die RGN, die SA Akademie, die Nasionale Vakterminologiediens van die Dept. Nasionale Opvoeding, die SABS, ens.)

(iii) woordbanke (o.a. die SAW, die RGN, die SAUK, Tafelberg-uitgewers, UPE, UP, Zululand). (Vgl. die DOCNET-verslag, RGN, vir verdere besonderhede.)

Benewens laasgenoemde databasisse is daar etlike ongerekenariseerde versamelings ook beskikbaar, vgl. die DOCNET-verslag, RGN.

8.3.4 Variasieondersoek

Die Departement Afrikaans by die Universiteit van Pretoria en die Departement Afrikaans en Nederlands by die Universiteit van Zululand is die enigste wat tans betrokke is by rekenaargesteunde variasieondersoeke. Enkele publikasies en 'n paar referate by internasionale kongresse, het reeds uit hierdie werk gevolg.

Afgesien van die PC-weergawe van VARBRUL wat by hierdie twee universiteite gebruik word, beskik die Universiteit van Pretoria ook oor 'n hoofraamweergawe daarvan wat deur hierdie universiteit se Buro vir Rekenaardienste in samewerking met sy Departement Afrikaans ontwikkel is op basis van 'n weergawe daarvan wat oorspronklik vir 'n mini-rekenaar geskryf is. Die verwerkte weergawe is in PL/I geskryf.

Die Universiteit van Pretoria beskik ook oor 'n kopie van VARBRUL 3. Hierdie kopie bestaan tans nog net op PC-disket, en moet nog op hoofraam geplaas en operasioneel gemaak word.

8.3.5 Psigolinguistiese ondersoek

Die enigste instansie in die land wat met rekenaargesteunde navorsing oor kindertaal besig is, is die RGN.

8.3.6 Tekskorpusse

Die volgende gerekenariseerde korpusse is na ons wete beskikbaar in die RSA:

(a) Die RGN se kindertaalkorpus.

(b) Masjienleesbare tekste van alle publikasies van Nasionale Pers sedert 1980 sowel as van die Nuwe Afrikaanse Bybelvertaling.

(c) Gerekenariseerde bandopnames (PU vir CHO, RAU, UP, US en Zululand) en teksdatabasisse (SAW, PU vir CHO).

(d) 'n Ongerekenariseerde lys taalgebruiksprobleme (US).

(e) Ongerekenariseerde tekskorpusse (UP, Zululand, die parlement, die WAT).

Vir meer inligting oor hierdie korpusse vergelyk die DOCNET-verslag, RGN.

Aangesien tekskorpusse so belangrik is as hulpbronne vir fonologiese, morfologiese, sintaktiese, leksikale en semantiese taalondersoek, is die Departement Afrikaans, UP, tans besig om 'n navorsingsprojek voor te berei

wat aandag sal skenk aan die ontwikkeling van 'n nasionale stelsel vir gerekenariseerde Afrikaanse tekskorpusse. Hierdie projek sal onder andere die volgende aspekte aanspreek:

(a) Die vereistes waaraan 'n tekskorpus moet voldoen ten opsigte van omvang en verteenwoordigendheid.

(b) 'n Eenvormige transkripsiestelsel.

(c) 'n Eenvormige stelsel vir die linguistiese prosessering van tekste en die analise daarvan.

(d) Effektiewe inligtingsherwinningsmetodes.

(e) Die apparatuur en programmatuur wat nodig is vir (c) en (d).

(f) Die vereistes waaraan die ontwikkeling van 'n nasionale rekenaarnetwerk waarmee navorsers by die tekskorpusse kan inskakel, moet voldoen. Aandag sal onder andere aan vrae soos die volgende gegee word: watter apparatuur en programmatuur is nodig, en hoe kan houers se regte (soos kopiereg) beskerm word.

8.4 REKENAARTOEPASSINGS IN TAALONDERSOEK IN SUID-AFRIKA: 'N EVALUASIE

Evalueer mens Suid-Afrikaanse rekenaartoepassings in taalondersoek aan die hand van die voorafgaande is dit duidelik dat ons 'n agterstand het, vergeleke met die vlak van aktiwiteit in die buitewêreld. Die vinnige tempo waarteen rekenaartoepassings in taalondersoek (en, meer nog, rekenaarlinguistiek) oorsee ontwikkel, kan meebring dat ons meer en meer geïsoleer kan raak van buitelandse vakkundiges op hierdie terrein. Indien ons dus wil voortgaan om deel te hê aan die internasionale taalkunde, sal ons ingrypende stappe moet neem om rekenaartoepassings in taalondersoek hier plaaslik te ontwikkel.

HOOFSTUK 9: AANBEVELINGS

Die aanbevelings hieronder is ten dele gebaseer op die aanbevelings wat in die oorspronklike werksdokument as deel van die onderskeie hoofstukke voorgekom het. Aangesien oorvleueling in heelparty gevalle voorgekom het, is daar besluit om die aanbevelings in hierdie slothoofstuk saam te vat.

9.1 DIE STIGTING VAN 'N NASIONALE WERKGROEP

(a) 'n Nasionale werkgroep vir rekenaarlinguistiek moet so gou moontlik tot stand gebring word. Voordat dit kan gebeur, moet oorsese institute vir rekenaarlinguistiek besoek word. Gevestigde institute van hierdie aard is die Linguistic Research Center van die Universiteit van Texas in Austin, wat sedert 1961 reeds bestaan, die Institut für deutsche Sprache en institute by Birmingham, Lancaster en Leeds in die VK, en Nijmegen in Nederland.

(b) So 'n werkgroep moet hom enersyds rig op die probleme van rekenaargesteunde vertaling en kunsmatige intelligensie, en andersyds op die stimulering, ondersteuning en koördinering van navorsing op hierdie twee gebiede en op die gebied van rekenaartoepassings in taalondersoek, by Suid-Afrikaanse universiteite, technikons, ander navorsingsinstellings en die privaatsektor.

(c) Die spesifieke opdragte van so 'n werkgroep kan die volgende behels:

(i) Gereelde skakeling met oorsese institute sodat die jongste inligting verkry kan word oor ontwikkelinge op die gebied van apparatuur, programmatuur en relevante taalkundige navorsing. Daar moet ook gepoog word om aktiewe deelname aan netwerke soos ICAME te bewerkstellig.

(ii) Die totstandbrenging en instandhouding van 'n nasionale gerekenariseerde kommunikasienetwerk waarby alle belanghebbende persone (linguiste, rekenaarkundiges, verbruikers) ingeskakel kan word.

(iii) Die totstandbrenging en uitbouing van 'n dokumentasiesentrum waarin alle beskikbare programmatuur (en inligting oor programmatuur), inligting oor apparatuur, en boeke, tydskrifte en navorsingsverslae oor die rekenaarlinguistiek en rekenaartoepassings in taalondersoek gehuisves word, en vryelik tot die beskikking van navorsers via die genoemde gerekenariseerde netwerkstelsel gestel word.

(iv) Die uitgee van 'n gereelde nuusbrief waarin die nuutste ontwikkelings hier te lande en ook oorsee bespreek word, sodat inligtingsuitruiling op 'n deurlopende basis kan plaasvind.

(v) Die werwing van navorsingsopdragte en -fondse by die staat en die privaatsektor, en die werwing van navorsers vir hierdie opdragte.

(vi) Die stimulering, ondersteuning, koördinering en kontrole van navorsing (byvoorbeeld met betrekking tot finansiering, hulpdienste - soos gesofistikeerde apparatuur, bibliografiese inligting, ensovoorts).

(vii) Die evaluering van nuwe programmatuur wat beskikbaar raak.

(viii) Die bevordering van die rekenaarlinguistiek en rekenaartoepassings in taalondersoek in die algemeen.

(d) Wat die plasing van die sekretariaat van so 'n werkgroep betref, kan aan die RGN gedink word. Die redes hiervoor is dat die RGN reeds informele maar effektiewe skakeling met alle universiteite, technikons en ander navorsingsinstellings opgebou het, sowel as met regeringsinstansies, reeds heelwat ter sake kundigheid opgebou het in verband met die stimulering en koördinering van navorsing; en verder ook nog sentraal geleë is ten opsigte van Suid-Afrika se belangrikste sake- en nywerheidskompleks.

(e) Ons wil voorstel dat spesifieke sentra geselekteer moet word vir opleiding en navorsing op die gebied van rekenaarlinguistiek en rekenaartoepassings in taalondersoek. Die beperking tot spesifieke sentra is finansieel van aard: Suid-Afrika kan die oormatige duplikasie van fasiliteite en mannekrag waarskynlik nie bekostig nie. Die basis vir die seleksie van hierdie sentra sou geografies, ekonomies en vakkundig kon wees.

9.2 DIE BEVORDERING VAN REKENAARLINGUISTIEK EN REKENAARTOEPASSINGS IN TAALONDERSOEK

Een van die redes vir die onbevredigende stand van rekenaarlinguistiek en rekenaartoepassings in taalondersoek in Suid-Afrika is waarskynlik dat linguiste so min kennis dra van die moontlikhede wat die rekenaar bied op hul gebied. Daarom moet eksplisiete pogings aangewend word om die nut van die rekenaar in teoretiese en praktiese terme aan alle belanghebbendes te demonstreer. Ons meen hierdie doelwit kan met die volgende bereik word:

(a) Uitnodigings aan oorsese deskundiges in die rekenaarlinguistiek en rekenaartoepassings in taalondersoek om Suid-Afrika te besoek en hier opleidingseminare vir linguiste te kom hou, en om saam met geselekteerde vakkundiges navorsingsprojekte te onderneem.

(b) Gereelde besoeke deur Suid-Afrikaanse linguiste aan oorsese kongresse oor die rekenaarlinguistiek en rekenaartoepassings in taalondersoek asook aan oorsese institute en universiteite, maar sonder die verpligting om referate by hierdie kongresse, universiteite of institute te moet lewer. Daarvoor is ons vlak van kundigheid nog nie voldoende ontwikkel nie. (Sien ook 9.3(c) hieronder.)

(c) Seminare oor die rekenaarlinguistiek en rekenaartoepassings in taalondersoek, te reël vir linguiste, rekenaarkundiges, elektroniese ingenieurs en persone uit die privaatsektor, sodat hulle van mekaar se kundighede en probleme kan verneem en die basis gelê kan word vir gesamentlike projekte.

(d) Die propagering van die rekenaarlinguistiek en rekenaartoepassings in taalondersoek by die jaarlikse kongresse van die Linguistevereniging van Suider-Afrika.

(e) Die instelling en gereelde verspreiding van 'n nuusbrief oor die rekenaarlinguistiek en rekenaartoepassings in taalondersoek (soos ICAME) aan geïnteresserde vakkundiges. (Sien ook 9.1(c)(iv).)

9.3 OPLEIDING

'n Voorvereiste vir sukses ten opsigte van die voorafgaande is 'n kennis van die rekenaarlinguistiek, rekenaartoepassings in taalondersoek en die rekenaar self, en 'n positiewe houding daarteenoor. Opleiding op hierdie terrein is dus van die allergrootste belang. Benewens die voorstelle wat in die vorige paragraaf gemaak is (veral punte (a) tot (c)), kan die volgende ook oorweeg word:

(a) Die indiensopleiding van 'n klein groepie geselekteerde linguiste deur kundige kollegas asook rekenaarwetenskaplikes tydens kort plaaslike opleidingsprogramme. Hierdie opleiding moet rekenaargeletterdheid, programmeringstale en die rekenaarlinguistiek en rekenaartoepassings in taalondersoek insluit. Kundige kollegas behoort buitelandse spesialiste in te sluit. 'n Vraag is natuurlik wie hierdie 'geselekteerde linguiste' moet wees. Onses insiens moet daar veral (maar nie uitsluitlik nie) aan belowende jong persone gedink word, sodat hul opleiding as 'n werklike langtermynbelegging kan dien.

(b) Daar moet gedink word aan die moontlikheid om beurse vir voltydse oorsese graadstudie in die rekenaarlinguistiek aan jong gegradueerdes/persone met 'n honneursgraad (of 'n magistergraad) in die taalkunde of die rekenaarwetenskap beskikbaar te stel.

(c) Daar moet deur die RGN en universiteite voorsiening gemaak word vir die finansiële ondersteuning van navorsers om op 'n gekoördineerde wyse opleidingskursusse, werkswinkels en kongresse op die gebied van die rekenaarlinguistiek en rekenaartoepassings in taalondersoek by te woon ten einde kundigheid alhier op te bou vir die ontwikkeling van plaaslike sisteme en die opleiding van belangstellende kollegas. Dit is wenslik dat finansiële ondersteuning vir die bywoning van geselekteerde opleidingsprogramme en kongresse sonder die huidige beperkende voorwaarde toegestaan word dat die navorser self ook 'n referaat aldaar moet lewer. Die punt is dat juis omdat 'n persoon nie op die huidige tydstip 'n bydrae te lewer het op 'n veld wat waarskynlik nuut vir hom is nie, hy daarom verhoed word om enige ervaring en akademiese kontak op die betrokke veld op te bou. Derhalwe bly die kans skraal dat hy ooit met die nodige selfvertroue aldaar 'n bydrae sal kan lewer. Dieselfde argument geld ook vir opleidingsprogramme van tot ses weke soos dié van die Linguistic Institute van die Linguistic Society of America. Langtermynbelegging op die gebied word bepleit.

(d) Departemente in die algemene taalwetenskap, tale, rekenaarwetenskap en elektroniese ingenieurswese aan universiteite en technikons moet bewus gemaak word van die voordele vir hul opleidingsprogramme van onderlinge skakeling en interdissiplinêre opleidings- of navorsingsprogramme, en studente moet gewys word op die sinvolheid van die kombinasie van rekenaarwetenskap en 'n taal/algemene taalwetenskap. Universiteite moet gevolglik ook versoek word om rekenaarwetenskap in kombinasie met 'n taal/algemene taalwetenskap as 'n kombinasie vir die B.A.-graad toe te laat.

(e) Dosente in die algemene taalwetenskap (asook in Afrikaans, Engels en die Afrikatale) moet aangemoedig word om in hul voorgraadse en nagraadse opleidingsprogramme meer aandag te gee aan aspekte van hul vakke wat aansluit by rekenaarlinguistiek en rekenaartoepassings in taalondersoek. Voorbeelde van dergelike aspekte is: die Generalized Phrase Structure Grammar (GPSG) en Lexical Functional Grammar (LFG), asook die instrumentele fonetiek. Daar kan selfs daaraan gedink word dat aspekte van die rekenaarlinguistiek as vaste gedeeltes in die voorgraadse leerplanne van algemene taalwetenskapdepartemente ingesluit word. Dosente wat hierdie vernuwings in hulle kursusse behoort aan te bring, besit oor die algemeen waarskynlik self nog nie die nodige kennis nie en daarom is opleiding so noodsaaklik.

(f) Alle taaldepartemente moet aangemoedig word om hul studente op nagraadse vlak 'n mate van rekenaargeletterdheid te gee en op te lei in die gebruik van die rekenaar vir navorsingsdoeleindes.

9.4 NAVORSING

(a) Basiese navorsing in die rekenaarlinguistiek en rekenaartoepassings in taalondersoek moet aangemoedig word; vanweë die huidige gebrek aan kundigheid in die RSA moet navorsing op hierdie stadium onses insiens egter hoofsaaklik gerig wees op die toetsing, aanpassing en toepassing van programmatuur wat oorsee ontwikkel is (byvoorbeeld stelsels vir die outomatiese analise van taalvorme se morfologiese en sintaktiese aspekte).

(b) Gesien ons genoemde agterstand moet navorsing in sowel rekenaarlinguistiek as rekenaartoepassings in taalondersoek liefs in die vorm van (interdissiplinêre) spanwerk aangepak word. Sodoende kan die kundighede van verskeie persone tot mekaar se voordeel ingespan word. Navorsing deur geïsoleerde klein groepies binne 'n departement of 'n universiteit lei nie tot optimale resultate nie. Die aard van die terrein en die tempo van ontwikkelings bring mee dat kundigheid interdissiplinêr, en oor departements- en universiteitsgrense heen, saamgesnoer moet word.

(c) Gesien die onontbeerlike rol wat die linguistiek in rekenaargesteunde vertaling en kunsmatige intelligensie speel, moet rekenaarwetenskaplikes en elektroniese ingenieurs aangemoed word om linguiste by hul projekte te betrek.

(d) Rekenaartoepassings in taalondersoek is direk afhanklik van die beskikbaarheid van uitgebreide, gerekenariseerde databasisse. Hierdie feit hou vir taalondersoek in Suid-Afrika onder meer die volgende in:

(i) dat (bestaande) datakorpusse vir Afrikaans, Suid-Afrikaanse Engels en die groot Afrikatale so gou as moontlik getranskribeer word en op rekenaar geplaas word. Daar sou byvoorbeeld op die CHiLDES-model 'n nasionale korpusbank opgebou kon word. Mits die deelnemers die materiaal in 'n ooreengekome formaat voorlê, hoef die bedryf van so 'n bank nie finansieel veeleisend te wees nie;

(ii) dat gesorg moet word dat die betrokke korpusse die hele spektrum van taalgebruiksfunksies en -genres (gesproke en geskrewe), sowel as verskillende registers, sosiale en geografiese dialekte, en style verteenwoordig;

(iii) dat 'n teksformateringstelsel verkry (of ontwikkel) word waarmee ortografies getranskribeerde tekste voorberei kan word sodat taalkundige analise regstreeks op die datakorpusse uitgevoer kan word; en

(iv) dat 'n stelsel tot stand gebring word wat meewerkende linguiste vrylik toegang gee tot die sentraal versamelde datakorpusse. So 'n stelsel kan een van die volgende behels: 'n nasionale rekenaarnetwerk, drukstukke van aangevraagde konkordansies, kompakteskyf plus die nodige afleesprogrammatuur.

(e) Gesien die vlak van Suid-Afrikaanse linguiste se kennis van die gebruik van die rekenaar vir taalondersoek en die geringe mate waarin die rekenaar vir linguistiese navorsing gebruik word, is dit moeilik om 'n geordende lys van navorsingsprojekte vir rekenaartoepassings in taalondersoek voor te

stel. Dit is waarskynlik verstandig om die gebiede waarop daar reeds rekenaargesteunde ondersoek plaasgevind het, uit te bou, terwyl daar ook aandag gegee word aan verdere terreine wat in die Suid-Afrikaanse situasie veral pertinent is, byvoorbeeld die korpuslinguistiek.

(f) Bestaande navorsing behoort ondersteun te word. In die lig van die geld wat reeds aan apparatuur en programmatuur bestee is, en die vordering wat reeds getoon word, is dit belangrik om te voorkom dat die huidige finansiële druk dalk lei tot die staking van navorsingsprojekte. Simpatieke oorweging van voorleggings vir die voortsetting van bestaande navorsing, en van voorleggings deur nuwe navorsers, is wenslik.

(g) Navorsing moet egter in alle gevalle goed en realisties beplan word, met goedomlynde doelwitte wat op die korttermyn haalbaar is en wat binne die raamwerk van doelgerigte langtermynbeplanning geplaas is. Oorambisie en 'n gebrek aan kundigheid kan tot groot terugslae lei, veral as verwagtinge wat by fondstoestaners geskep word, nie bevredig kan word nie. Kundigheid en ervaring vir groter projekte kan uitgebou word deur die aanpak van projekte van beperkte omvang.

9.5 FINANSIERING

(a) In die lig daarvan dat die moderne beskawing besig is om vanaf die nywerheidsera oor te beweeg na die inligtingsera, gaan sowel die rekenaar as die studie van natuurlike tale ongetwyfeld 'n al hoe belangriker rol speel in die mens se alledaagse bestaan. Die rekenaarlinguistiek sowel as rekenaartoepassings in taalondersoek is dus belangrike studiegebiede. Hieruit volg dat rekenaarlinguistiek en rekenaartoepassings in taalondersoek as prioriteitsgebiede in ons toekomstige nasionale navorsingsprogram geïdentifiseer moet word, en dat fondse vir taalgerigte navorsing van owerheidskant dáárvoor aangewend moet word.

(b) Die staat is natuurlik nie die enigste bron van sodanige navorsingsfondse nie want die privaatsektor het direkte belang by die gebied. Meganismes behoort dus geskep te word waardeur die staat, die RGN, die WNNR, die MNR, die universiteite en private maatskappye meer direk betrek word in die gesamentlike finansiering van veral navorsingsprojekte wat direk in die praktyk toepasbaar is en waarby alle partye uiteindelik kan baat. Gesamentlike navorsingsooreenkomste soos dié van die BMFT in Duitsland sou tot almal se voordeel kon strek. Ter operasionalisering word hier byvoorbeeld gedink aan:

- belastingverligting aan privaatinstansies wat hul industriële navorsing aan en deur universiteite en navorsingsrade doen of laat doen;
- spesifieke opdragte deur staatsdepartemente vir die ontwikkeling van byvoorbeeld kommunikasiesisteme en databasisse met koppelvlakke in natuurlike tale soos Afrikaans en die groot Afrikatale.

9.6 SLOT

In die lig van die tegnologiese, wetenskaplike, ekonomiese en strategiese belang van rekenaargesteunde taalontleding en -verwerking, en met inagneming van die snelle ontwikkeling op die gebied in die buiteland, is Suid-Afrika se huidige agterstand beslis kommerwekkend. Om dié agterstand in te haal, gaan dinamiese optrede nodig wees.

BIBLIOGRAFIE

AARTS, J., MEIJS, W. (eds.) 1983. *Corpus Linguistics.* Amsterdam: Rodopi

AARTS, J., VAN DEN HEUVEL, T. 1985. Computational tools for the syntactic analysis of corpora. *Linguistics* 23:303-335.

AHO, A.V., ULLMAN, J.D. 1972. *The Theory of Parsing, Translation, and Compiling, Vol. 1: Parsing.* Englewood Cliffs, NJ: Prentice-Hall.

AIJMER, K. 1986. Conversational phrases in the London-Lund corpus. *ICAME Journal* 10:40-43. Bergen: Norwegian Computing Centre for the Humanities.

AKKERMAN, E. et al 1985. Designing a computerized lexicon for linguistic purposes. *Ascot Report No.1 Costerius New Series* Vol. 51, Amsterdam: Rodopi.

ALAM, Y. 1983. A two level Morphology Analysis of Japanese. *Texas Linguistic Forum.* 22:229-252.

APPELT, D.E. 1983. Telegram: a grammar formalism for language planning. *21st Annual Meeting of the Association for Computational Linguistics.* Proceedings. 74-78. Pisa.

ATWELL, E.S. 1982. *LOB Corpus Tagging Project.* University Lancaster.

AUSTIN, J.L. 1962. *How to do things with words.* New York: Oxford University Press.

BACH, E., HARMS, R.T. 1968. *Universals in linguistic theory.* New York: Holt, Rinehart and Winston.

BARLOW, M., FLICKINGER, D.P., SAG, I.A. 1982. Developments in Generalised Phrase Structure Grammar. *Stanford Working Papers in Grammatical Theory,* Vol. 2. Bloomington: Indiana.

BARTON, G.E. 1985. On the complexity of ID/LP parsing. *Computational Linguistics* 11/4:205-218.

BARWISE, J. 1981. Some computational aspects of situation semantics. *Proc. 19th Annual Meeting of the ACL:* 109-111.

BATES, M. 1978. The theory and practice of augmented transition network grammars. *Natural Language Communication with Computers,* Berlin: L. Bolc, ed. Springer.

BAWRON, J.M., et al. 1982. Processing English with a generalized phrase structure grammar. *Proc. 20th Annl. Meeting Assn. Computational Linguistics,* (1982): 74-81.

BELLERT, S. 1974. On various solutions of the problem of presupposition. In: PETOFI, J.S., REISER, H. (eds.) (1974).

BERWICK, R.C. 1982. Computational complexity and Lexical Functional Grammar. *American Journal of Computational Linguistics* 8/3-4:97-109.

BERWICK, R.C. WEINBERG, A. 1982. Parsing efficiency, computational complexity, and the evaluation of grammatical theories. *Linguistic Inquiry* 13:165-191.

BERWICK, R.C., WEINBERG, A.S. 1983. Syntactic constraints and efficient parsability. *21st Meeting of the Annual Meeting of the Association for Computational Linguistics.* Proceedings. Cambridge, Ma.: MIT Press:119-122.

BIRNBAUM, L., SELFRIDGE, M. 1979. Problems in conceptual analysis of natural language. *Research Report* 168, Yale University: Department of Computer Science.

BIRNBAUM, L., SELFRIDGE, M. 1981. *Conceptual analysis of natural language.* In: SCHANK, RIESBECK, CHRISTOPHER (1981).

BLABERG, D. 1984. *Svensk Bojningsmorfologi: en tyaniva beskriyning.* Th. Helsinki.

BLAHA, H. 1980. Aufbau- und Nützungsmöglichkeiten einer Normen-Terminologie-Datenbank. *Fachsprache,* 2(4):146-155.

BOBROW, D.G., COLLINS, A. (eds.) 1975. *Representation and Understanding: Studies in Cognitive Science.* New York: Academic Press.

BOBROW, D., FRASER, B. 1969. An augmented state transition network analysis procedure. *Proc. Int. Joint Conf. on Artificial Intelligence,* Washington DC.

BOBROW, R.J., WEBBER, B.L. 1980. Knowledge representation for syntactic/semantic processing. *First Annl.Natl. Conf. on Artificial Intelligence,* AAAI, Stanford, CA, (1980):316-23.

BOBROW, R.J., WEBBER, B.L. 1980(a). PSI-KLONE - Parsing and Semantic Interpretation in the BBN Natural Language Understanding System. *Proceedings of the CSCSI/SCEIO Conference.*

BOBROW, R.J., WEBBER, B.L. 1980(b). Knowledge representation for syntactic/semantic processing. *First annl. natl. conf. on artificial intelligence,* AAAl, Stanford, CA:316-23.

BOBROW, D., WINOGRAD, T. 1977. An overview of KRL, a knowledge representation language. *Cognitive Science* I, 1(Jan. 1977):3-46.

BOBROW, D.G., WINOGRAD, T. 1977(b). Experience with KRL-O: One cycle of a knowledge representation language. *IJCAI* 5:213-222.

BOBROW, D. *et al.* 1977. GUS, a frame driven dialogue system. Artificial Intelligence 8.2 (April):155-73.

BOGURAEV, B.K. 1979. Automatic resolution of linguistic ambiguities. *Technical Report* 11. Cambridge: Computer Laboratory, University of Cambridge.

BOGURAEV, B.K. 1980. Automatic Resolution of Linguistic Ambiguities. *Technical Report* 11. Cambridge, England: Computer Laboratory, Cambridge University.

BOLC, L. (ed.) 1980. *Representation and Processing of Natural Language.* Munich: Carl Hanser Verlag.

BOOT, M. 1984. *Taal, tekst, computer.* Katwijk: Servire.

BRACHMAN, R.J. 1979. On the epistemological status of semantic networks. In: FINDLER (1979).

BRESNAN, JOAN W. (ed.) 1982. *The mental representation of grammatical relations.* Cambridge, Mass.: MIT Press.

BRILLINGER, P.C., COHEN, D.J. 1972. *Introduction of Data Structures and Non-numeric Computations.* Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

BRODDA, B. 1986. BetaText: an event-driven text processing and text analyzing system. *Coling* 86/ACH.

BRODDA, B., KARLSON, F. 1980. *An experiment with automatic morphological analysis of Finnish.* University Stockholm, Institute of Linguistics Pub. 40

BROWN, G., YULE, G. 1983. *Discourse analysis.* Cambridge: Cambridge University Press.

BRUCE, B.C. 1975. Discourse models and language comprehension. *Am. J. Computational Linguistics* 35:19-35.

BRUSTKERN, J., HESS, K.D. 1982. The BONNLEX Lexicon System. In: GOETSCHALCKX, ROLLING (eds.) (1982):35-40.

BRUSTKERN, J., W. REUDERS, G. WILLÉE. 1981. *Handbuch der Programbibliothek zur linguistischen und philologischen Textverarbeitung.* Westdeutsche Verlag.

BUNDY, A., BYRD, L., MELLISH, C.S. 1982. Special purpose but domain independent inference mechanisms. *ECAI* -82: Proceedings of the 1982 European Conference on Artificial Intelligence:67-74.

BURTON, R. 1976. Semantic grammar: An engineering technique for constructing natural language understanding systems. *BBN Report* 3453, Bolt, Beranek, Newman, Cambridge, MA.

BUTLER, C. 1985. *Computers in Linguistics.* Oxford : Basil Blackwell.

BYRD, J.R. et al. 1986. Computer Methods for Morphological Analysis in 24th *ACL Annual Meeting* - Proceedings.

CAMERON, K.C., DODD, W.S., RAHTZ, S.P.Q. (eds.) *Computers and modern language studies.* Chichester: Ellis Horwood Limited.

CAMPBELL, R., SMITH, T. (eds.) 1976. *Recent advances in the psychology of Language.* New York: Plenum.

CARBONELL, J.G. 1982. Metaphor: an inescapable phenomenon in natural-language comprehension. In: LEHNERT, RINGLE (1982):415-434.

CERCONE, N. 1974. Computer analysis of English word formation. *Technical Report* TR 74-6, Dpt. Computing Science, University Alberta in Edmonton (CND).

CERCONE, N. 1978. Morphological analysis and lexicon design for natural language processing. *Computers and the Humanities* 11.

CHARNEY, E.K. 1962. On the semantic interpretation of linguistic entities that function structurally. *NPL* :544-556.

CHARNIAK, E. 1981. Towards a model of children's story comprehension. *MIT AI Report TR* -266.

CHARNIAK, E. 1981(a). Passing markers: a theory of contextual influence in language comprehension. *Report TR* -80, Department of Computer Science, Brown University.

CHARNIAK, E. 1982. Context recognition in language comprehension. In: LEHNERT, W., RINGLE, M. (eds.).

CHARNIAK, E., WILKS, Y. 1975. (eds.) *Computational Semantics.* Amsterdam: North-Holland.

CHOMSKY, N. 1957. *Syntactic structures.* The Hague: Mouton.

CHOMSKY, N. 1965. *Aspects of the theory of Syntax.* Cambridge, Mass.: MIT Press.

CHOMSKY, N. 1981. Lectures on government and binding. *The Pisa lectures.* Studies in Generative Grammar 9. Dordrecht: Foris Publications.

CHOMSKY, N. 1982. *Some concepts and consequences of the theory of government and binding.* Cambridge, Mass.: MIT Press.

CHOMSKY, N. 1986. *Knowledge of language. Its nature, origin, and use.* New York: Praeger Publishers.

CHURCH, K. 1986. Morphological decomposition and Stress Assignment for Speech synthesis. *24th ACL Annual Meeting,* Proceedings.

CODD, E.F. 1978. How about recently? In: SHNEIDERMAN (ed.) (1978):3-28.

COHEN, P.R., PERRAULT, C.R. 1979. Elements of a plan-based theory of speech acts. *Cognitive science* 3:177-212.

COLE, P., MORGAN, J. (eds.) 1975. *Syntax and semantics 3: Speech acts.* New York: Academic Press.

COLLINS, A., QUILLIAN, M. 1969. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior* 8:240-248.

COMBRINK, J.G.H., DODDS, R.McD. 1984. *Retrograde Woordeboek van Afrikaans.* Stellenbosch: Universiteit van Stellenbosch.

COOPER, R. 1975. *Montague's semantic theory and transformational syntax.* University of Massachusetts at Amherst. (Ph.D. dissertation).

CORSTIUS, H.B. 1978. *Computer-Taalkunde.* Muiderberg: Coutinho.

CULICOVER, P., KIMBALL, J., LEWIS, C., LOVEMAN, D., MOYNE, J. 1969. An automated recognition grammar for English. *IBM Tech. Rep. FSC 69-5007.*

DAHL, V., SAINT-DIZIER, P. (eds.) 1985. *Natural language understanding and logic programming.* Amsterdam: Elsevier Science Publishers.

DAVIDSON, D., HARMAN, G. (eds.) 1972. *Semantics of natural language.* Dordrecht: Reidel.

DE BEAUGRANDE, R.A., DRESSLER, W.U. 1981. *Introduction to text linguistics.* London: Longman.

DE STADLER, L.G., COETZER, M.W. 1988. 'n Afrikaanse skrif-tot-spraak omsetter. *Die Tydskrif vir Taalkunde*:19-35.

DE TOLLENAERE, F. 1963. *Nieuwe wegen in de Lexicologie. Verhandelingen van de Koninklijke Nederlandsche Akademie van Wetenschappen, Nieuwe Reeks, 71:1.* Amsterdam.

DEVONS, N. 1986. Observation on ONE'S in contemporary American English. *ICAME Journal* 11:54-56. Bergen : Norwegian Computing Centre for the Humanities.

DEWAR, H., BRATLEY, P., THORNE, J.P. 1969. A program for the syntactic analysis of English sentences. *Comm. ACM 12,8 (Aug. 1969):476-9.*

DHONDT, Chantal n.d. Contribution au dossier Traduction Automatique. *Systran Doc EEC* (9747).

DOSTERT, B., THOMPSON, F. 1971. How features resolve syntactic ambiguity. *Proc. Symposium on Info. Storage and Retrieval.*

DOWTY, R.D., WALL, R.E., PETERS, S. 1981. *Introduction and Montague Semantics.* Dordrecht: Reidel.

DYER, M.G. 1982. In-Depth Understanding. *Report* 219, Dept. of Computer Science, Yale University.

EARLEY, J. 1970. An efficient context-free parsing algorithm. *Comm. ACM 13,2 (Feb. 1970):94-102.*

FAHLMAN, S.E. 1979. *NETL: A System for Representing and Using Real-World Knowledge.* Cambridge, Mass.: MIT Press.

FEIGENBAUM, E.A., FELDMAN, J. (eds.) 1964. *Computers and Thought.* New York: McGraw-Hill.

FILLMORE, C.J. 1968. The case for case. In: BACH, HARMS (eds.) (1968).

FILLMORE, C.J. 1971. Entailment rules in a semantic theory. In: ROSENBERG, TRAVIS (eds.) (1971).

FILLMORE, C.J., LANGEDOEN, D.T. (eds.). 1971. *Studies in semantics.* New York: Holt.

FINDLER, N.V. (ed.) 1979. *Associative Networks: Representation and Use of Knowledge by Computers.* New York: Academic Press.

FODOR, J.A., FODOR, J.D., GARRETT, M.F. 1975. The psychological unreality of semantic representations. *Linguistic Inquiry* VI:515-535.

FODOR, J.A., KATZ, J.J. (eds.) 1974. *The structure of language: Readings in the philosophy of language.* Englewood Cliffs, New Jersey: Prentice-Hall.

FREY, W. 1985. Noun phrases in Lexical Functional Grammar. In DAHL, SAINT-DIZIER (eds.) (1985):121-137.

FREY, W., REYLE, U. 1983. A Prolog implementation of Lexical Functional Grammar as a base for a natural language processing system. *1st Conference of the European Chapter of the Association for Computational Linguistics.* Proceedings. Pisa:52-57.

FRIEDMAN, J., MORGAN, D., WARREN, D. 1978. An interpretation system for Montague Grammar. *American Journal of Computational Linguistics.* Microfiche 94:23-96.

FUJIMURA, O. (ed.) 1973. *Three dimensions of linguistic theory.* Tokyo: TEC Co.

GAZDAR, G, KLEIN, E, PULLUM, G, SAG, I. 1985. *Generalised phrase structure grammar.* Cambridge, Mass.: Harvard University Press.

GAZDAR, G, PULLUM, G.K. 1982. *Generalised phrase structure grammar:* a theoretical synopsis. Bloomington: Indiana University Linguistics Club.

GAZDAR, G., KLEIN, E., PULLUM, G., SAG, I. 1985. *Generalized phrase structure grammar.* Oxford: Blackwell.

GOETSCHALCKX, J., ROLLING, L. (eds.) 1982. *Lexicography in the electronic age.* Proceedings of a symposium held in Luxemburg, 7-9 July 1981. Amsterdam: North Holland.

GOLDING, A.R., THOMPSON, H.S. 1985. A morphological component for language programs. *Linguistics* 23.

GOLDSTEIN, I.P., ROBERTS R.B. 1977. NUDGE: a knowledge-based scheduling program. *Proc., Fifth Int. Joint Conf. Artificial Intelligence:* 257-63.

GRIFFITH, R.L. 1982. Three principles of representation for semantic networks. *ACM Transactions on Database Systems* 7:417-442.

GRIMES, J.E. 1983. *Affix Positions and Cooccurrences: The PARADIGM Program.* Arlington: The Summer Institute of Linguistics, University of Texas.

GRISHMAN, R. 1973. Implementation of the string parser of English. In: RUSTIN (ed.) (1973).

GRISHMAN, R. 1976. A survey of syntactic analysis procedures for natural language. *American J. Computational Linguistics, microfiche 47.*

GRISHMAN, R. 1979. Response generation in question-answering systems. *Proc. 17th annl. meeting assn. Computational Linguistics:* 99-101.

GRISHMAN, R. 1980. Conjunctions and modularity in language analysis procedures. *Proc. 8th int. conf. Computational Linguistics:* 500-3.

GRISHMAN, R. 1986. *Computational Linguistics.* Cambridge: Cambridge University Press.

GRISHMAN, R., KITTREDGE, R. (eds) 1986. *Analysing language in restricted domains; sublanguage description and processing.* Hillsdale: Lawrence Erlbaum Association.

GRISWOLD, R.E., GRISWOLD, M.T 1973. *A SNOBOL4 Primer.* New Jersey: Prentice-Hall, Inc. Englewood Cliffs.

GRISWOLD, R.E., POAGE, J.F., POLONSKY, I.P. 1971. *The SNOBOL4 Programming Language.* Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

GUNJI, T. 1981. *Toward a computational theory of pragmatics - discourse, presupposition, and implicature.* Ohio: Ohio State University. (Ph.D. dissertation).

GUNJI, T., SONDHEIMER, N. 1980. The mutual relevance of model-theoretic semantics and artificial intelligence. SMIL. *Journal of Linguistic Calculus* 3:5-42.

GUS, a frame-driven dialog system. *Artificial Intelligence* 8, 2(Apr. 1977):155-73.

HAJI*COVA, E., SGALL, P. 1981. Towards automatic understanding of technical texts. *Prague Bulletin of Mathematical Linguistics* 36:5-23.

HALLIDAY, M.A.K. 1956. The linguistic basis of a mechanical thesaurus. *Mechanical Translation* 3(3):81-88.

HALLIDAY, M.A.K., HASAN, R. 1976. *Cohesion in English.* London: Longman.

HANN, M.L. 1974. Principles of automatic lemmatization. *ITL Review of Applieed Linguistics* 23.

HANN, M.L. 1978. *The application of computers to the production of systematic, multilingual specialised dictionaries and the accessing of semantic information systems.* Report University of Manchester UMIST Centre for Computational Linguistics 78/1.

HARRIS, Z. 1962. *String Analysis of Sentence Structure.* The Hague: Mouton and Co.

HARTMANN, R.R.K. (ed.) 1983. *Lexicography: Principles and Practice.* Academic Press.

HAYES, P.J. 1979. The logic of frames. In: METZING (ed.) (1979).

HAYS, D. 1967. *Introduction to Computational Linguistics.* New York: American Elsevier.

HEINZE, G., BRUSTKERN, J. 1983. Das Lexikonsystem ALEXSYS. *IKP Arbeitsberichte. Abteilung LDV,* nr. 6.

HELLBERG, S. 1972. Computerized lemmatization without the use of a dictionary. *Computers and the Humanities* 6/4.

HENDRIX, G.G. 1975. Expanding the utility of semantic networks through partitioning. *Proc. IJCAI* -75:115-121.

HENDRIX, G.G. 1977. Human engineering for applied natural language processing. *Proc. 5th int. conf. Artificial Intelligence.* Cambridge.

HENDRIX, G.G. 1978. Semantic knowledge. In: WALKER (ed.) (1978):121-226. 121-226.

HENDRIX, G.G. et al. 1978. Developing a natural language interface to complex data. *ACM Transactions on Database Systems* 3:105-147.

HESS, K., BRUSTKERN J., LENDERS, W. 1983. *Maschinenlesbare deutsche Wörterbücker.* Tübingen: Max Niemeyer.

HIRSCHMAN, L., SAGER, N. 1982. Automatic information formatting of a medical sublanguage. *Sublanguage: Studies of language in restricted semantic domains.*

HOBBS, J. 1982. Towards an understanding of coherence in discourse. *Strategies for natural language processing.* Hillsdale: Lawrence Erlbaum Association.

HOBBS, J., ROSENSCHEIN, S. 1978. Making computational sense of Montague's intensional logic. *Artificial Intelligence* 9:287-306. HOCKEY, S. 1980. *A guide to computer applications in the humanities.* Baltimore: The Johns Hopkins University Press.

HOEPPNER, W.H. 1982. A multi layered approach to the handling of word formation. *Coling* 82.

HOEPPNER, W.H. et al. 1984. Beyond domain independence: experience with the development of a German access system to highly diverse background systems. *Proc. IJCAI* 8.

JACKENDOFF, R. 1972. *Semantic interpretation in generative grammar.* Massachusets: MIT Press.

JAEPPINEN, H. et al. 1983. 1. Knowledge engineering approach to morphological analysis 2. Morphological analysis of Finnish: a heuristic approach. *Report B 26 ACL,* European Chapter, Pisa.

JAEPPINEN, H. et al. 1986. Associative model of morphological analysis: an empirical inquiry. *Computational Linguistics* 12/4.

JOHANSSON, S. 1982. Grammatical Tagging of the LOB Corpus: predicting word class from word endings. In: JOHANSSON (ed.) (1982).

JOHANSSON, S. (ed.) 1982. *Computer Corpora in English Language Research.* University Bergen.

JOHANSSON, S. 1987. *The new Oxford English Dictionary Project.* University Oslo.

JONES, K.S., WILKS, Y. (eds.) 1983. *Automatic Natural Language Parsing.* Chichester: Ellis Horwood Limited.

JONES, V. 1986. A query language for semantic network databases. *ICAME JOURNAL* 10:56-59. Bergen: Norwegian Computing Centre for the Humanities.

JOSHI, A.K. et al. (eds.) 1981. *Elements of Discourse Understanding.* Cambridge: Cambridge University Press.

KAPLAN, R., BRESNAN, JOAN 1982. Lexical functional grammar: a formal system for grammatical representation. In: BRESNAN (ed.) (1982):173-281.

KARTTUNEN, L. 1984. Features and values. *10th International Conference on Computational Linguistics.* Proceedings of Coling 84:28-33 Stanford.

KARTTUNEN, L., WITTENBURG K. 1983. A two-Level Morphological Analysis of English - *Texas Linguistic Forum* 22.

KARTTUNEN, L. et al. 1981. Texfin: morphological analysis of Finnish by Computer - *Proc. 21st Meeting of the SASS in Albuquerque NM.*

KATZ, J.J. 1966. *The philosophy of language.* New York: Harper.

KATZ, J.J. 1970. Interpretive semantics vs. generative semantics. *Foundations of Language* 6:220-259.

KATZ, J.J. 1971. Generative semantics IS interpretive semantics. *Linguistic Inquiry* II:313-331.

KATZ, J.J. 1972. *Semantic theory.* New York: Harper.

KATZ, J.J., FODOR, J.A. 1963. The structure of a semantic theory. *Language* 39:170-210.

KAY, M. 1970. Experiments with a powerful parser - *Coling* 67.

KAY, M. 1973. The Mind System. In: RUSTIN (1973).

KAY, M. 1973(a). Morphological analysis. In: ZAMPOLLI, CALZOLARI (eds.) (1973).

KAY, M. 1983. Unification grammar. *Technical report, Xerox Palo Alto Research Center.* Palo Alto, Calif.

KAY, M. 1984. Functional Unification Grammar: a formalism for machine translation. *10th International Conference on Computational Linguistics.* Proceedings of Coling 84: 75-78. Stanford.

KAY, M. 1985. Unification in grammar. In: DAHL, SAINT-DIZIER (eds.) (1985):233-240.

KAY, M. 1985(a). Two level Morphology with Tiers. *OSLI Workshop.* Stanford.

KAY, M. 1987. Non concatenative finite state morphology. *ACL european Chapter.* Copenhagen.

KEENAN, E. 1975. *Formal semantics of natural language.* Cambridge: Cambridge University Press.

KEYSER, S.J., PETRICK, S.R. 1967. Syntactic analysis. *Report AFCRL-67-0305.* Air Force Cambridge Research Laboratories.

KLOPPER, R.M. 1976. *Sosiaal gestratifiseerde taalgebruik in die Kaapstadse Kleurlinggemeenskap - 'n fonologiese ondersoek.* M.A.-verhandeling, Universiteit Stellenbosch.

KLOPPER, R.M. 1983. *Kaapse Afrikaans.* Proefskrif, Universiteit van Pretoria.

KOSKENIEMMI, K. 1983. *Two level morphology: A general Computational Model for word form recognition and production.* Univ. Helsinki Dept. of General Linguistics Publ. II.

KOTZÉ, E.F. *Variasiepatrone in Maleier-Afrikaans.* Proefskrif, Universiteit van die Witwatersrand.

KUHLEN, R. 1987. *Informationslinguistik.* Tübingen: Max Niemeyer.

KUNO, S., OETTINGER, A.G. 1962. Multiple-path syntactic analyzer. In: *Information Processing 1962,* Amsterdam: North-Holland.

LABOV, W. 1973. *The boundaries of words and their meanings. New ways of analyzing variation in English.* Georgetown.

LABOV, W., LABOV, TERESA, 1976. Learning the syntax of questions. In: CAMPBELL, SMITH (eds.) (1976).

LABOV, W., YAEGER, M., STEINER, R. 1972. A quantitative study of sound change in progress. *Report NSF* GS-3287. U. Penn: Philadelphia.

LAKOFF, G. 1971. On generative semantics. In: STEINBERG, JAKOBOVITS (eds.) (1971).

LANHAM, L.W, MACDONALD, C.A. 1979. *The standard in South African English and its social history.* Heidelberg: Julius Groos Verlag.

LEHNERT, W. 1982. Plot units: a narrative summarization strategy. In: LEHNERT, RINGLE (eds.) (1982).

LEHNERT, W., RINGLE, M.H. (eds.) 1982. *Strategies for natural language processing.* Hillsdale: Lawrence Erlbaum Associates.

LESMO, L, TORASSO, P. 1983. A flexible natural language parser based on a two-level representation of syntax. *1st Conference of the European Chapter of the Association for Computational Linguistics.* Proceedings. Pisa.

LEUNBACH, G. 1973. Morphological analysis as a step in automated syntactic analysis of a text. In: ZAMPOLLI, CALZOLARI, (eds.) (1973).

LUN, S. 1983. A two level analysis of French. *Texas Linguistic Forum* 22.

MARCHUK, Y.N. 1985. *Melody Modelirovanja Perevoda.* Moskva: Nauk.

MARCUS, M. 1980. *Theory of Syntactic Recognition for Natural Language.* Cambridge: MIT Press.

MARCUS, M., HINDLE, D., FLECK, M. 1983. D. Theory: Talking about trees. *Proc. 21st Meeting Assn. Computational Linguistics.* (1983):129-36.

MARSH, E., SAGER, N. 1982. Analysis and processing of compact text. *COLING 82,* North-Holland, New York:201-206.

MARSH, E. 1983. Utilizing domain-specific information for processing compact text. Proc. conf. appl. lang. processing. Santa Monica.

MARTIN, W.J.R., Al, B.P.F., VAN STERKENBURG, P.J.G. 1983. On the processing of a text corpus. In: HARTMANN (ed.) (1983):77-87.

McCAWLEY, J.D. 1968. The role of semantics in a grammar. In: BACH, HARMS (eds.) (1968).

McCAWLEY, J.D. 1971. Interpretive semantics meets Frankenstein. *Foundations of Language* 7:285-296.

McCAWLEY, J.D. 1973. Syntactic and logical arguments for semantic structures. In: FUJIMURA (ed.) (1973).

MCKEOWN, K. 1982. The TEXT System for natural language generation: an overview. *Proc. 20th Annl. Meeting Assn. Computational Linguistics, (1982):113-20.*

MCKEOWN, K. 1985. *Text Generation.* Cambridge, England: Cambridge Univ. Press.

METZING, D. (ed.) 1979. *Frame Conceptions and Text Understanding.* Berlin: de Gruyter.

MINSKY, M. (ed.) 1968. *Semantic Information Processing.* Cambridge, Mass.: MIT Press.

MINSKY, M. 1975. A framework for representing knowledge. *The psychology of computer vision.* New York: McGraw-Hill.

MODGIL, S., MODGIL C. (eds.) 1987. *Noam Chomsky: Consensus and Controversy.* London: Falmer Press.

MONTAGUE, R. n.d. Universal Grammar. *Theoria* 36:373-398.

MONTAGUE, R. 1974. Formal philosophy: Selected papers of Richard Montague. In: THOMASON (ed.) (1974).

MOORE, K. 1983. *Text processing in dictionary compilation.* In: SNELL (ed.) (1983):134-136.

NAGAO, M. 1986. Machine Translations from Japanese into English. *Proc. IEEE* 74/7 1986.

NPL. 1962. *1961 International Conference on Machine Translation of Languages and Applied Language Analysis: proceedings of the conference held at the National Physical Laboratory, Teddington, Middlesex, on 5th, 6th, 7th and 8th September.* 2 vols. London: HMSO.

OOSTDIJK, NELLEKE 1983. An extended affix grammar for the English noun phrase. In: AARTS, MEIJS (eds.) (1983).

O'SHEA, T., EISENSTADT, M. 1984. *Artificial Intelligence. Tools, techniques and applications.* New York: Harper and Row.

PARTEE, B. 1975. Montague grammar and transformational grammar. *Linguistic Inquiry* VI:203-300.

PARTEE, B. 1976. *Montague grammar.* New York : Academic Press.

PEREIRA, F.C., WARREN, H.D. 1983. Parsing as deduction. *21st Annual Meeting of the Association for Computational Linguistics.* Proceedings:137-144. MIT Press.

PEREIRA, F.C.N., SHIEBER, S.M. 1984. The semantics of grammar formalisms seen as computer languages. 10th International *10th International Conference on Computational Linguistics.* Proceedings of Colling 84:123-129. Stanford.

PERRAULT, C.R., ALLEN, J.F. 1980. A plan-based analysis of indirect speech acts. *American Journal of Computational Linguistics* 6,3-4. (July-Dec. 1980):167-82.

PETERS, S. 1972. *Goals of linguistic theory.* New York : Prentice Hall.

PETOFI, J.S., RIESER, H. (eds.) 1974. *Studies in text grammar.* Dordrecht: Reidel.

PETRICK, S.R. 1965. A recognition procedure for transformational grammars. MIT. (D. Phil. dissertation.)

PETRICK, S.R. 1966. A program for transformational syntactic analysis. *Report AFCRL-66-698.* Air Force Cambridge Research Labs.

PETRICK, S.R. 1977. Semantic interpretation in the REQUEST system. *Proc. International Conference on Computational Linguistics.* Pisa:585-610.

PIGOTT, J. 1979. Theoretical options and practical limitations of using semantics to solve problems of natural language analysis and machine translation. *DOC EEC* (118/79). PIGOTT, J. 1982. SYSTRAN - a reality of the Translation World of the 1980's. *DOC. EEC*

PLATH, W.J. 1974. String transformations in the REQUEST system. *IBM T.J.* Watson Research Center, Rep. RC 4947 (21963).

PLATH, W.J. 1976. REQUEST: a natural language question-answering system. *IBM J. of Research and Development* 20:326-335.

POLLARD, C. 1987. *Generalised phrase structure grammars, head grammars and natural language.* Cambridge.

POPLACK, S. n.d. *The care and handling of a mega-corpus: the Ottawa-Hull French project.*

POSTAL, P.M. 1972. The best theory. In: PETERS (ed.) (1972).

QUARTERLY REPORT Vol. 2 No. 1. September 1987. Language research laboratories. Washington D.C.: Georgetown University.

RADFORD, A. 1981. *Transformational syntax. A student's guide to Chomsky's Extended Theory.* Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press.

RALLI, A., GALLIOTOU, E. 1987. Morphological processor for Modern Greek *ACL European Chapter.*

RIESBECK, C.K. 1975. Conceptual Analysis. In: SCHANK (ed.) (1975).

RITCHIE, G.D., HANNA F.K. 1983. Semantic Networks: A General Definition and a Survey. *Information Technology: Research and Development* 2(3), 1983.

ROUSSEAU, P. 1983. A versatile program for the analysis of sociolinguistic data. *Tegniese verslag 1163 van die 'Centre de Recherche de Matematiques Appliquees'.*

ROUX, J.C. 1984: Fonetiek, Fonologie en die Rekenaar. In: SINCLAIR (red.). (1984):463-470.

ROUX, J.C. 1986: Computer-assisted phonetic research in African languages. *S.A. Journal of African Languages:* 42-45.

RUSTIN, R. (ed.) 1973. *Natural Language Processing.* New York: Algorithmics Press.

RYLE, G. 1949. *The concept of mind.* London: Hutchinson.

SAG, I. 1987. *Lectures on head driven phrase structure grammars.* Cambridge.

SAGER, N. 1978. Natural language information formatting: the automatic conversion of texts to a stuctured data base. In: YOVITS (ed.) (1978).

SAGER, N. 1981. *Natural Language Information Processing.* London: Addison-Wesley.

SAGVALL-HEIN, A-L. 1978. Finnish morphological analysis in the reversible grammar system. *Coling* 78.

SAGVALL-HEIN, A-L. 1983. A parser for Swedish. *Report Univ. Uppsala UCDL* R 83-2.

SANKOFF, D. 1976. Probability and linguistic variation. *Tegniese verslag 663 van die CRMA.*

SANKOFF, D. (ed.) 1978. *Linguistic variation: Models and methods.* New York: Academic Press.

SANKOFF, D. 1985. Statistics in Linguistics. In Kotz-Johnson: *Encyclopedia of statistical sciences,* vol. 5, John Wiley and Sons:74-81.

SANKOFF, D, POPLACK, S. 1981. A formal grammar for code-switching. *Papers in Linguistics* 14.2.3-46.

SANKOFF, D, SANKOFF, G. n.d. *Sample survey methods and computer assisted analysis in the study of grammatical variation.*

SCHANK, R.C. 1972. Conceptual Dependency: A Theory of Natural Language Understanding. *Cognitive Psychology* 3(4):552-630.

SCHANK, R.C. 1973. Identification of conceptualisations underlying natural language. In: SCHANK, COLBY (eds.) (1973).

SCHANK, R. 1973(a). Identification of conceptualizations underlying natural language. In: SCHANK, COLBY (eds.) (1973):187-247.

SCHANK, R.C. 1975. The Primitive ACTs of Conceptual Dependency. *TINLAP* 1:34-37.

SCHANK, R.C. 1975(a). *Conceptual Information Processing.* Amsterdam: North-Holland.

SCHANK, R., ABELSON, R. 1975. Scripts, plans and knowledge. *Advance papers 4th Intl. Joint Conf. Artificial Intelligence.*

SCHANK, R., ABELSON, R. 1977. *Scripts, Plans, Goals, and Understanding.* Hillsdale, N.J.: Lawrence Erlbaum Assoc.

SCHANK, R., COLBY, K.M. (eds.) 1973. *Computer models of thought and language.* San Francisco: Freeman.

SCHANK, R.C., JAIME, G.C.J.R. 1979. 'Re: the Gettysburg Address.' In: FINDLER (1979):327-362.

SCHANK, R.C., KOLODNER, JANET, DE JONG, G. 1980. Conceptual Information Retrieval. *Report* 190, Department of Computer Science, Yale University.

SCHANK, R.C., RIESBECK, C.K., CHRISTOPHER, K. 1981. *Inside Computer Understanding: Five Programs Plus Miniatures.* Hillsdale, N.J.: Lawrence Erlbaum Assoc.

SCHANK, R.C. et al. 1975. Inference and Paraphrase by Computer. *JACM* 22(3):309-328.

SCHMITZ, K-D. 1986. *Automatische Segmentierung natürsprachiger Sätze.* Hildesheim: George Olms.

SCHWARZE, C. WUNDERLICH, D. (reds.) 1985. *Handbuch der Lexikologie.* Athenaeum.

SEARLE, J.R. 1969. *Speech Acts.* Cambridge, England: Cambridge Univ. Press.

SEARLE, J.R. 1979. Indirect speech acts. In: COLE, MORGAN, (eds.) (1979).

SELLS, P. 1985. Lectures on contemporary syntactic theories: an introduction to Government-Binding theory, Generalised Phrase Structure Grammar, and Lexical-Functional Grammar. *CSLI Lectures Notes Number 3.* Stanford: Center for the Study of Language and Information.

SHAPIRO, S.C. 1971. A net structure for semantic information storage, deduction and retrieval. *Proc. IJCAI* -71:512-523.

SHAPIRO, S.C. 1982. Generalized Augmented Transition Network Grammars for generation from semantic networks. *American Journal of Computational Linguistics* 8/1:12-25.

SHIEBER, S.M. 1984. Direct parsing of ID/LP grammars. *Linguistics and Philosophy* 7:135-154.

SHIEBER, S.M. 1984a. The design of a computer Language for linguistic information. *10th International Conference on Computational Linguistics.* Proceedings of Coling 84: 362-366. Stanford.

SHIEBER, S.M. 1985. Criteria for designing computer facilities for linguistic analysis. *Linguistics* 23:189-211.

SHIEBER, S.M. 1986. *An introduction to unification-based approaches to grammar.* CSLI Lectures Notes Number 4. Stanford: Center for the Study of Language and Information.

SHIEBER, S.M., KARTTUNNEN, L. PEREIRA, F.C. 1984. Notes from the unification underground: a compilation of papers on unification-based grammar formalisms. *Technical Report 327.* Menlo Park, Calif.: Artificial Intelligence Center, SRI International.

SHIEBER, S., USZKOREIT, H.,, PEREIRA, F., ROBINSON,JANE, TYSON, MARY. 1983. The formalism and implementation of PATR-II. *Research on Interactive Acquisition and use of Knowledge:* 39-72. Menlo Park, Calif.: Artificial Intelligence Center, SRI International.

SHNEIDERMAN, B. (ed.) 1978. *Databases: Improving Usability and Responsiveness.* New York: Academic Press.

SIMMONS, R.F. 1973. Semantic networks: their computation and use for understanding English sentences. In: SCHANK, COLBY (eds.)(1973):66-113.

SIMMONS, R., CHESTER, D. 1982. Relating sentences and semantic networks with procedural logic. *Comm. Assn. Computing Machinery* 25,8 August 1982:527-47.

SIMMONS, R.F., SLOCUM J. 1972. Generating English discourse from semantic networks. *Comm. ACM* 15:891-905.

SIMON, H.A. 1969. *The Sciences of the Artificial* Cambridge: MIT Press.

SIMONS, G.F. n.d. *Powerful Ideas in Text Processing.* Arlington: The Summer Institute of Linguistics: University of Texas.

SIMONS, G.F. n.d. *The PTP Programmer's Reference Manual.* Arlington: The Summer Institute of Linguistics: University of Texas.

SINCLAIR, A.J.L. (red.) 1984. *G.S. Nienaber - 'n huldeblyk.* Bellville: U.W.K.

SLOCUM, J. 1981. A practical comparison of parsing strategies. *Proc. 19th Annl. Meeting Assn. Computational Linguistics,* Stanford, CA: 1-6.

SNELL, B. (ed.) 1983. *Term banks for tomorrow's world: Translating and the computer.* Aslib.

SOMERS, H.L. 1982. The use of verb features in arriving at a 'meaning representation'. *Linguistics* 20:137-265.

SOMERS, H.L. 1987. Valency and case in computational linguistics. *EDITS: Edinburgh Information Technology Series.*

SOWA, J.F. 1968. Semantics of conceptual graphs. *Proc. 17th Annual Meeting of the ACL:* 39-44.

SOWA, J.F. 1983. Generating language from conceptual graphs. *Computers and Mathematics with Applications* 8.

SOWA, J.F. 1984. *Conceptual Structures: Information Processing in Mind and Machine.* London: Addison-Wesley.

SPARCK JONES, K., WILKS, Y. 1983. *Automatic natural parsing.* Chichester: Ellis Horwood Limited.

STEINBERG, D.D., JAKOBOVITS, L.A. (eds.). 1971. *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology.* New York: Cambridge University Press.

STENSTRÖM, A-B. 1986. Carry-on signals in English conversation. *ICAME Journal* 11:62-64. Bergen: Norwegian Computing Centre for the Humanities.

STOCK, O., CASTELFRANCHI, C., PARISI, D. 1983. Wednesday: Parsing flexible word order languages. *1st Conference of the European Chapter of the Association for Computational Linguistics.* Proceedings. Pisa.

SWANEPOEL, P.H. 1986. *Leksikonstrukture vir gerekenariseerde linguistiese databanke.* Pretoria: UNISA.

THOMASON, R. (ed.) 1974. *Formal philosophy: selected papers of Richard Montague.* New Haven: Yale University Press.

THOMPSON, H. 1981. Chart parsing and rule schemata in phrase structure grammar. *Proc. 19th Annl. Meeting Assn. Computational Linguistics,* Stanford, CA: 167-72.

THOMPSON, H., RITCHE, G. 1984. Implementing natural language parsers. In: O'SHEA, EISENSTADT (eds.) (1984).

TSUJII, J.-I. 1980. Computer und semantische Repräsentation. *Sprache und Datenverarbeitung* 4(2), 142-152.

USZKOREIT, H. 1983. A framework for processing partially free word order. *21st Annual Meeting of the Association for Computational Linguistics.* Proceedings: 106-112. MIT Press.

VAN BAKEL, J. 1984. *Automatic Semantic Interpretation.* Dordrecht: Foris.

VAN DEN BROECKE, M.E. et al 1987. Type and Token Frequencies of Word Classes, Phonemes, and Phoneme pairs in Dutch. *PRIPU* 12(1):1-15, Amsterdam.

VAN DER ZWAN, DINA 1986. *Die taalgebruik in die dagboek van Hendrik Witbooi.* Verhandeling, UNISA.

VAN DIJK, T. 1973. Text Grammar and text logic. In: PETOFI, REISER, (eds.) (1973).

VAN STERKENBURG, P.G.J. 1986. *Algemene inleiding in de lexicologie.* Leiden.

WAITE, W.M. 1973. *Implementing Software for Non-numeric Applications.* New Jersey: Prentice-Hall, Inc. Englewood Cliffs. Inc.

WALKER, D.E. 1978. *Understanding spoken language.* New York: North-Holland.

WALTZ, D.L. 1975. An English language question answering system for a large relational database. *Comm. ACM* 21:7:526-539.

WALTZ, D.L. (ed.). 1978. Proceedings of the Second Workshop on Theoretical Issues in Natural Language Processing. *TINLAP* 2. Illinois: Champagne-Urbana.

WEBB, V.N. 1984. Die bestudering van linguistiese heterogeniteit met behulp van die rekenaar vanuit die perspektief van Labov se kwantitatiewe paradigma. In die *LVSA-kongresreferate.* Universiteit van Pretoria.

WEBB, V.N. 1987. Volgordevariasie in Afrikaans se afhanklike sin. *Tydskrif vir Geesteswetenskappe* 27.4. 283-304 5 April 1988.

WEHRLI, E. 1984. A Government-Binding parser for French. *ISSCO working papers no. 48.* Institut Dalle Molle.

WILENSKY, R. 1978. Why John married Mary: Understanding stories involving recurring goals. *Cognitive Science* 2:235-66.

WILENSKY, R. 1981. Pam and Micro-Pam. In: SCHANK, RIESBECK (eds.):136-96.

WILKS, Y. 1975. An intelligent analyzer and understander of English. *Comm. ACM* 18, 5(May 1975):264-74.

WILKS, Y. 1975(a). Preference Semantics. In: KEENAN (ed.).

WILKS, Y. 1977. Frames, scripts, stories and fantasies. *Pragmatics Microfiche.*

WILKS, Y. 1978. Making Preferences More Active. *Artificial Intelligence* 11(3):197-223.

WINOGRAD, T. 1972. *Understanding Natural Language.* New York: Academic Press.

WINOGRAD, T. 1976. Towards a Procedural Understanding of Semantics. *Revue Internationale de Philosophie* 3-4:260-303.

WINOGRAD, T. 1983. *Language as a Cognitive Process,* Volume I: Syntax. Reading, MA: Addison-Wesley.

WISBEY, R.A. (ed.). *The computer in literary and linguistic research.* London: Cambridge University Press.

WOODS, W.A. 1968. Procedural semantics for a question-answering machine. *Proc. 1968 Fall Joint Computer Conf.:* 457-71.

WOODS, W.A. 1970. Transition network grammars for natural language analysis. CACM 13/10:591-606.

WOODS, W.A. 1970(a). Transition network grammars for natural language analysis. *CACM* 13/10:591-606.

WOODS, W.A. 1975. What's in a link: foundations for semantic networks. In: BOBROW, COLLINS:35-82.

WOODS, W.A. 1978. Semantics and quantification in natural language question-answering. In: YOVITS (ed.) (1978).

WOODS, W.A. 1980. Cascaded ATN grammars. AM. J. *Computational Linguistics* 6,(1) (Jan. 1980):1-12.

WOODS, W.A. 1981. Procedural semantics as a theory of meaning. In: JOSHI, et al.: 300-334.

WOODS, W.A., KAPLAN R.M., NASH-WEBBER, B.L. 1972. The LUNAR Sciences Natural Language System. *Final Report, NTIS N* 72-28984.

YOVITS, M.C. (ed.) 1978. *Advances in computers* 17. New York: Academic Press.

ZADEH, L. 1978. A meaning representation language for natural languages. *Int. J. Man-Machine Studies* 10, 4 July 1978:395-460.

ZAMPOLLI, A., CALZOLARI, N. (eds.) 1973. *Computational and Mathematical Linguistics* Vol. II, Florence: Olschki.

ZWICKY, A., FRIEDMAN, J., HALL, B., WALKER, D. 1965. The MITRE syntactic analysis procedure for transformational grammars. *Proc. 1965 Fall Joint Computer Conf.* Washington, DC: Thompson Books.

BYLAE 1

LYS VAN INLIGTINGSTUKKE

1. **VERRIGTINGE: 10TH INTERNATIONAL CONGRESS OF PHONETIC SCIENCES 1983**
   Bibliografiese opsomming van referate gelewer by bogenoemde kongres (pp.6).

2. **PROGRAM: 11TH INTERNATIONAL CONGRESS OF PHONETICS SCIENCES 1987**
   Bibliografiese opsomming van referate gelewer by bogenoemde kongres (Slegs verrigtinge van eerste dag) (pp.11)

3. **PROGRAM: SPEECH TECH '87 NEW YORK 1987** (pp. 20)
   - Recent military application successes for speech recognition/synthesis
   - Busines outlook for speech recognition
   - Large vocabulary speech recognition
   - Cost justifiable products - speech recognition and the telephone
   - Large vocabulary speech recognition in personal computers
   - The kurzwiel voiceworks: A large vocabulary voice activated word processor
   - The IBM tangora: an experimental 20 000 word speech recognizer
   - A natural speech recognition system
   - Limitations to natural dialog: some problem areas
   - Integrating voice recognition and response into automobiles
   - Speech synthesis in educational products: yesterday, today and tomorrow
   - Speech recognition in the consumer market today
   - Artificial intelligence (speech technology) in the consumer product arena or this little thinker went to market
   - Speech processing at RADC
   - Speech technology for land-based systems
   - DARPA-advanced military speech systems
   - Military/goverment applications functional requirements
   - Integrated text-to-voice synthesis in a voice-store-forward and recognition system
   - Application software tools in speech technology
   - An integrated digital speech processing workstation
   - Voice recognition for 100% parts audit
   - Speech-driven factory control system
   - Use of voice technologies to solve real business problems in the office
   - The effectiveness of voice messaging in a sales and marketing application
   - Speech synthesis for remote text-mail retrieval: is current technology good enough?
   - Air force requirements for speech synthesis in tactical aircraft
   - The application of voice response technology to the process of approving customer orders

- A voice response system in a banking application
- Approaches to speech recognition for information acquisition tasks
- Advanced speech technology and the telephone network
- Military systems
- Quality assurance
- Aids for the handicapped
- Speech coding
- Performance assessment
- Academic institutions
- Medical applications
- Space systems

4. **EUROPEAN CONFERENCE ON SPEECH TECHNOLOGY 1987.**
Bibliografiese opsomming van referate gelewer by bogenoemde konferensie. (pp. 16).

5. **PROGRAM - EUROPEAN CONFERENCE ON SPEECH TECHNOLOGY: EDINBURGH 1987.** (pp. 21).
Bibliografiese opsomming van referate gelewer by bogenoemde konferensie oor onderstaande onderwerpe.
- Continuous speech systems
- Text-to-speech systems
- Speech analysis
- Formant analysis
- Speech modeling
- Phonetic decoding
- Large vocabulary systems
- Evaluation : synthesis
- Speech enhancement
- Prosodic analysis
- Knowledge based systems
- Morphology and phonology
- Formant tracking
- Medical applications
- Phonetic processing
- Linguistic processing
- Pitch tracking
- Aids to the death
- Acoustic analysis
- Prosody
- Codecs
- Telecommunications
- Phoneme-grapheme
- Industrial applicatians
- Speech production
- Voice mail
- Synthesisers
- Human factors
- Language processing systems
- Speaker verification

- Speaker adaptation.

6. **VOORDRAG: G. FANT: PHONETICS AND SPEECH TECHNOLOGY** (pp.10)
   - Computerized phonetics
   - Speech recognition and research needs
   - Perception
   - In quest of speech code variability and invariance

7. **AUDLAB : EDINBURGH UNIVERSITY 1987** (pp. 12)
   'A speech and signal analysis package (A UDLAB) has been developed by the CSTR at the University of Edinburgh. AUDLAB is the primary speech analysis tool used by researchers working on the Alvey Large Scale Demonstrator in automatic speech recognition, a project whose objective is to build a speech-driven word processor and workstation. In the context of this project, AUDLAB's primary application is in the interactive development and testing of acoustic and phonetic hypotheses about segmentation and labelling of speech waveforms.'

8. **USING PTPC AND PTP PROGRAMS** (pp. 22)
   'The Programmable Text Processor (PTP) is an interpreter for a special purpose computer language for processing text-files....'

9. **TEKSANA - VER. 12.0 - METAKEUSES** (pp. 5)
   - Analitiese metakeuses
   - Frekwensie-analise
   - Voorbeeld-analise
   - Voorbeeld- en frekwensie-analise
   - Variant-analise
   - Variantfrekwensie-analise
   - Variant- en variantfrekwensie-analise
   - Definisie-analise
   - Definisiefrekwensie-analise
   - Definisie- en definisiefrekwensie-analise
   - Letterfrekwensie-analise
   - Letterfrekwensie vergelyking
   - Glossarium opstelling

10. **LANGUAGE RESEARCH LABORATORIES.** Georgetown University. Quarterly Report Vol.2 No. 1 (pp.27)
   - Microcomputers and the Humanities
   - Hardware Reviews
   - Use of Fonts in the Grundy System
   - A Hindi Learning Package for the personal computer
   - Prospectus for a Spanish Vector Grammar
   - Specifications for a Russian Instructional Parser

11. **SUMMARIES OF PAPERS FROM THE INTERNATIONAL CONGRESS ON TERMINOLOGY AND KNOWLEDGE ENGINEERING 1987** (pp. 4)
- Ein grafischer editor für semantische netze
- Smart assistant for information retrieval
- Systems theory for knowledge bases
- TEGEN - a self-adaptive information retrieval system
- Construction and evaluation of an enviromental microthesaurus
- Sprachmatische aspekte der terminologie- und wissensgewinnung
- Spezification eines konzeptionellen schemas für terminologiedatenbanken
- Representation rules for scientific words in knowledge base systems
- What is a concept?
- Definition of terms, word meaning and knowledge structure
- Linguistic and domain knowledge sources for the universal parser architecture
- Storage and intelligent retrieving of information carried by natural language messages

12. **PROGRAMME OF THE 11TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS 1986** (pp. 11)
Bibliografiese opsomming van referate gelewer by bostaande kongres oor onderstaande onderwerpe:
- Semantics
- Dialogue
- Software
- Machine translation
- Parsing
- Discourse
- Knowledge
- Morphology
- Dictionaries

13. **EXTRACTS FROM ICAME NEWS 10, 1986** (pp. 9)
- Research at Lancaster University
- Research at Birmingham University
- Research at Nijmegen University
- Research at Lund University

14. **CELEX - CENTRE FOR LEXICAL INFORMATION** (pp. 9)
Newsletter No. 1, 1986
'This first issue of the CELEX Newsletter serves as an introducroty exposition of the CELEX lexical-database project and the service it will provide in the near future.'

15. **TRANSCRIPT ANALYSIS** Vol. No. 1 1984 (pp. 30).
- Reasons for promoting data exchange
- The need for greater efficiency in data sharing
- The need for greater precision in data collection and analysis
- The need for increased automation in analysis
- The formation of the system

- The organization of the system
- The shape of the database

16. **TRANSCRIPT ANALYSIS** Vol. 2 No. 1 1985 (pp. 45)
- Archiving data sets: issues and problems
- Two applications of computers to second language research
- Overview of adult speech error corpora
- Access to the CHILDES data base via telnet and telenet

17. **TRANSCRIPT ANALYSIS** Vol. 2 No. 2 1985 (pp. 65)
Coding manual for the description of conversation.
'This sceme is designed for the description of dyadic conversation between native speakers of English, where one of the speakers is a child. The purpose of the resulting description is to allow comparisons to be made between samples of conversation'

18. **TRANSCRIPT ANALYSIS** Vol. 4 No. 1 1987 (pp. 52)
CHAT MANUAL
- Getting set up for transcription
- The basic components of a CHAT transcript
- UBINET for English
- Error coding for normal and disordered speech
- Simplified speech act codes
- A system for interlineer Morphemic glosses
- Examples of transcribed data using the CHAT format

19. **TRANSCRIPT ANALYSIS** Vol. 4 No. 2 1987 (pp. 38)
- Phonascii newsletter
  - Goals of phonascii system
  - Constraints on the phonascii system
  - Coding phonemic strings
  - Coding phonascii segment strings
  - Coding phonascii suprasegmentals
- Using CHAT coding for interactional analysis
  - Computer transcription and coding in CHAT
  - Speech act coding schema

20. **TRANSCRIPT ANALYSIS** Vol. 3 No. 1 1986 (pp. 108)
A manual for classifying verbal communicative acts in mother-infant interactions in mother-infant interaction

21. **EXPERIMENTAL DESIGN** - Uittreksel uit Proefskrif van J. Vorster (pp. 20)
'...the method employed for obtaining the data, and for preparing it for analysis, is briefly explained. This is done in terms of the subjects, the sampling procedure and the coding procedure.'

22. **'n AFRIKAANSE SKRIF-TOT-SPRAAKOMSETTER** (De Stadler en Coetzer) (pp. 20)
- Verslag oor vordering wat gemaak is met die ontwikkeling van 'n skrif-tot-spraakomsetter vir Afrikaans
- Skrif na fonetiese voorstelling
- Fonologiese reëls
- Morfologiese reëls
- Morfeemwoordeboek
- Sillabifikasiereëls
- Klemtoekenningreëls

23. **BOBRA: AN EXPERT SYSTEM FOR AUTOMATATED LANGUAGE DESCRIPTION AS A BASIS FOR STATISTICAL STUDIES** (pp. 8)
- Automated corpus description
- Computational linguistics
- Artificial intelligence
- The BOBRA expert system: design and implementation

24. **CHARACTERISTICS OF THE METAL MACHINE TRANSLATION SYSTEM AT PRODUCTION STAGE** (p. 20)
- Machine translation
- Augmented phrase structure grammar
- Natural language processing
- Parsing
- Lexical databases
- Grammar component
- Semantic handling in METAL
- Analysis Component
- User interface

Doc no 206784
Copy no 208375

RGN BIBLIOTEEK | HSRC LIBRARY

R18,40