

PARLEVINK IN CONTEKST: Meer dan Language Engineering

1. Inleiding

Language engineering houdt zich bezig met het ontwerpen en realiseren van software (of software en hardware) voor de verwerking van natuurlijke taal. Meestal wordt aangenomen dat er sprake is van een zodanige verwerking dat gebruikmaking van linguïstische kennis de kwaliteit van verwerking kan verhogen. Deze kennis kan uiteenlopen van statistische informatie over lettervolgordes tot geavanceerde syntactische, semantische en pragmatische kennis van taal. De systemen die gebruik maken van dergelijke kennis lopen uiteen van OCR-systemen en spelling checkers tot natuurlijke taal interfaces, intelligente information retrieval en machinale vertaalsystemen. Parlevink is een onderzoeksproject van de Faculteit Informatica waarvan de kern wordt gevormd door onderzoek op het gebied van language engineering (taaltechnologie, linguistic engineering, etc.) ondersteund door meer fundamenteel onderzoek op het gebied van taalanalyse en taalbeschrijving. In dit artikeltje zullen we ons grotendeels tot de kern moeten beperken. Zijdelings zullen een aantal andere aspecten van het onderzoek aan bod komen. Om het onderzoek enigszins te kunnen plaatsen is een kennismaking met achtergronden van het onderzoeksterrein onontbeerlijk. Zodra die gegeven is kunnen we in het kort de onderzoeksterreinen binnen Parlevink verduidelijken.

Een niet-marginale onderzoeksinspanning op het gebied van de verwerking van natuurlijke taal door computers is ongebruikelijk bij informatica-faculteiten of vakgroepen aan Nederlandse universiteiten. Meer dan elders heeft men zich erg lang weten te beperken tot onderwerpen die tot de zogenaamde kerninformatica behoren. In het buitenland, in het bijzonder in de Verenigde Staten, is dat anders. Niet alleen de universiteiten, maar ook tal van bekende onderzoeksinstituten of onderzoekslaboratoria van grote bedrijven die wij meestal enkel kennen van hun onderzoek op het gebied van informatiesystemen, AI, telematica, etc., blijken vrijwel altijd ook veel aandacht te besteden aan onderzoek waarin natuurlijke taal centraal staat.

Afgezien van de aandacht voor onderzoek naar natuurlijke taal bij de grote telematica- en informatietechnologiebedrijven in de Verenigde Staten is het daar ook vooral DARPA (Defense Advanced Research Program Agency) die op grote schaal onderzoek naar taal en spraak (in relatie met de computer) van de grond trekt. Gezien de veelaligheid in Europa zou men ook van de Europese Gemeenschap een voortrekkersrol voor

dergelijk onderzoek kunnen verwachten en dat is ook inderdaad het geval. Het is de Europese Gemeenschap die een taalindustrie van de grond probeert te krijgen en het daarvoor benodigde ondersteunende onderzoek stimuleert. In het bijzonder de laatste jaren wordt in tal van rapporten geprobeerd tot een visie te komen op de taalproducten en taalsystemen van de toekomst en op welke wijze onderwijs en onderzoek gestimuleerd moet worden om dergelijke producten en systemen te realiseren. Het uitgangspunt daarbij is dat er binnen tal van deelgebieden van de (computer)linguïstiek relevante wetenschappelijke resultaten zijn bereikt, maar dat de toepassing ervan achterloopt bij de mogelijkheden die er zijn omdat de bereidheid tot een technologische benadering van taal nog niet voldoende wortel heeft geschoten.

Er zijn politieke redenen voor de Europese gemeenschap om veel aandacht te besteden aan de stimulering van een onderwerp als Language Engineering, er zijn ook economische argumenten. In een onlangs in opdracht van de Europese Commissie vervaardigd rapport wordt met behulp van door de OECD ontwikkelde macro-economische modellen de potentiële economische impact van Language Engineering toepassingen ingeschat. Grotere bewustwording van de mogelijkheden en verbetering van de technologie zouden volgens gehanteerde scenarios leiden tot tweehonderd-duizend nieuwe banen en een additionele 26 billion ECU als output in de verschillende lidstaten. Om deze redenen ziet men steeds vaker dat in de EG-rapporten gesproken wordt over de belangrijke triade telematics, information and language engineering en over de integratie van deze drie terreinen.

2. Politieke Adviseurs?

Dat taal politiek, economisch en technologisch benaderd kan worden mag al uit bovenstaande blijken. Er zijn tal van andere benaderingen mogelijk (biologisch, cognitief, linguïstisch, sociologisch, etc.) en een goede technologische benadering zoals wordt nagestreefd in het Parlevink project vereist aandacht voor die andere benaderingen. Daarbij komt bovendien nog dat mensen emotioneel en niet rationeel naar taal kijken, taal bepaalde eigenschappen toedichten en taal minder als gebruiksartikel dan als iets volkseigens zien, waarvoor men, zoals in Zuid-Afrika voor de van oorsprong Creolentaal Zuid-Afrikaans, een standbeeld wil oprichten. Sommige talen worden in een bepaalde situatie gezien als

elitetalen (bijvoorbeeld het Engels in Zuid-Afrika), anderen als ondemocratisch (bijvoorbeeld in Zuid-Afrika het Afrikaans). Ook binnen de Europese Gemeenschap, ondanks de beleden gelijkwaardigheid van talen, gaat men in de richting van het ook expliciet benoemen van belangrijke en minder belangrijke talen. Voor uitbreiding van de Gemeenschap is die gelijkwaardigheid een belangrijk gegeven. Bij de verdere eenwording is aan die gelijkwaardigheid niet te voldoen zoals mag blijken uit het geruzie in 1993 over de te gebruiken talen binnen een gezamenlijke legermacht (Eurokorps) of binnen het Europese Merkenbureau. Er wordt een onderscheid gemaakt tussen 'grote' en 'kleine' talen. Een dergelijk geruzie is overigens peanuts als we kijken naar wat er eveneens recentelijk aan de hand is in Centraal en Oost-Europa en in het bijzonder het voormalige Joegoslavië. Zo wordt in Slowakije in haar streven naar een eenzijdige identiteit de taal van de Hongaarse minderheid uit haar openbare leven en haar onderwijs verjaagd. In het Kroatisch streven tot een eigen taal wordt gesteld dat "De strijd voor onze taal en cultuur is een onderdeel van de oorlog." en "Taal is niet alleen een middel om te communiceren. De taal houdt de geschiedenis en cultuur van de natie in stand." Moslim-taalkundigen

Skopje en Sofia maken ruzie over taal

Reuter
SOFIA

De presidenten van Bulgarije en de Voormalige Joegoslavische Republiek Macedonië (FYROM) hebben dinsdag geweigerd hun standpunt in een taalkwestie te wijzigen.

Tijdens een bezoek van Gligorov aan Sofia bleef de ondertekening van een akkoord met president Zjelev achterwege. Gligorov wenste dat de documenten van het akkoord zou worden getekend in 'de Macedonische en Bulgaarse talen'.

Zijn Bulgaarse gastheren eisten dat de tekst zou worden gewijzigd in 'de officiële talen in Bulgarije en Macedonië'.

Volgens Gligorov kan er geen meningsverschil bestaan over de taal. 'Het is het recht van elk volk de naam van de taal die het spreekt zelf te bepalen.' Zjelev en andere Bulgaarse functionarissen weigerden de tolk van Gligorov te gebruiken omdat zij ook zonder hem met hun Macedonische gasten konden spreken.

Sofia erkende als eerste de afgescheiden Joegoslavische republiek Macedonië in 1992, maar weigerde het bestaan van een Macedonische natie te erkennen.

in Bosnië introduceren het Bosnisch als een taal die absoluut niet is afgeleid van het Servisch of Kroatisch en Serviërs pleiten voor een verbod van de letter U omdat dat het teken was van de Kroatische fascist. Dat dit geruzie soms leidt tot tragisch-komische tafereelen zoals beschreven in bijgaand Volkskrantartikeltje is dan een schrale troost. In dit tafereel zien we dat taal een subtiele rol kan spelen bij de verschillen tussen land, natie en volk en dat dit op zijn beurt kan leiden tot het opdringen van een tolk aan een gesprekspartner ook als die het zonder tolk uitstekend kan verstaan. Wanneer men in het huidige Europa als firma taalproducten wil slijten dan is een politiek adviseur geen luxe.

3. Taalgebruiker en Taaltechnologie

Van grotere invloed op een technologische benadering van taal is het gebruik en de gebruiksomgeving van het taalproduct dat men wil realiseren. Het centraal stellen van gebruik en gebruiksomgeving kan beschouwd worden als een stijlbreuk met de traditionele (psycho-)linguïstische en cognitieve benadering van taal. Daar is sprake van een universeel 'domain of discourse', en hoewel een dergelijk 'domain' object van studie kan zijn wordt er geen relatie gelegd tussen eigenschappen van syntaxis, semantiek en pragmatiek en een bepaald taaldomein. Om een voorbeeld te noemen, het dagelijkse weerbericht in de krant is geschreven in natuurlijke taal. Het zal echter duidelijk zijn dat er slechts in zeer beperkte mate van natuurlijke taal gebruik gemaakt wordt. Er is eerder sprake van een soort code en dergelijke codes gelden in meer of mindere mate ook voor tal van andere domeinen.

Het centraal stellen van gebruik en gebruiksomgeving betekent ook dat door linguïsten of cognitiewetenschappers gewenste (psycho-)linguïstische of cognitieve plausibiliteit van gebruikte of te ontwikkelen theorie en methoden niet noodzakelijkerwijs een rol speelt. In de technologische benadering hoeft men zich niet te bekommeren om die plausibiliteit, hetgeen overigens niet hetzelfde is als zeggen dat men geen aandacht hoeft te schenken aan dergelijke theorieën. Overigens, ook binnen de linguïstiek wordt wel onderscheid gemaakt tussen 'competence' enerzijds, het taalvermogen dat voor een deel is aangeboren en dat toestaat om zinnen te construeren en te begrijpen, en anderzijds 'performance', de uiterlijke vorm van dit vermogen die slechts een afreksel daarvan kan zijn aangezien men door concentratieverlies, afleiding, etc, slechts beperkt ingewikkelde zinnen kan spreken en begrijpen.

Deze dichotomie tussen performance en competence is niet onomstreden en is, voor alle duidelijkheid, niet aan een bepaald 'domain of discourse' gekoppeld. Het gaat weer om een algemene beginsel dat echter bij een technologische

benadering van taal meer aandacht kan krijgen dan de op competence georiënteerde computationele linguïstiek. Waarom zouden we iets modelleren dat in de praktijk toch niet voorkomt? 'Marie leert Piet zwemmen' is een Nederlandse zin. 'Jan hoort Marie Piet leren zwemmen' eveneens. Het is niet moeilijk de regels te vinden die aan deze zinsconstructies ten grondslag liggen en vervolgens puur mechanisch deze zinnen uit te breiden, bijvoorbeeld: 'Klaas ziet Jan Marie Piet horen leren zwemmen'. Inderdaad, in onderhoudshandboeken van de Fokker Friendship zullen we dergelijke constructies niet tegenkomen. Een syntactische component van een vertaalsysteem voor dergelijke handboeken hoeft niet in staat te zijn dergelijke zinnen te analyseren. Overigens, nu we toch met dit voorbeeld bezig zijn, met behulp van editors kan men schrijvers van handboeken dwingen tot of helpen met het schrijven van teksten die zodanig eenvoudig gestructureerd zijn en bovendien uitgaan van een beperkte woordenschat zodat een machinale vertaling tot de mogelijkheden gaat horen.

4. De Rauwe Werkelijkheid

In een gegeven kontekst, met gegeven actoren, vervullen taaluitingen bepaalde functies. Een voor de hand liggend uitgangspunt voor onderzoek dat op enigerlei wijze probeert te komen met resultaten die het mogelijk maken deze functies te simuleren is dan ook de taaluitingen zoals die in de praktijk, eventueel binnen een bepaald domein, voorkomen als onderzoeksmateriaal te nemen. Zeker in het geval van spraak wordt het er daardoor niet eenvoudiger op. Veel uitgesproken zinnen, zeker in een tweegesprek, zijn niet grammaticaal correct. Ook in teksten, of in ingetikte vragen of opdrachten komen tal van onvolkomenheden voor. Voor mensen is dit meestal geen enkel beletsel om het gesprek voort te zetten of een tekst te begrijpen. De taalfaculteit is robuust, er is meestal voldoende redundantie aanwezig om ook half uitgesproken zinnen of zinnen die halverwege een andere wending nemen zonder dat een grammaticaal correcte zin wordt verkregen te begrijpen. Toepassingen binnen de taaltechnologie vereisen eveneens robuustheid. Een grammatica-checker moet niet simpelweg stoppen op onjuiste zinnen, maar dient suggesties te geven hoe een fout te corrigeren. Een taalinterface zal een poging moeten doen om een zo goed mogelijke interpretatie te geven aan een vraag, opmerking of opdracht van een gebruiker, al was het maar om een zo goed mogelijke parafraze te geven. Teksten die ten behoeve van text retrieval linguïstisch geanalyseerd moeten worden kunnen niet uit de database verwijderd worden omdat ze grammaticale fouten bevatten. Een vertaalsysteem, eventueel als hulpmiddel voor een professionele vertaler, zal toch

een zo goed mogelijke suggestie voor vertaling dienen te geven.

Tot nu toe is het niet gebruikelijk in het onderzoek de werkelijkheid zo rauw te nemen als hierboven geschetst. Een geïdealiseerde 'werkelijkheid' leent zich beter voor het opsporen van dieperliggende eigenschappen en elegante theorieën waarop algoritmen kunnen worden gebaseerd. Wel is er sprake van de ontwikkeling van theorie gebruikmakend van een corpus van zinnen, dialogen of teksten die inderdaad uit de werkelijkheid geplukt zijn, maar waarbij meestal enkel de grammaticaal correcte taaluitingen in het corpus zijn opgenomen. Een corpus van zinnen kan bijvoorbeeld uitgangspunt zijn voor het ontwerpen van een grammatica voor een zekere taal. Uitgangspunt voor het ontwerpen van een natuurlijke taalinterface kan bijvoorbeeld zijn een verzameling dialogen die met behulp van een zogenaamd *Wizard of Oz* experiment zijn verkregen. Hierbij wordt gebruikers van een bepaalde dienst gesuggereerd dat men met een computer communiceert, waarbij de rol van de computer echter gespeeld wordt door een onderzoeker die volgens bepaalde voorschriften een dialoog voert met de gebruiker. In beide gevallen kan men hopen dat de ontwikkelde beschrijvingen en methoden inderdaad generaliseren tot het onbeperkte taaldomein waar een gebouwd taalsysteem in gebruik mee te maken krijgt. Representativiteit van het gebruikte corpus is belangrijk. Een corpus biedt ook de mogelijkheid tot het expliciet vergaren van statistische informatie die gebruikt kan worden in methoden voor, bijvoorbeeld, zinsinterpretatie. Statistische benaderingen van natuurlijke taal zijn na een periode van relatieve rust gedurende de jaren '60 en '70 opnieuw binnen het aandachtsveld van taalonderzoekers gekomen. Bij gesproken taal daarentegen ziet men dat pas in de jaren '90 er aandacht komt voor niet-statistische benaderingen en de integratie van taal en spraak. Dat wil zeggen dat men kennisbronnen als syntaxis en semantiek gaat aanwenden voor het herkennen van spraak.

Het vinden van regels uitgaande van een noodzakelijkerwijs beperkt aantal observaties is inherent aan linguïstisch onderzoek. Het construeren van theorie uitgaande van een eindig corpus is er een voorbeeld van. Hetzelfde geldt voor het afleiden van statistische kennis uit zo'n corpus. Het is een uitgangspunt dat inherent aanwezig is in Chomsky's kijk op taal en de taalfaculteit. Het uitgangspunt is dat het basisontwerp van ons taalvermogen is aangeboren. Door blootstelling van dit basisontwerp aan voorbeelden uit een taal (gedurende de kindertijd) wordt de grammatica van een bepaalde taal geleerd. Door sommigen wordt wel gesproken over het menselijke taalinstinct, het door evolutie tot stand gekomen basisontwerp van het taalvermogen van de mens. Men ziet dit vermogen in werking

wanneer kinderen in hun taalleertijd opgroeien met een vrijwel structuurloos jargon ('pidgin') en daaruit een taal met een grammatica ontwikkelen. Bij een modellering van taal en taalgebruik ten behoeve van een verwerking door computers kan men zich laten leiden door het gezichtspunt van een biologisch gefundeerd, evolutionair tot stand gekomen taalvermogen. In het bijzonder geldt dit bij het onderzoek naar het gebruik van (kunstmatige) neurale netwerken en genetische algoritmen voor het leren van (de regels van) natuurlijke taal. Hoewel er een gigantisch verschil bestaat tussen de aantallen neuronen en verbindingen die aanwezig zijn binnen de menselijke hersenen en de aantallen die men hanteert bij de kunstmatige neurale netwerken, biedt dit model toch voldoende de gelegenheid om te illustreren hoe evolutionair rekenschap gegeven kan worden van het taalvermogen van de mens, hoe sprake kan zijn van een leerproces op grond van voorbeelden en hoe taalrobustheid een natuurlijke eigenschap van dergelijke netwerken kan zijn.

5. Taal als Probleem

Dat taal veel benaderingen toelaat maakt al aannemelijk dat het moeilijk zal zijn om te komen tot een allesomvattende formalisering van taalbeschrijving en taalgebruik. Zonder formalisering van (aspecten van) taal is een computationele verwerking van taal onmogelijk. Begin jaren '50 vonden de eerste pogingen plaats om tot een formele beschrijving van natuurlijke taal te komen. Onderzoek naar machinaal vertalen kreeg om politieke redenen veel aandacht. Noam Chomsky keerde zich tegen een behaviouristische opvatting van taal en sprak over het 'aangeboren zijn' van het taalvermogen. De aandacht ging uit naar de beschrijving van de syntactische component van de daaruit voortvloeiende 'competence' van de taalgebruiker. De bekende Chomsky-hierarchie (reguliere, context-vrije, context-gevoelige talen) stamt uit deze tijd. Het optimisme van die tijd, ondanks dat er nog geen enkel formalisme was dat toeliet dat een essentieel deel van een natuurlijke taal syntactisch beschreven kon worden, werd door logicus en taalkundige Bar-Hillel de grond ingeboord met een simpel voorbeeldzinnetje zoals 'De bokser stond in de ring'. Met dit zinnetje vestigde hij de aandacht op de semantiek van woorden en zinnen en bovenal de dubbelzinnigheid van woorden (zoals 'ring') en zinnen. Het feit dat mensen meestal geen enkele moeite hebben om aan zinnen een eenduidige interpretatie te geven heeft te maken met semantische kennis, maar bovenal met extra-linguïstische kennis, dat wil zeggen kennis van hoe de wereld in elkaar zit, kennis van het domein, kennis van hoe mensen met taal omgaan, etc. En, volgens Bar-Hillel, het zal onmogelijk zijn om dergelijke kennis volledig te formaliseren.

Er werd op verschillende wijze gereageerd op Bar-Hillels observaties. Het vrijwel stopzetten van financiering van onderzoek op het terrein van machinaal vertalen was één van de reacties. Meer aandacht voor minder toepassingsgerichte computationele linguïstiek (bijvoorbeeld computationele semantiek) in het algemeen en voor artificiële intelligentie was een ander gevolg. Binnen de AI was het niet ongebruikelijk zich bij het simuleren van intelligentie tot een bepaald beperkt domein te beperken. Zoals, bijvoorbeeld, een gesprek tussen kapper en klant over het weer, het in natuurlijke taal kunnen informeren naar baseball-uitslagen gedurende een bepaald seizoen, of het simuleren van de rol van een Rogeriaanse therapeut in een tweegesprek met een klant. Een zeer bewuste inperking tot een afgesloten domein gebeurde in het SHRDLU-systeem van Terry Winograd dat toeliet dat vragen gesteld werden over een 'blokkenwereld'. De blokkenwereld kon veranderd worden door in natuurlijke taal opdrachten aan een (gesimuleerde) robot te geven. Het beperkte domein liet toe dat domeinkennis gebruikt kon worden om te komen tot eenduidige interpretaties van vragen en opdrachten. Bij Winograd was de syntaxis nog een belangrijke kennisbron bij het analyseren van zinnen. Bij ander onderzoek aan natuurlijke taalverwerking binnen de AI-wereld was dit veel minder het geval. Vaak werd geprobeerd, in overeenstemming met de succesvolle ontwikkeling van expertsystemen, kennis op het gebied van een domein te formaliseren met formalismen als scripts en frames. Syntaxis werd binnen de AI-wereld overgeslagen onder het motto dat die ten opzichte van de aanwezige domeinkennis te weinig bijdroeg aan de eenduidige interpretatie van zinnen. Domeingeoriënteerde natuurlijke taalverwerking leidde in de jaren '80 tot onderzoek op het gebied van overdraagbare natuurlijke taal interfaces. Eveneens in overeenstemming met wat in de wereld van expertsystemen gebeurde ging de aandacht uit naar het ontwerpen van shells die gevuld met domeinkennis voor een bepaalde (relationele) database het construeren van een interface voor die database vergemakkelijkten. Het zich beperken tot een domein was eerder gebeurd, bijvoorbeeld ook bij het machinaal vertalen. In feite zou een domeinexpert met het gereedschap ontwikkeld door onderzoekers in staat moeten zijn een natuurlijke taal interface te componeren. Het uitgaan van 'overdraagbaarheid' en het ontwikkelen van software daarvoor zou men het begin van de taaltechnologie (of language engineering) kunnen noemen. Begin jaren '80 begon ook een conferentie-reeks onder de naam 'Applied Natural Language Processing'.

Opgemerkt moet worden dat deze 'switch' naar toepassingsgericht onderzoek niet gepaard is gegaan met het oplossen van de fundamentele

problemen binnen de computationele taalkunde. Het fundamentele probleem is dat voor een eenduidige interpretatie van een taaluiting men tal van kennisbronnen moet kunnen aanwenden, dat deze kennisbronnen formeel gerepresenteerd moeten zijn op een zodanige wijze dat met deze kennis geredeneerd kan worden en dat het toekennen van een interpretatie een dynamisch proces is in die zin dat zowel voorgaande als volgende taaluitingen bijdragen aan die interpretatie. Daarbij komt dat bij een niet al te beperkt domein geen sprake kan zijn van het sequentieel 'afwerken' van kennisbronnen. Dit zou namelijk in het begin leiden tot een combinatorische explosie van mogelijke interpretaties. Wat betreft de verschillende kennisbronnen kan men denken aan: functie van het systeem, bedoelingen van de gebruiker, eigenschappen van de gebruiker (model van de gebruikersgroep en de individuele gebruiker), domeinkennis, wereldkennis (gezond verstand kennis), pragmatische kennis, semantische kennis, syntactische kennis, morfologische kennis en prosodische kennis (bij gesproken taal biedt de melodie van de uitgesproken zin aanknopingspunten voor de interpretatie). Aan formalisering van ieder van deze kennisbronnen wordt onderzoek verricht. Op geen enkele van deze onderwerpen kan gesproken worden van 'on-the-shelf' theorie of technologie.

Ter illustratie geven we een aantal simpele zinnestelsels die uitgesproken binnen een bepaalde context een menselijke interpretator geen moeilijkheden opleveren. Binnen een formeel kader waarin enkel naar een letterlijke betekenis wordt gekeken ontstaan problemen. Bijvoorbeeld,

Zij zagen het meisje met een verrekijker leidt tot de vraag wie de verrekijker in de hand heeft. Zij of het meisje? Verder kan men zich afvragen of het meisje wordt 'gezaagd' of 'gezien'. Wat te denken van:

Het gemeentebestuur verbood de demonstratie van de vrouwenorganisatie omdat ze het gebruik van geweld niet wilde uitsluiten

Waar slaat 'ze' op, het gemeentebestuur of de vrouwenorganisatie?

Jan wil een Noorse trouwen.

Heeft Jan al iemand op het oog of is het een uitspraak van Jan dat als hij ooit zal trouwen dat het dan met een Noorse zal zijn?

Wat zijn de gemiddelden over de afgelopen vijf jaar?

Tja, er is blijkbaar sprake van een tweegesprek waarin eerder tal van vragen zijn gesteld en waar deze vraag een logisch vervolg is.

Nee, de eerste.

Idem.

Het is koud hier.

Wat is dit, een constatering of een impliciet verzoek aan iemand die ook in de kamer aanwezig is om het raam te sluiten?

Het zal duidelijk zijn dat voor interpretatie van dergelijke zinsuitingen tal van kennisbronnen dienen te worden aangeroepen.

6. Parlevink Topics

Nadat we zo (relatief) uitvoerig een aantal invalshoeken op taal de revue hebben laten passeren kunnen we kort zijn wat betreft de onderwerpen die binnen het Parlevink project aandacht krijgen. De onderwerpen passen binnen een vijftal deelprojecten. AiO-onderzoek is ook onderdeel van deze projecten.

1. Taalspecificatie en Parsing in Language Engineering. Fundamenteel en praktisch georiënteerd onderzoek op het gebied van taalanalyse waarbij over het algemeen de syntactische invalshoek wordt gekozen. Wiskundige analyse van de ontworpen methoden behoort tot het onderzoek. Evaluatie van de methoden met betrekking tot praktische toepasbaarheid eveneens.

Toelichting: Er bestaan grammatica's die in meer of mindere mate delen van een natuurlijke taal beschrijven. Onderzoek leidt tot nieuwe en betere formalismen voor taalbeschrijving. Voor ieder domein en iedere toepassing moet een grammatica geschreven worden, formalismen aangepast worden en ontleedmethoden geschikt gemaakt worden. Binnen dit deelproject wordt gewerkt aan taalbeschrijving met behulp van stochastische grammatica's, efficiënte en robuuste parsing methoden en aan semantisch georiënteerde ontleedmethoden. Methoden en formalismen hebben vaak een context-vrije basis, maar laten toe dat ook niet-syntactische kennis gerepresenteerd en verwerkt kan worden. Een meer praktische component van dit deelproject is de ontwikkeling van een ontledingomgeving die toelaat grammatica's te editen en ontleedmethoden te evalueren.

2. Semantiek en Pragmatiek in Language Engineering. Gezocht wordt naar een uniform raamwerk voor de representatie van dialoogsemantiek, voor zover ze nodig is voor het onderscheiden van de functie van taaluitingen in mens-machine dialogen. Uitgangspunt is dat een volledige interpretatie niet alleen semantische, maar ook pragmatische en niet-linguïstische kennis vereist. Wat betreft de pragmatische invalshoek wordt in het bijzonder gekeken naar de toepassing van de semiotische theorie van Ch.S. Peirce op het gebied van taalverwerking.

Toelichting: De betekenis van een taaluiting wordt niet alleen bepaald door de semantiek, maar ook door pragmatische en niet-linguïstische aspecten.

Bovendien, in de context van een dialoog zal de betekenis voortdurend bijgesteld moeten worden. Het semantiek onderzoek binnen dit deelproject kent drie te integreren benaderingen. Een theoretische, waarbij vooral aandacht is voor het dynamische karakter van dialogen en waarbij theorieën als conversatie-analyse, 'speech act' theorie en plantheorie mee zullen helpen bij het ontwerpen van het representatiefomalisme. Een operationele, waarbij in de context van een natuurlijke taal interface aspecten als gebruikersvriendelijkheid (robuustheid) en 'on-line' interpretatie aan bod komen. Een empirische, waarbij mens-machine dialogen bestudeerd zullen worden om eigenschappen van conversatiegedrag te vinden. Het pragmatiekonderzoek binnen dit deelproject beschouwt pragmatiek als de theorie die ten grondslag ligt aan de dialoogcontrole functies van een natuurlijke taal interface. In het bijzonder biedt de semiotiek van Peirce de mogelijkheid een raamwerk te ontwikkelen waarbinnen verschillende niveaus van zinsrepresentatie en vooral de relaties tussen die niveaus kunnen worden georganiseerd.

3. *SCHISMA: Schouwburg Informatie Systeem.* Doel is het bouwen van een via natuurlijke taal toegankelijk systeem dat informatie kan geven over schouwburgvoorstellingen in een bepaalde stad. De gebruiker van het systeem kan ook voorstellingen reserveren. Verschillende andere deelprojecten van Parlevink leveren onderzoeksresultaten die binnen SCHISMA gebruikt en geëvalueerd dienen te worden.

Toelichting: Een dergelijk systeem biedt de gelegenheid theorie ontwikkeld op het gebied van taalanalyse, dialoogmodellering en taalgeneratie te integreren binnen een prototype van een praktisch taalverwerkend systeem. Via een aantal tussenstappen wordt de invoer van de gebruiker omgezet naar een semantische representatie die, zodra het mogelijk is, omgezet wordt naar een SQL-query voor een database. Het resultaat van de query wordt voor zover nodig omgezet in natuurlijke taal. Niet alle input van de gebruiker zal vertaald worden naar een query. Voor een deel zal die invoer niet relevant zijn, voor een ander deel maakt ze deel uit van een dialoog met het systeem om tot een uiteindelijke query te kunnen komen. Binnen het project is een corpus van dialogen geconstrueerd. Via een Wizard of Oz experiment (zie sectie 4) en via interviews met schouwburginformatrices wordt geprobeerd het waarheidsgetrouw-zijn-gehalte van de dialogen te toetsen en eventueel te verhogen. Het te ontwerpen systeem gaat (noodgedwongen) uit van in- en uitvoer via toetsenbord en scherm van een pc. Uiteindelijk mag men wat betreft het gebruik van een dergelijk systeem denken aan spraak via telefoon en/of huis-pc. SCHISMA is een samenwerkingsproject tussen de Parlevink

onderzoeksgroep en de spraak- en taalgroep van PTT Research te Leidschendam.

4. *Niet-Traditionele Benaderingen in Language en Information Engineering.* Kenmerkend voor het onderzoek binnen Parlevink is de aandacht voor nieuwe computationele modellen en berekeningswijzen binnen de informatietechnologie en de toepassing ervan binnen de taaltechnologie. Daarbij moet gedacht worden aan onderwerpen als neurale netwerken, cellulaire automaten, genetische algoritmen, fuzzy logic, etc. De belangstelling voor deze onderwerpen vloeit ook voort uit de gedachte dat onderzoek op deze terreinen resultaten en ideeën kunnen opleveren die gebruikt kunnen worden bij het ontwerpen van systemen waarbij sprake is van integratie van verschillende perceptuele processen, in het bijzonder taal (spraak) en beeld (vision).

Toelichting: De richting van het onderzoek in dit deelproject wordt niet direct bepaald door mogelijke toepassingen. De kern van het onderzoek is overigens wel taal-georiënteerd. Hoe kan een neurale netwerk beslissen of een bepaalde sequentie van symbolen (woorden) wel of niet acceptabel is binnen een bepaalde taal? Hoe kan met behulp van neurale netwerken of genetische algoritmen uitgaande van een verzameling voorbeeldzinnen de bijbehorende grammatica en/of automaat (ontleder) geleerd worden? Dergelijke fundamentele vragen behoren tot de kern van het onderzoek binnen dit deelproject. Via stages en D-opdrachten en via participatie in het Neuro-Fuzzy Centre van de regio Twente-Münster komen ook andere toepassingen aan bod. Meestal hebben deze toepassingen op een of andere wijze te maken met 'text retrieval', 'optical character recognition' (OCR), 'text recognition' en andere vormen van beeldverwerking ten behoeve van tekst processing. Op deze terreinen is sprake van een vrij hechte, maar niet geïnstitutionaliseerde samenwerking met de University of Texas in Austin en met Nederlandse bedrijven zoals Océ R&D (Venlo) en Sentient Machine Research (SMR, Amsterdam).

5. *Interactie tussen Sociale Wetenschappen, Filosofie en Language Engineering.* Language engineering kan ook op een minder wetenschappelijke en technologische wijze bron van onderzoek zijn. In dit deelproject wordt aandacht besteed aan maatschappelijke, filosofische en methodologische aspecten. Dit gebeurt in samenwerking met de Faculteit Wijsbegeerte en Maatschappijwetenschappen.

Toelichting: De interactie tussen language engineering en haar maatschappelijke context is een van de onderwerpen die hier nader wordt bestudeerd. Wat voor maatschappelijke effecten treden op bij de integratie van taalproducten in de maatschappij? Is er sprake van een specifieke maatschappelijke verantwoordelijkheid van de

'language engineers'? Welke vorm krijgt technology assessment indien gericht op language engineering? Een tweede onderwerp dat aandacht krijgt vloeit voort uit het blijkbaar grote verschil tussen taal en rekenen wanneer het gaat over het modelleren van intelligentie. In het bijzonder zal gekeken worden hoe een pragmatistische invalshoek meer licht kan werpen op het traditionele 'mind-body' probleem. Tenslotte, als derde onderwerp wordt het huidige informatietechnologische paradigma van toepassing van algemene en universele formalismen op objectieve representaties van informatie aan een onderzoek onderworpen. De vraag wordt gesteld in hoeverre er behoefte is aan een paradigma waarmee meer nadruk kan komen te liggen op relaties tussen formalismen, actuele representaties van kennis en de functionele of kontekstuele aspecten die daar een link tussen leggen.

7. Andere Parlevink Activiteiten

Gerelateerd aan het onderwijs en onderzoek binnen Parlevink zijn tal van andere activiteiten. Deze activiteiten variëren van ondersteuning van studentenactiviteiten (excursies, organisatie Interactief/Ideefiks conferentie *Linguistic Engineering in Context '93*) tot deelname aan organisatie van professionele activiteiten in het kader van de Special Interest Group on Parsing Technologies (SIGPARSE) van de Association of Computational Linguistics (ACL). Speciale vermelding verdient de participatie voor een deel van het onderzoek in het NeuroFuzzyCentre (een samenwerking in het kader van Euregio), het CTIT (Centre for Telematics and Information Technology), de onderzoekscholen i.o. op het gebied van Programmatuurkunde en Algoritmiek en die op het aandachtsgebied van het CTIT. Internationaal trekken vooral de door Parlevink georganiseerde Twente Workshops on Language Technology de aandacht. Het gaat om zesmaandelijks workshops, steeds op een ander thema binnen de taaltechnologie en met elke keer een aantal uitgenodigde buitenlandse onderzoekers die hun sporen op dat thema hebben verdiend.

8. Studenten en Parlevink

Een groeiend aantal studenten verricht een stage en/of D-opdracht binnen het project. Stages worden aangemoedigd, bijvoorbeeld bij Océ R&D (Venlo), Sentient Machine Research (Amsterdam), TNO-Documentaire Informatiesystemen (Rijswijk), CAP Volmac Lingware (Utrecht) en het Nijmeegs Instituut voor Cognitie en Informatica (NICI). Stages en D-opdrachten vinden ook plaats in de Verenigde Staten, bijvoorbeeld bij Carnegie Mellon University (Pittsburgh), University of Texas (Austin) en Indiana State University (Bloomington). Om eens een voorbeeld te noemen,

bij Carnegie Mellon wordt gewerkt aan een meertalig spraaksysteem dat helpt bij het maken van afspraken. Het systeem moet ook zinnen genereren, zoals, 'Maandag om 10.00 uur 's ochtends is goed'. Het genereren van dit soort zinnen vanaf een semantische representatie is onderwerp geweest van een recente stage van een student. Het is niet ongebruikelijk dat stage en/of D-opdracht leidt tot een wetenschappelijk artikel dat op een internationaal congres gepresenteerd wordt. In 1994 worden door Parlevink-studenten artikelen gepresenteerd in Ankara, Seattle (2) en Kyoto. Tal van afgestudeerden vinden een plaats als AiO (Assistent in Opleiding) aan een Nederlandse universiteit of onderzoeksinstituut.

9. Parlevink 2000

Zoals in de inleiding gesteld gaat Parlevink uit van de triade telematics, information en language engineering. De nadruk ligt op dit moment uitdrukkelijk op language engineering. In de toekomst zou op een of andere wijze onderzoek en ontwikkeling gericht op integratie een meer nadrukkelijke rol moeten krijgen. Hierin kan uitdrukkelijk de kontekst van het CTIT met haar aandacht voor integratie van verschillende invalshoeken bij systeemontwerp een belangrijke rol spelen. Het is een dergelijke kontekst die meestal bij projecten op het gebied van language engineering volledig ontbreekt. De mogelijkheid tot het aanwenden van toekomstige 'state of the art' methoden op het gebied van telematics en information engineering in combinatie met het huidige Parlevink onderzoek op het gebied van language engineering is een aantrekkelijk vooruitzicht. Een onderwerp dat nu nog vrijwel volledig buiten beschouwing blijft in het onderzoek is spraakverwerking. Aangezien tal van toepassingen slechts mogelijk worden als taalsystemen van een spraakcomponent zijn voorzien is dit een onderwerp waarvan bekeken moet worden hoe in de toekomst hier onderzoekscapaciteit aan kan worden toegekend.