# VAN REKENMACHINE TOT TAALAUTOMAAT

# Van rekenmachine tot taalautomaat

Leonoor van der Beek

© 2010 Leonoor van der Beek
ISBN: 978-90-816570-1-3

Binnenwerk: Zink Typografie (www.zinktypografie.nl)
Omslag: Uncinc (www.uncinc.nl)
Foto binnenflap: Bas Hiemstra
Drukwerk: Ipskamp Enschede

# Voorwoord

Soms denk je een goed idee te hebben, en dan moet zo'n idee op stel en sprong uitgevoerd worden, je kunt niet meer wachten. Zo herinner ik me dat enkele maanden voordat mijn opa 79 jaar oud zou worden, het idee ontstond om bij zijn tachtigste verjaardag een grote familiereünie te organiseren. Iedereen leek enthousiast over dit idee, totdat iemand de opmerking maakte dat er in een jaar veel kon gebeuren, en dat opa nú nog goed gezond was. Kortom: waarom moesten we eigenlijk tot zijn tachtigste jaar wachten? Konden we die reünie niet beter onmiddellijk organiseren, want je kon immers niet weten… (de reünie werd inderdaad op opa's 79e verjaardag georganiseerd, en uiteindelijk is opa in goede gezondheid 96 jaar oud geworden).

Hetzelfde gevoel van urgentie beleefde ik een kleine twee jaar geleden, toen op de receptie van de CLIN in Groningen in gesprek met Jan Odijk de gedachte ontstond dat de geschiedenis van ons vakgebied in ons deel van de wereld eigenlijk zou moeten worden opgeschreven, 'voor het te laat was' – en de helden van het eerste begin er niet meer zouden zijn of althans het verhaal niet meer konden vertellen.

De geschiedenis van het vakgebied – het leek me een prima idee, maar hoe voer je zoiets uit? Ik sprak erover met een aantal mensen, en ik had onmiddellijk ook bedacht dat ik Leonoor van der Beek

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



om advies wilde vragen. Leonoor is in Groningen gepromoveerd als één van de AiO's in mijn PIONIER-project. Daarom was ik ervan op de hoogte dat ze – naast Nederlandse taalkunde – ook journalistiek had gestudeerd. Haar afstudeerscriptie had ik gelezen als onderdeel van de sollicatie. De scriptie ging over de receptie van de Montague-grammatica in Nederland, heel interessant en heel leesbaar geschreven – hoe vaak lees je een afstudeerscriptie vrijwillig van begin tot eind? – en enigszins vergelijkbaar met onze plannen. Ik wilde haar dus eens vragen hoe we zoiets moesten aanpakken.

Toen ik Leonoor vertelde wat we wilden was haar reactie: "Wat een leuk project! Dat wil ik zelf doen!". Die reactie had ik niet verwacht, maar was precies wat het project nodig had. Leonoor wist Q-go, het bedrijf waar ze werkzaam is, te overtuigen om haar een jaar lang voor de helft van de tijd onbetaald verlof te geven. Gedurende die periode kreeg ze bij ons in Groningen een part-time aanstelling. Er waren natuurlijk wel enkele knopen die moesten worden doorgehakt, bijvoorbeeld over de afbakening in tijd en plaats van het onderwerp, de preciese aanpak, met wie moet een afspraak gemaakt worden voor een interview, enzovoorts, maar het project had momentum gekregen en al deze vragen kregen vanzelf een antwoord – waarbij de titel van het uiteindelijke boek nog de meeste hoofdbrekens kostte.

Het project is mede mogelijk gemaakt door de Faculteit der Letteren van de Rijksuniversiteit Groningen, in het bijzonder de afdeling *computational linguistics* van het CLCG onderzoeksinstituut; het meerjarige onderzoeks- en stimuleringsprogramma STEVIN; de Stichting Taaltechnologie in Utrecht, en de Nederlandse Taalunie. We zijn ook Q-go dankbaar dat ze Leonoor de ruimte voor dit project hebben geboden. Maar het meest dankbaar ben ik Leonoor zelf, die met haar enthousiasme dit project tot een heel goed einde heeft gebracht, waardoor we nu vol trots het resultaat aan alle collega's kunnen aanbieden!

Gertjan van Noord

Zink Typografie Sheet 6 of 213 - Page ii of 205

# Inhoudsopgave

**Inleiding** *1*

**1 Pioniers** *5*
1.1 Het sprookje van Rekenen en Taalkunde *5*
1.2 Icoon van de computerlinguïstiek *8*
1.3 Rekenaars *13*
1.4 Kwantitatieve taal- en letterkunde *17*
1.5 PHLIQA *22*
1.6 Een aanloop van een kwart eeuw *35*

**2 De vertaalmachine** *41*
2.1 Wat eraan voorafging *41*
2.2 Eurotra *45*
2.3 METAL *59*
2.4 Distributed Language Translation *65*
2.5 Rosetta *71*
2.6 Computational Linguistics in the Netherlands *83*

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



**3 Spraak** *97*
3.1 Dinosaurussen *97*
3.2 De techniek *105*
3.3 Internationale netwerken *107*
3.4 1984 *111*

**4 Industrie** *121*
4.1 Vrijheid, blijheid *122*
4.2 Op zoek naar tastbaar resultaat *127*
4.3 De grote belofte *141*
4.4 Ondertussen in Nederland *162*

**5 Taal en spraak in dialoog** *171*
5.1 Openbaar Vervoer Informatie Systeem *172*
5.2 Corpus Gesproken Nederlands *181*
5.3 Samen werken aan technologie? *184*

**6 Epiloog** *191*

**Bibliografie** *193*

**Index** *201*

Zink Typografie Sheet 8 of 213 - Page iv of 205

# Inleiding

Dit boek is geen wetenschappelijke taal- of spraaktechnologische publicatie. Dit boek vertelt de Nederlands-Vlaamse geschiedenis van het vakgebied aan de hand van de persoonlijke verhalen van enkele pioniers uit het vak. Ik heb een aantal van deze mensen opgezocht en geïnterviewd. Bij hen thuis, in hun werkkamer of in een stationsrestauratie. Eén keer zelfs op het strand, zodat ik bij het uitwerken van de geluidsopnames begeleid werd door het rustgevende geluid van de golfslag tegen de Noordzeekust.

Ik realiseer mij dat mijn verhaal de geschiedenis van de Nederlandse en Vlaamse taal- en spraaktechnologie niet uitputtend behandelt. Dat had van dit boek een saaie opsomming gemaakt van proefschriften, oraties en afscheidscolleges – de drie ijkpunten in het professionele leven van een wetenschapper. In plaats daarvan heb ik getracht het boek te schrijven dat ik zelf graag had willen lezen toen ik vanuit de Neerlandistiek binnentrad in het vak. Het boek dat mij vertelde in welke traditie ik belandde, waarom de dingen zijn zoals ze zijn. En dan graag niet te zwaar. Een boek dat je in de trein, op weg naar congres of projectbespreking, lekker wegleest.

Of dat gelukt is, is aan de lezer om te beoordelen. Aan alle hulp die ik gekregen heb zal het niet liggen. Jan van Bakel, Peter de Bie, Martin

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat

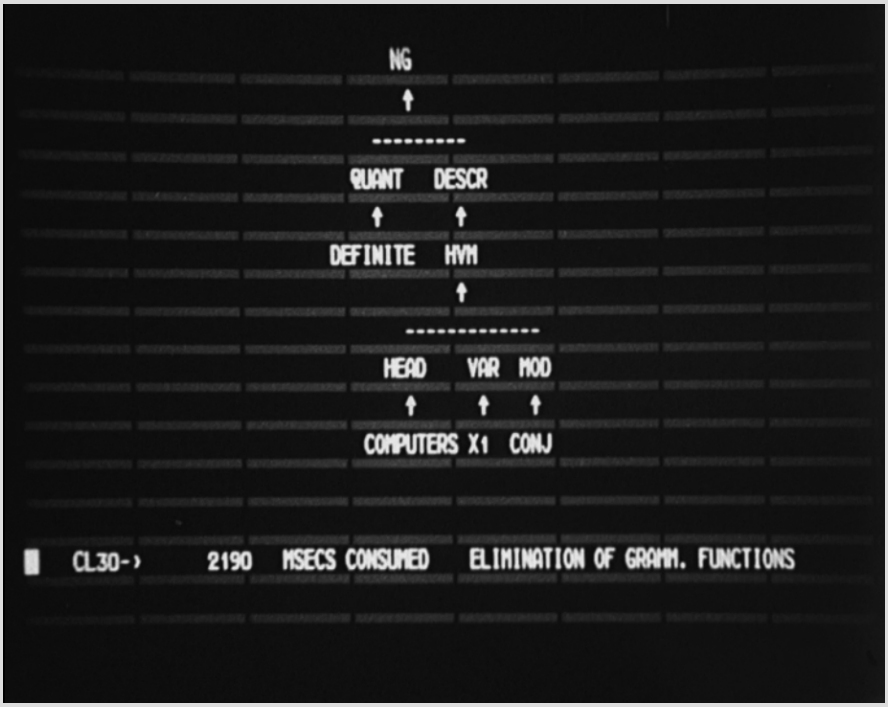


Boot, Gosse Bouma, Lou Boves, Hugo Brandt Corstius, Harry Bunt, Stan van de Burgt, Herman Caeyers, Dirk van Compernolle, Walter Daelemans, Alice Dijkstra, Frank van Eynde, Pius ten Hacken, Steven Krauwer, Franciska de Jong, Gerard Kempen, Jan Landsbergen, Jean-Pierre Martens, Norbert Mergen, Martin Spitholt, John Nerbonne, Anton Nijholt, Sieb Nooteboom, Jan Odijk, Els den Os, Louis Pols, Jan van Sas, Remko Scha, Bondi Sciarone, Tigran Spaan, Luc Steels, Toon Witkam, Jakub Zavrel en Frans Zwarts maakten tijd voor mij vrij in hun vaak drukke agenda's, waarvoor ik hun erg dankbaar ben. Wanneer een quote in dit boek *niet* voorzien is van een referentie, is het gezegde afkomstig uit een van deze interviews. En dan zijn er nog al die anderen, die mij geholpen hebben door een vraag te beantwoorden, een artikel naar me op te sturen, mij boeken of een filmrol te lenen of op enige andere wijze informatie naar mij door te spelen die bruikbaar was voor het verhaal.

Het initiatief kwam van Gertjan van Noord en Jan Odijk. Wat een goed idee! Beiden zijn een belangrijke rol blijven spelen in het project. Gertjan als formele opdrachtgever, klankbord en editor – zijn input zit overal in dit boek. Jan als eerste lezer van het volledige – dacht ik toen – manuscript. Het boek is beter geworden door hun inbreng. De tweede lezer was Eveline Meurs, die ontelbaar veel fouten en foutjes voorkomen heeft. Ze heeft goed voor mijn kindje gezorgd.

Q-go ben ik erg dankbaar dat ze mij niet hebben laten kiezen tussen mijn leuke baan bij hen en het mooie project waarvan het resultaat nu voor je ligt. En ook mijn dierbare oude afdeling bij de Rijksuniversiteit Groningen, STEVIN, de Stichting Taaltechnologie in Utrecht en de Nederlandse Taalunie ben ik dank verschuldigd voor hun bijdrage aan dit project.

Zink Typografie Sheet 10 of 213 - Page 2 of 205




PHLIQA-analyse van de vraag “Hoeveel computers zijn er?”

# 1 Pioniers

## 1.1 Het sprookje van Rekenen en Taalkunde

Bestuurders van de Universiteit Amsterdam, hoogleraren van de faculteit Wis- en Natuurkunde, medewerkers van het Mathematisch Centrum en studenten kwamen op 27 oktober 1952 samen om te luisteren naar de inaugurele reden van de nieuwe bijzonder hoogleraar wiskunde Adriaan van Wijngaarden. "U kent allen het sprookje van Assepoester," (van Wijngaarden 1952) begon hij. "Het bescheiden meisje, dat met blijdschap het huishoudelijk werk verrichtte in het gezin met de schone zusters, die haar zo harteloos behandelden. Een vriendelijke petemoei zag echter kans haar met behulp van enige tovenarij zo aantrekkelijk te maken, dat zij en niet de zusters het hart won van de prins." De Assepoester in het sprookje van Van Wijngaarden is Rekenen, het ondergeschoven kindje van de wiskunde. Maar dankzij de toverkracht van de moderne computer wordt zij zo aantrekkelijk, dat zij de schone prins Taalkunde weet te bekoren. Samen, zo stelt Van Wijngaarden, kunnen zij een nieuw koninkrijk stichten, dat Automatisch Vertalen heet.

We spreken 1952. Nederland bezit nog geen werkende computer. Eerder dat jaar was de eerste Nederlandse computer ARRA (Auto-

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



matische Relais Rekenmachine Amsterdam) officieel en in bijzijn van minister Rutten van Onderwijs in gebruik genomen op het Mathematisch Instituut, maar direct na de demonstratie gaf de computer de geest – het apparaat heeft nooit meer gewerkt. In de taalkunde werd nog slechts beschrijvend gewerkt. Structuralisme was alles wat er was. *Syntactic Structures* (Chomsky 1957), de eerste taalkundige publicatie die natuurlijke taal beschrijft als een algoritme, een wiskundig systeem, moest nog geschreven worden.

Nee, Adriaan van Wijngaarden was niet de eerste die de potentie van automatisch vertalen verwoordde. Warren Weaver, directeur van de Natural Sciences Division van de Amerikaanse Rockefeller Foundation, schreef al in 1949 een memorandum (Weaver 1949), waarin hij zijn ideeën voor automatische vertaling uit de doeken deed. En enkele maanden voor Van Wijngaardens oratie, in juni 1952, werd al de eerste conferentie over automatisch vertalen georganiseerd door Yehoshua Bar-Hillel, die een baan had gekregen aan het Massachusetts Institute of Technology (MIT), speciaal om te onderzoeken wat de potentie was van automatisch vertalen. Weaver's memo werd pas in 1955 gepubliceerd, en van het MIT-congres bestonden geen proceedings. Ook andere publicaties op het gebied van machinaal vertalen waren er nog niet. Toch is het waarschijnlijk dat Van Wijngaarden vrij precies op de hoogte was van de ontwikkelingen in de Verenigde Staten. De professor reisde veel, en was in 1947 op een grote reis door de VS geweest. Ook had hij contact met de enige niet-Amerikaan op het congres in 1952 in MIT, de Brit Andrew Booth. Zelfbedacht zullen de ideeën over automatisch vertalen dan waarschijnlijk ook niet zijn geweest. Maar de grondlegger van de informatica in Nederland was wel visionair genoeg om de mogelijkheden van de computer en de taalkunde te herkennen en deze nieuwe richting als onderzoeksgebied in de schijnwerpers te zetten, ondanks het feit dat noch de computers, noch de taalkundigen klaar waren voor deze samenwerking.

Het inzicht van Van Wijngaarden ging verder. Hij signaleert in zijn rede dat zelfs de simpelste aanpak van automatisch vertalen, woord voor woord vertalen, moeilijkheden oplevert, namelijk bij woorden die ambigu zijn. Over dit probleem zegt hij: "De fraaiste oplossing is natuurlijk de machine te leren de passende vertaling te kiezen op grond van andere woorden in de tekst. Dit is evenwel niet eenvou-

Zink Typografie Sheet 14 of 213 - Page 6 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



dig." Een understatement. Zijn voorbeelden laten zien dat hij begrepen heeft dat hiervoor geen deterministische regels op te stellen zijn, maar alleen waarschijnlijkheden te bepalen zijn. Nadat hij kort het probleem van de zogenaamde multi word units (woorden met spaties) aanstipt, stelt hij bovendien dat woord voor woord vertalen slechts een "pover" resultaat oplevert, en dat daarom grondige kennis van de syntaxis nodig is van de talen waarin en waaruit men wil vertalen. Maar dat is kennis die op dat moment nog ontbreekt: "Nu is de taalwetenschap op dit punt van uit een wiskundig standpunt gezien nog in haar kinderschoenen. Weliswaar hebben wij zogenaamd intuïtief een idee van syntaxis, maar de regels die wij toepassen, missen ieder spoor van strengheid en efficiëntie, zoals voor het gestelde doel nodig is." Rekenaars en taalkundigen worden opgeroepen samen te werken om dit gat te vullen.

De rede van Van Wijngaarden verklapt niet hoe het sprookje eindigt. Hij voorspelde slechts dat "binnen afzienbare tijd wezenlijk resultaat valt te verwachten." Hij kon niet weten welke enorme ontwikkelingen Taalkunde en Rekenen door zouden maken, hoe Taalkunde door Rekenen overschaduwd zou worden en hoe ze uiteindelijk toch samen zouden gaan werken. Het vak computerlinguïstiek, waar computerwetenschappen en taalkunde elkaar kruisen, bestond nog niet. Er moest nog ontdekt worden dat er naast Automatisch Vertalen nog andere toepassingsdomeinen bestonden, en dat de computerlinguistiek op haar beurt weer onderdeel uitmaakt van wat wij taal- en spraaktechnologie zijn gaan noemen.

De hoogleraar realiseerde zich niet dat op het moment dat hij deze rede uitsprak, er in zijn directe omgeving twee mensen rondliepen die een belangrijke rol zouden gaan spelen in het vervolg van het verhaal. Het is goed mogelijk dat beiden in de zaal zaten toen hij zijn rede uitsprak. De eerste speelde een faciliterende rol in dit verhaal: Evert Willem Beth. De Amsterdamse logicus had als hoogleraar onder meer zitting in de Faculteit Wis- en Natuurkunde van de Universiteit van Amsterdam (UvA), waarin Van Wijngaarden toegetreden was. Eerder dat jaar had Beth zijn *Instituut voor grondslagenonderzoek en philosophie der exacte wetenschappen* opgericht in een kamertje op het Mathematisch Instituut van de UvA op de Roetersstraat. Er is echter geen enkel bewijs van enige interactie tussen de twee wetenschappers. Toch was het Beth die elf jaar later een lans brak voor

Zink Typografie Sheet 15 of 213 - Page 7 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



de Generatieve Grammatica door middel van zijn rede 'Constanten van het mathematisch denken' ten overstaan van Koninklijke Nederlandse Akademie van Wetenschappen (KNAW). Hij bood opvang aan de eerste generativisten, die elders nergens terecht konden, door ze bij hem te laten promoveren (van der Beek 2001). En juist deze taalkundigen waren actief op zoek naar een syntaxis met de "strengheid en efficiëntie" die volgens Van Wijngaarden nodig was voor automatisch vertalen. Ook was het Beth, die Montague naar Nederland haalde. De logicus Richard Montague legde de basis voor de formele semantiek die grote invloed zou hebben op de computerlinguïstiek. Zijn ideeen inspireerden onder meer Jan Landsbergen, de architect van Neerlands meest gerespecteerde vertaalsysteem Rosetta. Tenslotte is Beth de opleider van de tweede centrale figuur uit de begindagen van het vak, misschien wel de bekendste Nederlandse computerlinguïst ooit, promovendus van Van Wijngaarden en degene die de wereld vertelde dat automatisch vertalen niet meer dan een sprookje is: Hugo Brandt Corstius.

## 1.2 Icoon van de computerlinguïstiek

Er bestaan maar een paar wetenschappelijke publicaties van de hand van Hugo Brandt Corstius. Een school of stroming heeft hij niet opgericht, niet meer dan een handvol computerlinguïsten zijn bij Brandt Corstius gepromoveerd, en zelfs doctoraalstudenten computerlinguistiek heeft hij nauwelijks opgeleid. Toch wordt zijn naam met eerbied genoemd. Ook wie het niet met hem eens was of het met hem aan de stok had – zijn scherpe pen is berucht – erkent zijn positie als grondlegger van het vak in Nederland – en in mindere mate in Vlaanderen. Waarom? "Omdat ik de eerste was", antwoordt hij zelf. "Maar dat stelt niets voor. Eigenlijk heb ik niet zo veel verdiensten voor het vak. Behalve dat ik zei dat automatisch vertalen niet kon."

Brandt Corstius begon zijn carrière als computationeel taalkundige in de volle overtuiging dat hij persoonlijk het levenslicht zou schenken aan machinaal vertalen. "Dat varkentje zou ik wel eens wassen. Jammer dat er geen Nobelprijs voor dat soort zaken bestond", aldus Brandt Corstius over zijn eigen ongebreidelde optimisme in

Zink Typografie Sheet 16 of 213 - Page 8 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



die tijd (Brandt Corstius 2003). Het onderwerp van de oratie van Van Wijngaarden fascineerde hem direct[1], maar toen hij in 1953 bij Evert Willem Beth wiskunde ging studeren, werkte hij niet aan machinaal vertalen. Wel zocht hij binnen de wiskunde ("omdat het vak informatica niet bestond" (Brandt Corstius 2003)) naar toepassingen van dat vak in de taalkunde en de computerwetenschappen. Brandt Corstius leerde programmeren. In ieder geval in theorie, want er was voor studenten geen computer beschikbaar om programma's op te draaien. En met zijn programmeerkennis probeerde hij antwoord te vinden op de vraag of het mogelijk is om de computer de grens tussen twee lettergrepen te laten bepalen. Kun je een computer leren dat het *woes-te* is, maar *toe-stand*? Dat het *ges-te* is, maar *ge-ste-gen* (voorbeelden uit (Brandt Corstius 1970))? Op papier werkte Brandt Corstius een programma uit dat hiertoe in staat was, maar testen kon hij het nog niet. In een poging de militaire dienst te ontwijken vertrok Brandt Corstius in 1961 naar de Verenigde Staten. Beth had hem een introductie bezorgd om op Berkeley bij de beroemde wiskundige Tarski te werken. En Berkeley had wel een computer waar studenten gebruik van mochten maken. Voor het eerst kon Brandt Corstius zijn lettergrepenprogramma draaien. 's Avonds leverde hij zijn stapel ponskaarten in, en de volgende ochtend haalde hij de uitvoer op. Tot verbazing van de directeur daar. "Wat doet u nou eigenlijk?" vroeg die hem na enige tijd, "uw uitvoer is altijd hetzelfde als uw invoer!" Bijna hetzelfde. Alleen stonden er soms streepjes tussen. Nu noemt Brandt Corstius dat werk "kinderachtige programma's." "Hij heeft die paar dingen gedaan die echt kunnen", nuanceert vak-

[1] Het is niet duidelijk of Brandt Corstius aanwezig was bij de inaugurele rede van Van Wijngaarden. In zijn keynote lecture op CLIN 2002 zegt hij van wel: "Ik liep op 29 November 1952 door de Oudemanhuispoort in Amsterdam omdat ik van plan was daar het volgend jaar, na mijn eindexamen, wiskunde te gaan studeren. Ik viel met mijn neus in de boter. Een nieuwe hoogleraar hield zijn inaugurele rede. Ik wist niet hoe hij heette en welk vak hij ging geven. Ik hoorde de rede met groeiende opwinding aan. Het ging over het rekenen op computers." (Brandt Corstius 2003) In 2009 ontkent hij echter daar geweest te zijn. "Daar ben ik nooit geweest. Ik kwam in '53 in Amsterdam. Misschien zei ik dat ik er geweest ben, maar dat is niet zo." In ieder geval is de eerder genoemde datum 29-11-1952, de dag waarop Brandt Corstius naar Amsterdam zou zijn gekomen, ruim een maand *na* de oratie, die op 27 oktober 1952 plaatsvond.

Zink Typografie Sheet 17 of 213 - Page 9 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



genoot Jan Landsbergen. In ieder geval is het afbreken van woorden nog ver verwijderd van machinaal vertalen. Dat Brandt Corstius toch ging werken aan die droom van de vertaalmachine komt door een misverstand dat inmiddels een bekende anekdote is geworden.

Hugo Brandt Corstius was niet gelukkig op Berkeley. Tarski wilde zijn nieuwe Hollandse student laten promoveren op een onderwerp dat hij helemaal niet zag zitten. En toen zat Brandt Corstius op een goede dag in de bus van Berkeley naar San Francisco met de Amsterdamse hoogleraar Adriaan Van Wijngaarden, die op dat moment enige tijd aan Berkeley doceerde. Brandt Corstius was niet vergeten hoe de professor ooit het droomscenario schetste waarin wiskunde en taalkunde samen machinaal vertalen zouden realiseren. Dus toen Van Wijngaarden hem vroeg om in Amsterdam computertaalkunde te komen doen, zei hij direct ja. Eenmaal aan de slag bij het Mathematisch Centrum bleek echter al snel dat Van Wijngaarden, een van de architecten van programmeertaal Algol, wilde dat hij zou gaan werken aan computertalen. Computertaal-kunde dus. Brandt Corstius biechtte op dat hij dacht computer-taalkunde te gaan doen en herinnerde de professor aan zijn inaugurele rede. Van Wijngaarden gaat overstag: Hugo Brandt Corstius krijgt als opdracht uit te zoeken of machinaal vertalen mogelijk is.

Naar eigen zeggen kostte het hem een half jaar om tot de conclusie te komen dat automatisch vertalen een sprookje is; dat zijn generatie het in ieder geval niet meer mee zou maken dat het realiteit zou worden. Hij bleef het wel nog enige tijd proberen, en toen hij in 1970 bij Van Wijngaarden promoveerde bevatte zijn proefschrift (Brandt Corstius 1970) zowel een hoofdstuk over het automatisch vertalen van getalsnamen, als een duidelijke stellingname tegen het idee dat machinaal vertalen op korte termijn tot bruikbare resultaten zou leiden. Stelling 3 bij het proefschrift luidt: "Mechanische vertaling, ofschoon in de nabije toekomst niet te verwezenlijken, verschaft een nuttig gezichtspunt voor vele problemen in *computational linguistics*" – mogelijk de eerste vindplaats van de Engelse term voor computerlinguïstiek in een Nederlandse tekst. De in de bijzin verpakte conclusie dat machinaal vertalen op korte termijn niet haalbaar was, werd overigens in Vlaanderen in 1969 ook al getrokken door Flip Droste in zijn boek *Vertalen met de computer* (Droste 1969). Teleurge-

Zink Typografie Sheet 18 of 213 - Page 10 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



steld wendde Brandt Corstius zich af van het automatisch vertalen, en richtte zich meer en meer op de algemene taalwetenschap. De wiskundige doceerde Transformationele Grammatica aan het instituut voor Neerlandistiek van de UvA, en later bij Algemene Taalwetenschappen. Maar in deze colleges bleef hij wel computationele vraagstukken aan de orde stellen. Hij onderwees zelfs de programmeertaal Algol68 aan zijn studenten, waaronder Frans Zwarts, die vele jaren later de vakgroep alfa-informatica aan de Rijksuniversiteit Groningen mede zou oprichten en nog weer later rector magnificus zou worden van die universiteit. "Hij onderwees ons dat er abstracte computationele begrippen zijn waarmee je taal te lijf kunt. […] Hij kon onconventioneel naar taal kijken en problemen definiëren. Het was een wonderlijke man, maar ik heb veel van hem geleerd."

Brandt Corstius' desillusie op het gebied van automatisch vertalen betekende niet dat hij de computertaalkunde helemaal opgaf. Hij bleef zeer betrokken bij het vak, stimuleerde het gebruik van de computer in de Letterenfaculteit in het algemeen en de kwantitatieve taal- en letterkunde in het bijzonder, en speelde een rol in de totstandkoming van het vak computerlinguïstiek, aspecten die in de volgende paragrafen aan bod zullen komen. Hij pleitte verder voor transparantie in artikelen over computerlinguïstiek: het moest duidelijk zijn hoe de gerapporteerde resultaten precies behaald werden. Dit was een reactie op enkele gevallen van "sjoemelen" uit die tijd: programma's die Kunstmatige Intelligentie leken te bevatten, maar eigenlijk voor één specifiek domeintje of trucje geprogrammeerd waren. Bekende voorbeelden waren het programma Eliza, dat leek te converseren als een mens, en daarmee zou slagen voor de beroemde Turing test,[2] maar eigenlijk met een heel simpel mechanisme werkte. En SHRDLU, het programma van Winograd dat via natuurlijke taal opdrachten kon uitvoeren in een blokkenwereld, maar bij nadere beschouwing feitelijk niet meer kon dan de voorbeelden uit de demo. Een artikel van Ed Hoenkamp over het computerprogramma FRUMP, dat krantenarti-

[2] "The idea of the test is that the machine has to try and pretend to be a man, by answering questions put to it, and it will only pass if the pretence is reasonably convincing"(Newman, Turing, Jefferson & Braithwaite 1952). Een recente publicatie in NRC Handelsblad laat zien dat wetenschappers niet overtuigd zijn dat enig programma tot op heden de Turingtest doorstaan heeft (Mols 2010).

Zink Typografie Sheet 19 of 213 - Page 11 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



kelen zou kunnen lezen (NRC, 20 februari 1980), classificeerde Brandt Corstius in Battus (1983) als bedrog: emeritus hoogleraar Harry Bunt van de Universiteit van Tilburg rapporteerde destijds een *success rate* van slechts 10% voor FRUMP (Bunt 1980). Tussen neus en lippen door schaarde hij overigens ook Martin Boot's argumentatie dat automatisch vertalen wèl mogelijk is wanneer kennis van de wereld in de vorm van frames gecodeerd wordt, en ook zijn eigen programma voor het oplossen van vierkantsvergelijkingen onder dezelfde categorie. Volledige helderheid over de totstandkoming van resultaten wordt maar op één manier bereikt: als de volledige broncode wordt meegeleverd. En zo worden vele pagina's van Brandt Corstius' proefschrift gevuld met de code van zijn programma's. "Alleen het punt was", zo merkt Landsbergen op, "dat kan wel met zijn programma's, die waren een paar pagina's lang, maar als een programma honderden pagina's wordt, dan heeft dat weinig zin." De behoefte aan mogelijkheden om resultaten te kunnen reproduceren bleef wel bestaan. Ondanks de introductie van formele evaluatieprogramma's in het vakgebied schrijft Ted Pedersen in 2008 nog in *Computational Linguistics* "If we believe in empirical methods and the value of comparisons and experimental studies, then we must also believe in having access to the software that produced those results as a necessary and essential part of the evidentiary process" (Pedersen 2008).

Brandt Corstius schreef zelf nog een paar kleine programma's na zijn proefschrift. Bijvoorbeeld in 1971, toen de overheid voorstelde om persoonsnummers in te voeren, en dat op veel weerstand stuitte bij de bevolking. Brandt Corstius schreef een script dat de nummers vertaalde naar woorden. Daarmee zou het plan meteen een stuk minder controversieel worden, want in plaats van een nummer, zouden mensen gewoon een extra naam krijgen. En dat is toch heel wat anders. Landsbergen herinnert zich dat hij zijn Philips personeelsnummer invoerde in het programma, dat per twee cijfers een combinatie van een medeklinker en een klinker toekende. De uitkomst? Jobidobi. Het feit dat Landsbergen dat bijna dertig jaar na dato nog weet, illustreert nog een voordeel van namen boven nummers: ze zijn zeer gemakkelijk te onthouden.

Het ongeloof van Brandt Corstius (en van Droste) in de haalbaarheid van automatisch vertalen betekende ook niet dat de onderne-

Zink Typografie Sheet 20 of 213 - Page 12 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



ming door anderen gestaakt werd. Verderop in dit boek zal een heel hoofdstuk gewijd worden aan de verschillende pogingen in de jaren tachtig om een happy end te schrijven aan het verhaal dat Van Wijngaarden begon, en dat Hugo Brandt Corstius trachtte de nek om te draaien. In de jaren negentig verdween machinaal vertalen wat naar de achtergrond, maar in 2010 was met 49 bijdragen op het gebied van *machine translation* bijna twintig procent van het programma van het mondiale congres van de *Association of Computational Linguistics* aan het onderwerp gewijd. Automatisch vertalen is inderdaad nog geen opgelost probleem, maar geen computerlinguïst laat zich meer afschrikken door de pessimistische woorden van Brand Corstius.

## 1.3 Rekenaars

Om het werk uit de begintijd van de computerlinguïstiek op waarde te schatten, is het noodzakelijk stil te staan bij de technologische stand van zaken in die periode. Van Wijngaarden voorziet mechanische vertaling op een moment dat er in Nederland nog geen werkende computer bestaat. Zoals gezegd had de eerste computer van het Mathematisch Instituut, ARRA, na één demo het leven gegeven, waardoor Shell begin 1953 met haar Miracle van Britse makelij de primeur had van de eerste echt werkende computer in Nederland. Een jaar na Van Wijngaardens oratie, in december 1953, voerde ARRA II op het Mathematisch Centrum haar eerste programma uit – waarna zij 2,5 jaar onafgebroken gerekend heeft (Dijkstra 1980). Het was Van Wijngaarden zelf, die zich hard maakte voor de ontwikkeling van computers op het instituut. In 1956 werd echter besloten dat computers bouwen te ver afstond van wiskunde bedrijven, en dat deze activiteiten bij een bedrijf ondergebracht dienden te worden. Philips werd gepolst, maar weigerde (Smeiers 2007), en het bedrijf Electrologica werd opgericht. Adriaan van Wijngaarden bleef zelf nauw betrokken bij het bedrijf, waar de voor die tijd zeer moderne X1 en X8 computers geproduceerd werden. Philips had deze wel willen produceren, maar was gebonden aan een contract met IBM, dat vastlegde dat IBM alleen bij Philips onderdelen zou kopen, als Philips geen concurrerende computers zou

Zink Typografie Sheet 21 of 213 - Page 13 of 205

produceren. Maar toen IBM het contract brak, nam Philips in 1968[3] Electrologica over, dat toen verhuisde naar Apeldoorn en samen met het daar gevestigde Philips Computer Industrie het nieuwe Philips-Electrologica vormde – een bedrijf dat later in dit hoofdstuk nog een prominente rol zal gaan spelen.

Deze eerste computers stonden niet zomaar tot ieders beschikking. De meeste mensen hadden nog nooit een computer gezien, en ook aan de universiteit bleef de computer lange tijd iets zeer exclusiefs. Voor studenten waren er geen computerfaciliteiten. Hugo Brandt Corstius ontwikkelde zijn woordafbreekprogramma met pen en papier, omdat hij geen beschikking had over een computer om zijn programma op te draaien. In de Verenigde Staten was de situatie wat beter, en kon hij eindelijk zijn programma draaien. Toen hij terug in Nederland in dienst kwam bij het Mathematisch Centrum, was de situatie wel iets verbeterd: op een gegeven moment stonden er maar liefst drie werkende computers. Maar voor Brandt Corstius bleef het schipperen, want de computers werden de hele dag gebruikt voor berekeningen ten behoeve van de nieuwe dijken die in de toekomst een watersnoodramp als die in 1953 moesten voorkomen. De computerlinguïst mocht alleen 's nachts gebruik maken van de rekenkracht van het instituut.

Net als in Amerika werkten ook de computers van het Mathematisch Instituut met ponskaarten. En niet alleen daar: de hele jaren zeventig nog gebruikten de meeste computers in Nederland ponskaarten of ponsband. Pas in de jaren tachtig werden terminals gemeengoed. Professor Franciska de Jong over het vak Formele Methoden en Technieken van Bill Dotson-Smith in Utrecht in de tweede helft van de jaren zeventig: "We moesten programmeren in Algol68. Ik weet nog dat dat een bijvak was van 400 uur. Ik denk dat ik de helft van die tijd op de fiets heb gezeten met ponskaarten op weg naar de Uithof. En achter de ponsmachine om de fouten eruit te halen." Al doende debuggen zat er niet in: de ponskaarten leverde je in, en een dag later kwam je de uitvoer ophalen. Bleek er een fout in de code gezeten te hebben, dan betekende dat een volle dag verlies.

[3] Sommige bronnen noemen 1966 als het jaar waarin Philips Electrologica overnam, zie bijv. `www.computer-erfgoed.nl/.../Tijdbalk_Comp_Hist_conc_080424_voor_website.pdf`.

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Behalve de beperkingen door in- en uitvoer, lag er ook een grote beperking in de hoeveelheid beschikbaar geheugen. Geheugen was schaars en duur. Emeritus hoogleraar computerlinguïstiek Jan van Bakel schat dat de IBM computer die de Katholieke Universiteit Nijmegen in 1965 bezat een geheugen had van ongeveer 350kB. In 1982 betaalde Jan Landsbergen van Philips 80.000 gulden (36.000 euro) voor een harde schijf van 256MB (Landsbergen 2009, pc) – tegen 40 euro voor 250GB in 2010. Dit gebrek aan geheugencapaciteit betekende bijvoorbeeld dat grote woordenboeken niet in geheugen ingelezen konden worden, en dat veel van de ontwikkeltijd ging zitten in het comprimeren en zo efficiënt mogelijk maken van programma's en resources. Het vraag-antwoordsysteem PHLIQA, hoogtepunt uit de pionierstijd en uitgebreid beschreven in paragraaf 1.5, had eind jaren zeventig slechts 480kB geheugen nodig. Maar daarvoor was dan wel de centrale computer van het Natuurkundig Laboratorium van Philips, 'Natlab', nodig. Een demo nam vrijwel het hele systeem in beslag, zodat in het hele bedrijf niemand anders op dat moment gebruik mocht maken van de computer. Eind jaren tachtig nam vertaalprogramma Rosetta zo'n 160MB in beslag. Het compileren van een nieuwe versie nam een hele nacht, en later zelfs een heel weekend in beslag op de VAX-11/780 computer.

Veel van de benodigde hulpprogramma's bestonden nog niet. Softwarebeheerpakketten en regelcompilers moesten zelf gemaakt worden. En de vraag was dan: in welke taal moesten de programma's geschreven worden? In het prille begin waren er nog helemaal geen machine-onafhankelijke programmeertalen, en moesten programma's direct in de machinetaal van de betreffende computer geschreven worden. In de zestiger jaren deden tweede generatie programmeertalen als Algol60, COBOL en PL1 hun intrede. Bedrijven als Philips ontwikkelden hun eigen talen. Een standaard was er niet, en iedereen verdedigde met overtuiging de taal van zijn keuze. Veel programmatuur uit die tijd is verdwenen, omdat de talen waarin ze geschreven waren uit de mode raakten, en er geen compilers meer voorhanden zijn om de broncode opnieuw te compileren.

De afwezigheid van internet en email maakte bovendien de communicatie tussen wetenschappers een stuk lastiger dan nu. Het duurde vaak lang voordat artikelen in proceedings of boeken verschenen. Om toch al iets te hebben dat verspreid kon worden, werden wor-

Zink Typografie Sheet 23 of 213 - Page 15 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



king papers geschreven. Deze werden niet officieel gepubliceerd, maar simpelweg gestencild en uitgedeeld. Het belang van conferenties was groot. Waar nu vooral de persoonlijke interactie, het netwerken en de discussie vooropstaan, waren conferenties en zomerscholen destijds ook werkelijk de manier om informatie te verzamelen over onderzoek dat elders gedaan werd. Luc Steels was een van de eerste computerlinguïsten van Vlaanderen, en had dus geen professoren, docenten of collega's van wie hij onderwijs of informatie kon krijgen "Voor mensen als ik, die ergens zaten waar verder geen andere computerlinguïsten waren, waren de zomerscholen een geschenk uit de hemel." In het bijzonder herinnert hij zich de eerste zomerschool in 1972 in Pisa. "Die werd georganiseerd door Antonio Zampolli. Alle hoofdrolspelers van dat moment had Zampolli uitgenodigd. Winograd [MIT], Woods [BBN], Kay [Xerox], Fillmore [Berkeley] en Wilks [Cambridge]. Voor mij is dat bepalend geweest." Vanuit Nederland waren er in ieder geval ook Hugo Brandt Corstius en wiskundige Steven Krauwer uit Utrecht, die elkaar op dat congres voor het eerst ontmoetten. "Raar genoeg ontmoette ik ook de Nederlandse wetenschappers vooral in het buitenland op congressen", aldus Krauwer. Ondanks het feit dat email en internet het inmiddels veel gemakkelijker hebben gemaakt om toegang te krijgen tot elkaars werk, zijn dit soort bijeenkomsten volgens Steels ook nu nog steeds heel belangrijk. "Mijn strategie is altijd geweest om je eerst af te vragen wie de interessante mensen zijn. En als je die gevonden hebt, dan ga je erheen of je nodigt ze uit. Persoonlijk contact, niet alleen maar papers lezen."

Wat betreft computerfaciliteiten liepen Nederland en Vlaanderen duidelijk achter op de Verenigde Staten. Hugo Brandt Corstius merkte dat al toen hij naar Berkeley ging en eindelijk zijn programma kon testen. Maar ook Luc Steels heeft het verschil direct ervaren. Eerst maakt hij een grote sprong vooruit: "Het werd mij snel duidelijk dat als ik iets wilde doen in dit vak, ik computerwetenschappen moest studeren. Maar dat kon niet in België. Er was wel een cursus hier en daar, maar geen studie. Dus toen heb ik de sprong gemaakt van de ponskaarten hier naar het AI lab van MIT. Dat was een tijdmachine. Ik was plots twintig jaar vooruit." Maar begin jaren tachtig zet hij de stap ook weer achteruit, wanneer hij naar Vlaanderen terugkeert. "Dat was moeilijk, ik ging terug in de tijd. Wij hadden in de VS internet en schermen, en toen moest ik in Brussel weer met ponskaarten en een centrale

Zink Typografie Sheet 24 of 213 - Page 16 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



machine gaan werken. Ik heb toen veel technologie binnengebracht, zoals locale netwerken en internet."

## 1.4 Kwantitatieve taal- en letterkunde

Misschien wel de meest voor de hand liggende toepassing van de computer in de taalkunde was de kwantitatieve taalkunde. Computers waren tenslotte grote rekenmachines, en in de kwantitatieve taalkunde moet veel geteld worden. Nog steeds is voor veel taalkundigen de kwantitatieve taalkunde dan ook de eerste kennismaking met computertoepassingen in het vak. Het is zeker niet het geval dat de kwantitatieve taalkunde *ontstaan* is door de opkomst van de computer. Al in 1934 besloot de Nederlandse overheid dat een commissie moest onderzoeken hoe op meisjesscholen in de Nederlandse kolonien in Indonesië het best Nederlands als vreemde taal onderwezen kon worden. Deze commissie concludeerde al snel dat hiervoor een frequentielijst gemaakt moest worden van de meest frequente woorden en collocaties van het Nederlands. J.F.H.A. de la Court voerde dit onderzoek uit, en de resultaten werden in 1937 gepubliceerd (de la Court 1937). De lijst is samengesteld op basis van één miljoen woorden, handmatig geteld en met een kaartsysteem gesorteerd. Jarenlang werd met deze lijst gewerkt, en nog in 1963 werd hij opnieuw uitgegeven (Linschoten 1963).

Toch was het nut van de oude woordenlijst enigszins beperkt. Enerzijds omdat hij verouderd was, en toegespitst op het tweedetaalonderwijs in Nederlands-Indië, anderzijds door de gevolgde methode. Om het enorme handwerk iets in te perken, worden woorden ingedeeld in frequentieklassen. Dat betekent dat van de meest frequente woorden geen exacte frequentie bepaald is: zodra vastgesteld kon worden dat ze tot de hoogste frequentieklasse behoorden, is opgehouden met tellen. In 1964 schrijft Brandt Corstius over deze telmoeheid: "Een computer kent zo'n vermoeidheid niet: de Elektrologica X1 kan het verlangde miljoen tekstwoorden in twintig nachten makkelijk tellen" (Battus 1983). Kortom: er is behoefte aan een nieuwe lijst, en er zijn nieuwe methoden voorhanden om deze automatisch te genereren.

Zink Typografie Sheet 25 of 213 - Page 17 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Van 1955 tot 1959 komt in Amsterdam de Werkcommissie voor Taalstatistiek bijeen om de mogelijkheden voor kwantitatief onderzoek naar het Nederlands te doen. In deze commissie hebben ook De la Court zelf en professor Van Wijngaarden zitting. De commissie bestaat verder uit een bont gezelschap van taalkundigen, fonetici, psycholinguisten, statistici, informatietheoretici en elektroengineers.[4] De commissie besluit opnieuw een telling van één miljoen Nederlandse woorden te doen, maar ditmaal beter verdeeld over verschillende typen teksten en elektronisch verwerkt. Hierdoor kunnen niet alleen woordfrequenties, maar ook letters, bigrammen, trigrammen, woordlengtes en woordklassen berekend worden. Het lettergrepenprogramma van Hugo Brandt Corstius wordt gebruikt om frequenties van lettergrepen te bepalen. Het onderzoek wordt uiteindelijk uitgevoerd door J.A.Th.M. van Berckel, H. Brandt Corstius, R.J. Mokken en A. van Wijngaarden. In 1965, zes jaar na opheffing van de werkcommissie, worden de resultaten over krantentekst (zo'n 45.000 woorden) gepubliceerd, voorzien van een voorwoord door Adriaan van Wijngaarden (van Berckel, Brandt Corstius, Mokken & van Wijngaarden 1965). De rest van het corpus is er nooit gekomen.

In andere landen en in andere talen komen langzaamaan grotere datasets ter beschikking. Professor L.K. Engels van het Instituut voor Toegepaste Linguïstiek in Leuven raakt geïnteresseerd in computertoepassingen voor taalonderwijs, en zijn enthousiasme brengt hij over op zijn student Willy Martin. Martin doet frequentieonderzoek op twee samples van 100.000 woorden Nederlands, één van kranten- en weekbladentekst, één van romantekst (Martin 1967). In Utrecht doet Mario Alinei van het Italiaans Instituut baanbrekend werk op relatief grote Italiaanse corpora: in 1962 publiceerde hij al een omgekeerd woordenboek van het Italiaans (Alinei 1962), in 1965 een frequentielijst van de Divina Commedia (Alinei 1965), en vanaf 1968 een serie uitgaven met de lexicale inventaris van het vroeg Italiaans (van 960–1321) (Alinei 1968/1977). Alinei beweegt zich vanuit de kwantitatieve taalkunde niet richting computerlinguïstiek, maar richting dialectologie, en verhuist bovendien vrij snel terug naar Italïe en is

[4] De volledige commissie bestond uit J. Bakker, J.F.H.A. de la Court, P. Guiraud, L. Kaiser, J.H.C. Lisman, G.L. Meinsma (secretaris), J.L. van Soest (voorzitter), F.L. Stumpers, M. de Vries, D. Vuysje en A. van Wijngaarden.

Zink Typografie Sheet 26 of 213 - Page 18 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



daardoor in de Nederlandse computerlinguïstiek wat in de vergetelheid geraakt. Maar werk en methodiek (Alinei 1963) waren zijn tijd vooruit (Burton 1981).

Het meest invloedrijke werk op het gebied van corpuscollectie vindt plaats in de Verenigde Staten. Aan Brown University wordt het beroemde Brown Corpus (Kučera & Nelson Francis 1967) bijeengebracht. Meer dan een miljoen woorden en bovenal: representatief, met samples uit verschillende teksttypen. Precies zoals de Werkcommissie voor Taalstatistiek van De la Court en Van Wijngaarden ooit bedacht had. Nederland kan niet achterblijven, en in 1968 wordt dan ook een nieuwe poging ondernomen: de Werkgroep Frequentie-Onderzoek van het Nederlands wordt opgericht. Adriaan van Wijngaarden is opnieuw van de partij, en ditmaal heeft ook Hugo Brandt Corstius zitting in de commissie.[5] Ditmaal wordt bijna direct met het feitelijke onderzoek gestart, en in 1975 is daar dan het resultaat: *Woordfrequenties in geschreven en gesproken Nederlands* (Uit den Boogaart 1975), beter bekend als het Eindhovencorpus (naar de Technische Hogeschool te Eindhoven, waar het meeste werk verricht is). Het corpus bestaat uit 720.000 woorden, gelijkelijk verdeeld over zes subcorpora (dagbladen, opiniebladen, gezinsbladen, romans en novellen, populaire-wetenschappelijke boeken en gesproken taal) en handmatig geannoteerd met vrij gedetailleerde "grammaticale codes", Part-of-Speech (POS) tags. Representativiteit, toegankelijkheid en beschikbaarheid zijn de belangrijkste doelstellingen van het project. En die zijn gehaald. Jarenlang is het Eindhovencorpus *het* standaardwerk voor kwantitatieve-taalkundigen, corpuslinguïsten en taaltechnologen, maar ook voor toegepaste-taalkundigen, psycholinguïsten en neurolinguïsten. Een generatie taalwetenschappers heeft gewerkt met de korte tekstfragmenten uit 1969 en 1970. Jaren waarin Feyenoord de Europacup won, de Apollo 12 op aarde terugkeerde, en koningin Beatrix ging skiën in een jack van zeehondenbont. Het corpus is bovendien verder verwerkt in andere resources: de Alpino Treebank (van der Beek, Bouma, Malouf & van Noord 2002), een volledig

[5] De overige leden van de commissie waren B.Th. Tervoort, J. Daan, D. Zwama, J.J.M. Bakker, D.M. Bakker, B. van den Berg, J.P.M. Eggermont, A. Feitsma, J. de Rooij en A.M. Zwaneveld. P.C. Uit den Boogaart, E.D. de Jong en A.M. van Mierlo waren belast met de uitvoering van het project.

Zink Typografie Sheet 27 of 213 - Page 19 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



syntactisch geannoteerd corpus, besloeg het CDBL (kranten-) subcorpus van het Eindhovencorpus. Pas met de komst van zeer grote (kranten)corpora en kwalitatief goede automatische annotatie verloor het corpus zijn unieke waarde.

Toen Hugo Brandt Corstius op 29 november 2002 het eerste exemplaar van de syntactisch geannoteerde Alpino Treebank in ontvangst nam, merkte hij op dat hiermee een langgekoesterde wens van Van Wijngaarden postuum in vervulling ging: de hoogleraar had er destijds al voor gepleit dat het corpus verrijkt zou worden met syntactische annotatie, maar dat plan had het niet gered in de Werkgroep. Het Eindhovencorpus is helemaal handmatig geannoteerd met een platte structuur op woordniveau.

Het corpus had wel geannoteerd kunnen worden met syntactische structuren. In Nijmegen hield Jan van Bakel zich bezig met kwantitatieve taalkunde. Van Bakel was in het bijzonder geïnteresseerd in de frequentie van syntactische constructies, en had een alternatief systeem ontwikkeld om teksten van syntactische meta-informatie te voorzien. Dat werk was begonnen met de wens om het middenstuk van de Nederlandse zin te beschrijven. Omdat daarover heel weinig literatuur bestond, ging hij zelf op zoek naar data. Met de hulp van de studentassistenten die bij de bewakingsdienst van de universiteit werkten (toen een gratis bron van arbeidskracht) verknipte Van Bakel enkele zaterdagkranten en ordende de data tot een overzichtelijk kaartensysteem. Op basis hiervan konden samenvattingen en analyses gemaakt worden van volgordeverschijnselen in dat middenstuk. “Dat zou je door een computer kunnen laten doen!”, dacht Van Bakel bij het zien van het resultaat. Het eerste werk dat hieruit voortvloeide, Syntanal (van Bakel 1970), vereiste nog steeds handmatige annotatie op woord(grens)niveau, op basis waarvan de computer de constituenten kon samenstellen en tellen. De volgende grote stap in het werk van Van Bakel volgde nadat hij kennis had gemaakt met formele grammatica’s, zowel vanuit de Transformationele Grammatica (TG) als vanuit de informatica. In 1975 publiceerde hij AMAZON, het eerste computerprogramma, geschreven in de programmeertaal SNOBOL, dat Nederlandse zinnen volautomatisch van een syntactische analyse kon voorzien (van Bakel 1975). De Werkgroep Frequentie-Onderzoek van het Nederlands koos echter niet voor Syntanal, en AMAZON kwam voor het Eindhovencorpus te laat. Van Bakel zelf raakte teleurgesteld

Zink Typografie Sheet 28 of 213 - Page 20 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



in de kwantitatieve taalkunde. Een van de aanleidingen was dat hij met zijn studenten probeerde een kwantitatieve theorie over "moeilijk" en "makkelijk" taalgebruik te formuleren. De enige conclusie die hij uit zijn onderzoek kon trekken is dat alle verschillen op een of andere manier samenhangen met de lengte van de zin. "Kwantitatieve uitspraken hebben geen enkele relevantie, kwantiteiten zijn geen onderdeel van een wetenschappelijke notie", zegt hij nu. En: "het taalgevoel is leidend." Van Bakel keerde zich af van de kwantitatieve taalkunde en stortte zich op een computermodel van TG.

Niet alleen in de taalkunde werd de rekenkracht van de computer ingezet om kwantitatieve gegevens te verzamelen. Willy Martin publiceerde in 1970 een kwantitatieve analyse van twee romans van de Vlaamse schrijver Ivo Michiels (Martin 1970). Dit is het onderwerp van zijn proefschrift. Een poging de "esthetische beoordeling van het vocabularium van een letterkundig werk" te objectiveren. Ofwel: een poging om tot exactere uitspraken te komen dan "het taalgebruik is eenvoudig" of "op sommige plaatsen lyrisch". Dit was de eerste keer dat iemand dit deed (tenminste in het Nederlands taalgebied) en zeer controversieel. Hugo Brandt Corstius herinnert zich de promotie nog goed. "Ik heb nog nooit zoiets meegemaakt. Dat gaat in België heel anders dan hier. Ik moest hem echt 4 uur lang verdedigen tegen zijn eigen promotor en de andere mensen, want die zeiden allemaal dat het niets voorstelde. En ik zei: 'nee het stelt niets voor, maar het is wel voor het eerst dat dit gebeurt'." Het is dan ook Hugo Brandt Corstius die het Woord Vooraf schrijft voor de publicatie van het boek. Hierin vergelijkt hij de gangbare, inexacte literatuurkritiek met astrologie, en de exacte literatuurkritiek met astronomie. "De astrologie geeft zeker belangrijkere informatie dan de astronomie, maar wat heeft men aan informatie waarvan niet is na te gaan of hij juist is? Toen de astronomen kijkers gingen gebruiken kregen zij te horen dat dit de "echte" waarneming van de sterren in de weg stond. Zij trokken zich er niets van aan. Ik hoop dat velen in de toekomst het moedige voorbeeld van Dr. Martin zullen volgen, zich van zulke vooroordelen niets aan te trekken, zodat dit eerste werk niet het enige zal blijven" (Martin 1970). Het werk van Martin in de kwantitatieve literatuurkritiek heeft echter weinig navolging gekregen, en Martin zelf heeft zich vrij snel hierna toegelegd op de lexicologie, waar de kwantitatieve benadering *wel* voet aan de grond gekregen heeft: hij kreeg de leerstoel lexicografie

Zink Typografie Sheet 29 of 213 - Page 21 of 205

aan de Vrije Universiteit Amsterdam. Ook Leiden was een centrum van lexicografische interesse en deskundigheid, en hier werd in 1967 het Instituut voor Nederlandse Lexicografie (INL) opgericht. De opdracht van het instituut was expliciet om met behulp van de modernste technieken de Nederlandse woordenschat in kaart te brengen. Bovendien moest het Woordenboek der Nederlandsche Taal zo snel mogelijk afgemaakt worden, waar al sinds 1851 aan gewerkt werd. Het werd in 1998 opgeleverd, en is het grootste woordenboek ter wereld.

## 1.5 PHLIQA

### Naar Apeldoorn

In 1968 zag Adriaan van Wijngaarden "zijn" Electrologica overgaan in de handen van Philips, dat direct het hele handeltje naar Apeldoorn verhuisde om het samen te voegen met Philips Computer Industrie. Met deze combinatie, zo'n 2000 man bij elkaar, wilde Philips de concurrentie aangaan met IBM. Behalve een heleboel computerbouwers en softwareontwikkelaars werkten er ook een aantal onderzoekers. "Al mocht dat niet zo heten", vertelt Jan Landsbergen, een van die onderzoekers. "Want onderzoek hoorde thuis op het Natlab in Eindhoven". Dus werd het clubje GSO genoemd, Geavanceerde Systemen Ontwikkeling.[6] Dat groepje onderzoekers in die grote computerfabriek zou aan de basis staan van het eerste grootschalige industriële ontwikkelproject op het gebied van de taaltechnologie. Weer kwam er geen taalkundige aan te pas: na de wiskundigen Van Wijngaarden en Brandt Corstius, die in Amsterdam de basis legden, waren het in Apeldoorn opnieuw exacte jongens die geschiedenis schreven: één wiskundige ditmaal, en twee natuurkundigen.

[6] Dit is de naam die Jan Landsbergen zich herinnert. Remko Scha noemt de naam Geavanceerde Systemen Onderzoek, wat in strijd zou zijn met het idee dat het geen research mocht heten. Harry Bunt herinnert zich de naam Advanced Software Projects. Hard bewijs voor de ene of de andere naam is niet gevonden. De versie van Landsbergen is aangehouden, onder meer omdat het verhaal dat het geen onderzoek genoemd mocht worden ook al 9 jaar eerder in Coppen (2001) opgetekend staat.

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Jan Landsbergen studeerde wiskunde in Delft, waar hij zich specialiseerde in formele talen. Zijn afstudeeropdracht was om te kijken naar de semantiek van programmeertalen, maar via een omweg kwam Landsbergen uit bij de syntaxis van natuurlijke talen. Hij studeerde af op een formalisering van Chomsky's Aspects (Chomsky 1965). Zonder het te weten – in Delft waren geen taalkundigen, laat staan generativisten – begaf Landsbergen zich daarmee in de frontlinie van de opmars van de generatieve taalkunde. Er waren in de jaren zestig wel al leesgroepen geweest in Amsterdam, waar onder meer Hugo Brandt Corstius deel van uitmaakte (van der Beek 2001), en in 1966 was de dissertatie van Albert Kraak de eerste in het generatieve framework, maar Chomsky's taalkunde was nog allerminst breed geaccepteerd of onderwezen. Maar de interesse van autodidact Landsbergen was gewekt. Na zijn afstuderen wilde hij verder op het snijvlak van wiskunde en taalkunde, maar niet in de academische wereld. "Ik had al te lang gestudeerd", zegt hij zelf. En een baan vinden was geen probleem: in 1970 lagen de banen voor ingenieurs voor het oprapen. Hij stuurde een briefje van een paar regels naar Philips in Eindhoven, en zij brachten hem in contact met Philips Apeldoorn. Daar ontmoette hij een paar onderzoekers die, gestimuleerd door Herman Schweigmann, onderzoek deden naar de potentie van Kunstmatige Intelligentie (KI) voor Philips. Nou ja, onderzoek... "Het was eigenlijk meer een soort leesclubje", aldus Landsbergen. Het was de tijd waarin de verwachtingen voor de computerindustrie bij Philips nog hoog gespannen waren. Schweigmann was vooruit gestuurd naar de allereerste conferenties over KI en was daar enthousiast geraakt. In Apeldoorn zocht hij een aantal mensen bij elkaar met als doel een project te definiëren op het gebied van KI. Jan Landsbergen was de laatste die erbij kwam. Remko Scha en Harry Bunt waren net een paar maanden bezig[7]. Scha wilde na zijn afstuderen vooral weg uit Eindhoven, en weg uit de natuurkunde. Een vage interesse in taal en logica – hij had tijdens zijn studie al stage gelopen bij het Instituut voor Perceptie-onderzoek (IPO) – had hem naar Apeldoorn gebracht. "Mijn vrienden, die aan de TU scheikunde gestudeerd hadden gingen allemaal naar Curaçao of Venezuela, dat leek mij ook wel wat. Maar ja, de content gaat voorop, ik

[7] De complete groep was groter. Scha noemt Jaap Calis en Gerhard Daalenoord, Bunt heeft het over een totaal van 6 mensen, zonder namen te noemen.

Zink Typografie Sheet 31 of 213 - Page 23 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



wilde niet zomaar een baantje op een leuke plek zoeken. Dan maar Apeldoorn, als het maar geen Eindhoven was." Bunt was nog tijdens zijn studie aangestoken door het enthousiasme van Schweigmann, die hij tijdens een excursie bij Philips ontmoet had.

Achteraf lijkt het vanzelfsprekend dat de drie mannen die hoofdrolspelers werden in de Nederlandse computerlinguïstiek zich bezig gingen houden met taaltechnologie. Maar in 1970 waren het twee natuurkundigen en één wiskundige, die in een computerfabriek een project op het gebied van Kunstmatige Intelligentie zochten. Wat was KI op dat moment? De grote ster in het vakgebied op dat moment was Shakey, de robot die ontwikkeld was bij onderzoeksinstituut SRI in Californië. Shakey was uitgerust met een camera, een wereldmodel, en een soort van snorharen voor het "voelen" van objecten. Hij kon zich verplaatsen zonder ergens tegen aan te botsen, en een reeks commando's uitvoeren voor het verplaatsen van objecten. Life Magazine verwees naar Shakey als "the first electronic person". De robot opende een wereld van mogelijkheden en deed mensen dromen van nieuwe toepassingen en electronische vrienden. Naast robotics was ook beeldverwerking een hot topic eind jaren zestig, begin jaren zeventig. Visuele patroonherkenning was een goede kandidaat voor een eerste KI-project bij Philips. "Maar ja, dat was meteen zo ingewikkeld", aldus Remko Scha. "Wij hadden het idee dat een project met taalinput redelijk overzichtelijk zou zijn, en relatief snel toepasbaar." Landsbergen was zich wel degelijk zeer bewust van de complexiteit van werken met natuurlijke taal. "Ik was de enige van de drie die zich tijdens zijn studie al met taalkunde had beziggehouden, en ik zag het als mijn eerste taak erop te wijzen dat je niet te naïef moest denken over het gebruik van natuurlijke taal"(Coppen 2001). Maar hij zag ook de toepasbaarheid. "Computers waren toen groot en duur. [...] Er waren wel al databases, maar die hadden vaak een heel ondoorgrondelijke structuur, heel onlogisch, bepaald door de hardware met als doel zo efficiënt mogelijk informatie op te slaan. Wanneer je informatie uit zo'n database wilde halen, dan moest daar een specialist aan te pas komen, die dan een programma ging schrijven om het eruit te halen. Als je op een wachtlijst terechtkwam, dan kon je soms wel weken wachten op een antwoord. Ons idee was om een natuurlijketaalinterface te maken voor heel concrete vragen aan de database." Scha vult aan: "Het was nog de tijd van de grote mainframes, maar

Zink Typografie Sheet 32 of 213 - Page 24 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



je kon je wel al voorstellen dat als het zou doorzetten met die computers, dat ze dan zo groot zouden worden dat je er heel veel informatie op kon zetten, en dat mensen via terminals daarbij zouden kunnen. Visuele interfaces waren er nog niet, dus je kon je voorstellen dat als je mensen in staat wilde stellen bij die informatie te komen zonder een programmeertaal te leren, dat natuurlijke taal dan een goed idee zou zijn." Tel daarbij op dat Landsbergen graag verder wilde werken aan taal-gerelateerd onderzoek, en dat Scha en Bunt zich gezamenlijk al wat georiënteerd hadden in de logica en semantiek, en de de keuze van het onderzoeksonderwerp is geen verrassing meer. In het voorjaar van 1971 werd besloten dat er gewerkt zou gaan worden aan een Question Answering (QA) project, namelijk een natuurlijke-taalinterface voor databases. Landsbergen, Scha en Bunt nemen het voortouw, Calis en Daalenoord verdwijnen naar de achtergrond. Het Philips Question Answering project, beter bekend als PHLIQA, was geboren.

Terwijl Landsbergen, Scha en Bunt de focus van hun onderzoek bepaalden, werd het in de top van Philips duidelijk dat het niet goed ging met de verkoop van computers. De fabriek in Apeldoorn werd eens goed onder de loep genomen, en men kwam tot de schrikbarende conclusie dat er feitelijk een volledige onderzoeksafdeling van 25 man in de fabriek zat. En dat terwijl onderzoek was voorbehouden aan Natlab in Eindhoven! De hele groep werd dan ook snel opgeheven. De hele groep? Nee, een klein aantal onderzoekers bleef gespaard. Philips besloot dat het KI-project interessant genoeg was om voort te zetten. Landsbergen, Scha en Bunt werden samen met de manager van de hele GSO-groep, Piet Medema, overgeplaatst naar het Natlab in Eindhoven, waar ze bovendien versterking kregen van de informatici Eric van Utteren en Wijnand Schoenmakers. En zo kwam Remko Scha, die er alles aan had gedaan om na 25 jaar Eindhoven te ontvluchten, na slechts twee jaar in Apeldoorn weer terug in de stad waar hij geboren en getogen was.

## Het project PHLIQA

Het is 22 oktober 1975. Op een computerterminal in Eindhoven knippert de Engelse vraag: "What is your question?". Een gebruiker typt in: "What computers are there?". Dan begint het boomstructuren te

Zink Typografie Sheet 33 of 213 - Page 25 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



regenen, en processingtijden. Na enige tijd maken de bomen plaats voor lijsten van antwoorden: een PDP15 bij Akzo in Arnhem, een X1 bij Philips in Eindhoven en een IBM 360/20 bij Shell in Rotterdam.

Het is een demo van PHLIQA-1, het tastbare resultaat van de onderzoeksgroep van Piet Medema. Een systeem in het Engels, want Philips was internationaal georiënteerd, en Engels was de facto de voertaal van het bedrijf. En een systeem met informatie over computergebruik in Europa. "Om Philips gunstig te stemmen – en omdat het ons verder niet uitmaakte", aldus Landsbergen (Coppen 2001). PHLIQA-1 kon vragen beantwoorden als "How high is the price of Akzo's PDP15?", "Where are IBM configurations used?" en "What configuration did Shell buy from IBM?". In het geval van verkeerd gespelde woorden of anderszins niet herkende termen vraagt het systeem dat specifieke woord aan te passen, waarna het alsnog de interpretatie begint. Altijd biedt het systeem aan om te kijken of er nog een andere interpretatie mogelijk is. Voor de vraag waar IBM computers staan bijvoorbeeld, is een antwoord in termen van steden mogelijk, maar ook in termen van landen of bedrijven waar de computer in kwestie staat. En aangezien Shell meer dan één computer van IBM kocht, is de response op de laatstgenoemde vraag "False presupposition". Wordt de vraag herhaald als "What configuration*s* did Shell buy from IBM", dan geeft het systeem zonder morren de gevraagde computers.

De hoofdlijnen van het ontwerp van PHLIQA kwamen uit de koker van Remko Scha. De man van de grote ideeën, een prater en een debater. "Iemand die al discussiërend dingen bedenkt", aldus Landsbergen, die zich hem heel goed herinnert. "We zijn heel lang kamergenoten geweest op het Natlab. Dat was wel handig, want Remko begon pas om half twaalf – als enige op het Natlab. Dan heb je toch de halve dag de kamer voor je alleen." Landsbergen vervult zelf de rol van het precies uitwerken en kloppend maken van de grote, maar nog wat vage ideeën. In de woorden van Scha: "Jan Landsbergen heeft ervoor gezorgd dat een zinnig idee een mathematische precisie kreeg. [...] Als ik nu naar het ontwerp kijk van hoe het systeem geworden is, dan zie ik daar heel erg de hand van Jan Landsbergen in. Zorgen dat die architectuur strak wordt, dat het hele systeem een werkelijk modulaire opzet krijgt. [...] Dat het zo strak geworden is als het geworden is, dat komt door Jan. Hij heeft in die zin misschien in het tot stand

Zink Typografie Sheet 34 of 213 - Page 26 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



komen van dat systeem als geheel nog wel een belangrijker rol gespeeld dan ikzelf." Harry Bunt richtte zich voornamelijk op specifieke semantische problemen, in het bijzonder mass terms. Alledrie de taaltechnologen geven echter aan dat er vooral heel veel samengewerkt werd. "We deden eigenlijk alles samen. Er was wel een taakverdeling, maar we overlegden heel intensief over alles", vertelt Landsbergen. En Scha vult aan: "we werkten met wekelijkse vergaderingen waar mensen voorstellen indienden. Die werden dan besproken, en iedereen moest het ermee eens zijn, of op zijn minst consent geven. Pas op het eind, toen er modules gebouwd moesten worden, werden mensen echt verantwoordelijk voor een stuk software. Maar die fase van het echt implementeren hebben we heel lang uitgesteld." Toch was het niet allemaal alleen maar harmonieuze samenspraak. De complexe semantiek die nodig was om mass terms te kunnen verwerken paste niet goed in het overall design van PHLIQA, en dit leidde nog wel eens tot discussie.

Het systeem is ontworpen op een manier die niet standaard was voor de KI van dat moment. Roger Schank en Terry Winograd behaalden aan Yale en MIT grote successen met hun aanpak van KI. De onderzoekers waren er helemaal op gericht om werkende systemen te bouwen, en maakten zich minder druk om de vraag of de aanpak nu wel of niet "mooi"was. Formele systemen zagen zij als een beperkende factor; empirisch resultaat was het belangrijkst. En empirisch resultaat kwam er: Winograd ontwikkelde SHRDLU, een programma dat het mogelijk maakte om via natuurlijke taal opdrachten te geven over objecten in een beperkte ruimte, zoals "Find a block which is taller than the one you are holding and put it into the box". Net als Shakey, zorgde SHRDLU voor enorm optimisme in het vak. Het PHLIQA-team was sceptisch over deze aanpak. "Zo kreeg je programma's die niemand begreep", vindt Landsbergen nog steeds. "Daar waren wij tegen. Wij waren een beetje calvinistisch, hadden een Nederlandse stijl om dingen te doen. Je moest kunnen bewijzen dat het antwoord correct was." Daarmee werden de drie van PHLIQA duidelijke vertegenwoordigers van het kamp van de "neats", en zetten ze zich af tegen de "scruffies". De strijd tussen beide kampen is tot op de dag van vandaag niet beslecht.

Inspiratie voor de opzet van PHLIQA kwam dan ook zeker niet uit de hoek van de scruffy's, hoewel de onderzoeksgroep goed op

Zink Typografie Sheet 35 of 213 - Page 27 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



de hoogte was van de ontwikkelingen daar. Dat kwam onder meer doordat er heel goede contacten waren met het Stanford Research Institute SRI. Remko Scha vertelt: "Er was een soort mantelcontract gesloten tussen Philips en SRI, waarin stond dat Philips voor een bepaald non-triviaal bedrag aan onderzoeksconsultancy zou afnemen. En dat werd ingezet voor ons project. We hebben verschillende keren mensen van SRI laten komen. Don Walker is met een aantal medewerkers in Eindhoven bij het Natlab op bezoek geweest." Ook in de fabriek in Apeldoorn hadden de onderzoekers al beschikking over deze resources: "Burt Rayfeld, die is in de zomer of het najaar van 1970 al in Apeldoorn geweest om ons een cursus theorem proving te geven." Hierdoor hadden de PHLIQA architecten toegang tot kennis ("die mensen waren bijzonder goed ingevoerd in wat er op dat moment speelde in de VS") en een platvorm voor discussie ("ik herinner mij heftige discussies over het werk van Winograd. Dan kwamen zij ons vertellen "dat moet je zo doen". Maar wij waren eigenwijs, wij gingen helemaal niet doen wat zij zeiden. Dat kan ook anders."). Scha ziet nog een derde voordeel: "Wat ook helpt, je kan je ideeën meteen ergens kwijt. Je kent die mensen persoonlijk, dus je weet hoe je papers in elkaar moeten steken om ze geaccepteerd te krijgen."

Inspiratie voor de opzet van PHLIQA kwam veeleer uit de hoek van de formele semantiek. Amsterdam was in de jaren zeventig (en nog steeds) een sterk centrum van formele semantiek. De geschiedenis van de formele semantiek in Amsterdam begint in hetzelfde jaar als deze geschiedenis van de taal- en spraaktechnologie: in 1952. Beth richt dan zijn *Instituut voor grondslagenonderzoek en philosophie der exacte wetenschappen* op, beter bekend als het Grondslageninstituut. Beth was zelf een begenadigd logicus, met bovendien een enorm netwerk en de wil tot het onderwijzen en begeleiden van anderen. "After Beth there was logic in the Netherlands", wordt er dan ook gezegd (van der Beek 2001). Uit de school van Beth komen de semantici Johan van Benthem en Hans Kamp, die elk een grote nieuwe stroming in de formele semantiek oprichtten. Ten tijde van PHLIQA werd de formele semantiek in Nederland grotendeels bepaald door van Benthem en zijn studenten – Hans Kamp was inmiddels al naar het buitenland vertrokken: eerst naar de VS, later naar Duitsland. En het waren Van Benthem en zijn studenten die verantwoordelijk waren voor de verspreiding van de heel invloedrijke Montague-grammatica

Zink Typografie Sheet 36 of 213 - Page 28 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



in Nederland. Beth had Montague wel al in 1962 naar Amsterdam gehaald voor een lezing, maar toen begreep nog niemand de impact van zijn verhaal. Want hoe groot zijn wiskundig talent ook was, didactische vaardigheden waren *niet* zijn sterkste kant. Ook drie jaar later, als hij de dan overleden Beth tijdelijk vervangt, viel zijn compositionele modeltheoretische semantiek nog in dorre aarde. Henk Verkuyl was bij de colleges die hij gaf. "Hij schreef eigenlijk alleen maar formules op het bord. Dan stond hij te schrijven en niemand, nou ja heel weinigen begrepen de portee van die formules." Maar Alice ter Meulen, Jeroen Groenendijk en Martin Stokhof hadden hun leermeester Van Benthem vrij dringend gevraagd zijn colleges modale logica in te ruilen voor onderwijs in Montague-grammatica. En vanaf dat moment is Montague-grammatica niet meer uit Amsterdam verdwenen.

De informele Montague Colloquia, die de Amsterdamse semantici vanaf 1976 iedere twee weken organiseerden, werden ook door Scha, Bunt en Landsbergen goed bezocht. Toch is het niet zo dat PHLIQA een implementatie is van Montague-grammatica. Dat kan ook niet, simpelweg omdat PHLIQA al bestond toen de drie onderzoekers kennis maakten met het werk van de logicus. Landsbergen: "Ik geloof dat het zo ging, dat toen dit project al een hele tijd liep, in 1976 of 1977, iemand tegen ons heeft gezegd "Goh, wat jullie doen dat lijkt wel een beetje op Montague-grammatica". Dus toen zijn we dat boek gaan lezen." Wel was er een grote overlap in manier van werken. Scha: "Stokhof, Groenendijk, Janssen, Veltman, Van Emde Boas, al die mensen kenden wij uit die tijd. Zij waren op een heel degelijke manier bezig met hetzelfde soort theorieën als waar wij mee bezig waren. Met nog grotere mathematische precisie. Want wij waren in vergelijking met wat er in de KI gebruikelijk was wel heel degelijk bezig, maar van de professionele formele semantici konden wij nog wat leren. Dat vonden wij bijzonder inspirerend."

Halverwege de jaren 70 komt er een omslagpunt in het project. De lang uitgestelde implementatie wordt dan toch ter hand genomen, wat leidt tot een werkend systeem, getuige de hierboven beschreven demonstratie in het najaar van 1975. En dat was zeker niet de enige demonstratie die de PHLIQA-groep verzorgde. "Op het Natlab kwamen nogal eens mensen langs, die dan rondgeleid moesten worden", herinnert Landsbergen zich. "En het zijn natuurlijk vaak heel technische zaken die daar ontwikkeld werden. Dus als nu de vrouw van de

Zink Typografie Sheet 37 of 213 - Page 29 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



minister kwam, en die moest beziggehouden worden, dan dachten ze "nou, natuurlijke taal, dat snapt iedereen", en dan kwamen ze dus bij ons." Die demo's gingen altijd prima, zolang ze gedaan werden door iemand uit het ontwikkelteam, die wist wat er gevraagd kon worden. Want zodra het publiek zelf een vraag bedacht, ging het vaak mis. "Dan zat een woord niet in het woordenboek, of de informatie niet in de database." Want PHLIQA was heel duidelijk een *limited domain* systeem.

Niet lang na voltooiing van de eerste implementatie veranderde het team vrij drastisch. Oorzaken: een kentering in de support vanuit de top en frictie binnen het team. Philips Electrologica in Apeldoorn had zijn interesse al enige tijd verloren en was uit het project gestapt. Scha en Landsbergen kwamen in dienst van Natlab, maar Piet Medema moest tegen zijn zin elders in de organisatie aan de slag. Zijn programmeerwerk werd overgenomen door Wim Bronnenberg. Niet lang daarna verliet Harry Bunt het team: hij werd overgeplaatst naar het IPO. Bunt: "Ik ben op zeker ogenblik door het Natlab op een ander project gezet, op het IPO. Ze wilden in die tijd meer aan Kunstmatige Intelligentie gaan doen, en hebben mij daarvoor aangetrokken." De directeur van het IPO had een lezing van Bunt gehoord, en herkende in hem de juiste, meer technische, aanvulling op het team Cognitie en Communicatie, waar vooral psychologen en psycholinguisten zaten. "Ik vond het ook leuk, een uitdaging" zegt Bunt, "maar het [PHLIQA] project had van mij ook langer mogen duren, daar niet van." Enkele jaren later promoveerde Bunt vanuit het IPO op de semantiek van *mass terms*, en meer algemeen op het type semantiek zoals dat door het driemanschap Landsbergen-Scha-Bunt voor PHLIQA ontwikkeld was (Bunt 1981). Met name het algemene stuk over de multi-level semantics van PHLIQA heeft wenkbrauwen doen fronsen, omdat Bunt niet gezien werd als de hoofdontwerper van deze theorie, maar er wel ruim aandacht aan besteedde in zijn proefschrift. Hugo Brandt Corstius was een van de promotoren.

De nieuwe leider van de groep werd niet Scha, niet Landsbergen, maar Van Utteren. Tot grote verbazing van Remko Scha. "Ik had zelf zo'n idee van dat moet ik zelf doen of Jan." In eerste instantie was hij een beetje ontstemd. "Eric was een jongen die nogal op de achtergrond was en nuttig werk deed bij de software-engineeringkant van het project. Helemaal ok, maar moet die nu het project leiden?"

Zink Typografie Sheet 38 of 213 - Page 30 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Later begrijpt het het beter. "Het ging toen om de implementatie. Dan maakt het niet uit wie er allerlei geweldige ideeën heeft, dan gaat het er gewoon om dat het werkt." Scha zegt het zelf niet, maar het is ook niet ondenkbaar dat de verantwoordelijke adjunct-directeur zich afgevraagd heeft of promotie van dan wel Landsbergen, dan wel Scha de dynamiek tussen de twee onderzoekers niet te veel zou schaden.

Ondertussen werkten Scha en Landsbergen door aan PHLIQA. Vooral onder de motorkap. Landsbergen: "In die tweede fase hebben we zeg maar PHLIQA-1+ gebouwd. Die was eigenlijk van buiten niet veel anders, maar zat veel mooier in elkaar." Hoeveel mooier het ook werd, de naam PHLIQA-2 zou het nooit krijgen. Want die naam was al vergeven. "Dat heeft ons altijd dwars gezeten", vertelt Landsbergen. "We hadden een ontwerp voor PHLIQA-2. Dat was zeer ambitieus. En toen is besloten dat we eerst maar eens PHLIQA-1 zouden maken. Het punt was, PHLIQA-2 was veel te ambitieus, en dat is er dus nooit gekomen. Met als gevolg dat wanneer wij weer een nieuw systeem maakten, we dat PHLIQA-1.1 noemden, etc. We konden nooit zeggen dit is PHLIQA-2, want het ontwerp daarvan stond op papier en dat konden we niet maken."

In de tweede helft van de zeventiger jaren kwam langzaam de discussie op gang hoe het verder moest met PHLIQA. Hoewel onderzoek op het Natlab redelijk de vrije hand kreeg – "vrijheid blijheid" noemt Landsbergen het – kwam zo langzamerhand toch wel de vraag op wat Philips er nu precies aan zou gaan hebben. En die vraag was niet gemakkelijk te beantwoorden. Want Philips wilde apparaten maken, en was eigenlijk niet geïnteresseerd in software. Daar kwam bij dat er vanaf 1980 ook grafische interfaces ontworpen werden voor databases, die hetzelfde doel hadden: niet-technische gebruikers toegang geven tot databases. Hoewel natuurlijke taal nog wel gezien werd als een interessant gebied om in te werken, werd het project PHLIQA toch steeds minder populair. Er is wel gezocht naar andere interessante toepassingsgebieden. Scha verwijst graag naar deze tijd als "het moment waarop wij het internet *niet* hebben uitgevonden". Of eigenlijk: de *twee* momenten waarop ze het internet niet hebben uitgevonden. De PHLIQA-club deed verkenningen bij zowel de Philips-afdeling die zich bezighield met Teletekst, als bij de telefonieafdeling. Beide boden aanknopingspunten om het internet te ontdekken. Teletekst was een breed gedeelde informatiebron, die het grote publiek vanuit de

Zink Typografie Sheet 39 of 213 - Page 31 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



huiskamer kon raadplegen. Als mensen daar via een PHLIQA systeem vragen aan konden stellen, had het een groot potentieel. Maar Teletekst bleek geen tekstdatabase te zijn, maar een collectie plaatjes, waar de taaltechnologen verder niets mee konden. Ook de gesprekken met de telefonie-afdeling leidden nergens toe. Package-switching, de techniek die het mogelijk maakt om heel veel mensen van dezelfde telefoonlijn gebruik te laten maken en die cruciaal is voor de ontwikkeling van internet, was net ontdekt. Nu is Information Retrieval op een database plus een zich snel ontwikkelende telefoontechnologie nog geen internet. "Maar", zegt Scha, "als je heel briljant bent, dan kan je een paar van die stappen tegelijk maken. En dat is wat wij toen niet gezien hebben, en onze gesprekspartners bij Philips in Hilversum ook niet." De ultieme toepassing van PHLIQA werd niet gevonden, en zo eindigde eind jaren zeventig het eerste grote industriële taaltechnologieproject.

Taaltechnoloog Franciska de Jong kijkt met de kennis van nu terug op het vroege taaltechnologische onderzoek bij Philips: "Je zou kunnen zeggen dat ze daar een interessant volume hadden aan onderzoekers op dat terrein, met een enorme potentie, maar dat de ideeën te vroeg zijn gekomen. Neem PHLIQA, het idee dat je een question answering systeem maakt voor digitale content. Een mooi idee, maar ja, op een moment dat er nog geen digitale content was, alleen nog maar databases met tekstvelden... Als dat tien jaar later was begonnen dan was dat zonder meer door Philips gehandhaafd gebleven. Dan had dat zeker tot interessante toepassingen of spinoffs geleid. Nu is het afgeserveerd omdat het te vroeg gebloeid heeft."

Er was weinig onderzoek in Nederland waar PHLIQA mee vergeleken kan worden. Wel had de Nederlandse PTT in het Dr. Neherlab in Leidschendam een project met soortgelijke doelstelling: REIMS. "Het doel van REIMS (REsearch Informatie Manipulatie Systeem) is het verkennen van de eisen en de mogelijkheden van systemen waarbij een of meer gebruikers zonder tussenkomst van programmeurs of operateurs interactief een computersysteem gebruiken, in het bijzonder gegevensbestanden manipuleren."[8] Het project had echter enerzijds een grotere scope, omdat de gebruiker niet alleen vragen moest

[8] Document gedateerd 29-11-1971, afkomstig uit het PTT archief in het Nationaal Archief te Den Haag, dossier 3920.

Zink Typografie Sheet 40 of 213 - Page 32 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



kunnen stellen aan de database, maar ook gegevens moest kunnen invoeren. Anderzijds was de doelstelling heel veel kleiner, omdat interactie met de databank weliswaar op basis van gebruikerstaal moest kunnen plaatsvinden, maar deze gebruikerstaal zeker geen natuurlijke taal was: het project voorzag in de specificatie van een nieuwe, artificiële gebruikerstaal, die weliswaar simpel in het gebruik moest zijn, maar speciaal voor deze toepassing ontworpen zou worden, en best operatoren mocht bevatten zoals die bekend waren uit de logica of programmeertalen. Een voorbeeld stelt dat de geboortedatum van de chef van Piet opgevraagd moet kunnen worden door de functie "geb. datum (chef (Piet))" – wat nog wel iets anders is dan de vraag beantwoorden op welke dag de baas van Piet geboren is.

De twee overgebleven PHLIQA-voormannen gaan elk zijns weegs wanneer het project op zijn einde loopt. Remko Scha promoveert op zijn werk aan PHLIQA (Scha 1983), werkt tijdelijk bij de Universiteit van Amsterdam en vertrekt al snel naar de Verenigde Staten, waar hij een baan krijgt bij Bolt, Beranek and Newman (BBN Technologies) in Cambridge. Eindelijk ver van Eindhoven. Landsbergen blijft bij het Natlab en start een nieuw project, naar eigen idee, geïnspireerd door Montague-grammatica. "Wat me bij PHLIQA altijd dwars gezeten had, dat was die database. Ik wilde gewoon met taal bezig zijn. En wat is dan het mooiste? Dat is vertalen. Dan heb je alleen nog maar met taal te maken." Het automatisch-vertaalproject Rosetta is geboren.

## Het systeem

PHLIQA is een QA-systeem dat destijds beschouwd kon worden als state-of-the-art. Het beantwoordt Engelse vragen over computersystemen in Europa – in de zeventiger jaren nog niet zo veel – op basis van (verzonnen) data in een database. Behalve het goede antwoord geven wanneer dat in de database gevonden wordt, kan het systeem ook aangeven wat er aan schort wanneer het geen antwoord kan vinden, en kan het op verzoek meerdere antwoorden geven wanneer een vraag ambigu is. De beschrijving hieronder is gebaseerd op enkele publicaties (Bronnenberg, Bunt, Landsbergen, Scha, Schoenmakers & van Utteren 1980, Scha 1983) en mondelinge toelichting door Landsbergen en Bunt. PHLIQA is geschreven in een interne programmertaal

Zink Typografie Sheet 41 of 213 - Page 33 of 205

van Philips, waarvoor geen compiler meer beschikbaar is, en kan dus niet meer gedraaid worden. Wel is er een 16mm film bewaard gebleven van de eerder beschreven demo van PHLIQA-1 op 22 oktober 1975, die bovendien nu ook in gedigitaliseerde vorm beschikbaar is op `http://www.let.rug.nl/~vannoord/TST-Geschiedenis`.

PHLIQA beantwoordt vragen door de Engelse vraagzinnen in een aantal stappen te vertalen naar databasequeries, en het resultaat van de databasequeries vervolgens te verwerken tot het gevraagde antwoord. Globaal zijn er drie niveaus van semantische analyse, die telkens resulteren in een expressie in een formele taal die de semantische informatie van dat niveau uitdrukt. Dit principe wordt *multilevel semantics genoemd*.

De eerste analysestap is van een Engelse vraagzin naar een expressie in EFL, *English-oriented Formal Language*. Net als alle andere tussentalen van PHLIQA is EFL een wiskundig gedefinieerde formele taal, een vorm van typed lambda-calculus. In EFL wordt syntactische en morfologische informatie expliciet gemaakt in een logische vorm. Is de inputzin ongrammaticaal, bevat hij spelfouten of onbekende woorden, dan wordt dat gemeld. Typfouten kunnen eventueel nog verbeterd worden door de gebruiker. De referentiële constanten, grofweg de inhoudswoorden, worden op dit niveau echter nog niet semantisch geïntepreteerd. Motivatie hiervoor is dat op dit niveau wel de kennis van de taal beschikbaar is, maar nog geen informatie over het wereldmodel waarbinnen de vraag geïnterpreteerd moet worden. En juist dat wereldmodel bepaalt de betekenis, of de extensie, van inhoudswoorden.

Het wereldmodel wordt geïntroduceerd in de tweede analyse stap: van EFL naar WML, of *World Model Language*. Het wereldmodel bevat informatie over de concepten en relaties (functies) tussen concepten. Een computerinstallatie heeft bijvoorbeeld wel een prijs, een land niet. En een installatielocatie bestaat uit een straatadres en een stad. Op dit punt kan PHLIQA al een vraag verwerpen als onbeantwoordbaar, bijvoorbeeld als iemand naar de prijs van Duitsland vraagt. Omdat landen geen prijs hebben, zal PHLIQA de vraag verwerpen.

De derde stap is een vertaling van WML naar DBL of *Database Language*. De constanten verwijzen hier direct naar entries in de database. Op dit niveau wordt dus kennis van de precieze structuur van de database geïntroduceerd. Hierna volgt nog de evaluatie van

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



de resultaten en de formulering van een antwoord (zo wordt bijvoorbeeld het resultaat FALSE omgezet in het antwoord "No"). Wanneer na evaluatie blijkt dat een presuppositie uit de vraagstelling niet waar blijkt te zijn, geeft PHLIQA geen antwoord, maar meldt simpelweg dat de vraag een false presuppositie bevat. Een voorbeeld is de vraag "How expensive is the installation at Shell?". De vraag vooronderstelt dat Shell precies 1 computer heeft, maar dit is niet het geval in het wereldmodel van PHLIQA-1. De logische expressies op de verschillende niveaus kunnen behoorlijk complex worden. Er vinden tussen de verschillende analysestappen dan ook simplificaties plaats. Deze vervangen complexe expressies in eenvoudiger, maar logisch equivalente, expressies van dezelfde logische taal.

Een bijzonderheid van het systeem is dat alle expressies in de tussentalen boomstructuren zijn. Landsbergen: "We hadden bedacht dat als je zit te vertalen van expressie naar expressie, dat het dan natuurlijk heel dom is om daar eerst een string van te maken, en dan moet je die string eerst weer ontleden om er wat mee te kunnen doen. Dus intern hadden we allemaal bomen."

PHLIQA was heel veel netter dan de meeste KI toepassingen van die tijd, die volgens de scruffy methode werkten. Tegelijkertijd was PHLIQA heel veel toegepaster dan het formeel semantische werk van de academische logici. In Nederland was nooit eerder een dergelijk uitgebreide toepassing gebouwd – ook al was het maar een demosysteem. Internationaal gezien was PHLIQA state-of-the-art. Terecht zegt Jan Landsbergen dan ook: "Wij waren hier heel trots op, en eigenlijk nog wel een beetje."

## 1.6 Een aanloop van een kwart eeuw

Midden jaren vijftig zijn de computers in Nederland nog op één hand te tellen, en lijkt een romance tussen wiskunde en taal een sprookje. Eind jaren zeventig schrijven wiskundigen grammatica's, en taalkundigen algoritmes. Het landschap is onherkenbaar veranderd op het gebied van de taalkunde en op het gebied van de rekenkracht.

Van Wijngaarden merkte in 1952 in zijn oratie op dat de taalkunde wiskundig gezien nog in de kinderschoenen stond. Linguïstiek op de

Zink Typografie Sheet 43 of 213 - Page 35 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Nederlandse universiteit werd nog steeds voor een groot gedeelte bepaald door *Het woord* van Anton Reichling (Reichling 1935). Structuralisme was troef. De publicatie van Chomsky's *Syntactic Structures* (Chomsky 1957) veranderde daar aanvankelijk weinig aan. Het werk markeert het beginpunt van de generatieve taalkunde, die een grammatica ziet als een algoritme dat alle mogelijke zinnen van een taal kan genereren, maar zeker niet direct geaccepteerd werd. Goed, taalkundige en journalist Henk Schultink schreef in de *Nieuwe Rotterdamse Courant* dat Chomsky "een verdienstelijke steen bijdraagt" aan de methodiek van het blootleggen van de principes van zinsvorming. Maar daar stond tegenover dat zijn hoogleraar Reichling de theorie in een vlammende lezing verwierp als een 'kryptanalytisch formalisme'. Slechts een paar Amsterdamse studenten namen de moeite om het werk aandachtig te bestuderen. Onder de geuzennaam *Kryptanalytisch Genootschap* kwamen onder meer Albert Kraak, Wim Klooster en Pieter Seuren op een zolderkamer bijeen. Chomsky stuurde ze zelfs een microfilm van zijn nog ongepubliceerde *The Logical Structure of Linguistic Theory* (Chomsky 1975). Onder leiding van Frits Staal, die in 1961 in Amsterdam hoogleraar Filosofie werd, ontwikkelde de groep zich van een geheim clubje op de studentenkamer van Kraak, via een informele groep bij Staal thuis tot een meer officieel forum met wel veertig bezoekers op het Instituut in de Roetersstraat. Onder luid tumult volgde in 1966 de eerste generatieve promotie, en wel van Kraak, die bij Staal promoveerde. Vanaf de jaren zeventig kregen taalkundestudenten overal in Nederland onderwijs in de nieuwe "mathematische linguïstiek", of "algebraïsche taalkunde", zoals Hugo Brandt Corstius prefereert (Brandt Corstius 1974), en al snel was het de dominante stroming in taalkundig Nederland.

Al vrij vroeg waren er ook critici die opmerken dat ook het generatieve kader wiskundig gezien nog wat haken en ogen had. Landsbergen schreef zijn afstudeerscriptie over Chomsky's *Aspects* (Chomsky 1965). "Eigenlijk deed ik de formalisering ervan, want het bleek minder formeel te zijn dan het leek: er ontbrak van alles aan. Eigenlijk was helemaal niet gedefinieerd wat een transformatie was." Volgens Brandt Corstius was de Amerikaanse wetenschapper niet gecharmeerd van kritisch commentaar: "Twee keer heb ik Noam Chomsky mijn bezwaartjes voorgelegd, twee keer deed hij of ik gek was." (Brandt Corstius 2003). In de jaren zeventig ontwikkelden Joan

Zink Typografie Sheet 44 of 213 - Page 36 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Bresnan en Ron Kaplan een alternatieve theorie: Lexical Functional Grammar. Dat is ook een generatieve grammatica, maar zonder de transformaties van Chomsky, en veel preciezer gedefinieerd. Nog weer later kwamen er andere alternatieven, zoals Head-Driven Phrase Structure Grammar en Categoriale Grammatica.

Ook op het gebied van de rekenkracht had de tijd niet stilgestaan. Was er aan het begin van ons verhaal nog niet één werkende computer in Nederland, eind jaren zeventig hadden alle universiteiten computers. De meesten werkten nog met ponskaarten, maar langzamerhand begon ook de terminal zijn intrede te doen, waarmee het in de jaren tachtig mogelijk werd om hands-on computerpractica te geven. In 1972 richtten een paar pioniers een informeel overleg op: het Landelijk Overleg Programmeer Onderwijs Letteren. Hierin zaten in ieder geval Steven Krauwer (Utrecht), Hugo Brandt Corstius (Amsterdam), Gert van der Steen (Amsterdam), Piet Rolf (Nijmegen) en Kwee Tjoe Liong (Amsterdam). De groep kwam tot 1975 bij elkaar, waarna het programmeeronderwijs in de verschillende opleidingen ingebed werd. In Utrecht werd bijvoorbeeld het bijvak Formele Technieken geïntroduceerd, waarin programmeren, maar ook mathematische linguïstiek, logica en statistiek onderwezen werd.

Langzaam maar zeker begon de computerlinguïstiek binnen de universiteit een eigen plek te veroveren. Gezien de voorgeschiedenis was het verrassend dat dit niet gebeurde binnen de faculteit Wis- en Natuurkunde, maar bij de Letterenfaculteit. Jan van Bakel kreeg in 1975 een onderwijsopdracht computerlinguïstiek, die eerst verlengd en in 1979 omgezet werd in een leeropdracht. In Utrecht waren er rond 1980 zelfs twee stromingen computerlinguïstiek die onderwezen werden: Steven Krauwer en Louis des Tombe doceerden bij Algemene Taalwetenschappen mainstream computationele taalkunde, zoals augmented transition networks en grammaticamodellen. In deze colleges werd formele taalkundige informatie gebruikt om een automatisch taalverwerkingssysteem te maken. Heel anders was de insteek van Martin Boot bij Toegepaste Taalkunde. “Mijn invalshoek was de invalshoek van een letterkundige. Gericht op teksten en het oplossen van problemen die te maken hebben met communicatie”, aldus Boot, die inmiddels geen computerlinguïst meer is, maar astroloog. Voor het oplossen van deze problemen mocht alle beschikbare informatie gebruikt worden, taalkundig interessant of niet. Dus een Duitse POS-

Zink Typografie Sheet 45 of 213 - Page 37 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



tagger mag best gebruik maken van de hoofdletters aan het begin van een woord om zelfstandig naamwoorden te herkennen. Nu een veelgehoord standpunt, toen vloeken in de kerk. Het boterde dan ook niet tussen Krauwer en Boot. Alice Dijkstra volgde bij allebei college. "Volgens Steven Krauwer had Boot de term 'Computerlinguïstiek' gekaapt. Vandaar dat zij zichzelf Taalkunde en Automatisering noemden." Krauwer zelf legt uit: "Ik vond en vind statistiek niet interessant. Ok, het werkt. So what? Dat is ook de reden dat ik het werk van Boot niet leuk vond, dat vond ik suf programmeren. Daar zat geen linguistiek in, dus was het ook geen computerlinguïstiek. Ik geloofde heilig dat je taalkunde nodig had, maar dat is fout gebleken. Nu zijn mijn standpunten wel een beetje veranderd. De taak van de computerlinguistiek is niet meer het modelleren van een cognitief proces, maar het oplossen of reduceren van een probleem."

In Amsterdam onderwees Hugo Brandt Corstius algebraïsche taalkunde, maar na enige tijd werd hij aangesteld tot lector en later hoogleraar automatische informatieverwerking aan de Erasmus Universiteit Rotterdam. "Daar zouden ze een Letterenfaculteit krijgen", vertelt Brandt Corstius. "Nou, daar wilde ik wel bij zijn. Ik mocht vast beginnen. Maar er was geen student die mijn verhaal wilde horen. Het waren allemaal economiestudenten. En die hele Letterenfaculteit is uiteindelijk niet doorgegaan." Hij vertrok al snel naar Amerika. Een jaar later, na terugkomst, gaf hij min of meer gedwongen toch weer een jaar les, ditmaal in Leiden. "Ik weet zeker dat ik bij de rector in Rotterdam gezegd heb dat ik ermee ophield. 'Ik neem ontslag'. Maar toen kwam ik na een jaar terug, en toen zeiden ze dat ik weer moest lesgeven."Rotterdam bleek het contract met Hugo Brandt Corstius niet verbroken te hebben, en hem het jaar in Amerika gewoon doorbetaald te hebben. "Ja, ik vond ook wel dat ik heel veel geld op die rekening had staan, maar ik dacht: 'dat zoek ik later wel uit'. Ik ben niet zo goed met geld." Via een uitwisselingsovereenkomst tussen de beide universiteiten kwam Brandt Corstius bij de Letterenfaculteit in Leiden terecht, waar hij opnieuw 1 jaar computerlinguïstiek onderwees. Ook de drie technische mannen van PHLIQA kwamen uiteindelijk alledrie in Letterenfaculteiten terecht, al zaten ze eerst nog bij het Natlab (Landsbergen), BBN (Scha) en het IPO (Bunt).

Ook in Vlaanderen ontwikkelde computerlinguïstiek zich binnen de Letterenfaculteiten. In Leuven hadden Martin en Engels enkele

Zink Typografie Sheet 46 of 213 - Page 38 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



studenten, en in Antwerpen zat Luc Steels. Steels was de eerste die er het licentiaatsprogramma Taal en Letteren volgde, en de jonge groep professoren in Antwerpen stond er voor open om nieuwe ontwikkelingen in het programma op te nemen. En dat betekende niet alleen de nieuwste taalkundige theorieën, maar ook computerlinguïstiek. Professor Jacques Noël, die in Canada aan de University of Victoria een Masters Degree had behaald in Computational Linguistics, en later betrokken raakte bij het Franstalige Eurotraproject voor automatisch vertalen, doceerde begin jaren zeventig in Antwerpen korte tijd computerlinguïstiek. Wellicht de eerste cursus in het Nederlands taalgebied. Steels was een van zijn studenten en raakte gebiologeerd door het vak. Via conferenties en zomerscholen ontwikkelde hij zich verder, en uiteindelijk promoveert hij in 1977 bij Hugo Brandt Corstius. In Koenraad De Smedt vond Steels een opvolger, maar zelf vertrok hij naar de VS. Steels was nog lang niet uitgeleerd, en begreep dat kennis van de informatica voor zijn vakgebied cruciaal zou zijn. Hij kreeg een plek in het moderne AI-lab van MIT, waar Winograd zijn experimenten met SHRDLU uitvoerde. Daar ging de gepromoveerde computerlinguïst weer studeren: computerwetenschappen op masterniveau. Bijna twintig jaar zou het duren voordat Steels terugkeert bij taal als onderzoeksthema.

De enige die *niet* in een Letterenfaculteit opereerde, was de toegepaste taalkundige Bondi Sciarone. Zijn vakgroep huisde samen met de wiskundigen in de Faculteit Algemene Wetenschappen in Delft. Maar hoe succesvol Sciarone ook was met zijn werk, hij bleef een eenling. Computerlinguïstiek was niet langer onderdeel van de informatica en dus de wiskunde, maar begon een zelfstandig vakgebied te worden, gebaseerd op een taalkunde die niet langer uitging van de intuïtie, maar de strengheid en efficiëntie nastreefde die Van Wijngaarden als voorwaarde noemde voor een gezamenlijke toekomst voor Rekenen en Taalkunde. Zou het sprookje van Automatisch Vertalen dan toch werkelijkheid worden?

Zink Typografie Sheet 47 of 213 - Page 39 of 205

De Eurotra workshop van 1980 in Bangor. Hugo Brandt Corstius ontbreekt – hij was al vertrokken.

# 2 De vertaalmachine

"Aangezien alle belangrijke woorden slechts aanduidingen van voorwerpen zijn, ligt de oplossing [...] onmiddellijk voor de hand. De menschen zullen die dingen met zich meedragen, die kunnen dienen om de gedachten ten aanzien van eigen noodwendigheden uit te drukken. [...] Intusschen passen reeds velen der geleerdsten en verstandigsten het nieuwe spreeksysteem toe, waarbij zij zich uitsluitend bedienen van voorwerpen. Hieraan is slechts één nadeel verbonden en dat is de noodzaak om een grootere voorraad dingen mee te dragen, naarmate men een veelzijdiger spreker is."

*(Swift 1726).*

## 2.1 Wat eraan voorafging

Taaltechnologie heeft een haat/liefdeverhouding met automatisch vertalen. Aan de ene kant is het vooruitzicht van die ultieme toepassing een heel belangrijke drijfveer geweest voor computerlinguïsten en engineers, en heeft dat perspectief heel veel geld losgemaakt voor onderzoek in de computerlinguïstiek. Anderzijds is automatisch vertalen

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



besmet geraakt door het grote aantal teleurstellingen en niet waargemaakte beloften.

Ondanks de oratie van Van Wijngaarden, die zijn tijd ver vooruit was, is Nederland pas laat begonnen met onderzoek naar automatisch vertalen. Het eerste Nederlandse vertaalproject dateert uit 1963 en werd gefinancierd door Euratom. Evert Willem Beth werkte sinds 1960 aan Euratom-projecten. De Europese gemeenschap voor atoomenergie Euratom had in Ispra in Italië een enorme computer laten neerzetten voor het doorrekenen van resultaten van een atoomreactor. Alleen werkte die atoomreactor niet zoals hij zou moeten werken. Om die computer niet te laten verstoffen, besloot Euratom andersoortig onderzoek te subsidiëren dat op een zinnige manier van de rekenkracht gebruik zou maken. In 1960 werkte Beth aan onderzoek op het gebied van de logica, in 1962 werden toepassingen van de logica in de mathematische linguïstiek opgenomen (Chomskiaanse taalkunde), en in 1963 wordt het contract uitgebreid met de volgende doelstellingen: "afsluiting van het onderzoek naar woordvolgorde in het Nederlands, en uitbreiding daarvan naar andere Euratom-talen en het Russisch. Ontwikkeling van een generatieve en relationele grammatica voor het Nederlands en andere talen als grondslag voor het vertaalsysteem."[1] Een doelstelling die ambitieus genoemd mag worden. De resultaten die aan het eind van het programma in 1965 behaald zijn: enkele voorbereidende studies in generatieve grammatica door Pieter Seuren, een grammatica voor Franse telwoorden door Rudolf de Rijk, en een grammatica voor niet-samengestelde bevestigende Nederlandse zinnen door Harm Smits, aldus de Euratomverslagen van Henk Visser. Over het vertaalprogramma en de computer in Ispra geen woord. Was het dus weggegooid geld? Misschien wel voor de Euratomleiding die in 1965 de subsidie stopzette. Maar het colloquium voor de taalkundige afdeling van Euratom bestond uit onder meer Hans Kamp, Hugo Brandt Corstius, Wim Klooster, Albert Kraak, Pieter Seuren en Rudolf de Rijk. Mensen die de generatieve

[1] Brief Beth-Commissie van de Europese Gemeenschap voor Atoomenergie, programmaleiding, Directoraat Onderzoek, 15 mei 1963, Onderzoeksvoorstel bijlage I, 1. Onderzoek betreffende toepassingen van de mathematische logica. Brief K.L. de Bouvère-Beth, 30 september 1963, programma-aanvullingen. Brieven in het Rijksarchief in Noord-Holland te Haarlem.

Zink Typografie Sheet 50 of 213 - Page 42 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



grammatica, de formele semantiek en de computerlinguïstiek in Nederland een enorme duw vooruit gaven. Was dat vertaalprogramma eigenlijk ooit wel een oprecht en realistisch doel van Beths Euratom-project? Hans Kamp over zijn werk voor het project: "Nou ja, het was eigenlijk meer om mij de gelegenheid te geven te studeren" (van der Beek 2001).

Op het moment dat Nederland instapte in het onderzoek naar mechanisch vertalen, heeft het in de Verenigde Staten al een roemruchte geschiedenis doorgemaakt, die in 1966 abrupt tot een voorlopig einde komt. Want waar Van Wijngaardens oratie in dorre grond viel, kon het pionierswerk van Weaver in Amerika wèl wortel schieten. In 1954 konden IBM en Georgetown University al een demonstratie geven van een automatische vertaling van Russisch naar Engels. Het woordenboek bevatte slechts 250 woorden, en van grammaticale analyse was nauwelijks sprake, maar de algemene opinie was dat het een kwestie van drie tot vijf jaar was voordat een volledige automatische vertaalmachine beschikbaar zou zijn. Die opinie werd niet gedeeld door andere wetenschappers in het vak. Yehoshua Bar-Hillel verklaarde op het door hem georganiseerde congres op MIT in 1952 al dat volautomatische vertaling van hoge kwaliteit voorlopig onhaalbaar zou zijn door het grote probleem van semantische ambiguïteit, een boodschap die hij nog vaak zou herhalen. Een vertaalhulp die met pre- en/of postediting een goede vertaling zou opleveren, was volgens hem wel haalbaar. Maar het Georgetown-verhaal was een stuk aantrekkelijker. En dus werd er grof geïnvesteerd in automatisch vertalen, met de belofte van een werkend programma binnen vijf jaar in het achterhoofd. Maar in 1964 was er nog voor geen enkel talenpaar een goede, volautomatische vertaalmachine voor een groot domein ontwikkeld. De Amerikaanse overheden die hun geld in onderzoeksprojecten hadden gestoken richtten de Automatic Language Processing Advisory Committee (ALPAC) op, die in 1966 in haar rapport concludeerde: "we do not have useful machine translation. Further, there is no immediate or predictable prospect of useful machine translation" (ALPAC 1966). De redenen: menselijke vertaling is goedkoper, er zijn genoeg vertalers om al het vertaalwerk te doen, de kwaliteit van automatisch vertalen is te laag en het ziet er niet naar uit dat die op korte termijn goed genoeg wordt. Dat maakte in een klap een einde aan de subsidiegelden en daarmee aan het onderzoek naar automatisch vertalen in

Zink Typografie Sheet 51 of 213 - Page 43 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



de Verenigde Staten. In Canada werd wel nog een succesvol systeem ontworpen: METEO, dat weerberichten uit het Frans naar het Engels vertaalde en vice versa. De kwaliteit was goed, het domein zeer beperkt. In Europa werd Systran gebruikt, een product dat voortkwam uit het team van Georgetown. [2]

Na het Euratomproject gebeurde er in Nederland en Vlaanderen een tijdlang niets aan automatisch vertalen. En toen, in de jaren tachtig, zagen we bijna tegelijkertijd ineens vier initiatieven: het Europese Eurotra, een METAL-groep in Leuven, het industriële DLT in Utrecht en Rosetta bij Philips in Eindhoven. Wat was er veranderd? In 1980 gaf Bondi Sciarone in Delft in zijn inaugurele rede het antwoord. Hij was benoemd tot hoogleraar Toegepaste Taalkunde, en had interesse en ervaring in computerlinguïstiek, onder meer in de kwantitatieve taalkunde. De voorzitter van het vertaalinstituut had hem gevraagd of hij Nederland niet wilde vertegenwoordigen bij een Europees vertaalproject met de naam Eurotra. Sciarone verdiept zich in het onderwerp, inclusief de bezwaren van ALPAC, en gaat in zijn oratie in op elk van de argumenten. De situatie is in Europa anders dan in de VS, en de omstandigheden zijn nogal veranderd in de vijftien jaar die intussen verstreken zijn. De kosten van computergebruik zijn spectaculair gedaald, computerkracht is toegenomen, en arbeid is heel veel duurder geworden. Daarmee zijn de kosten van automatische vertaling ten opzichte van menselijke vertaling enorm gedaald. Bovendien is de vraag naar vertalingen in Europa veel groter dan in Amerika, waar Engels de enige taal van betekenis is, en er dus altijd maar naar één taal vertaald hoeft te worden. De Europese Gemeenschap is aan het uitbreiden: Spanje, Portugal en Griekenland staan op het punt om toe te treden, waarmee het aantal leden op 12 komt. Alle documenten moeten in alle talen vertaald worden. Wat betreft de kwaliteit van automatische vertaling verwacht Sciarone dat die altijd achter zal blijven bij de kwaliteit van menselijke vertaling. Tegelijkertijd stelt hij: "dat automatische vertaling, hoewel fundamenteel beperkt, een aantal toepassingsmogelijkheden heeft" (Sciarone 1980).

[2] John Hutchins heeft heel uitgebreid geschreven over de geschiedenis van het automatisch vertalen wereldwijd, maar in het bijzonder ook over die eerste periode in de VS. Veel van deze informatie is te vinden op zijn website `http://www.hutchinsweb.me.uk/`

Zink Typografie Sheet 52 of 213 - Page 44 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Sciarone neemt namens Nederland deel aan Eurotra, en geeft daarmee het startsein voor een periode waarin het taaltechnologisch onderzoek en onderwijs bloeit als nooit tevoren, en bijna volledig gericht is op automatisch vertalen. "Ik hoop daarom", zo eindigt Sciarone, "dat het in Europees verband genomen initiatief om te komen tot een nieuw veeltalig vertaalsysteem werkelijkheid zal worden. Daarmee is niet alleen de oplossing van een praktisch communicatieprobleem gediend, maar ook de samenwerking tussen onderzoekers op Europees niveau."

## 2.2 Eurotra

### Een Europees initiatief

Toen Sciarone in zijn inaugurele rede een pleidooi hield voor het Europees initiatief om een automatisch vertaalsysteem te ontwikkelen, was het project in kwestie, Eurotra, informeel al begonnen. De Europese Commissie had besloten dat de informatieoverdracht tussen sprekers van de verschillende Europese talen beter moest, en had daarvoor de adviescommissie CETIL opgericht. De eerste documenten uit het Eurotra-archief zijn dan ook documenten van het adviesorgaan CETIL, en dateren van 1977. In het Frans, een taal die veel gebruikt werd in de begintijd van Eurotra. Verder had de Commissie het Amerikaanse vertaalsysteem Systran aangekocht, waar men eigenlijk niet zo blij mee was. Waarom moest de Europese Commissie een Amerikaans product kopen, dat eigenlijk helemaal niet zo goed was, terwijl er in Europa wel op hoog niveau research werd gedaan op het gebied van automatisch vertalen? Systran had wel wat taalkundige kennis, maar leek toch nog heel sterk op het directe vertaalsysteem uit Georgetown. Terwijl in Grenoble aan GETA-Ariane gewerkt werd, en in Saarbrücken SUSY ontwikkeld werd, twee systemen van wat de tweede generatie genoemd werd: met uitgebreide taalkundige analyse en transfermodules tussen de verschillende abstracte representaties. En ook in Pisa en Manchester waren onderzoeksgroepen met het onderwerp bezig.

Het idee was er. Er moest een programma komen dat zou leiden

Zink Typografie Sheet 53 of 213 - Page 45 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



tot de toekomstige vervanger van Systran. Maar hoe moest dat georganiseerd worden? Moest de ontwikkeling centraal plaatsvinden of juist niet, en wie zou de hele onderneming moeten gaan leiden? De universiteiten die voorop liepen in het onderzoek vonden dat ze een grote rol moesten krijgen in het project. Kon er niet gewoon verder gebouwd worden op de basis van het systeem in Grenoble? Wat te doen met de onderzoeksresultaten uit de andere groepen? Besluiten hierover werden genomen in de stuurgroep, de 'Groupe de Coordination', een club van een man of tien, die het uiteindelijke voorstel aan de Europese Gemeenschap voorbereidde. In deze stuurgroep had ook computerlinguïst Dirk Geens uit Leuven zitting. Leuven was op dat moment de meest actieve Belgische onderzoeksgroepgroep op het gebied van computationele taalkunde, en had bovendien goede contacten bij de overheid. Via contacten bij het Ministerie van Wetenschapsbeleid en het CETIL werd het team uit Leuven uitgenodigd zitting te nemen in de stuurgroep. Dirk Geens werd namens de groep afgevaardigd en praatte mee over de vormgeving van het project tot 1982, toen hij een leerstoel kreeg aangeboden aan de Franstalige Université libre de Bruxelles. Voorzitter van de stuurgroep was Maghi King uit Genève. Zij was het die in de aanloop de leiding had, een rol die later overgenomen werd door de leider van het uiteindelijke project, Sergei Perschke uit Ispra.

Perschke gaf leiding aan het MT-onderzoek van Euratom in Ispra. Hetzelfde centrum van nucleaire fysica, dat eerder het Euratomproject van Beth gefinancierd had. Ook in Ispra zelf was naast het onderzoek naar atoomenergie een aparte poot opgezet voor onderzoek naar automatisch vertalen, eerst op basis van het Georgetownsysteem, later met Systran. In tegenstelling tot de Nederlandse poot, had het lokale onderzoek wel stand gehouden. Perschke pleit al in december 1977 voor een decentrale opzet van Eurotra: "the only construct which in my view has realistic chance of success, is a co-operative venture where each participant contributes to the development of "this project" in the sense that he can fully identify himself with it".[3] Daarbij is volgens hem een cruciale rol weggelegd voor een centrale unit,

[3] Doc. CETIL 12/77, EUROTRA/2/78, *Considerations about a European System of Machine Translation* door Sergei Perschke, kopie uit het persoonlijk archief van Frank Van Eynde.

Zink Typografie Sheet 54 of 213 - Page 46 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



"which must be strong enough to exercise leadership and to maintain momentum. The function of this unit is not so much administrative (although this aspect is important) as intellectual. It must dispose of expertise both in linguistics and informatics at a sufficiently high level to be recognized as an authority by the different teams [...]." En wie is geknipt voor het voorzitterschap over dit belangrijke centrale orgaan? Juist, Sergei Perschke zelf. De Europese Commissie haalt hem naar Luxemburg om hoofd te worden van het Eurotraproject. Geheel volgens het idee van Perschke wordt Eurotra decentraal opgebouwd. Uiteindelijk met teams en co-financiering in alle deelnemende landen, maar daaraan voorafgaand met een aantal voorbereidende teams per taal.

Nederland deed in dit stadium nog niet mee, maar er was wel aanzienlijke Nederlands*talige* participatie. Dirk Geens praatte in de stuurgroep op het hoogste niveau mee. En als hij vertrekt, kan hij zijn werk gemakkelijk overdragen aan Frank Van Eynde, die hij al vroeg bij het project betrokken heeft. Van Eynde was nog student, toen hij voor het eerst in aanraking kwam met Eurotra: "De hoogleraar toegepaste linguïstiek was professor Engels. En die had een assistent, Dirk Geens, die het college computerlinguïstiek gaf in de tweede helft van de jaren zeventig. Ik heb dat college zelf ook gevolgd in '78 of '79. Geens en Engels waren toen al bij de voorbereidende activiteiten betrokken. Geens kreeg opdrachten mee vanuit Genève, waar de coordinatievergaderingen plaatsvonden. Dan vroeg hij enkele van zijn studenten, waaronder mijzelf, om mee te werken aan die rapporten." Door zijn betrokkenheid bij het project belandt hij op de allereerste Eurotraworkshop in 1979 in Aix-en-Provence. Geens: "In de stuurgroep hadden we besloten dat we al vroeg een kern van onderzoekers wilden trainen in het basismodel dat we aan het ontwikkelen waren. Dus alle leden van de stuurgroep brachten een man of tien mee naar de zomerschool, die een paar weken duurde." Frank Van Eynde was een van de studenten die Geens meenam. Een enorme kans voor de jonge wetenschapper: "Tweeëntwintig was ik, en toen ontmoette ik mensen als Antonio Zampolli, Yorick Wilks, Sergei Perschke, Maghi King, Dieter Maas, Bernard Vauquois, Christian Boitet."

Dirk Geens heeft geen idee waarom Nederland in de aanloop van Eurotra niet betrokken was. "Iemand als Jan Landsbergen had zeker een waardevolle bijdrage kunnen leveren." De Vlaming probeerde zelf

Zink Typografie Sheet 55 of 213 - Page 47 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



wel om zijn Nederlandse collega's te betrekken bij het project. Maar die staan niet direct te springen. Geens kende de wiskundige Steven Krauwer, die bij ATW in Utrecht computerlinguïstiek deed. Hoogleraar Schultink had daar het idee opgevat om zijn club van taalkundigen te verrijken met een psycholoog, Louis des Tombe, en een wiskundige, Steven Krauwer. De twee buitenbeentjes bij ATW konden goed met elkaar opschieten, en vonden elkaar inhoudelijk in de computerlinguïstiek: eind jaren zeventig zetten de twee gezamenlijk een cursus Taalkunde en Automatisering op. Maar toen Geens aan Steven Krauwer vroeg of hij mee wilde doen met Eurotra, kreeg hij nul op rekest. "Ik vond het disrespectvol", zegt Krauwer. "Machine Translation vonden we een wereldje van charlatans. Met weinig taalkunde, en veel hackers." Maar toen Bondi Sciarone hem in 1980 opnieuw vroeg, zei hij wel 'ja'. "Ik heb het besproken met Louis. We vonden het nog steeds wel raar, maar het leek ons ook wel leuk om te kijken wat er zou gebeuren. We zagen mogelijkheden voor meer taalkunde in machine translation. Bovendien zochten we naar toekomstperspectieven voor onze studenten." En zo vormden Leuven, Delft en Utrecht gedrieën het Nederlandstalige Eurotrateam.

Bondi probeert ook andere computerlinguisten in Nederland te interesseren voor het project. Jan Landsbergen en Hugo Brandt Corstius worden uitgenodigd voor de tweede Eurotra workshop, ditmaal georganiseerd in Bangor, Wales. Hugo Brandt Corstius is sceptisch en wil eigenlijk niet, maar Sciarone dringt aan. "Ik vertelde hem dat hij dan juist mee moest, dan kon hij het allemaal eens van dichtbij bekijken." En hij gaat, net als Jan Landsbergen, van wie Brandt Corstius naar eigen zeggen een lift krijgt in het Philips vliegtuig. Coördinator Maghi King probeert de workshop strak te regisseren, maar dat is buiten Brandt Corstius gerekend. Waar de coördinator uit Genève een grote stap voorwaarts wilde zetten in het opzetten van het systeem, wilde de wiskundige uit Amsterdam een principiële – en naar alle waarschijnlijkheid kritische – discussie voeren over de mogelijkheid of onmogelijkheid van het te bouwen systeem.[4] Brandt Corstius verlaat de workshop voortijdig, gebelgd over de manier waarop met geld

[4] Jan Landsbergen herinnert zich een ander conflict, namelijk dat de Nederlandstaligen graag een eigen sessie over het Nederlands wilden organiseren in plaats van gezamenlijk met de Engelsen op te trekken. King zou dit niet hebben wil-

Zink Typografie Sheet 56 of 213 - Page 48 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



gesmeten wordt, over de manier waarop diplomatie en bureaucratie voorrang krijgen boven gedegen onderzoek, en de manier waarop hem de mond gesnoerd wordt en hij geacht wordt te doen wat op het programma staat. Hij verwijst naar Eurotra als "dat bespottelijke plan van die vrouw uit Genève", en naar de initiatiefnemers als "Margaret King en een paar andere oplichters" (Brandt Corstius 2003). Na twee dagen houdt hij het voor gezien. "Ik vond het echt te gek worden. Dus halverwege ben ik heerlijk gaan wandelen in Wales en daarna ben ik naar huis gegaan." Landsbergen blijft. Hij leert er het Eurotraformalisme kennen, maar is weinig enthousiast. Hij mist de strakke architectuur die hij kende vanuit PHLIQA en ook in zijn nieuwe project Rosetta weer zou nastreven. "Het idee was om te vertalen met wat een transfersysteem heette. Je vertaalt een zin eerst in een boom met allerlei labels en daar hangt alle informatie aan die je denkt nodig te hebben voor vertaling in de target taal. Maar iedereen interpreteert die informatie anders. Er zit geen formele semantiek achter, dat moest je zelf maar invullen. Daar zag ik helemaal niets in. Dat wordt chaos." Landsbergen wordt geen onderdeel van het Nederlandse team, maar bijft wel contact houden met Steven Krauwer en Louis des Tombe.

## Het systeem

Er bestaan grofweg drie typen vertaalsystemen: direct vertalen, vertalen via een interlingua, en vertalen via transfer. Het Georgetown-systeem, en in zekere zin ook Systran, behoren tot de eerste categorie. Direct vertalen betekent dat er geen tussenliggende representaties bestaan van de betekenis van de te vertalen zin. Het systeem wordt helemaal opgezet met een specifiek taalpaar in gedachten, en het tweetalig woordenboek is de belangrijkste component. Taalkundige analyse wordt tot een minimum beperkt. In de praktijk gaat het bij die eerste vertaalsystemen vooral om woord-voor-woord-vertalingen, die gecombineerd worden met wat regels om de woordvolgorde aan te passen aan de doeltaal. Dit wordt wel de eerste generatie van automatisch vertaalsystemen genoemd.

len toestaan, om het idee van soepele internationale samenwerking in stand te houden.

Zink Typografie Sheet 57 of 213 - Page 49 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



De tweede generatie vertaalsystemen vertaalt ofwel via een interlingua, ofwel via transfer. Bij vertalen via een interlingua wordt een extra stap geïntroduceerd. Eerst wordt de brontaal vertaald naar de interlingua, en vervolgens wordt de representatie in de interlingua vertaald naar de doeltaal. Dit systeem is bedacht als oplossing voor het explosief stijgende aantal talenparen in een multilinguale omgeving zoals de Europese Gemeenschap. De rekensommetjes zijn bekend: in 1980 was het aantal talen in de EG gestegen van vier naar zes (Frans, Duits, Engels, Deens, Nederlands, Italiaans), en op korte termijn zouden er nog eens drie talen bijkomen (Grieks, Spaans, Portugees). Dit betekent een stijging van 4x3=12, via 6x5=30 tot 9x8=72 taalparen. Wanneer met een interlingua gewerkt wordt, dan hoeven er voor de negende taal geen acht vertalingen bij te komen, maar slechts 1 analysemodule (van brontaal naar interlingua), en 1 genereermodule (van interlingua naar doeltaal). Het nadeel van interlinguasystemen is bijna net zo bekend als het voordeel: de analysecomponenten worden complexer, omdat de interlingua alle informatie moet bevatten voor vertaling naar alle talen. Dus ook al vertaal je *De rivier is ruim 500 km lang* van het Nederlands naar het Engels, je zal toch moeten aangeven of je het hebt over een rivier die naar zee stroomt of niet, want in het Frans zijn dat twee verschillende woorden. Voor de interlingua kan een formele taal (bijvoorbeeld een logische taal), een natuurlijke taal (bijvoorbeeld Engels) of een kunstmatige gebruikstaal (bijvoorbeeld Esperanto) gebruikt worden. In Nederland ontwikkelde het Buro voor Systeemontwikkeling (BSO) een systeem op basis van een interlingua: DLT, dat later uitgebreid aan bod komt.

Vertalen via transfer betekent twee extra stappen. Eerst vindt de taalspecifieke analyse plaats van de zin in de brontaal. Dit levert een abstracte, maar taalspecifieke, representatie op. Dan vindt de eigenlijk transfer plaats van de abstracte representatie van de bronzin naar de abstracte representatie van de doelzin. Op basis van deze representatie produceert een taalspecifieke generatiemodule de vertaling van de oorspronkelijke zin in de doeltaal. In eerste instantie lijkt dit systeem weinig efficiënt, aangezien voor iedere extra taal er een transfermodule naar elk van de bestaande talen gebouwd moet worden, plus een analyse- en generatiemodule. Maar het idee is dat wanneer analyse en generatie goed zijn opgezet, transfer relatief simpel kan zijn. Bovendien maakt deze opzet het mogelijk om per taal een ei-

Zink Typografie Sheet 58 of 213 - Page 50 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



gen analysemodel te gebruiken, zolang de output maar voldoet aan de specificaties van de representatietaal. Voor een decentraal opgezet samenwerkingsverband als Eurotra is dat een belangrijke eigenschap. Eurotra is dan ook opgezet als een transfersysteem, net als ARIANE, SUSY en METAL, die later in dit hoofdstuk beschreven worden.

De abstracte representatie op het tussenniveau is in het geval van Eurotra een dependency-boom, waaraan extra informatie is toegevoegd in de vorm van labels. Deze labels moeten bijvoorbeeld voor een Nederlands werkwoord in de tegenwoordige tijd aangeven of het werkwoord iets aanduidt dat nu bezig is, een stelling is over de wereld op dit moment, of over de toekomst gaat. Die drie verschillende dingen kunnen in het Nederlands allemaal met de tegenwoordige tijd aangeduid worden, maar in het Engels worden daarvoor drie verschillende tijden gebruikt: *Ik lees een boek, ik lees vaak boeken* en *ik lees morgen dat boek* worden vertaald als *I'm reading a book, I often read books* en *I'll read that book tomorrow*. Binnen Eurotra wordt door middel van labels dat onderscheid al in de analysefase gemaakt, waardoor de transfer zo simpel mogelijk blijft. De verrijkte boomstructuur is het resultaat van analyse in een aantal verschillende stappen: normalisatie, morfologische analyse, oppervlakkige syntactische analyse, diepe syntactische analyse en tot slot semantische analyse. Voor elk van deze niveaus is een representatietaal ontworpen, en de laatste representatietaal is de input voor transfer (gevolgd door generatie in de doeltaal).

Eurotra is volledig gebaseerd op taalkundige regels, en verschilt daarmee van de latere corpusgebaseerde systemen, maar niet van bijvoorbeeld SUSY en ARIANE. In de opzet van het programma was heel duidelijk meegenomen dat Eurotra in staat moest zijn om recente ontwikkelingen op taalkundig niveau te integreren. Dat was een relevant gegeven, want er gebeurde veel op taalkundig gebied. Generatieve grammatica was inmiddels gemeengoed geworden, het was duidelijk het systeem daar gebruik van moest maken. En daarmee onderscheidde Eurotra zich wèl van ARIANE. Frank Van Eynde: "Theoretische taalkunde had een minder belangrijke rol in Grenoble. Vauquois zelf was een astronoom. Die wou een systeem bouwen, maar hij was geen taalkundige van achtergrond. En ook Boitet, zijn rechterhand, was geen taalkundige. Zij baseerden zich op dependentiegrammatica à la Tesnière uit de jaren vijftig, dus linguïstisch gesproken een beetje

Zink Typografie Sheet 59 of 213 - Page 51 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



gedateerd, ook toen al: dit was de tweede helft van de jaren zeventig, en Tesnière, dat was jaren vijftig. Dus er was helemaal niets te merken van transformationele grammatica, of generatieve grammatica."

De opzet van Eurotra was dus wèl gebaseerd op transformationele grammatica, al werd er een eigen variant bedacht, die beter voldeed aan de eisen van het project. Maar het programma werd ingehaald door de taalkundige realiteit: al snel nadat het Eurotra officieel van start was gegaan, deed weer een nieuw type syntax zijn intrede: unificatiegrammatica's. In Amerika introduceerden Bresnan en Kaplan Lexical Functional Grammar (LFG) (Kaplan & Bresnan 1982), dat snel aan populariteit won. Ook andere unificatiegrammatica's waren in opkomst: eerst Generalized Phrase Structure Grammar (GPSG) (Gazdar, Klein, Pullum & Sag 1985) en later Head-Driven Phrase Structure Grammar (HPSG) (Pollard & Sag 1994). Eurotra besloot uiteindelijk om niet een van deze nieuwe frameworks over te nemen, maar een eigen kader te ontwikkelen, dat echter wel een aantal eigenschappen van deze nieuwe taalkundige stroming overnam. Dit framework werd het <<C,A>,T> of CAT framework (Debille 1986). CAT is het geesteskind van een groep centrale linguïsten, die zich sterk maakte voor een solide taalkundig-theoretische basis binnen Eurotra. Onder deze linguisten ook Steven Krauwer en Louis des Tombe uit Utrecht. Het CAT-framework is in verschillende varianten in spin-offs van Eurotra geïmplementeerd. Niet in het officiële Eurotrasysteem, tot teleurstelling van de bedenkers.

## Het Nederlands/Vlaamse team

De Vlaams-Nederlandse invloed in Eurotra is vanaf het begin relatief groot geweest. Er waren geen grote vertaalprojecten gaande, zoals wel in Duitsland en Frankrijk, maar Dirk Geens wist desondanks een plek te veroveren in de stuurgroep en de semantiekgroep, en overlegde zodoende met de hoofdrolspelers over specificaties van het systeem. Het Eurotra-archief maakt bijvoorbeeld al melding van een vergadering in 1979 met Boitet uit Grenoble, Maas uit Saarbrücken, centrale coördinator King en Geens uit Leuven.

Ook Steven Krauwer en Louis des Tombe kwamen al snel in centrale teams terecht. Krauwer: "Louis ging naar een cursus voor nieuwe

Zink Typografie Sheet 60 of 213 - Page 52 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



deelnemers, waar hij Doug Arnold van Essex en Lieven Jaspaerts uit Leuven ontmoette, met wie hij het goed kon vinden, en met wie hij veel heeft samengewerkt. Louis concludeerde al snel dat het taalkundig anders moest. Dat werd dus zijn focus." Krauwer zelf hied zich voornamelijk bezig met de systeemspecificaties, maar ook hij ijverde voor een meer solide basis. "Er moest op een abstracter niveau nagedacht worden over het systeem, voordat er van alles geïmplementeerd werd. Wij voerden oppositie, probeerden research te doen in plaats van development, omdat we het idee hadden dat dat nodig was." En met twee Nederlanders en een Belg in de groep van 8 mensen die zich bezig hielden met de taalkundige specificaties is het niet verbazingwekkend dat de documentatie doorspekt is met Nederlandse voorbeelden. De correcte vertaling van *Jan zwemt graag* naar het Engelse *John likes to swim* werd een lakmoesproef voor vertaalsystemen. Een indirect effect van de bemoeienis van Des Tombe en Krauwer in de centrale commissies is dat Frank Van Eynde terugkeerde bij Eurotra. Hij had het project verlaten om aan zijn doctoraat te werken. "Ik zag eigenlijk geen verband tussen de theoretische taalkunde waar ik toen mee bezig was voor mijn proefschrift en de taalkunde die toen binnen Eurotra werd bedreven", aldus Van Eynde. "Maar vooral Louis en Steven hebben ervoor gezorgd dat binnen Eurotra voldoende aandacht werd besteed aan het gebruik maken van theoretisch verantwoorde monolinguale componenten. Op die manier heb ik er weer interesse in gekregen. Dus vanaf '82 was ik er weer helemaal bij." Zelf werd Van Eynde actief in de werkgroep die zich richt op de vertaling van Tense en Aspect.

Lidmaatschap van de centrale commissies betekende niet alleen invloed op het systeemdesign, maar ook extra inkomsten: voor deze taken werden aparte contracten afgesloten. Zeker voor de groep in Utrecht is dat belangrijk gebleken. Het had namelijk nogal wat voeten in de aarde voordat Nederland werkelijk Eurotra-partner was. Officieel begon Eurotra in 1982. Voor die tijd gebeurde er wel al heel veel, en dat werd dan gefinancierd via studiecontracten tussen de Commissie en het deelnemende instituut. Veel van deze contracten waren taalgebaseerd, maar er werden ook aparte contractenopgesteld voor specifieke taalkundige problemen of voor systeemdesign. In het begin was het alleen Leuven dat meedeed. Later, van 1981 tot 1984, werkten Leuven, Delft en Utrecht op basis van deze studiecontracten samen

Zink Typografie Sheet 61 of 213 - Page 53 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



aan het Nederlands. Werk aan andere onderwerpen was op persoonlijke basis. Krauwer en Des Tombe deden dat wel, Sciarone niet. In 1984 tekende België het zogenaamde *Contract of Association*. Tegen die tijd hadden de meeste andere landen ook getekend, en hielden de taalgerelateerde studiecontracten op te bestaan. Maar Nederland had nog niet getekend. Dit betekende dat er voor Delft geen financiële middelen meer waren, waardoor Sciarone steeds minder betrokken raakte bij het project. "Utrecht had genoeg mensen om eraan te werken, en ik was hier maar alleen", aldus de voortrekker in Nederland. Uiteindelijk besluit hij om er helemaal uit te stappen, en als de Nederlandse overheid in september 1986 via een omslachtige constructie dan toch tekent, zijn het Utrecht en Leuven die samen doorgaan.[5]

De verdeling van taken en geld tussen Leuven en Utrecht was vrij simpel: de 2:1 ratio van de Nederlandse Taalunie werd overgenomen. Deze zegt dat waar het gaat over Nederlandse taal- (of letter-)kunde, Nederland tweederde van het werk en de financiering verzorgt, en België eenderde. Ook inhoudelijk viel er weinig te verdelen, omdat België twee jaar eerder getekend had en de Vlaamse groep het Engels en Duits al hadden uitgekozen als vreemde talen om mee te werken. Voor Utrecht bleven het Frans, Deens, Spaans, Italiaans, Grieks en Portugees over. Voor alle vreemde talen moesten de transfermodules naar het Nederlands opgeleverd worden, en daarnaast zouden de beide teams werken aan de monolinguale componenten (analyse, synthese en de woordenboeken).

Om al deze taken uit te voeren, werden veel medewerkers aangesteld. In Utrecht werkten in 1986, toen eindelijk het Contract of Association getekend werd, vier mensen aan Eurotra. Eind 1990 waren dat er meer dan 45.[6] Onder hen Gertjan van Noord, van wie de enorme uitbreiding wel iets minder had gemogen: "De eerste paar jaar had ik het uitstekend naar mijn zin in Utrecht, totdat die Eurotra-groep zo enorm uitgebreid werd met mensen die juist in het vak begonnen.

---

5 Zie voor details over de Nederlandse contracten Steven Krauwers verslagen over Eurotra op `http://www-sk.let.uu.nl/eurotra/act80-90rev.pdf` en `http://www-sk.let.uu.nl/eurotra/act91-93.pdf`.

6 In Leuven waren het er in 1986 eveneens 4, maar er zijn geen gegevens over hoeveel het er in 1990 waren. De groei zal hier minder sterk zijn geweest, omdat de groep al twee jaar bezig was. Bovendien hoefde Leuven minder verschillende vreemde talen af te dekken, en had dus minder verschillende taalkundigen nodig.

Zink Typografie Sheet 62 of 213 - Page 54 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Louis des Tombe en Steven Krauwer waren er natuurlijk nog, maar die hadden door de groei te weinig tijd om zich om alle mensen te bekommeren. Zij moesten het hele project *runnen*. Op een gegeven moment was het echt een groep van meer dan twintig mensen. Er was dus een hele grote groep, waar je weinig aan had vanwege alle beginners." (Mineur 1993). De sterke groei was echter nodig, omdat Nederland door de late start meer werk in minder tijd moest doen. En meer geld in minder tijd moest opmaken.

Het geld van Eurotra heeft een enorme impact gehad op de computerlinguïstiek in Leuven en Utrecht, en daarmee de computerlinguistiek in Vlaanderen en Nederland. Van Eynde: "Het is met Eurotramiddelen dat wij de eerste PCs hebben gekocht, dat was begin jaren tachtig. Mijn proefschrift was het eerste dat met een computer geschreven was." Krauwer schetst eenzelfde beeld van zijn groep: "Met Eurotra begon ook het computeronderwijs in Utrecht. Van Eurotrageld kochten we onze eerste computer, een Osborne. En verder gingen onze studenten in het buitenland bij Eurotra-deelnemers op stage van Eurotrageld." Krauwer schat dat hij samen met Des Tombe tussen de zestig en tachtig studenten heeft opgeleid. Alice Dijkstra, nu programma-officer bij NWO, was er een van. Ze begon bij Engelse Taal- en Letterkunde, had bij Martin Boot al een bijvak computerlinguïstiek gevolgd, en hoorde via via dat er bij ATW ook een soort computerlinguïstiek gegeven werd. "Toen begon ik pas echt in de taaltechnologie. Toen dacht ik voor het eerst in mijn studie "dit is hartstikke leuk"." Van Noord en Dijkstra gaan beide via het Eurotra-netwerk op stage, de een in Essex, de ander in Manchester. Dijkstra: "Als je de basisvakken had afgerond zeiden Steven en Louis "En nu eruit, ergens anders wat opsnuiven." Via het Eurotra-netwerk hadden ze wel wat contacten, waaronder Essex, Manchester, Denemarken. Drie maanden ging je. Want dat was gezond." Ook voor Krauwer en Des Tombe zelf kwam het extra geld heel goed van pas. "Luxemburg gaf contracten voor specifieke taken en geld daarvoor. Dat geld was salaris, maar wij werden gewoon betaald door de faculteit – dat hadden ze eerst niet in de gaten. Dus wij konden met dat geld onze onderwijsverplichting vrijkopen. En reizen. Vanuit de faculteit was daar vrijwel geen budget voor, en met al die vergaderingen reisden wij heel veel", aldus Krauwer.

De budgetten van Eurotra lieten Utrecht ruimte om naast de voor-

Zink Typografie Sheet 63 of 213 - Page 55 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



geschreven takenpakketten een aantal andere dingen te doen, die direct in het verlengde lagen van het Eurotraproject. Dit werd oogluikend toegestaan, zolang de afgesproken taken maar naar behoren uitgevoerd werden. Des Tombe en Krauwer hadden zich in hun centrale rollen hard gemaakt voor het CAT-systeem, waar elementen van unificatiegrammatica in verwerkt zaten. In samenwerking met Essex werd er gewerkt aan een implementatie van deze ideeën. Gertjan van Noord, die op dat moment student-assistent is, doet namens Utrecht het meeste werk. Uit deze samenwerking met Essex ontstaat MiMo, een implementatie van een vertaalsysteem op basis van kleine grammatica's die gebruik maakten van unificatie. Er was hoop dat MiMo de opvolger zou worden van het officiële programma, maar dat is nooit gebeurd. Sergei Perschke was kritisch. "Toen wij aan het systeem begonnen, had de baas van het Eurotra-project, Perschke, kleinerend opgemerkt dat het een leuk systeem was voor wetenschappelijke doeleinden, een Mickey Mouse, maar niet geschikt voor het echte werk", aldus Gertjan van Noord (Mineur 1993). MiMo werd een geuzennaam en is gebleven. Dat dit systeem nooit in zijn geheel geïmplementeerd is in het officiële Eurotrasysteem, stelde de Krauwer en Des Tombe teleur. Vanaf dat moment investeerden ze minder in centrale teams, en meer in MiMo. In Utrecht werd een nieuwe versie van MiMo ontwikkeld. Een zijlijn, die strikt genomen niet onder Eurotra viel, maar er wel uit voortkwam en gedoogd werd. Krauwer probeerde zelfs wat extra geld los te peuteren om MiMoII parallel aan het officiële systeem door te blijven ontwikkelen, maar dat was te veel gevraagd. Toch werd er doorontwikkeld, en uiteindelijk levert dat een systeem op voor een subset van het Nederlands, Engels en Spaans (van Noord, Dorrepaal, van der Eijk, Florenza & des Tombe 1990). Gertjan van Noord over MiMoII: "Het begon weliswaar als een verbetering van het eerste systeem, maar uiteindelijk was er nauwelijks iets meer hetzelfde. [...] Het was gebaseerd op unificatiegrammatica's. Daardoor verliep een en ander niet alleen wat makkelijker, het werd ook veel leuker, omdat andere mensen geïnteresseerd waren in wat je deed. Een paper over de oude Eurotra-spullen trok niet zo veel aandacht, omdat het uit een erg idiosyncratisch wereldje kwam. Voor problemen met unificatie-grammatica's waarover ik schreef, was wel belangstelling, omdat anderen ook unificatiegrammatica's gebruikten. Op die manier was er veel meer communicatie mogelijk, naar beide kanten,

Zink Typografie Sheet 64 of 213 - Page 56 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



want wat andere mensen deden, was nu ook interessant voor ons." En het nieuwe MiMo-systeem heeft succes: "[Wij hebben] het ook wel eens gedemonstreerd in Luxemburg bij officiële evaluaties. Dat was heel aardig, omdat het met het originele Eurotra-systeem nog niet zo lukte. Op twintig verschillende plaatsen in Europa werd aan het systeem gewerkt; dan schroef je het niet zo makkelijk in elkaar. Ons systeem werd op één plaats gemaakt, door een paar man. Dat werkte: er was altijd wel een versie die wat kon, waar daadwerkelijk een zin inging en een zin in een andere taal uitkwam." (Mineur 1993). Een mooi resultaat, misschien wel het belangrijkste van het Utrechtse team. Krauwer: "Ik kan wel zeggen dat Louis en ik altijd trots zijn geweest op Gertjan."

Het Nederlandse team organiseerde in 1990 de laatste editie van de jaarlijkse Eurotra workshop. Het evenement trok 160 bezoekers uit de verschillende landenteams. Hans Kamp treedt op als een van de gastsprekers.

## Eurotra(uma?)

Op 22 november 1980 vraagt Hugo Brandt Corstius zich onder het pseudoniem Dr. M. Schip in de Volkskrant af: "Zal er in 1986 tot ieders tevredenheid een Europese vertaalmachine staan te snorren, of zal men in 1986 tot dezelfde droeve conclusie komen als in 1966 in Amerika? Gaat Europa Amerika inhalen, of gaan ze een Amerikaanse fout van twintig jaar geleden herhalen?" Die Europese vertaalmachine is er nooit gekomen. Betekent dat dat Europa dezelfde fout heeft begaan als Amerika twintig jaar eerder?

De uitgangssituatie was, zoals Sciarone in zijn oratie duidelijk maakte, veel beter. De computers waren beter, de taalkunde was een stuk verder ontwikkeld, en de business case was een stuk sterker dan in 1966. Hugo Brandt Corstius stelt daartegenover dat de fundamentele moeilijkheid van automatisch vertalen, het heel precies bepalen van de bedoelde betekenis van de te vertalen zin, nog steeds niet opgelost was. Dat zou overkomelijk zijn, wanneer er wel een systeem opgeleverd zou zijn dat voor alle Europese taalparen wel alle mogelijke vertalingen gaf. In dat geval zou later onderzoek het probleem van desambiguatie aan kunnen pakken. Maar ook dat systeem is er niet

Zink Typografie Sheet 65 of 213 - Page 57 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



gekomen. Het is gemakkelijk een aantal factoren aan te wijzen die grote invloed hebben gehad op de uitkomst van het project.

Allereerst is Eurotra opgezet als een ontwikkelproject, terwijl er nog heel veel research nodig was om te kijken of en hoe zaken aangepakt moesten worden. Of, zoals Steven Krauwer het zegt in de presentatie waaraan de titel van deze paragraaf ontleend is: je kunt niet iets bouwen dat nog niet uitgevonden is.[7] Tijdens het interview vraagt hij zich hardop af: "Waarom zijn er niet eerst een paar parallelle onderzoekstrajecten georganiseerd, waarna de beste uitkomsten geselecteerd kunnen worden en toegepast in een later ontwikkeltraject?"

Dan de decentrale opzet. De meeste betrokkenen zien dat als een enorme handicap voor het ontwikkelen van een goed werkend systeem. Het betekende dat over alles overlegd moest worden. Dat bracht om te beginnen een enorme vergaderlast met zich mee – al zijn ook daar wel positieve noten bij te plaatsen: "Ik vond dat zelf overigens niet zo erg", aldus Louis des Tombe (Mineur 1992). "Ik vond het wel leuk om bij allerlei buitenlandse activiteiten op bezoek te gaan. Door Europa reizen en zien hoe men daar bezig was vond ik erg leuk. Maar productief voor het maken van een vertaalsysteem was het niet." Verder leverde de decentrale opzet een afhankelijkheid op. Als er één groep te laat opleverde, moesten anderen groepen daarop wachten. Plus: integratie van de componenten was niet triviaal. Doordat zeer veel groepen min of meer onafhankelijk van elkaar en elk op hun eigen wijze aan het systeem werkten, konden componenten niet zomaar eventjes met elkaar gecombineerd worden. Van Eynde gelooft echter niet dat de decentrale opzet de reden is dat Eurotra geen werkend systeem heeft opgeleverd. "De moeilijkheden van zo'n opzet werden opgelost door de centrale teams en door uitvoerige documentatie. Volgens mij was dat wel opvangbaar."

Van Eynde noemt wel een andere oorzaak voor het uitblijven van een product: er is te weinig gewerkt aan softwareontwikkeling. Veruit de meeste deelnemers waren taalkundigen, en die hadden niet altijd het grote doel voor ogen. Jan Odijk van het concurrerende Rosettasysteem zag het zo: "Ze waren allemaal geïnteresseerd om hun eigen taaltheorietjes te toetsen of te implementeren. Waardoor het vertaalaspect naar de achtergrond verdween. En dat kon ook mooi,

[7] http://www-sk.let.uu.nl/ond/eurotrauma_files/v3_document.htm

Zink Typografie Sheet 66 of 213 - Page 58 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



want ze hadden een analysesysteem monolinguaal, en ze hadden een generatiesysteem, monolinguaal, en dan moest er ook nog ergens een transfersysteem zijn, waar alles wat moeilijk was in terecht kwam, maar dat stelden ze telkens maar uit." Pius ten Hacken, die vanuit het Utrechtse team in de Taskforce Lexicografie van Eurotra zat, is het met Odijk eens: "Vertalen werd gebruikt als demonstratie van taalbegrip. Het was bijna een excuus om mooie taalkundige dingen te doen." Behalve softwareontwikkelaars, had hij ook graag meer vertalers betrokken gezien bij het project. Een laatste oorzaak, die samenhangt met de voorgaande, is het management. Steven Krauwer was daar niet erg van onder de indruk: "Sergei Perschke van ISPRA was de leider van Eurotra, zowel administratief als inhoudelijk. Hij dacht dat hij wel wist hoe het moest, en gebruikte zijn hierarchische positie om wetenschappelijke disputen te beslissen." Een principiële fout volgens Krauwer, die er voorstander van is dat de administratieve leiding en de inhoudelijke leiding gescheiden worden.

Iedereen is het erover eens dat Eurotra wel degelijk wat opgeleverd heeft. Een uitgebreide documentatie, opbouw van kennis over automatisch vertalen, theoretisch taalkundig werk en vooral: een Europees netwerk. Krauwer: "De impact van Eurotra was dat er ineens een taaltechnologische gemeenschap ontstond, waar wel samengewerkt werd. Een netwerk. En dat netwerk bestaat nog steeds. Dat was de investering best waard. Maar als dat de opzet was had dat ook voor minder gerealiseerd kunnen worden."

In Europees verband was dat netwerk heel erg gericht op Eurotra. Maar binnen Nederland en Vlaanderen was er nog een netwerk: dat tussen de verschillende projecten op het gebied van automatisch vertalen. In de volgende paragrafen komen achtereenvolgens METAL, DLT en Rosetta aan bod.

## 2.3 METAL

De Vlaamse Eurotra-groep huisde op de tweede verdieping van de Maria-Theresiastraat 21 in Leuven. Een verdieping lager werkte vanaf eind 1985 een onderzoeksgroep aan METAL, het vertaalsysteem van Siemens. Die verdeling van werkplekken illustreert de onderlinge ver-

Zink Typografie Sheet 67 of 213 - Page 59 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



houdingen in het begin van het project zoals Herman Caeyers, de eerste projectleider van het METAL-onderzoek in Leuven, die beschrijft. In zijn ogen werd er door Eurotra namelijk neergekeken op hun werk. "Dat was niet academisch genoeg", vertelt hij. "Zo'n commercieel project werd als minderwaardig beschouwd."

Dat METAL daar aan de Maria-Theresiastraat terecht gekomen was, lag niet voor de hand. De vertaalsoftware is gebaseerd op het werk van Jonathan Slocum en Winfield Bennett (Bennett & Slocum 1985) aan de University of Texas in Austin. Zij ontwikkelden een vertaalsysteem dat net als Eurotra gebaseerd was op het transfermodel, hoewel transfer en generatie in METAL enigszins in elkaar overliepen. Op een gegeven moment is Siemens dit werk gaan sponsoren. Er werd in Austin vooral gewerkt aan het taalpaar Duits-Engels. Caeyers: "Misschien had dat te maken met de sponsoring van Siemens. Maar ik denk dat het ook van doen had met het feit dat het taalkundige gedeelte vanuit de Duitse groep in Austin gedaan werd. En omdat men dacht dat het vanwege de naamvallen het gemakkelijkst was om vanuit het Duits te vertrekken." In 1984 besloot Siemens dat alle research gecentraliseerd moest worden in München. Vanaf dat moment werden er nog wel wat zaken in Austin gedaan, maar verhuisde verder alle ontwikkeling naar Duitsland. Een onderzoekscentrum in België, waar op dat moment alleen een verkoopkantoor gevestigd was, paste absoluut niet in de plannen voor centralisatie. Maar slechts één jaar later sloot Siemens een groot contract met de Belgische overheid. "Ettelijke miljoenen Belgische franken" waren er volgens Caeyers mee gemoeid. En die overheid stelde een voorwaarde: Siemens moest investeren in Research en Development in België. En dus moest er een onderzoeksproject gevonden worden dat enerzijds paste in een bestaande onderzoekslijn van het bedrijf, en anderzijds om een of andere reden niet in München gedaan kon worden. Toen is besloten om het automatisch vertalen naar België te verhuizen, waar dan de taalparen Nederlands-Frans en Frans-Nederlands voor METAL ontwikkeld zouden worden. De talenkennis die daarvoor nodig was kon niet gemakkelijk gevonden worden in Duitsland, maar juist wel heel gemakkelijk in het drietalige België. Naar goed Belgisch gebruik moest de investering wel verdeeld worden over Wallonië en Vlaanderen, en zo kwam de grammaticaontwikkeling terecht in Leuven, waar de Eurotragroep al enige tijd bezig was, en werd de woordenboekontwikkeling ondergebracht in

Zink Typografie Sheet 68 of 213 - Page 60 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Mons (Bergen), bij de tolkenopleiding *Ecole des Interprêtes Internationaux*. In het eerste jaar was ook het *Institut Montefiori* in Luik betrokken. Hier zaten enkele specialisten op het gebied van Lisp, de programmeertaal waarin METAL geschreven was. Zij zorgden voor ondersteuning op het gebied van de parse-software en trainden de computerlinguïsten in Leuven in het gebruik van Lisp. Caeyers zelf werkte direct voor Siemens. Hij was tijdens zijn proefschriftverdediging gespot door een Siemensdirecteur, die geïnteresserd was in Artificiële Intelligentie (AI) en hem vroeg projectleider te worden. "Ik had natuurkunde en wiskunde gedaan, en doctoreerde op een onderwerp over muziek en AI, over wiskundige modellen voor de perceptie van muziek." Op 1 oktober 1986 begint Caeyers bij Siemens, en niet lang daarna startte het METAL-project.

Caeyers was in eerste instantie niet erg gelukkig met de ploeg mensen die het project moesten uitvoeren. "De mensen die het moesten uitvoeren, zaten feitelijk bij Eurotra. Ze namen dan wat net afgestudeerde mensen in dienst om voor Siemens te werken." aldus Caeyers. "Want dat leverde geld op." De projectleider werkt er in het eerste jaar hard aan om meer ervaren krachten te vinden die uitsluitend aan het METAL-project zouden werken. Die vond hij in de personen van Rudi Gebruers en Geert Adriaens. Daarnaast moest er hardware komen. Want METAL was geschreven in Lisp, en dat draaide op speciale machines. "Dat waren Symbolics machines", herinnert Caeyers zich. "De aankoop van de eerste machine in die tijd was 14 miljoen Belgische frank [350k euro]. Dat was ontzettend veel geld. Die machine woog 200 kg en maakte ontzettend veel lawaai." De machine werd zeer intensief gebruikt: "Wij werkten dag en nacht in shifts, want er kon maar één persoon tegelijk aan werken." Ondanks de enorme kosten was het geheugen met 4MB RAM geheugen zeer beperkt naar moderne maatstaven. "Maar toen was iedereen wel jaloers dat wij zo'n krachtige machine hadden."

Het team in Leuven werkt aan grammatica's voor het Nederlands en het Frans. Dat was een heel belangrijk gegeven, omdat METAL ontworpen was voor het Duits. De Vlamingen liepen op tegen aspecten van het systeem die niet toepasbaar waren voor andere talen. Caeyers: "Ik herinner mij dat er nog excuses geweest zijn achteraf, omdat men in Amerika op dat moment dacht dat elke vreemde taal naamvallen had." Toen het team eenmaal goed ingewerkt was, heeft het zich

Zink Typografie Sheet 69 of 213 - Page 61 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



dan ook toegelegd op veranderingen aan het systeem die het breder toepasbaar maakten. De groep ontwierp een valentiesysteem dat de syntactische rol van verschillende phrases herkent zonder daarvoor afhankelijk te zijn van naamvallen. Hierdoor had het team invloed op de ontwikkeling van het kernsysteem.

Hoewel METAL net als Eurotra een transfersysteem is, zijn er fundamentele verschillen in de opzet van de beide systemen. Om te beginnen is METAL vanaf het begin bedoeld als een systeem dat ingebed zou worden in een werkomgeving voor vertalers, waarin de automatisch vertaalde tekst gepostedit kon worden. Ook aan de voorkant voorzag het systeem een interactieve stap waarin tolken de woordenboeken konden aanvullen met woorden uit de tekst die niet in de reeds aanwezige elektronische lexica gevonden konden worden. Waar binnen Eurotra een uitgangspunt was dat de vooruitgang in de theoretische taalkunde een plek zou krijgen in het praktische systeem, wat leidde tot uitgebreide discussie over het te hanteren theoretisch kader, is binnen METAL vanaf het begin gewerkt met relatief simpele contextvrije grammaticaregels, die verrijkt werden met expressies die de applicatie van de regel beperkten of de te bouwen structuur mede definieerden. Deze expressies konden condities zijn waaraan voldaan moest worden, of instructies over de te vormen structuur (Gebruers 1986). Deze verschillen illustreren de fundamenteel andere basis waarop gewerkt werd: METAL was een commercieel project en moest een werkend systeem zijn, zo snel mogelijk. Hoewel Eurotra in principe ook gericht was op het ontwikkelen van een werkend vertaalsysteem, zorgden de decentrale opzet, de focus op theoretische taalkunde, en het inefficiënte projectmanagement ervoor dat het toch vooral een academische onderneming werd. Deze heel andere insteek vertaalde zich naar heel andere resultaten: waar Eurotra documentatie en publicaties opleverde en een groot Europees netwerk van taalkundigen en computerlinguïsten, maakte METAL een werkend systeem. Niet perfect, maar wel werkend, zodat reviews en vergelijkend onderzoek gedaan konden worden (Bernhard 1991). Caeyers: "We hadden beschikking over corpusdata en gebruikten evaluatietechnieken om vooruitgang te meten. De grammaticaregels konden elkaar beïnvloeden, dus het kon zijn dat je iets veranderde om een probleem op te lossen, maar daarmee elders iets stuk maakte. Dus na iedere verandering werd het corpus automatisch opnieuw getest."

Zink Typografie Sheet 70 of 213 - Page 62 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Over deze evaluatietechnieken is weinig gepubliceerd, maar er is wel nog een paper uit 1985, waarbij de originele brontekst en ruwe output van het systeem zijn meegeleverd (Bennett, Slocum, Whiffin & Norcross 1985). De praktische insteek van METAL werd prominent geëtaleerd. "Vermits het projekt een industrieel prototype beoogt, kan het Belgisch METAL-team zich niet veroorloven om een klein aantal problemen in teksten van beperkte omvang te onderzoeken in het kader van een of andere ambitieuze semantische theorie, maar moet het zich richten tot problemen die voortkomen uit de toepassing van een minder ambitieuze linguïstische theorie op een grote hoeveelheid taalmateriaal in een meer realistische omgeving"(Gebruers 1986). Hiermee lijkt het team zich bewust af te zetten tegen de gangbare praktijk binnen Eurotra. En inderdaad, waar Caeyers enerzijds aangeeft dat er vanuit Eurotra wat werd neergekeken op METAL, was de resultaatgerichte METAL-groep anderzijds ook behoorlijk kritisch over Eurotra. "*Eurotravel* noemden we dat", vertelt Caeyers. "Zaten meer in het vliegtuig dan op kantoor." De verhoudingen tussen beide groepen normaliseerde zich later overigens wel. Eurotra-medewerker Van Eynde: "We wisten van elkaars aanpak, we lazen elkaars artikelen, en we hebben gezamenlijk een lezingenreeks georganiseerd, waar bijvoorbeeld ook Jan Landsbergen van Rosetta, Toon Witkam van BSO, en Theo Janssen uit Amsterdam zijn geweest." Zo staat METAL ook in contact met de andere vertaalprojecten in Nederland en Vlaanderen. Verregaande samenwerking is er echter nooit geweest. "Op een gegeven moment zijn Rosetta en METAL wel naast elkaar gelegd om te kijken of een technisch huwelijk mogelijk was", herinnert Caeyers zich. Maar er is nooit wat van gekomen. Jan Landsbergen van het Rosetta-project bij Philips was daar niet rouwig om: "Wij deden ons ding en zij deden hun ding. Ik heb nog aangeboden een Nederlandse versie te maken van het METAL-systeem, dan konden we daaraan werken, maar ik wilde veel liever aan Rosetta werken."

Na enige tijd verhuisde Siemens het METAL-project van Leuven naar Luik. Ook deze beslissing was gemotiveerd door een grote overheidsbesteding bij Siemens. Als compensatie moest het bedrijf een onderzoeksinstelling oprichten in Luik. Toen zijn alle onderzoeksactiviteiten overgeplaatst naar Luik. Geert Adriaens had het projectmanagement van METAL op dat moment al overgenomen van Herman

Zink Typografie Sheet 71 of 213 - Page 63 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Caeyers, die manager van het AI-lab geworden was. Voor de onderzoeksgroep kwam er nog steeds geen einde aan de veranderingen. In 1993 richtte Caeyers Lant op, dat de ontwikkeling van METAL weer overnam. Maar Siemens besloot op een gegeven moment om alle rechten voor METAL te verkopen aan GMS. Een beslissing waar Caeyers zo zijn bedenkingen bij heeft. Het *Gesellschaft für multilinguale Systeme mbH* was een bedrijf dat zich toegelegde op de ontwikkeling van vertaalsystemen van en naar het Russisch. Grote man van GMS was Stephan Bodenkamp. Deze Bodenkamp werd later, op 20 december 2000, door de rechtbank in München veroordeeld wegens valsheid in geschrifte: hij had onder een valse naam geopereerd als coördinator van het Sensusproject, een Europees initiatief dat tot doel had een informatie-uitwisselingsprogramma te ontwikkelen voor de Europese politie- en inlichtingendiensten. Bodenkamp bleek eigenlijk Christoph Klonowski te heten en een agent te zijn van de Duitse geheime dienst, de Bundesnachrichtendienst. Overigens melden René De Witte, Luc van Aelst en Luc van Peteghem dat Bodenkamp de naam is van de echtgenote van Klonowski (de Witte, van Aelst & van Peteghem 2002), en sinds 1976 is het in Duitsland voor een man mogelijk om bij een huwelijk de naam van zijn vrouw aan te nemen. Bodenkamp/Klonowski zou later ook bij Lernout en Hauspie een belangrijke rol spelen. Volgens Caeyers is Siemens door de Duitse overheid onder druk gezet om ervoor te zorgen dat de METAL-technologie in Duitsland bleef, en was dat de reden dat GMS alle rechten voor het systeem kreeg. Met zijn bedrijf Lant spande hij een rechtzaak aan en uiteindelijk werd besloten dat GMS alle rechten kreeg over de C++ versie van het systeem, die inmiddels gemaakt was met het oog op de consumentenmarkt, maar dat Lant de rechten kreeg over de oorspronkelijke Lisp-versie – en deze niet mocht vertalen naar C++. “Wij wilden integreren in workflowsystemen voor grote vertaalbureaus, en waren dus helemaal niet van plan om de consumentenmarkt op te gaan.” GMS is uiteindelijk opgekocht door Lernout en Hauspie, en met de val van het moederbedrijf mee de afgrond in getrokken. Lant is opgegaan in Xplanation, dat in haar product Tstream, inderdaad een workflowsysteem voor vertalers, nog een verre afstammeling van METAL in productie heeft.

Zink Typografie Sheet 72 of 213 - Page 64 of 205

## 2.4 Distributed Language Translation

> ...als je Engels gebruikt als tussentaal, dan hoeft er helemaal niet meer vertaald te worden, want iedereen kent Engels!
>
> *(IJlander 1990)*

Distributed Language Translation (DLT) van het Buro voor Systeemontwikkeling (BSO) is een vreemde eend in de vijver van de projecten op het gebied van Machine Tranlation (MT) uit de jaren tachtig. In tegenstelling tot Eurotra en METAL, maar ook SUSY en ARIANE, is DLT geen transfer-systeem, maar een vertaalsysteem via een interlingua. En hiervoor wordt geen logische taal gebruikt, en ook niet een van de brontalen, maar Esperanto. Een keuze die nogal wat kritiek uitlokte, bij voorbeeld van Hugo Brandt Corstius: "Zoals elke goochelaar weet, moet je bij een truc altijd iets totaal onverwachts en onzinnigs doen om de aandacht af te leiden. Bij BSO was de truc dat het Esperanto als middeltaal tussen brontaal en doeltaal gebruikte." (Brandt Corstius 2003) En Jan Odijk, die zelf bij het concurrerende vertaalsysteem Rosetta werkte: "BSO werd door mij en volgens mij iedereen op universiteiten niet serieus genomen." MT-expert John Hutchins noemt DLT in 1988 echter één van "Two of the most innovative MT projects at present time" (Hutchins 1988) – de ander was Rosetta, dat hierna aan bod komt.

Het verhaal van DLT begint in 1979, en wederom niet bij een linguïst. Vliegtuigbouwkundige Toon Witkam werkt voor BSO, het softwareservicebedrijf van ICT-guru Eckart Wintzen. Hij wordt als automatiseringsdeskundige verhuurd aan bedrijven, en brengt zodoende veel tijd door in zijn auto, waar hij naar de radio luistert. "En toen heb ik een ingeving gehad", vertelt Witkam. "Ik hoorde op de BBC iets over een Frans werk dat vertaald werd. Op een of andere manier heb ik toen een verbinding gelegd tussen vertalen en Esperanto." Met een interlingua kan het aantal taalparen waartussen vertaald moet worden verkleind worden van bijvoorbeeld 72 naar 18 voor een groep van 9 brontalen. Daar staat tegenover dat niet geprofiteerd kan worden van overeenkomsten die bestaan tussen de talen van sommige taalparen, dat de interlingua alle semantische onderscheidingen moet coderen die in alle doeltalen samen voorkomen, en dat het grote struikelblok, ambiguïteit in de brontaal, er niet mee opgelost

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



wordt. De ingeving van Witkam was dat deze interlingua Esperanto moest zijn. Geen goed idee, vindt Jan Landsbergen: "dat was het meest kenmerkende [van DLT], en ook wat mij meest tegenstond. Omdat Esperanto natuurlijk niet de oplossing is voor het vertaalprobleem. Je hebt dan twee vertaalproblemen: van en naar Esperanto." Witkam stelt daar een aantal argumenten tegenover. De taal is zeer regelmatig opgebouwd. Hierdoor kan de taal compacter gecodeerd worden dan andere talen, namelijk door morfemen te coderen in plaats van letters. Witkam heeft een stagiair dit uit laten zoeken, en de uitkomst daarvan was dat dit 19 bits per woord zou kosten, tegenover vijf tot zes bytes voor de meeste andere talen. Mooi meegenomen, want Witkam had videotex als applicatie op het oog, wat betekende dat de boodschap over een netwerk verzonden moest worden. En dataverkeer was kostbaar in de tachtiger jaren. Maar compressie was niet het cruciale punt. Witkam: "Esperanto had niet die ballast van uitzonderingen en alles wat daarmee samenhangt, wat je voor het Frans en het Engels en het Duits op school geleerd had. Dus het ontbreken van uitzonderingen, de regelmatigheid, wat het programmeren makkelijker maakt. Dat was de ingeving." Over de mate van ambiguïteit in Esperanto is Witkam vrij nuchter. "Je hebt natuurlijk heel veel soorten ambiguïteit. Bijvoorbeeld het zinnetje "zwangere vrouwen en kinderen". Zijn ook die kinderen zwanger? Dat heb je dus ook bij Esperanto, datzelfde probleem. Maar de mate van woordambiguïteit bij Esperanto is wat minder dan bij andere talen." Belangrijk aspect was ook dat de interlingua een toegankelijke taal moest zijn: "De hoofdzaak is dat een het een interlinguasysteem is met als interlingua niet een of ander broedsel van theoretische linguïsten, maar een echte taal, want Esperanto is een echte taal." Een laatste argument voor het Esperanto vinden we in *Een fabelachtig uitzicht* van Gijs IJlander, een boek dat nauwelijks verholen vertelt over het wel en wee bij BSO, al dan niet op de werkelijkheid gebaseerd. In de verbeelding van de vertellereen opgezette eekhoorn – beargumenteert de hoofdrolspeler: "Het is voor een tussentaal belangrijk dat hij een volkomen onafhankelijke positie inneemt tussen de te vertalen teksten: een vertaling in de tussentaal moet een volwaardig halfproduct zijn dat naar iedere natuurlijke taal even gemakkelijk is over te brengen. Er was eigenlijk maar één tussentaal die in aanmerking kwam en wel: het Esperanto."

Zink Typografie Sheet 74 of 213 - Page 66 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



DLT verschilde op meer punten van andere MT-projecten. Waar Eurotra bedoeld was voor batch-processing van documenten die een 'redelijke' vertaling moesten krijgen, was DLT gericht op interactieve vertaling in een netwerkopstelling: een terminal vertaalt een tekst in of van het Esperanto, en de tekst wordt over het netwerk verzonden in Esperanto. Dit moet resulteren in een hoge kwaliteit vertaling, waaraan niet meer hoeft te worden gedaan (geen post-editing). De interactie met de auteur van de originele tekst is erop gericht om mogelijke ambiguïteiten op te lossen. Een bijzondere component van DLT was het 'semantic word expert' systeem (SWESIL). Deze component zorgde na vertaling van de tekst van brontaal naar interlingua voor de desambiguering. Hiervoor werd een database gebruikt met semantische informatie in Esperanto, bijvoorbeeld acceptabele dependenties tussen woordparen, en taxonomiën van woorden.

Toon Witkam kan geen gebruik maken van BSO-middelen voor zijn project. Het eerste jaar studeert hij zelf op de materie, en trekt hij een stagair aan die het compressie-onderzoek uitvoert. In 1980, "een jaar na mijn ingeving", zoals Witkam het zelf zegt, roept hij een klein groepje van tien collega's inclusief Eckart Wintzen bij elkaar om zijn plan te presenteren. Witkam vertelt dat Eckart erg enthousiast was. Maar nog steeds kan hij niet in BSO-tijd aan het project werken. Dus gaat hij vier in plaats van vijf dagen per week werken, en doet hij DLT in de weekeinden en de extra vrij dag. Een aanvraag voor subsidie bij de Nederlandse overheid wordt niet gehonoreerd. Maar dan keert het tij, en in de zomer van 1982 krijgt Witkam van de Europese Gemeenschap een subsidie van 250.000 gulden toegewezen om een feasibility studie te doen. Witkam gaat weer voltijds in dienst en krijgt een assistent, Alex Olde Kalter. In oktober 1983 is de studie af: er ligt een 350 pagina's tellend rapport. Witkam kampeert de laatste weken in een camper op het parkeerterrein van BSO (Witkam 2006).

De studie wordt goed beoordeeld, er wordt zelf geadviseerd om vervolgfinanciering te geven. Voor BSO een enorme opsteker. Het bedrijf presenteert de plannen met grootse beloftes aan de pers. En daar was niet iedereen even blij mee. Landsbergen: "Automatisch vertalen had altijd al slechte naam doordat in de jaren 60 geweldige claims gemaakt werden, waarbij zelfs wel fraude werd gepleegd, en BSO maakte weer die fout. Dan stond er een stuk in de krant 'fantastisch idee, BSO heeft vertaalprobleem opgelost". En daar waren ze ook nog heel trots

Zink Typografie Sheet 75 of 213 - Page 67 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



op, dat dat in de krant stond. Dat wekte ergernis bij ons bedaarde academici." "PR bij BSO was uitstekend", beaamt Witkam. Maar hij nuanceert het beeld van frauduleuze claims: "Er zal natuurlijk best in die artikelen af en toe wel eens iets overdreven zijn of iets niet zo duidelijk vermeld zijn wat er wel had moeten staan. Maar in het algemeen was het wel juist."

De academische wereld staat op zijn achterste benen wanneer duidelijk wordt dat er vervolgfinanciering komt voor DLT. Niet de Europese Gemeenschap, maar het Ministerie van Economische Zaken gaat DLT subsidiëren, en wel met een subsidie van acht miljoen gulden, door BSO gematcht met nog eens acht miljoen, voor een project van zes jaar. Zestien miljoen gulden voor een vertaalproject bij BSO. Jan Landsbergen was net op tijd getipt over de aanstaande subsidie. "Ik geloof dat Sciarone toen een brief naar ministerie heeft gestuurd waarin hij uiteenzette dat het jammer zou zijn als zij subsidie kregen en dat wij een veel beter plan hadden. Intussen hadden wij snel een plan gemaakt met Philips, Utrecht en Delft, wat dan voortbouwde op Rosetta. Een project met 15 mensen, dat was het plan." Maar Landsbergen en consorten vissen achter het net: BSO krijgt de subsidie van het Ministerie, en investeert nu voor het eerst ook zelf serieus in DLT.

Het project loopt van 1984 tot 1990. Witkam neemt een team van taalkundigen, programmeurs en Esperantisten aan en gaat aan de slag. Met name twee Esperantisten hebben een grote invloed op de architectuur van het systeem: Klaus Schubert neemt een groot gedeelte van de syntactische analyse voor zijn rekening, en Victor Sadler tekent voor de semantiek. Vanuit het Ministerie wordt een toezichthoudende commissie aangesteld, waaraan DLT tweemaal per jaar rapporteert.[8] De commissie wordt ABK genoemd, naar de achternamen van de leden: Bernard Al, Harry Bunt en Gerard Kempen. En ondanks het feit dat Kempen aanvankelijk sceptisch is over het idee om Esperanto als interlingua te gebruiken, is hij uiteindelijk toch positief over het project: "Wij waren door het ministerie opgedragen dat project te volgen. Er waren gesprekken en er werden presentaties gegeven over de voortgang. Daar zat Eckart ook vaak bij. Ik vond het best een goed

[8] Gebaseerd op het verhaal van Toon Witkam. Gerard Kempen heeft het over driemaandelijkse bijeenkomsten.

Zink Typografie Sheet 76 of 213 - Page 68 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



project, even goed als Eurotra en Rosetta, omdat er heel nauw door allerlei mensen werd samengewerkt."

In de loop van het project verandert de opzet van DLT vrij radicaal. Volgens Witkam is 1988 het jaar van de doorbraak. Dan hoort hij op COLING-88 in Boedapest voor het eerst over het werk van IBM op het gebied van statistical machine translation (Brown, Cocke, Pietra, Pietra, Jelinek, Mercer & Roossin 1988), en overtuigt Junichi Tsuji in Japan hem ervan dat MT-wetenschappers moeten werken met echte zinnen uit bestaande teksten, in plaats van met de standaard voorbeeldzinnetjes, zoals *John loves Mary*. Waar DLT aanvankelijk helemaal regelgebaseerd was, wordt dan een corpusgebaseerde aanpak geïntroduceerd. En daarmee was het in Nederland het eerste vertaalsysteem dat corpusgebaseerde technieken gebruikte. Die gewijzigde aanpak lokt de vraag uit wat de rol van het Esperanto nog was in de nieuwe opzet. Tenslotte waren er voor Esperanto toch aanzienlijk minder corpora voor handen dan voor het Engels en Frans, de bron- en doeltaal van het DLT demosysteem. Zelf stelt Witkam: "At the end of the DLT pilot project I tended to conclude that Esperanto was in fact not the key issue in our technology any more." (Witkam 2006). In 2005 presenteert hij nog een plan voor een statistisch MT systeem voor het Esperanto (Witkam 2005), maar onduidelijk is of de rol van Espeanrto daarin nog die van interlingua is, of die van bron- en/of doeltaal. In zijn persoonlijke visie op de geschiedenis van DLT zegt hij hierover: "In a multilingual system, using an IL will reduce the required number of BKBs [bilingual knowledge banks] in the same way as it would limit the number of SL-analysis and TL-synthesis modules in conventional MT. Limiting the number of different BKBs is not completely uninteresting, because each BKB requires maintenance and a lot of preparatory work." (Witkam 2006).

In 1990 is het geld op. In december 1987, dus nog voor de introductie van corpusgebaseerde technieken, was er al een demo, die *Simplified English*, afkomstig uit technische documentatie uit de luchtvaartsector, naar Frans vertaalt. Hij wordt door Witkam 'slow' genoemd (Witkam 1988), en heeft slechts een woordenboek van 2000 woorden en een beperkte grammatica. Verder is er heel veel documentatie, Witkam schat wel 1800 pagina's, voornamelijk gepubliceerd in een reeks boeken, de DLT serie, uitgegeven door Foris. Witkam werkt aan het eind plannen uit om met de corpusgebaseerde aanpak

Zink Typografie Sheet 77 of 213 - Page 69 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



verder te gaan, en daar een product van te maken. "Dat betekende wel, om realistisch te zijn, dat we het tienvoudige aan inzet moesten hebben, aan geld en aan mensen."

BSO kan en wil die extra last zeker niet dragen, dus wordt er gezocht naar alternatieven. Onderdeel van de zoektocht naar nieuwe investeerders is een media-offensief. ANP-journalist Carel Wiemers, interviewt Klaus Schubert, de feitelijke leider van de researchgroep op dat moment. Het artikel wordt door tientallen kranten overgenomen. Het Parool kopt op 17 juli 1990: "Computer spreekt elke taal". De Volkskrant plaatst het stuk zelfs op de voorpagina onder de kop "Computer zet elke taal in andere taal om". Job van Zuijlen was als computationeel taalkundige aan het project verbonden. Via email vertelt hij vanuit de Verenigde Staten dat de BSO-leiding niet echt blij was met de lawine aan media-aandacht, maar dat er geen feitelijke onjuistheden in het originele artikel staan. "Maar de krantekoppen benadrukken natuurlijk het meest sappige".

Witkam en Wintzen gaan praten met IBM, DEC en verschillende grote uitgeverijen. Consultancybureau McKinsey voert een haalbaarheidsstudie uit. Ze concluderen dat er wel een markt is, maar ze durven geen oordeel te vellen over de tijdsplanning van Witkam. Die voorspelde minimaal vijf jaar nodig te hebben voor de ontwikkeling van een product, en dan nog twee tot drie jaar tot break-even, wat zou resulteren in een *return-of-investment* van 8 jaar. Er wordt geen investeerder gevonden, en het project houdt op te bestaan. Tot ongenoegen van het Ministerie van Economische Zaken, dat 8 miljoen geïnvesteerde guldens in rook ziet opgaan. Kempen: "En die gaven voor een deel de begeleidingscommissie de schuld, zo van jullie hebben niet goed opgelet."

Kempen zelf neemt een van de ontwikkelaars van het project, Theo Vosse, over. "Ik had een heel hoge pet op van Theo. Ik had een baan te vergeven, en ik wist dat hij goed was, doordat ik hem jarenlang had meegemaakt als externe reviewer van dat project." Witkam zelf blijft bij BSO in een andere rol, en is nog steeds actief in de wereld van het Esperanto. Terugkijkend op zijn tijd bij DLT is hij nog steeds trots. Hij heeft wel een verklaring voor de kritiek die er was: "Kijk, BSO sprong er een beetje uit als een glitterend bedrijf, en dat valt niet bij iedereen goed." In de jaren tachtig schreef hij de advertenties voor het blad *Language Technology/Electric World*, waarin hij voorspelt dat in

Zink Typografie Sheet 78 of 213 - Page 70 of 205

2000 het vertaalprobleem is opgelost, en eigenlijk staat hij daar nog steeds achter. "Als we na 1990 door hadden kunnen gaan – dan moet ik eigenlijk zeggen door hadden kunnen gaan op grotere schaal – dan hadden we het vertaalprobleem voor 2000 opgelost."

## 2.5 Rosetta

Hugo Brandt Corstius is beroemd geworden door zijn scherpe pen, en een van zijn favoriete onderwerpen van kritiek was automatisch vertalen. Zelf vindt hij het een van zijn grootste bijdragen aan het vak, dat hij als eerste in Nederland beweerde dat automatisch vertalen onmogelijk was. Des te uitzonderlijker zijn oordeel over het vertaalonderzoek bij Philips in Eindhoven: "Niet alle werk op vertaalgebied is bedrog. Wat Jan Landsbergen en zijn groep deden in Rosetta was verantwoord en prachtig." (Brandt Corstius 2003). Het vierde vertaalproject in het Nederlandse taalgebied in de jaren tachtig staat boven alles bekend om zijn schoonheid.

Jan Landsbergen zat aan het eind van de jaren zeventig bij het Philips Natlab, maar de leiding wilde het project rondom het vraag-/antwoordsysteem PHLIQA opheffen. "Dat viel te betreuren, maar kwam mij persoonlijk wel goed uit", zegt Landsbergen twintig jaar later in zijn afscheidsrede (Landsbergen 2000). "Ik had namelijk een droom [...]." Die droom was een vertaalsysteem op basis van de op Montague-grammatica gebaseerde M-grammatica, die hij voor PHLIQA ontwikkeld, maar nooit geïmplementeerd had. Landsbergen wist heel goed dat automatisch vertalen met grote problemen te kampen had. In 1977 schreef hij nog: "... tot 1965 was het onderzoek van computerlinguïsten vooral gericht op één, ongelukkig gekozen, toepassing: het machinaal vertalen." (Landsbergen 1977). Maar hij had een idee in zijn hoofd, en daar wilde hij mee aan de slag. En dus opperde hij zijn idee bij de Philipsleiding, die positief reageerde. "Toen ik het onderwerp vertalen liet vallen, riep dat heel veel enthousiasme op. Dat was het grote gevaar van vertalen. Dat komt natuurlijk vooral door onwetendheid." Gevaarlijk of niet, hij krijgt toestemming om één jaar lang samen met de jonge informaticus Joep Rous aan het onderwerp te werken.

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



M-grammatica was een aanpassing van Montague-grammatica door Jan Landsbergen. De grammatica van Montague, zoals bekend geworden uit zijn artikel PTQ (The Proper Treatment of Quantification) (Montague 1973) beschouwt zinnen als intensioneel-logische expressies die compositioneel zijn opgebouwd uit een eindige set basisexpressies. De grammatica is generatief: ze beschrijft regels die basisexpressies met elkaar combineren tot grotere eenheden. Maar Landsbergen wilde ook het omgekeerde: zinnen ontleden tot de basisexpressies waaruit ze opgebouwd zijn. Bovendien ontdekte hij dat het grammaticafragment uit PTQ in sommige gevallen naast de juiste interpretatie ook incorrecte analyses produceerde. Beide problemen loste hij op door twee veranderingen aan te brengen in het grammaticaformalisme. Allereerst maakte hij de regels krachtiger. Waar de meeste regels van Montague alleen concatenatie van expressies toepassen, stond Landsbergen ook transformaties toe. Deze krachtiger regels waren bovendien reversibel, dus zowel geschikt voor analyse als voor generatie. Maar ze overgenereerden nog wel. Dit onderving Landsbergen door de grammaticaregels niet, zoals in Montague's PTQ, direct op ongestructureerde zinnen toe te passen, maar in plaats daarvan ze te herformuleren als regels op constituentstructuren. Om dit te kunnen doen, moesten de mogelijke syntactische structuren van de zinnen eerst door een contextvrije grammatica geïdentificeerd worden, waarna de nieuwe regels van de M-grammatica op deze kandidaatstructuren konden worden toegepast en alleen de juiste analyse(s) overbleven (Landsbergen 1981).

De kern van het op M-grammatica gebaseerde vertaalsysteem is dat voor elk van de te vertalen talen reversibele grammatica's geschreven worden die isomorf zijn: ieder woord of phrase uit het woordenboek en iedere regel uit de grammatica in de brontaal is gelinkt aan een woord, phrase of grammaticaregel met dezelfde betekenis in alle andere talen. Twee zinnen zijn dan een mogelijke vertaling van elkaar als ze een analyse hebben die helemaal uit corresponderende basisexpressies en een opeenvolging van corresponderende grammaticaregels bestaat. De grootste ontwikkellast komt hierbij te liggen op het schrijven van de parallelle grammatica's: wanneer de grammatica van de brontaal een analyse geeft voor een zin, dan is er gegarandeerd een corresponderende boom in elk van de andere talen, en kan de bijbehorende zin gevonden worden door simpelweg de regels uit de

Zink Typografie Sheet 80 of 213 - Page 72 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



derivatie toe te passen. Of in Landsbergen's woorden: "Je schrijft de grammatica's parallel, en dan is vertalen een fluitje van een cent."

Rosetta past niet naadloos in de driedeling in vertaalsystemen (direct, transfer en via een interlingua). Waar in een zuiver interlinguasysteem de grammatica's voor de verschillende talen onafhankelijk van elkaar ontwikkeld kunnen worden, was dat in Rosetta niet het geval: de grammatica's waren op elkaar afgestemd. Toch plaatsen zowel Hutchins als Landsbergen het systeem tussen de interlinguasystemen: wanneer de woorden en regels in de derivatiebomen voorzien worden van hun logische betekenis, kunnen deze gezien worden als expressies in een logische tussentaal. Odijk noemt het systeem hybride: interlinguaal voor de regels, maar met taalpaarspecifieke transfer voor de basisexpressies (p.c.). Het uiteindelijke doel van Rosetta, genoemd naar Rosetta's Stone, was hetzelfde als van DLT: interactieve vertaling zonder postediting. De auteur kon via de interactie ambiguïteiten oplossen. De grammaticaliteit van de resulterende vertaling was in het systeem ingebakken: deze werd gegenereerd door de reversibele grammatica, een van de bijzondere kenmerken van het systeem. Overigens was interactieve vertaling een doel op de lange termijn: Landsbergen ging er van uit dat in eerste instantie gewerkt zou worden aan een applicatie die vertalers zou helpen bij hun werk, een verbeterd elektronisch woordenboek, dat niet alleen van woordjes, maar ook van grotere stukjes tekst alle mogelijke vertalingen zou kunnen geven.

Joep Rous implementeerde een M-parser, Jan Landsbergen schreef grammaticaregels, en samen leverden ze binnen een jaar een werkend mini-systeem op: Rosetta-1. Het vertaalt zinnetjes van en uit het Engels, Nederlands en Italiaans. "Waarom in godsnaam drie talen?", vraagt Landsbergen zich nu hardop af. "Het is al moeilijk genoeg." En over het antwoord: "Ik weet het niet zeker. Nederlands en Engels, daar konden we zelf goed mee uit de voeten, maar daar was geen behoefte aan, want iedereen bij Philips kon zelf beter naar het Engels vertalen dan de computer. Dat sprak niet aan. Dus er moest iets exotischers bij." Bovendien moest de groep aantonen dat het concept uitbreidbaar was, op zijn minst naar de Germaanse en Romaanse talen. "Want het grote twistpunt was natuurlijk altijd: is het niet heel erg moeilijk om parallel voor verschillende talen die grammatica's te schrijven?" Landsbergen kiest voor het Italiaans als vertegenwoordiger van de Ro-

Zink Typografie Sheet 81 of 213 - Page 73 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



maanse talen. "Dat was een goede aanleiding voor mij om een beetje Italiaans te leren."

Het management van Philips was onder de indruk van de demonstratie van Rosetta-1 in oktober 1980, en Landsbergen mag doorgaan met zijn project. Maar zijn verzoek om een paar taalkundigen aan te mogen nemen voor het verder ontwikkelen van de grammatica's, wordt niet gehonoreerd. "Philips had geen tijdelijk personeel, dat was toen heel moeilijk, de vakbonden waren daar tegen", legt hij uit. "Maar taalkundigen in vaste dienst was ook lastig, op een lab waar een wiskundige al een vreemde eend was. Het waren allemaal natuurkundigen en scheikundigen bij Philips." En dus begint Landsbergen zelf aan de opvolger van het eerste demosysteem: Rosetta-2. Via Jan van Bakel krijgt hij toch een tijdelijke kracht, Ellen van Loevezijn, die in een half jaar een set morfologische regels voor het Italiaans schrijft. In dienst van de universiteit, maar betaald door Philips, een constructie die later nog vaak ingezet zou worden. Bovendien heeft Landsbergen goede contacten met de Eurotragroep in Utrecht, geleid door Steven Krauwer en Louis des Tombe. Via hen komen er vele stagiairs naar Eindhoven, waaronder Lisette Appelo en André Schenk. Beiden volgden de cursus computerlinguïstiek van Krauwer en Des Tombe, Appelo bij haar hoofdvak Spaans, Schenk bij zijn hoofdvak Engels. In deze periode van het Rosetta-project worden kleine uitbreidingen en verbeteringen op het systeem gerealiseerd, maar met enkel stagiairs was de slagkracht beperkt. "Het is met een stage altijd zo, dat het grootste gedeelte van de tijd opgaat aan het uitleggen hoe alles werkt en dat er dan nog een klein beetje tijd over blijft om iets nuttigs te doen. Dat weegt eigenlijk niet tegen elkaar op," aldus Landsbergen. Maar deze periode is volgens hem wel heel belangrijk voor de beeldvorming bij de Philipsleiding. "Het stelde inhoudelijk nog niet zoveel voor, maar gaf de omgeving van het Natlab, de directie, de indruk dat er met onze groep te werken viel. Dat het wat opleverde, dat ik dat aankon, dat ik kon delegeren. Ik was natuurlijk geen manager." Uiteindelijk krijgt Landsbergen in 1983 toch de linguisten waar hij al een tijd om vroeg: Philips besluit een promotieplek bij de Universiteit Utrecht te financieren, op voorwaarde dat er in Eindhoven en onder Landsbergen's supervisie aan Rosetta gewerkt wordt. Krauwer en Des Tombe krijgen het op hun beurt voor elkaar dat de universiteit een tweede plek ter beschikking stelt. Appelo en Schenk

Zink Typografie Sheet 82 of 213 - Page 74 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



bemachtigen de twee plekken en zijn vanaf 1984 voltijds bezig met Rosetta.

Dan hoort Landsbergen over de contacten die BSO voorman Eckart Wintzen heeft bij het Ministerie van Economische Zaken, en die zouden kunnen leiden tot een grote subsidie voor het vertaalproject DLT. Hij besluit een alternatief voorstel in te dienen, samen met de Universiteit Utrecht en de Universiteit Leiden. Volgens dit plan zou elk van de partijen vijf man leveren. De groep van Krauwer zou voor de grammatica's zorgen en de groep van Bondi Sciarone zou voor de woordenboeken zorgen. Zoals eerder beschreven gaat de subsidie niet naar Rosetta, maar naar Toon Witkam en zijn DLT. Landsbergen over de reden: "Onderhandelingen over subsidies vonden vrij hoog in organisatie van Philips plaats, daar hadden wij helemaal geen vat op, dat was een apart clubje. [...] In diezelfde periode hebben die mensen met het ministerie van Economische Zaken afgesproken dat Philips gewoon ieder jaar, ik zeg maar wat, honderd miljoen gulden zou krijgen, zonder verder gezeur. Subsidie voor onderzoek. Philips spendeert een paar miljard, dus dat mag wel. Maar dan moest Philips verder niet zeuren over kleine projectjes als Rosetta. Dus ze hebben helemaal niet naar de inhoud gekeken. BSO was een leuk klein bedrijf, Philips kreeg al genoeg, dus ging dit niet door." Toch betekende dit niet het einde van Rosetta, en zelfs niet het einde van het uitbreidingsplan. De Philipsleiding oordeelde dat het eigenlijk wel goed zou zijn om wat grotere informaticaprojecten op te starten en besloot het project zelf te financieren, zij het in wat afgeslankte vorm: er was plek voor tien mensen, waarvan vijf extern. De externen zouden uit Utrecht komen, opnieuw in dienst van de universiteit, maar betaald door Philips. En, niet onbelangrijk, ze zouden werkzaam zijn op het Natlab. "Dat was iets wat ik zelf graag wou, omdat ik bij Eurotra, dat toen ook al liep, zag dat als je niet bij elkaar op de gang zat, dat dat grote ellende gaf in de samenwerking." Delft was niet langer betrokken. Dat kwam gedeeltelijk doordat de omvang van het project was afgenomen. Maar Landsbergen ziet ook een andere reden: "Daar kwam bij dat Sciarone niet enthousiast was over het idee dat zijn mensen in Eindhoven zouden zitten. Dat vond hij niet leuk, wat ik best kan begrijpen."

Halverwege 1984 begon Jan Landsbergen met het werven van personeel voor Rosetta-3. "Dat was nog helemaal niet gemakkelijk. Te weinig mensen wisten iets van computers. Ze hoefden geen specialist

Zink Typografie Sheet 83 of 213 - Page 75 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



te zijn, maar ze moesten wel formeel kunnen denken." Mede door de moeite die het kostte om geschikt personeel te vinden, switcht Landsbergen met lichte tegenzin van het Italiaans als derde taal naar het Spaans. "Het was vrijwel onmogelijk om mensen te vinden die Italiaans konden en die überhaupt wisten wat een computer was." En hij had al een spaansspreker in huis: Appelo had Spaans gestudeerd. Minstens even belangrijk was, dat de Philipsleiding Engels-Spaans een commercieel interessant talenpaar vond.

Landsbergen slaagt erin een groep te formeren. Geen standaard groep op het Natlab, met vijf informatici en maar liefst zeven taalkundigen. "En taalkundigen waren ook nog vaak vrouw, ook een merkwaardig fenomeen. Als je als vrouw daar binnenkwam, dan werd je als secretaresse gezien, door de portier enzo", herinnert Landsbergen zich. Een van die vrouwen was Franciska de Jong. Ook zij ervoer grote verschillen tussen Rosetta en de rest van het Natlab. "Wij, de taalkundigen, waren een bijzondere groep op Natlab. We waren we niet in dienst van Philips, we woonden geen van allen in Eindhoven, we kwamen allemaal met de trein op en neer. Bijna allemaal gingen we met de fiets van het station naar Natlab. Jan Odijk, Lisette Appelo, Harm Smit, Agnes Mijnhout, Elly van Munster."[9] De Jong vervolgt: "Wij hadden geen vaste aanstelling, waren dus officieel geen Philips medewerker, wat allerlei implicaties had: wij mochten niet overal aanwezig zijn en wij volgden bepaalde cursussen niet die anderen wel volgden. Wij waren een beetje een apart clubje. Bovendien waren wij een clubje met drie mannen en vier vrouwen en hadden we een andere cultuur. Wij zaten vaak bij elkaar aan tafel in de kantine en ik had altijd het gevoel dat bij ons harder en vaker gelachen werd, dat we luidruchtiger waren, dat er minder stiltes vielen. Ik geloof dat dat nooit opgehouden is, dat wij in veel opzichten anders waren."

De groep[10] werkt precies volgens Landbergen's plan zeer nauw samen. De Jong: "Het bijzondere was, we zaten met elkaar op de gang, werkten met elkaar aan dat systeem, dat was een *va et vien*, je kon

[9] André Schenk behoorde ook tot de groep taalkundigen bij Rosetta, maar hij kwam met de auto vanuit Geldrop.

[10] In het voorwoord van het boek *Compositional Translation* (Rosetta 1994) worden naast de reeds genoemde namen verder nog genoemd: Elena Pinillos Bartolomé, Petra de Wit, René Leermakers, Carel Fellinger, Jeroen Medema, Joep Rous, Margreet Sanders, Jan Stevens, Frank Uittenbogaard en twintig studenten.

Zink Typografie Sheet 84 of 213 - Page 76 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



alles bij elkaar vragen of checken. Toen we zagen hoe bepalend dat was voor hoe ver we kwamen, begrepen we ook hoe ondoenlijk het was zoiets als Eurotra te doen als je zo ver uit elkaar werkt." Het samenwerken vergde echter ook aanpassingen in de manier van werken. Waar Landsbergen eerst zijn regels met de hand schreef en vervolgens vertaalde in Pascal, moesten er nu compilers geschreven worden voor de grammatica's. Verder moest er softwarebeheer komen. Omdat goede software hiervoor nog niet bestond, werd die geschreven door de informatici van Rosetta. Tot frustratie van Landsbergen: "dat was zonde van de tijd, want die hadden we ook voor andere dingen nodig. Dat schoot niet op."

De omstandigheden leidden tot een vrij grote koerswijziging. Het oorspronkelijke plan was om eerst Rosetta-2 uit te breiden met grotere woordenboeken en grammatica's. Op deze manier konden alle nieuwe teamleden ervaring opdoen met het formalisme, voordat fundamentele wijzigingen in het formalisme werden aangebracht, die tot een nieuw systeem, Rosetta-3, zouden leiden. Landsbergen: "Als we dat gedaan hadden, dan hadden we op dat moment het beste vertaalsysteem ter wereld gehad." Maar dat deden ze niet. In de eerste plaats was er een probleem met de woordenboeken. Doordat het team uit Delft was afgevallen, waren de taalkundigen in Eindhoven zelf verantwoordelijk voor de ontwikkeling van woordenboeken. Dit werd onder meer de verantwoordelijkheid van Harm Smit, die bij Hugo Brandt Corstius was afgestudeerd. De groep had contacten met Van Dale, dat net bezig was de vertaalwoordenboeken Nederlands-Engels en Engels-Nederlands te maken. Zij stelden tegen betaling de tape ter beschikking, maar die bleek niet zomaar toepasbaar. Landsbergen: "Er waren veel praktische problemen. Het was wel computertape, maar zat vol met inconsistenties. Het was moeilijk daar een database van te maken. Van Dale was erg enthousiast daarover, maar die onderschatte dat probleem ook volgens mij." Odijk beaamt dat: "Een triviaal voorbeeld: om uit te drukken of een werkwoord vervoegd wordt met *hebben* of *zijn* heb je drie mogelijkheden: *hebben*, *zijn*, of allebei. In Van Dale waren er 32 verschillende waardes om die drie opties aan te duiden. Gewoon omdat het een papieren woordenboek was. Dus je hebt wel of niet afgekorte hulpwerkwoorden, variatie tussen *of* en *en*, en typografische varianten." Bovendien, en dat was moeilijker oplosbaar, bevatte de woordenboeken niet alle syntactische informatie die een

Zink Typografie Sheet 85 of 213 - Page 77 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



wat grotere grammatica nodig heeft. Vaak ook bevatte de woordenboeken de informatie wel, maar was deze niet formeel gerepresenteerd of was er geen expliciete link tussen gerelateerde eigenschappen, zoals tussen subcategorisatiekenmerken en werkwoordsbetekenis. Duidelijk was dat traditionele woordenboeken niet geschikt waren voor natuurlijke-taalverwerking door de computer. Voor Jan Odijk reden om in later jaren deel te nemen aan verschillende projecten die tot doel hadden om elektronische lexica te ontwikkelen.

De korte-termijnoplossing voor Rosetta was: twee verschillende woordenboeken. Landsbergen: "Eén woordenboek was zelfgemaakt en klein, maar heel precies. Daarnaast hadden we Van Dale, met zo'n 70.000 woorden. Dat leidde tot moeilijke keuzes bij demonstraties: gebruikte je het kleine woordenboek, dan gebruiken mensen altijd woorden die niet in het kleine woordenboekje staan, gebruikte je Van Dale, dan wist je bij God niet hoe het zou aflopen, dan konden er de meest verschrikkelijke dingen gebeuren. Dat is een zwak onderdeel van het geheel gebleven." Dat zou Landsbergen heel anders aanpakken als hij het over kon doen: "Nu zou ik gewoon een slaaf inhuren. Iemand een baan geven, een paar natuurlijk, die het woordenboek heel precies vullen. [...] In Japan deden ze dat wel. Daar zaten meisjes vlijtig te typen. Maar hier nam niemand zulke mensen in dienst, en niemand hier wilde dat doen. Jammer, want je kan nog zulke mooie grammatica's maken, uiteindelijk zijn de woorden toch nog belangrijker."

Een tweede omstandigheid was dat de informatici nodig waren om randsoftware te ontwikkelen, en zolang die er niet was, de taalkundigen niet aan de uitgebreide grammatica's konden werken. Hierdoor kon de geplande groei van woordenboeken en grammatica's van Rosetta-2 niet plaatsvinden zoals gepland. "En het lastige is dan", aldus Landsbergen, "dat de linguïsten allemaal fantastische nieuwe ideeën gingen ontwikkelen, waardoor ook de structuur van Rosetta-2 veranderde, ingewikkelder werd. Waardoor op een gegeven moment die planning is omgegooid. Dan gaan we eerst maar het ontwerp voor Rosetta-3 maken, dat heel anders is dan Rosetta-2, een veel geavanceerder systeem, mooier en ingewikkelder. Dan gaan we dat daarna wel echt vullen met regels en een toepassing zoeken." Een belangrijke wijziging in het formalisme wordt doorgevoerd. Landsbergen: "Er kwam een onderscheid tussen enerzijds regels die inhoudelijk

Zink Typografie Sheet 86 of 213 - Page 78 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



iets veranderden, dus echt iets betekenden, die in beide talen parallel moesten lopen en anderzijds transformaties die je tussendoor kon toepassen, die de betekenis niet veranderen, en die je per taal vrij kon kiezen. Daardoor kreeg je meer vrijheid per taal. Dat was wel een belangrijke stap." Met deze koerswijziging wordt het mogelijk om steeds moeilijker taalkundige problemen aan te pakken. En die kans grijpen de linguïsten aan. Odijk: "Wij keken erop neer om alleen maar kleine phrasen te doen. Dat was development. Wij waren research, dus wij deden de moeilijke dingen. Dat was de attitude. En dat had in zich dat het net als bij Eurotra, wel kon leiden tot een leuk laboratoriumsysteem, maar niet iets waar je een echt product mee kon maken." Met vereende krachten implementeren de taalkundigen oplossingen voor werkwoordsclusters, *er*-constructies en de beroemde *graag-like to* alternantie. In 1988 is Rosetta-3 voor Nederlands en Engels klaar. Een werkend systeem, met bijzondere kenmerken. Er zat geavanceerde taalkunde in. "Als er maar *er* in voorkwam dan ging het wel goed", aldus Landsbergen. Waar Odijk tegenover stelt: "Maar neem gewoon een zin uit de krant en hij doet het niet."

Deze eigenschappen van het systeem hadden consequenties voor de demonstraties die de groep van Natlab gaf. Aan de ene kant konden ze veel indruk maken met de ingewikkelde dingen die het systeem wél aankon. "Wij hadden fantastische demonstraties," vertelt Odijk. "We stopten de meest verschrikkelijke zinnen erin, we deden zelfs wedstrijdjes tussen de computer en de mensen, dan namen we een zin die nooit voorkomt, maar met acht werkwoorden achter elkaar in zo'n verb raising constructie, en dan lieten we de computer berekenen wat de vertaling was. Die deed er dan wel 5 minuten over, maar we lieten de mensen ook berekenen wat het eigenlijk betekende dan, en dat was zo moeilijk dat geen mens dat kan processen. Maar de computer kwam op een gegeven moment wel met iets terug, en de mensen konden niet eens controleren of het goed was." Aan de andere kant was het altijd gevaarlijk als mensen uit het publiek een zin voorstelden. "Je moest heel goed nadenken of alle woorden er wel inzitten, alle constructies afgedekt zijn. De enige die dat op een gegeven moment nog wist, was ik. Dus ik moest altijd demonstreren of bij de demonstratie zijn."

Langzamerhand groeit het systeem zo hard, dat Jan Landsbergen niet langer alles meer weet van alle onderdelen. En dat schuurt. Lands-

Zink Typografie Sheet 87 of 213 - Page 79 of 205

bergen: "Ik was intussen ook groepsleider geworden, dus er zat nog een andere activiteit in mijn groep, en ik was één dag per week hoogleraar in Utrecht. Ik had het gewoon erg druk. Het waren allemaal voortreffelijke mensen die voortreffelijke dingen deden, maar als er dan iets niet zo liep zoals ik zou willen, dan kon ik niet meer precies zeggen waaraan het lag. Dat vond ik jammer." Hij moet zich bovendien steeds meer bezig gaan houden met mogelijke commerciële toepassingen van Rosetta. Volgens het plan zou in de laatste fase van het project, dat tot 1991 zou lopen, Rosetta-4 ontwikkeld worden, een systeem voor een concrete applicatie. Wat die toepassing precies zou worden, zou in de loop van het project besloten worden. Vanaf 1987 groeit echter al de druk van de Philipsleiding om op zoek te gaan naar concrete toepassingen. Landsbergen is hetzelfde lot beschoren als Toon Witkam van DLT: hij moet zijn project zien te verkopen aan een geïnteresseerde partij. Hij praat met allerlei afdelingen binnen Philips. Even leek het erop dat vertaalsoftware onderdeel zou gaan uitmaken van het Work Station of the Future (WSF), dat computers voor het midden- en kleinbedrijf wilde ontwikkelen. Ook de elektronische schrijfmachine is onderzocht als mogelijke toepassing. Woordenboeken vormden in deze verkenningen een bottleneck: de ontwikkeling ervan zou veel tijd en geld kosten.

Deze meer commerciële inslag leidt ertoe dat het taalkundig team voor het eerst naar echte data kijken, in plaats van naar de constructies die in de grammaticaboeken beschreven staan. "Gedreven door die business waren we bezig met een vertaalsysteem voor reserveringen", vertelt Odijk. "Mensen wilden een brief schrijven en naar het buitenland versturen, maar die moest dan in die taal aankomen. Die brieven wilden we automatisch vertalen. Dus we namen een corpus van die brieven, en dan keken we wat we daarmee konden. Nou bijna niets. Geen enkele zin kwam er doorheen. Dus we gingen het woordenboek uitbreiden, en constructies toevoegen. Ik weet nog wel de constructie "gelieve...". Van die dingen waar je als theoretisch taalkundige nooit tegenaan loopt en niet in geïnteresseerd bent, waarvan je nu ineens moet bedenken hoe je die gaat analyseren in het Nederlands, en hoe je dat gaat vertalen." Nogal een verandering voor het team: "Wij keken puur vanuit de theorie, we keken nooit naar data. Dat deed je toen niet."

Plannen worden geschreven om via spelling- en grammaticachec-

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



kers uiteindelijk ook vertaalsoftware te ontwikkelen voor de schrijfmachine, maar uiteindelijk wordt besloten helemaal geen software voor het apparaat te maken. Het plan om te integreren met CD-I valt af omdat het apparaat uiteindelijk geen keyboard krijgt. Met Philips gaat het ondertussen slecht, en de contracten met de Universiteit van Utrecht lopen af. Het project wordt niet verder verlengd, er komt geen Rosetta-4. "En degene die het uiteindelijk afgeschoten heeft", vertelt Odijk, "was ene Gaston Bastiaans. Beroemd om drie dingen: hij is de grote man achter de compact disc, hij heeft Rosetta afgeschoten, en hij is CEO van Lernout en Hauspie geweest."

Van Rosetta rest haast niets meer. Toen het project op zijn eind liep zijn de grammaticaregels nog wel gebrand op cd – toen een gloednieuwe technologie. Maar het is niet meer demonstrabel. Landsbergen: "Het is in Pascal geschreven, wat op zichzelf een heel nette taal is, maar het is dan wel Pascal voor het operating system VHS, misschien later weer voor Unix, ik weet niet precies, maar we hadden dingen nodig die niet in de basis-Pascal zaten. Het woordenboek werkte dan weer met databasesysteem dat ook weer vrij exclusief was. Het is moeilijk weer aan de gang te brengen." De enige die er misschien nog iets mee zou kunnen, denkt hij, is Jan Odijk. En die heeft inderdaad in het kader van een veel recenter onderzoeksproject geprobeerd om Rosetta opnieuw tot leven te wekken. Hij liep daarbij aan tegen het feit dat software niet het eeuwige leven heeft. De versie van Pascal waarvoor Rosetta geschreven was, bestaat niet meer, waardoor de files omgezet moesten worden naar een nieuwe versie. Ook de commerciele software voor het indexeren van de lexica bestaat niet meer. "Mijn inschatting is dat een reanimatie met nog maar beperkte inspanning mogelijk zou moeten zijn", aldus Odijk – dit in tegenstelling tot de inschatting van taaltechnologiebedrijf Gridline, dat de inspanning op één manjaar schatte. Vooralsnog rest er in ieder geval van Rosetta niets meer dan enkele videofragmenten van demo's. Landsbergen: "Wat dat betreft is er een trieste analogie met PHLIQA, waar iets soortgelijks gebeurd is."

Het is niet zo dat Rosetta niets opgeleverd heeft. Alleen is de opbrengst meer indirect. Lisette Appelo, André Schenk en Jan Odijk zijn gepromoveerd op hun werk bij Philips. En zelfs Louis des Tombe van het Eurotra-project geeft aan dat Rosetta invloed heeft gehad op zijn (nooit afgeronde) promotieonderzoek: "Mijn dissertatie is redelijk

Zink Typografie Sheet 89 of 213 - Page 81 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



sterk op Rosetta geïnspireerd. Nee, nee, niet op EUROTRA", aldus de computerlinguïst uit Utrecht. "Wat ik feitelijk zit te doen, is een soort gecompliceerd model opbouwen waar Rosetta-achtige elementen inzitten: derivatiebomen die isomorf kunnen zijn, met de mogelijkheid om kleine betekenisverschillen tussen expressies of basisexpressies te accepteren." Op aandringen van de Amsterdamse semanticus Theo Janssen, een groot fan van Rosetta, heeft het team een boek geschreven over het project (Rosetta 1994). In *Compositional translation* staat het vertaalsysteem in detail beschreven. Hoewel Landsbergen klaagt over het feit dat de uitgever het erg duur gemaakt heeft, en het niet leesbaar genoeg is om aan studenten te geven, is het wel een compleet overzicht van het werk en een belangrijk onderdeel van de erfenis van Rosetta. Een ander belangrijk onderdeel van die erfenis zit in de mensen die op het project zijn opgeleid. De projectleider schat dat hij wel dertig tot veertig stagiairs en afstudeerders heeft gehad. "Dat was misschien wel een van onze nuttigste activiteiten", zegt hij nu, ondanks dat ze aan het systeem niet altijd even veel bijdroegen. En de uitstraling van Rosetta was groter dan alleen naar mensen die direct voor het project werkten. Alice Dijkstra herinnert zich de zomercolleges van Landsbergen in Utrecht over het Rosettasysteem. "Dat vond ik zó mooi. Dat was een reden om door te gaan met het vak. De schoonheid ervan. Of het werkte was een tweede, maar mooi was het wel."

Als de contracten via de Universiteit Utrecht niet verlengd worden, moeten de meeste taalkundigen op zoek naar ander werk. Franciska de Jong wordt hoogleraar aan de universiteit Twente. Odijk en Appelo waren inmiddels in vaste dienst gekomen van Philips en worden samen met de informatici en Jan Landsbergen overgeplaatst naar het Instituut voor Perceptieonderzoek (IPO) in Eindhoven, dat een samenwerking was tussen Philips en de Technische Universiteit Eindhoven. Hier werd zowel onderzoek gedaan naar beeld als naar spraak, en user interfaces waren een belangrijk onderdeel daarvan. Binnen het IPO werd een nieuwe taalgroep geformeerd, waarin de Rosettagroep terechtkwam. Voor Landsbergen een onaangename overgang: "Het IPO heeft een andere cultuur. Op het Natlab werken doeners, het IPO is – een beetje gechargeerd – een academisch clubje, met veel meer nadruk op publiceren dan op bouwen." Terugkijkend op zijn tijd bij Rosetta nuanceert Jan Odijk de praktische inslag van het Natlab: "Philips Natlab, dat was research. Het was wel industrie, maar het leek

Zink Typografie Sheet 90 of 213 - Page 82 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



eigenlijk meer een academische omgeving in die tijd. Pas vanaf 1990 toen Philips bijna failliet ging en die verplichting kwam om 70% van je research door de business te laten ondersteunen, is het veel zakelijker geworden en was het geen academische omgeving meer, waar je aan *Spielerei* kon doen. Daarvoor nog wel. Dat was een fantastische tijd, hoor. Ik ben zo blij dat ik dat meegemaakt heb. Maar je moet niet verwachten dat er dan een werkend vertaalsysteem uitkomt."

## 2.6 Computational Linguistics in the Netherlands

### Computerlinguïstiek verankerd in Academia

In de jaren tachtig leek het wel alsof computerlinguïstiek in Nederland en Vlaanderen gelijk stond aan automatisch vertalen, met vier grote vertaalprojecten. En in Nijmegen vond eind jaren tachtig en begin jaren negentig nòg een vertaalproject plaats, zij het op wat kleinere schaal: onder Jan van Bakel promoveert Albert Stoop op TRANSIT, een automatisch-vertaalsysteem gebaseerd op de AMAZON-parser van Van Bakel. Het systeem is bedoeld om vanuit het Nederlands naar het Turks te vertalen. Jan Landsbergen, die zelf na zijn emeritaat jarenlang Nederlands als tweede taal onderwezen heeft: "Dat was uit de tijd dat we dachten: er wonen zo veel Turken hier, wij moeten Turks leren."

De totstandkoming van dat project had nogal wat voeten in de aarde. Het project was ingediend bij de Nederlandse Organisatie voor Zuiver-Wetenschappelijk Onderzoek (ZWO) via de Stuurgroep Computerlinguïstiek. Deze stuurgroep was in 1981 opgericht. ZWO was destijds ingedeeld in een aantal stichtingen, waaronder de Stichting Taalwetenschap. Elke stichting bestond dan weer uit een aantal werkgemeenschappen, die elk een bepaald subdomein van de stichting afdekten. Een stuurgroep was een voorbereidende stap op weg naar een werkgemeenschap. De oprichting van de Stuurgroep Computerlinguïstiek was dan ook een duidelijk signaal dat het vak een gerespecteerd, zelfstandig onderdeel was van de taalwetenschap. De stuurgroep werd voorgezeten door Jan van Bakel, en verder hadden de toegepast taalkundige Bondi Sciarone en de wiskundigen Hugo

Zink Typografie Sheet 91 of 213 - Page 83 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Brandt Corstius en Jan Landsbergen zitting in het overlegorgaan. Via de stuurgroep konden aanvragen voor subsidie gedaan worden. Voorzitter Van Bakel werd boos, toe de Stichting Taalwetenschap besloot het TRANSIT-project niet te subsidiëren, terwijl een project dat door de stuurgroep als minder was beoordeeld wel financiering kreeg. "En toen heb ik op een vergadering van ZWO een toespraakje gehouden, dat ik op papier had staan. En toen heb ik gezegd: "ik treed af als voorzitter van het bestuur van de werkgemeenschap computerlinguistiek [in oprichting] op grond van het feit dat deze wetenschappelijke bevoegdheden door het bestuur van de Stichting Taalwetenschap worden overgenomen"." Van Bakel herinnert zich dat de rest van de stuurgroep toen ook terugtrad.[11] In 1985 werd alsnog de Werkgemeenschap Computerlinguïstiek en Mathematische Linguïstiek opgericht, onder voorzitterschap van Jan Aarts, die ook uit Nijmegen kwam, maar een corpusgerichte koers voer waar Van Bakel niets van wilde weten. Ook het Nederlands-Turkse vertaalproject kwam er uiteindelijk toch: het werd overgeheveld naar de Werkgroep Informatica, en alsnog gehonoreerd. In 1995 promoveerde Albert Troost op *TRANSIT: a linguistically motivated Dutch-Turkish MT-system* (Stoop 1995).

De oprichting van stuurgroep en later de werkgemeenschap geven aan dat computerlinguïstiek een eigen plek veroverd had binnen de Nederlandse onderzoekswereld. Steeds meer wetenschappers in steeds meer groepen hielden zich bezig met taaltechnologie, en de stuurgroep vormde het overlegorgaan waarin zij zich organiseerden. Een aantal van de taaltechnologiegroepen hing heel nauw samen met de golf van vertaalprojecten: in Utrecht hadden Steven Krauwer en Louis des Tombe een heel grote groep opgebouwd rondom Eurotra, en had Witkam de groep rond DLT, in Leuven zaten zowel medewerkers van Eurotra als van METAL en in Eindhoven zat de groep van Rosetta op het Natlab. Maar er gebeurde meer: op verschillende plekken in Nederland ontstonden leerstoelen op het gebied van taaltechnologie.

Harry Bunt was in 1983 van het IPO naar de Katholieke Universiteit Tilburg verhuisd. Twee jaar eerder had die universiteit een Letterenfaculteit opgericht, en er was ruimte voor een paar nieuwe

[11] Sciarone, Landsbergen en Brandt Corstius herinneren zich ofwel heel de stuurgroep niet meer, ofwel slechts een weinig succesvolle onderneming, zonder de precieze gang van zaken te kunnen terugroepen.

Zink Typografie Sheet 92 of 213 - Page 84 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



leerstoelen, waaronder een voor Taal en Informatica. Bunt had vanuit het IPO zitting in de Commissie die de plannen voor deze leerstoel moest uitwerken. Het was namelijk de bedoeling dat de nieuwe vakgroep zo veel mogelijk moest gaan samenwerken met het IPO. Bunt: "In het IPO werd veel aan spraakonderzoek gedaan, dat moest niet gedupliceerd worden in Tilburg. Aan de andere kant zou dan in Tilburg meer taalkundig, inclusief taaltechnologisch, onderzoek gedaan worden, en dat moest dan weer niet gedupliceerd worden in Eindhoven." Vanuit zijn positie in die commissie had Bunt ruim de gelegenheid om zware accenten te leggen op semantiek en mens-computerinteractie. Volgens Bunt was Hugo Brandt Corstius de beoogde hoogleraar, maar toen die gepolst werd bleek dat hij niet geïnteresseerd was in de positie in Tilburg. "En dat kwam", aldus Bunt, "omdat Hugo helemaal niets moest hebben van het katholieke karakter van deze universiteit." Hoewel Bunt zich naar eigen zeggen nog wat te jong voelde voor een hoogleraarschap, heeft hij toen toch zelf gesolliciteerd. "Ik kan mij herinneren dat ik op zeker ogenblik tegen Simon Dik zei: ik voel er niets voor om te solliciteren, ik voel mij veel te jong hiervoor. Simon Dik zei: 'dat kan ik mij voorstellen, maar als Hugo Brandt Corstius niet wil, noem dan eens een andere Nederlandse kandidaat die jij denkt dat beter zou zijn.' Die kon ik toch eerlijk gezegd niet noemen." Volgens Remko Scha was bij voorbaat al besloten dat Bunt het moest worden. In zijn ogen zinspeelde Hugo Brandt Corstius daar zelfs al op nog voordat de sollicitatieprocedure gestart was. Hierdoor geïrriteerd besluit Scha om zelf ook te solliciteren. De twee oud-collega's blijven samen over als laatste kandidaten. Maar voorbestemd of niet, Bunt krijgt de leerstoel.

Scha vertrekt naar de Verenigde Staten, waar hij bij Bolt, Beranek and Newman (BBN) aan de slag gaat. Hier deed hij in eerste instantie op PHLIQA geïnspireerd werk, paste hij toe wat hij in Eindhoven ontwikkeld had. Hiermee kon hij snel vooruitgang boeken. Maar toen hij begon aan wat later Data Oriented Parsing (DOP) zou gaan heten, de door Scha ontwikkelde parseertheorie, botste dat met de druk om zeer snel demonstrabele systemen te ontwikkelen. Dit had te maken met de financierder van het onderzoek bij BBN, de Amerikaanse defensieorganisatie DARPA, die bedrijven en kennisinstellingen subsidieerde om een kandidaatsysteem te ontwikkelen voor de jaarlijkse wedstrijden op het gebied van onder meer automatisch vertalen en

Zink Typografie Sheet 93 of 213 - Page 85 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



spraakherkenning. "Maar als ik binnen 1 jaar de wedstrijd wil winnen", vertelt Scha, "dan hebben de cynische ingenieurs gelijk, die zeggen doe maar PCFGs [Probabilistic Context Free Grammars] in combinatie met Markovmodellen. Want mijn visie dat je moet parseren met boompjes, dat is niet iets waarvan je kan plannen dat je dat binnen één jaar rond hebt, dat je dat geïmplementeerd hebt zodanig dat je daar ook nog wedstrijden mee wint." Een academische positie begint steeds aantrekkelijker te klinken, en in 1988 wordt Scha hoogleraar aan de Universiteit van Amsterdam.

## CLIN

Terwijl de officiële vertegenwoordiging aan de universiteiten in de vorm van leerstoelen langzaam gestaag groeit, ontmoet de uitdijende onderzoeksgemeenschap elkaar op congressen van aanpalende vakgebieden. Op de jaarlijkse TIN-dag (Taalkunde in Nederland) bijvoorbeeld. Of op de congressen van de formele logici in Amsterdam en Groningen. In 1989 organiseren Frans Zwarts en Jan de Vuijst in Groningen een zomerschool met de naam *European Summer School in Logic, Language and Information* (ESSLLI). Zo'n 350 studenten komen naar Groningen voor trainingen op het snijvlak van de verschillende disciplines. Ruim twintig jaar later is ESSLLI hèt event voor iedereen die zich bezighoudt op het snijvlak van logica, taal en informatica. Zwarts had destijds niet durven dromen dat hij een lange traditie startte. "We hadden een subsidie voor Intensive Programmes, onderdeel van het Erasmus programma", vertelt hij. "Ik ben in Edinburg gaan informeren of er interesse was, en had er al in Amsterdam over gesproken. Iedereen was enthousiast, dus toen zijn we begonnen." En Zwarts had zo zijn eigen redenen voor de organisatie van een dergelijk groot event. De alfa-informaticagroep in Groningen, waar taaltechnologie onderdeel van was, bestond uit een klein groepje en trok ook niet veel studenten. "Het was te moeilijk voor Letterenstudenten", verklaart Zwarts dit gegeven. "En de beste manier om te verhinderen dat de groep weer opgeheven zou worden, is dit event te organiseren. Want dan trekken we veel aandacht, en dat zal de faculteit zeker leuk vinden."

Internationaal ontmoetten computerlinguïsten elkaar op de bij-

Zink Typografie Sheet 94 of 213 - Page 86 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



eenkomsten van de van oorsprong Amerikaanse *Association for Computational Linguistics* (ACL). Tijdens de achtentwintigste ACL-bijeenkomst in Pittsburgh besluit de Eurotragroep uit Utrecht dat het tijd is voor een eigen, Nederlandse bijeenkomst voor computerlinguisten. Was het Steven Krauwer, die in het vliegtuig het idee opperde? Of toch Lisette Appelo, die in de Cathedral of Learning een goede ingeving kreeg? Dat blijft onderwerp van discussie. Maar feit is dat op 26 oktober 1990 gastspreker Mark Johnson de allereerste editie opende van CLIN, *Computational Linguistics in the Netherlands*. De organisatie was in handen van Dirk Heylen, Heleen Hoekstra en Gertjan van Noord, allen uit de Eurotragroep van Krauwer en des Tombe. Het actief promoten van de bijeenkomst door Van Noord heeft succes: 23 presentaties en twee demo's (Rosetta en MiMoII) vullen het programma. De affiliaties van de sprekers geven een goed beeld van de verspreiding van de computerlinguïstiek over heel Nederland: naast de al genoemde vertaalgroepen en groepen van nieuwe hoogleraren, waren er ook presentaties uit Leiden, Nijmegen, Twente, het IPO en Groningen.

Vanuit Leiden was er de informaticus Alexander Ollongren, die een presentatie gaf over de $\lambda$-calculus in Rosetta. Van een taaltechnologische groep was in Leiden echter geen sprake, dat zou pas starten met de komst van psycholinguïst Gerard Kempen in 1992. In 1990 zat die nog in Nijmegen, waar twee groepen geheel afzonderlijk van elkaar onderzoek deden op het gebied van automatische taalverwerking. Enerzijds waren daar de taalkundigen, bestaande uit een groep rond Jan Aarts met zijn corpusgebaseerde werk en een groep rond Jan van Bakel met zijn grammaticawerk. Op die eerste CLIN waren er maar liefst vier presentaties uit deze hoek, waaronder één van Peter-Arno Coppen, de opvolger van Van Bakel. Anderzijds was daar Kempen, die als experimenteel psycholoog begonnen was, en via geheugenmodellen en het zogenaamde *memory for sentences* terecht gekomen was in de taalproductie. Hij had namelijk ontdekt dat het reproduceren van zinnen beter gemodelleerd kan worden als een (re)generatieproces dan als het ophalen van kant-en-klare zinnen uit het geheugen. De groep van Kempen had helemaal geen contact met de groep van Van Bakel. Walter Daelemans promoveerde midden jaren tachtig bij Kempen. "Op een of andere manier was er geen samenwerking tussen het psychologisch lab van Gerard Kempen en de

Zink Typografie Sheet 95 of 213 - Page 87 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



computerlinguïsten aan de andere kant van de campus. Die waren toen ook wel heel actief, Van Bakel enzo, maar dat werd toch als iets anders beschouwd. Wij deden meer AI, en zij deden meer formele taalkunde, of toegepaste taalkunde. [...] Jan van Bakel zat er wel, maar ik heb hem nooit gezien in de drie jaar dat ik daar rondgelopen heb." De groep van Kempen deed het goed, de psycholinguïst keek buiten de grenzen van zijn eigen vak en dat had effect. "Hij combineerde psycholinguïstiek, AI, computerlinguïstiek, taaltechnologie, spraak", aldus Daelemans. En op die manier bouwde hij ook zijn onderzoeksgroep op. Alice Dijkstra maakte begin jaren negentig in Nijmegen en Leiden onderdeel uit van die onderzoeksgroep. "De kracht van Kempen was dat hij zag dat je allerlei mensen bij elkaar moest gooien, en als die maar een beetje met elkaar konden communiceren, dat je daar dan hele leuke dingen uit kreeg. Hij heeft toen zijn nek uitgestoken. Dat was toen niet gebruikelijk, om zulke verschillende mensen bij elkaar te zetten", aldus Dijkstra. Natuurlijk kan het gemakkelijk misgaan, met zo heel verschillende mensen bij elkaar. Maar ondanks dat de mensen niet alleen inhoudelijk, maar ook wat persoonlijkheid betreft erg van elkaar verschilden, wordt de sfeer in de groep van Kempen alom geroemd. De mensen vulden elkaar aan, en er waren zeer weinig conflicten. "Omdat alles uitgepraat en voorzichtig behandeld werd", denkt Daelemans. Die behoedzaamheid uitte zich ook in de manier waarop Kempen proefschriften begeleidde. Daelemans herinnerde zich dat Kempen moeite had met de titel *Studies in Language Technology*, omdat het te veel zou lijken op *Exercises in Computational Linguistics* van Brandt Corstius. En Theo Vosse heeft zelfs de achterkant van zijn proefschrift moeten afplakken met wit papier, omdat Kempen problemen verwachtte bij de verdediging. "Er was geen probleem met de tekst", haast Kempen zich te zeggen, "maar je mag niet in je proefschrift je eigen werk aanprijzen." En Theo Vosse had zijn proefschrift vormgegeven als een Duitse Krimi, inclusief de recensies op de achterflap. Iedereen begreep dat het satire was, maar Kempen wil het risico niet nemen, en dus plakken Theo Vosse en zijn paranimfen Alice Dijkstra en Stan van de Burgt een avond lang proefschriften af. De neiging van Kempen om alles helemaal netjes te willen doen, was overigens niet altijd even bevorderlijk voor de efficiëntie. Dijkstra: "Het kon altijd mooier en beter, zodat promovendi het nooit afkregen, tenzij ze het zelf afkapten."

Zink Typografie Sheet 96 of 213 - Page 88 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Op CLIN 1990 presenteerde een psycholinguïst uit Nijmegen, namelijk H. Wilfred Cambell, maar hij is eigenlijk geen goede vertegenwoordiger van de groep van Kempen. Cambells presentatie over de representatie van betekenis in een ongetypeerde lambdacalculus is een voorbeeld van het werk van de *neats* in Kunstmatige Intelligentie, terwijl Kempen nu juist tot het kamp van de *scruffies* behoorde. Daelemans: "Ik weet nog dat we een keer op bezoek waren bij Jan Landsbergen, toen had hij net PTQ van Montague in 20 exemplaren gekocht en uitgedeeld aan al de mensen van zijn groep. Dat was toen de bijbel, *The Proper Treatment of Quantification*. Kempen geloofde daar niet in, want formele systemen lossen ambiguïteit niet op, daar heb je context voor nodig en associaties. Toen had je al in Nederland die oppositie." Daelemans zelf is altijd een scruffy gebleven: na militaire dienst en een tijdelijke post in Brussel kwam hij in 1989 in dienst van de Universiteit Tilburg, waar hij een nieuwe onderzoeksgroep ILK (Inductie van Linguïstische Kennis) oprichtte, die zich helemaal richt op het toepassen van machine-learningtechnieken voor het oplossen van vraagstukken op het gebied van natuurlijke-taalverwerking. Kempen verruilt na ruim dertig jaar Nijmegen voor Leiden. Hij neemt een aantal mensen mee, waaronder Theo Vosse, Alice Dijkstra en Koenraad de Smedt. Maar al snel krijgt hij te maken met bezuinigingen en reorganisaties, en valt de groep een beetje uit elkaar. "Achteraf denk ik wel eens: 'Had ik wel weg moeten gaan?'." Vanaf 1999 is hij behalve aan de Universiteit Leiden ook verbonden aan het Max Planckinstituut in Nijmegen.

Jan Landsbergen en zijn groep zaten in 1990 nog op het Natlab bij Philips (op CLIN vertegenwoordigd met een presentatie door Jan Odijk en een demo van Rosetta door Odijk, De Jong en Appelo). De Rosetta-groep zou pas in 1991 naar het IPO verhuizen. Maar dat betekende niet dat het IPO daarvoor nog niet meedeed: Kees van Deemter, uit de *Cognition and Communication*-groep van het IPO, presenteerde op de eerste editie van CLIN een verhaal over de logica van ambiguïteit, een onderwerp dat in de daaropvolgende jaren een centraal thema in zijn werk zou blijven.

Tenslotte presenteerden op CLIN 1990 ook onderzoekers van twee instituten die we in dit verhaal tot nu toe nog niet of nauwelijks zijn tegengekomen: de Universiteit Twente (vertegenwoordigd door Anton Nijholt en Klaas Sikkel) en de Rijksuniversiteit Groningen (vertegen-

Zink Typografie Sheet 97 of 213 - Page 89 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



woordigd door Gosse Bouma). Anton Nijholt studeerde theoretische informatica aan de VU in Amsterdam en doceerde het aan de VU Brussel. Maar tijd en vrijheid om eigen onderzoek te doen, en een interesse in taal dreven Nijholt richting natuurlijke-taalverwerking. Net als de mannen van het eerste uur werkt Nijholt zichzelf in in de taalkunde, via literatuur en conferenties. Parsing is een natuurlijk ontmoetingspunt voor computerlinguïsten en informatici zoals Nijholt. Internationaal vindt hij aansluiting bij onder meer Martin Kay en Masaru Tomita. Vanuit de parsing volgt dan weer de interesse in taalkundig gemotiveerde grammaticaformalismen. Wannneer Nijholt in 1989 hoogleraar wordt aan de Universiteit Twente, is zijn opdracht opnieuw theoretische informatica. "Maar ik had het idee dat het allemaal niet zo erg zinvol was wat er in die theoretische informatica gebeurde. En ik had hier een groep die wel stond open voor verandering, voor vernieuwing. Dus toen zijn we langzamerhand meer en meer naar natuurlijke taalverwerking gegaan." Vlak na de eerste CLIN krijgt Nijholt de kans om zelfs een deeltijdhoogleraar aan te stellen in de Taaltechnologie en computerlinguïstiek. Nijholt: "Ik was decaan van de faculteit informatica en via het college van bestuur kreeg ik de gelegenheid om een vrouwelijke deeltijdhoogleraar aan te stellen." Via deze stimuleringsmaatregel kon Franciska de Jong direct na haar promotie in Enschede beginnen. Hiermee werd zij de eerste hoogleraar computerlinguïstiek binnen een informaticafaculteit in Nederland. Later zouden functies bij TNO en de Erasmusuniversiteit volgen.

Gosse Bouma van de Rijksuniversiteit Groningen presenteerde een verhaal over chart parsing voor categoriale unificatiegrammatica. Bouma was naar Groningen gehaald door professor Neerlandistiek Frans Zwarts, bij wie hij eerder al Nederlands gestudeerd had, en die zich hard had gemaakt voor een nieuwe afdeling met de naam alfa-informatica. Deze afdeling was in principe gewijd aan informatietechnologie voor de gehele Letteren, en was dan ook begonnen met zowel een historicus, George Welling, als een taalkundige, Jan de Vuijst. De komst van Gosse Bouma, die tot die tijd enigszins geïsoleerd had gezeten bij het Instituut voor Kennissystemen in Maastricht, betekende dat het profiel een meer taalkundige kleur kreeg. Maar Zwarts wilde meer. "Mijn idee was: als we als groep willen overleven, dan moeten we activiteiten ontplooien zoals de organisatie van ESSLLI, maar ook krachten van buiten halen die onze signatuur versterken." Hij

Zink Typografie Sheet 98 of 213 - Page 90 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



was dan ook blij dat het lukte om Gertjan van Noord, die inmiddels van Utrecht naar Saarbrücken was verhuisd, naar Groningen te halen. "Dat was fantastisch, want daarmee was de zaak beklonken." Alfa-informatica Groningen werd voornamelijk een taaltechnologiegroep. Dit werd nogmaals bevestigd in de keuze voor de taalkundige John Nerbonne als hoogleraar. De Amerikaan was in Ohio gepromoveerd in de taalkunde, en had zowel industriële als academische ervaring. Hij zat tegelijkertijd met Van Noord in Saarbrücken, kende Bouma vanuit workshops en congressen, en Zwarts vanuit ESSLLI, waar hij een cursus had gegeven. Frans: "De perfecte kandidaat." Toch was zijn benoeming niet onomstreden. De vertegenwoordigers van de andere stromingen binnen alfa-informatica (geschiedenis, fonetiek, letterkunde) die bij de groep betrokken waren, hadden heel andere ideeën over invulling van de leerstoel. Sommigen wilden een andere signatuur aan de groep geven, en sommigen wilden zelf die baan. Maar het werd Nerbonne. Frans: "Sommige mensen zeiden wel dat ik die taalkundige kant promootte omdat ik zelf taalkundige was, maar de waarheid is dat er geen duidelijke andere signatuur wàs. Alfa-informatica is geen op zichzelf staand vak. Computerlinguïstiek ontwikkelde zich in die tijd wel ineens heel snel tot een zelfstandig vakgebied. [...] Historicus George Welling is met zijn goede werk toch altijd een roepende in de woestijn gebleven." Formeel was er overigens nòg een hoogleraar in de groep. Nerbonne over de professor fonetiek Don Graham Stuart: "Een van de eerste zaken die ik deed was vragen of ik een begroting mocht zien. En daar stond het salaris op van een mij onbekend persoon. Ik zei: "dat is wel heel interessant, moet ik die man niet leren kennen?" Maar ik heb hem nooit leren kennen, want hij is nooit hier geweest."

De bijdragen van al deze instituten maakten van de eerste editie van CLIN een groot succes. De opkomst was hoog en gaf een goed beeld van de ontwikkeling van het vakgebied aan de verschillende instituten in Nederland. Aan het eind van de conferentie werd aangekondigd dat het CWI in Amsterdam het volgende jaar opnieuw een CLIN zou organiseren. Daarmee was een traditie geboren: in de twintig jaren die erop zouden volgen, was er ieder jaar een congres. Vanaf dat moment vormde CLIN hèt platvorm waar Nederlandse computationeel taalkundigen elkaar ontmoetten en hun meest recente werk presenteerden. Het congres koos bewust voor een laagdrempelige in-

Zink Typografie Sheet 99 of 213 - Page 91 of 205

stap zonder zware selectieprocedures, waardoor ook promovendi en afstudeerders hun werk konden presenteren, en meer ervaren onderzoekers hun lopend onderzoek konden laten zien, ook als daar nog geen eindresultaten van waren. Sprekers konden na de conferentie een artikel indienen voor de *conference proceedings*, die wèl uitgebreid gereviewd werden.[12] In 2010, twintig jaar na het eerste congres, is CLIN opnieuw in Utrecht. Voorzitter Linde van den Bosch van de Nederlandse Taalunie spreekt de zaal toe in het Nederlands, maar projecteert een Engelse vertaling op de televisieschermen. Ze rondt af met een wens: "Namelijk dat wanneer over tien jaar de Taalunie 40 wordt en CLIN 30, de correcte vertaling van mijn toespraak niet van tevoren hoeft te worden gemaakt, maar ter plekke door automatisch vertalen kan worden gegenereerd."

De grote afwezigen op de eerste CLIN waren de Vlamingen. En hoewel er vanaf de tweede editie ook ieder jaar ingezonden presentaties waren uit het buitenland, was Vlaanderen pas vanaf de zesde editie steevast vertegenwoordigd. Dat wil niet zeggen dat er niets gebeurde aan Vlaamse universiteiten. In Leuven zaten natuurlijk de Eurotra- en METAL-groepen. In Brussel zat Luc Steels, die inmiddels was teruggekeerd vanuit de Verenigde Staten en in 1983 in Brussel het Artificial Intelligence Laboratory had opgericht. De focus lag in het lab echter niet op taalverwerking. In Antwerpen was Willy Martin midden jaren tachtig komen lesgeven, nog steeds met een focus op kwantitatieve taalkunde en lexicografie. Vlak na CLIN versterkte ook Walter Daelemans het front in Antwerpen: vanaf begin jaren negentig verdeelde hij zijn tijd tussen Tilburg en Antwerpen. Vlaanderen had verder ook een eigen event: de tweede editie van ESSLLI vond in 1990 plaats in Leuven, georganiseerd door Frank Van Eynde, Ineke Schuurman en de Eurotra-staf. In 1995 organiseerden de Vlaamse taaltechnologen zich formeel in CLIF (Computational Linguistics in Flanders) op. In zekere zin is CLIF de Vlaamse tegenhanger van de Nederlandse werkgemeenschap, die bijna tien jaar eerder werd opgericht. Maar in tegenstelling tot die werkgemeenschap, die verdwenen is toen NWO (zoals ZWO later is gaan heten) het systeem met de

[12] Informatie over CLINdagen en proceedings is online beschikbaar op `www.clin.nl`

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



vele stichtingen verruilde voor een organisatie in veel bredere wetenschapsgebieden, bestaat CLIF nog steeds.

Utrecht zelf bleef een hotspot voor computerlinguïstiek. In 1993 organiseerde het Utrechtse Instituut voor Linguïstiek OTS de zesde editie van EACL, *the European chapter of the Association for Computational Linguistics*. Des Tombes partner, Susan Armstrong, zit op dat moment in het bestuur van de organisatie. Louis Des Tombe, Steven Krauwer en Michael Moortgat treden op als conference chairs en verschillende Eurotra-leden helpen mee in de organisatie, zoals Joke Dorrepaal en Anne-Marie Mineur. De conferentie heeft twee speciale thema's. Het eerste is de relatie tussen logica en computerlinguïstiek, en is karakteristiek voor de gemeenschap in Nederland, waar de rol van logici van Evert Willem Beth tot aan Theo Janssen altijd vrij groot is geweest. Het tweede thema is het gebruik van data-geörienteerde methoden in de automatische taalverwerking. Dit thema illustreerde de nieuwe beweging die vanaf 1988 op gang gekomen was binnen het vak, en die het vak radicaal van karakter zou doen veranderen.

## Rekenen neemt de leiding over van Taalkunde

Aan het eind van de jaren tachtig gebeurde er iets opvallends. Waar alle vertaalprojecten tot dit punt, niet alleen de Nederlandse en Vlaamse projecten die in dit hoofdstuk beschreven worden, maar ook die in de rest van de wereld[13] gebaseerd zijn op het coderen van taalkundige kennis, kwam een onderzoeksgroep van IBM met een totaal nieuwe aanpak. Ze experimenteerden met een systeem voor automatisch vertalen dat helemaal gebaseerd was op statistiek, in combinatie met een groot tweetalig, parallel corpus, namelijk de debatten uit het Canadese parlement, in het Frans en Engels. Het onderzoek werd voor het eerst gepresenteerd op COLING-88 (Brown et al. 1988), en kort daarna op verschillende andere conferenties. Overal zorgde het voor veel ophef. Niet zozeer omdat iedereen het zag als een enorme doorbraak, maar omdat het gezien werd als iets krankzinnigs. Steven Krauwer herinnert het zich nog goed: "In 1989 of 1990 was ik op het TMI congres. Peter Brown van IBM presenteerde daar *statistical machine transla-*

[13] Behalve wellicht het ontwerp voor de tweede fase van DLT.

Zink Typografie Sheet 101 of 213 - Page 93 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



*tion*. We lachten ons dood. We vonden het hilarisch! Dat kon toch niet waar zijn? Ik moest een rapport maken over die sessie, maar ik heb de opdracht teruggegeven, want dit kon ik alleen maar belachelijk maken."

Maar het was wel waar. De onderzoekers van IBM waren bloedserieus, en ze waren niet de enige die de mogelijkheden van datagedreven methoden onderzochten. In Japan werd ongeveer tegelijkertijd gewerkt aan wat later *example-based* vertaling is gaan heten, een methode die eveneens gebaseerd is op grote hoeveelheden parallelle corpusdata (Takeda, Nasukawa, Uramoto & Tsutsumi 1992). Vier jaar na de eerste presentatie van statistische automatische vertaling ging al bijna de helft van alle presentaties op TMI Montreal 1992 over corpus- of voorbeeldgebaseerde methoden, al dan niet gecombineerd met regelgebaseerde methodologieën. Het was niet voor het eerst dat statistische methoden werden voorgesteld voor automatisch vertalen. Weaver filosofeerde al in zijn oorspronkelijke memo (Weaver 1949) over de toepasbaarheid van informatietheoretische statistiek voor het vertaalprobleem. Door de slechte resultaten van de eerstegeneratiesystemen was het onderwerp echter van de agenda verdwenen. Eind jaren tachtig probeerden Amerikaanse onderzoekers het opnieuw, gemotiveerd door de goede resultaten die binnen spraakherkenning behaald werden met pure statistiek. "Met behulp van Hidden Markov Models (HMMs) werden op het gebied van spraak binnen vijf jaar dezelfde resultaten behaald als rule-based systemen in 15 jaar", herinnert Krauwer zich. Dat hielp enorm om de kracht van statistische methoden te onderkennen.

De data-driven aanpak werd bovendien enorm gestimuleerd door DARPA, dat grote subsidies verstrekte voor onderzoeksprojecten op basis van een competitie. Remko Scha maakte het vanuit BBN van dichtbij mee. "DARPA had het idee dat het computerlinguïstiek onderzoek dat zij subsidieerden niet opschoot. Mensen hadden mooie verhalen, op grond waarvan ze subsidies kregen, om weer leuke dingen te doen, en dan werden er interessante papers gepubliceerd. Er werden nieuwe representatietalen uitgevonden en discoursetheorieën. [...] Maar dat was niet wat DARPA wou. DARPA zag een soort proliferatie van leuke dingetjes terwijl zij eigenlijk liever resultaten wilden zien." Op het gebied van spraak had DARPA de wedstrijdgebaseerde aanpak geïntroduceerd, waarbij een competitie georganiseerd wordt

Zink Typografie Sheet 102 of 213 - Page 94 of 205

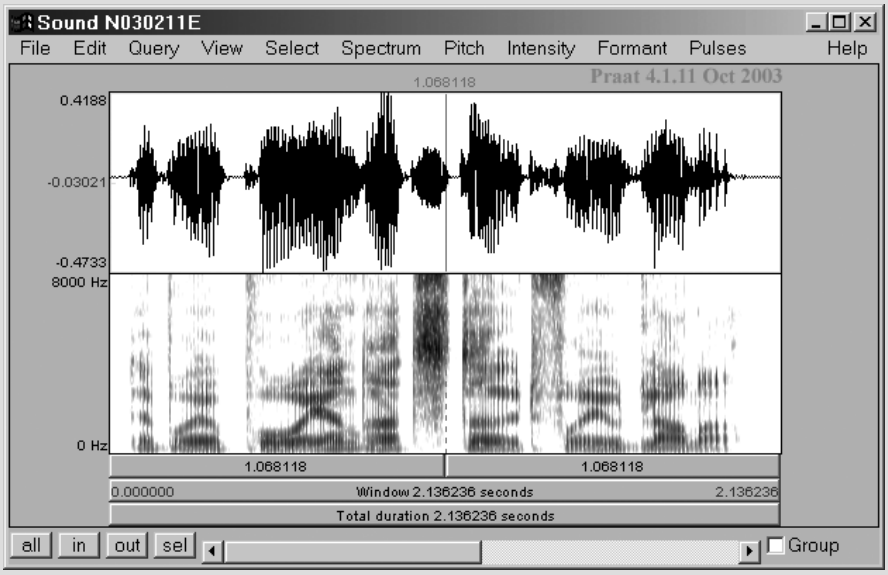
29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



voor een bepaalde taak, en de winnaar een forse subsidie ontvangt. Deze resultaatgerichte aanpak had veel succes gehad: Hidden Markov Models waren naar boven gekomen als de beste aanpak van dat moment, en iedereen was aan de slag gegaan om het HMM-model verder te verbeteren. DARPA hoopte eenzelfde succes te behalen op het gebied van taal. "Maar waar dat dus op neerkwam", aldus Scha, "was dat de taalmensen ook de kracht van de statistiek moesten leren kennen, en afgerekend moesten worden op wedstrijden." Om dat te bereiken, gaat DARPA *spoken language systems* organiseren. Scha vertelt dat BBN de eerste pilotsubsidie krijgt. Taal- en spraaktechnologie gaan voor het eerst samenwerken, en dat zal een grote impact hebben op de ontwikkeling van de taaltechnologie in het algemeen en automatisch vertalen in het bijzonder.

Walter Daelemans constateert dat die ontwikkeling, ingezet eind jaren tachtig, ons anno 2009 veel gebracht heeft: "Kijk naar Google Translate, in vijftig talen. Misschien is het voor de meeste talen nog steeds op het niveau van METAL en Systran, maar misschien voor sommige taalparen al wel beter. Het argument van Systran is altijd dat zij kunnen lokaliseren naar specifieke domeinen, maar dat kan met Google Translate ook, als je daar werk van maakt, door bijvoorbeeld feedback te verwerken in het vertaalproces, en door lexica toe te voegen. Dat kan je in een statistische aanpak ook doen. En nu heb je Hindi-Nederlands enzo, dat zie ik met technologie van twintig jaar geleden niet gebeuren. Je zou Google translate elk jaar moeten evalueren, om de verbetering door betere corpora te meten. Je ziet ook niet-letterlijke vertalingen, parafrasen, dat kan bij de statistische methode. De vraag is: topt het af net iets beter dan Systran, of wordt het beter met meer data? De volgende vijf jaar is heel cruciaal voor het vakgebied. Als het aftopt, dat zitten we weer terug, in een iets hogere boom dan twintig jaar geleden, maar nog altijd heel ver van de maan af. Maar ik denk van niet. Het bereikt ooit het niveau van vertalers. Mensen kunnen ook niet zo heel goed vertalen, die maken ook fouten. Over tien jaar hebben we *human level machine translation*."

Zink Typografie Sheet 103 of 213 - Page 95 of 205




Spraaktechnologie van eigen bodem: het softwarepakket Praat.

# 3 Spraak

## 3.1 Dinosaurussen

De geschiedenis van de spraaktechnologie gaat veel verder terug dan die van de computer. In de tweede helft van de achttiende eeuw bouwde Wolfgang von Kempelen de eerste spraakmachine. De machine, die met twee handen bediend werd en waar met behulp van een blaasbalg lucht in gepompt werd, kon woordjes en zelfs korte zinnetjes produceren, al vergde bediening wel enige training. De mechanische stem is bewaard gebleven en doet het nog steeds.[1] De opkomst van de elektrotechniek bood nieuwe kansen voor de spraaktechnologie: in 1939 presenteerde Homer Dudley op de Wereldtentoonstelling in New York zijn VODER, een handmatig te bedienen spraaksynthesizer. Van deze VODER zijn de oudste opnames van gesynthetiseerde spraak afkomstig.[2]

Wie zich met spraak bezig hield, moest zelf de hulpmiddelen bouwen om het geluidssignaal te analyseren of te genereren, want er waren geen kant-en-klare machines, laat staan softwarepakketten. Het is

[1] Zie `http://www.ling.su.se/staff/hartmut/kemplne.htm`
[2] Zie (hoor) `http://www.ling.su.se/staff/hartmut/ljud/voder.wav`.

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



dan ook niet verwonderlijk, dat de meeste pioniers op het gebied van de spraaktechnologie technische wetenschappers waren, te beginnen met de fysioloog Von Kempelen, en de elektrotechnicus Dudley. Hoewel het spraakgeluid gebruikt wordt om talige boodschappen over te brengen, en spraaktechnologie nauw verbonden is met de fonetiek en de fonologie, waren het bij de grote laboratoria, zoals Bell Labs in New Jersey in de Verenigde Staten, toch geen taalkundigen die spraakonderzoek deden, maar natuurkundigen en elektrotechnici.

Alleen in Nederland liep het anders. Spraaktechnologie in Nederland begint bij Antonie – Toon – Cohen. In 1952 promoveert hij bij hoogleraar Algemene Taalwetenschap Anton Reichling in Amsterdam op een proefschrift over de structuur van Engelse fonemen. Voor Cohen is het duidelijk dat de studie van spraak en de studie van taal onlosmakelijk met elkaar verbonden zijn. "For him, this not only meant that the study of speech should be firmly based in linguistics, but also that the study of language should find its emperical basis in the study of speech", schrijft Sieb Nooteboom in zijn necrologie over de foneticus bij wie hij zelf promoveerde (Nooteboom 1996) en met wie hij jarenlang samenwerkte bij het IPO. Een taalkundige dus, niet alleen door zijn training en achtergrond, maar ook in zijn overtuigingen. Maar door een aantal gelukkige omstandigheden kwam deze taalkundige terecht in een omgeving, die hem in staat stelde om spraaktechnologisch onderzoek te verrichten, te stimuleren en te faciliteren.

Het Natuurkundig Laboratorium van Philips, de Technische Hogeschool Eindhoven (nu Technische Universiteit Eindhoven) en ZWO (nu NWO) richtten in 1957 het Instituut voor Perceptieonderzoek (IPO) op. "Dat was vooral een plek voor lastpost Schouten van Philips", denkt Nooteboom. "Dat was een man met wilde ideeën en die moest ergens heen." Jan Schouten werd de eerste directeur van het IPO. In die functie besloot hij dat het instituut onderzoek moest doen naar spraak, en in 1959 trok hij dan ook Cohen aan om dat spraakonderzoek op te zetten. De foneticus belandde daarmee in een omgeving van louter technici, die alles konden bouwen wat de foneticus bedacht. In 1962 presenteerde de gloednieuwe spraakgroep haar eerste resultaat: een spraaksynthesesysteem dat later IPOVOX-1 is gaan he-

Zink Typografie Sheet 106 of 213 - Page 98 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



ten (Cohen & 't Hart 1963). De kwaliteit van de gegenereerde spraak kan beluisterd worden op internet.[3]

Cohen ontwikkelde al vrij snel een heel eigen onderzoeksprogramma. Het spraakonderzoek hield zich tot dan toe voornamelijk bezig met de synthese en de zogenaamde resynthese van spraak. Bij resynthese wordt gebruik gemaakt van een vocoder, die het spectrum van een spraaksignaal analyseert en daar de karakteristieken van een klank uit haalt. Daarna worden die specifieke eigenschappen van het geluid gebruikt om een kunstmatig signaal te bewerken tot een op het originele spraakgeluid gelijkende klank. Cohen richtte zich op een nog weinig bestudeerd fenomeen, namelijk intonatie. Doel was om uiteindelijk de taalkundige functies van intonatie te onderzoeken, maar daarvoor moesten eerst de formele eigenschappen van het verschijnsel in kaart gebracht worden (Nooteboom 1996). Dit onderzoek leidde in 1966 tot de INTONATOR (Willems 1966). Dit instrument leek op een vocoder – er ging natuurlijke spraak in, die vervolgens geanalyseerd en gehergenereerd kon worden, met dat verschil, dat intonatiepatronen geïsoleerd konden worden van de rest van het signaal, en zodoende weggelaten en vervangen konden worden door andere intonatiepatronen. Op deze manier konden Cohen en zijn collega's bestuderen wat het effect was van een intonatiepatroon op de perceptie en interpretatie van een gesproken zin. De INTONATOR was het werk van ingenieur Lei Willems. Willems heeft tot aan zijn pensioen de apparatuur gebouwd voor de spraakwetenschappers van het IPO. Zijn belang voor de ontwikkeling van het IPO tot een vooraanstaand instituut op het gebied van spraakonderzoek wordt door velen benadrukt, zowel van binnen als buiten het IPO. Professor Lou Boves uit Nijmegen: "Naar mijn idee was één van de belangrijke redenen waarom het IPO een lange tijd voorop heeft kunnen lopen, de aanwezigheid van Lei Willems. Dat was een goede ingenieur die kon toveren met in het begin harde elektronica, en hij is op relatief late leeftijd in staat geweest om de switch te maken van de oorspronkelijke analoge elektronica naar digitale computers. Hij heeft een heel centrale rol gespeeld in het onderzoek daar, door de apparatuur te

[3] Zie `http://www.let.uu.nl/~Sieb.Nooteboom/personal/HistoryIPOspeechsynthesis/IPO_speechsynthesis.htm`.

Zink Typografie Sheet 107 of 213 - Page 99 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



ontwikkelen – hij heeft het echt zelf van de grond af aan ontwikkeld – waar mensen als Johan 't Hart en Sieb Nooteboom en René Collier hun onderzoek mee konden doen." En Nooteboom bevestigt: "Hij was *de* spraakingenieur."

Het Eindhovense instituut, waar taalkundig geïnspireerd experimenteel spraakonderzoek werd gedaan, had een aantrekkingskracht op taalkundigen die graag onderzoek deden aan de hand van experimenten. Nooteboom was na zijn afstuderen in 1967 het liefst in Amsterdam blijven wonen, waar hem bovendien een baan was aangeboden. Maar Eindhoven leek hem wat betreft het vak veel interessanter. "Omdat daar een grotere groep zat, veel beter toegerust, met veel internationale contacten. Dat leek me aantrekkelijk en daar heb ik ook nooit spijt van gehad." Overigens werd die keuze in het geheel niet ingegeven door een interesse in spraaktechnologie per se. Hij begint met onderzoek naar versprekingen en naar de verwerking van geschreven tekst. Maar toen hij één dag in de week fonetiek ging geven in Leiden veranderde dat. "Nu zit ik weer in die fonetiek en moet ik daar les over geven", dacht hij toen, "dan kan ik maar beter dat lezen even laten zitten en ook weer onderzoek gaan doen in spraak." Nooteboom gooide zijn onderzoek over een andere boeg, en daar deed niemand moeilijk over: "Je kon doen en laten wat je wou op het IPO in die tijd. Later is dat wel veranderd." Ook de Vlaamse germanist René Collier kwam in deze tijd naar Eindhoven. Hij had in 1967 bij professor Engels in Leuven zijn licentiaat behaald met een analyse van melodie in spraak en wilde graag verder in de spraaktechnologie, maar in Vlaanderen was simpelweg de apparatuur en de kennis niet voorhanden om dit te kunnen doen. Met een beurs kon Collier uiteindelijk in Eindhoven het onderzoek doen voor zijn doctoraat, dat hij in 1972 in Leuven behaalde. Maar ook daarna bleef hij, ondanks een baan aan de Universiteit van Antwerpen, onderzoek doen bij het IPO. Het pionierswerk op het gebied van spraakonderzoek dat bij het IPO werd verricht vond ook zijn weg naar de universitaire wereld. In 1967 werd Cohen hoogleraar Engels in Utrecht, en vijf jaar later hoogleraar fonetiek aan dezelfde universiteit. Nooteboom gaf vanaf 1968 college in Leiden, waar hij enkele jaren later hoogleraar werd. Zo ontstond een sterke as van op taalkunde gebaseerd spraakonderzoek in Eindhoven, Utrecht en Leiden.

Onafhankelijk hiervan ontwikkelde zich ook een tweede as van

Zink Typografie Sheet 108 of 213 - Page 100 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



spraakonderzoek, die zijn wortels niet heeft in de taalkunde, maar in de natuurkunde. Aan de basis van deze beweging staat Louis Pols. De natuurkundige Pols werd tijdens zijn diensttijd, direct na zijn afstuderen in 1964, gedetacheerd bij het Instituut voor Zintuigfysiologie in Soesterberg, dat voor een groot gedeelte gefinancierd werd door Defensie. "Ze wisten daar niet wat ze met me aan moesten", vertelt hij nu. "Dus lieten ze me in het begin gewoon maar een beetje lawaaimetingen doen rondom geweren. Ik was tenslotte militair." Maar Pols had geen interesse in lawaaimetingen. Hij had wèl interesse in het werk van Reinier Plomp, die in Soesterberg op de afdeling Audiologie werkte. Plomp deed onderzoek naar de waarneming van geluid. Wanneer kunnen mensen nog een verschil waarnemen tussen twee tonen, en wanneer niet? En toen Plomp hiervoor klinkers ging gebruiken in plaats van stationaire geluiden en muzieknoten, kwam spraak om de hoek kijken. Pols begon met het synthetiseren van klinkers, en in 1967 publiceerde hij zijn eerste artikel in het *Journal of Acoustical Society of America* (JASA), samen met zijn baas Plomp en J.P. van der Geer, over het meten van klinkerspectra met behulp van bandfilters (Plomp, Pols & van der Geer 1967).

Pols had geen enkele taalkundige achtergrond toen hij aan dit onderzoek begon. Een klinker was voor hem een geluidssignaal niet fundamenteel anders dan andere geluidssignalen. Hij dacht niet in termen van formanten, de karakteristieke frequenties van klinkers waarbij resonantie optreedt in de verschillende holtes van het spraakkanaal, maar simpelweg in geluidsspectra waarin metingen gedaan konden worden. Deze metingen deed hij met behulp van bandfilters, die elk alleen het geluid van een specifieke frequentie doorlaten. Door bijvoorbeeld twintig van deze bandfilters te combineren konden spectra goed gerepresenteerd worden. En hoewel deze representatie niet gebaseerd was op kennis van formanten, leek de uiteindelijk representatie van de spectra (na datareductie) wel heel sterk op een formantrepresentatie. Het gebrek aan taalkundige achtergrond heeft Pols nooit gehinderd. "Dat beetje fonetiek, dat leer je er wel bij", aldus de wetenschapper. "Mijn wiskundige en natuurkundige achtergrond was veel belangrijker."

Tussen het IPO en het Instituut voor Zintuigfysiologie in Soesterberg was sprake van samenwerking, en er was ook wel enige overlap in onderzoeksfocus. Pols: "Alleen al de naam. In het Engels heette het

Zink Typografie Sheet 109 of 213 - Page 101 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



[Instituut voor Zintuigfysiologie] *Institute for Perception*. En het IPO heette *Institute of Perception Research*." De onderzoekers van beide instituten gingen elk half jaar bij elkaar op bezoek, en bezochten elkaars evenementen. Nooteboom: "Dat was onder meer om doublures in onderzoek te voorkomen." Maar de verschillen tussen de beide instituten waren groot. Pols: "Het IPO was rijker, dat had zowel universitaire als industriële middelen tot zijn beschikking." Het grootste verschil zat hem echter in de benadering. En dat leverde soms spanning op. Nooteboom: "Maar dat was puur inhoudelijk. Soesterberg had meer een ingenieursbenadering, wij meer een taalkundige benadering. Dat zou je niet zo snel verwachten bij Philips, maar was wel zo. Dat kwam door de invloed van Cohen. Maar die tegenstelling heeft nooit tot animositeit geleid."

De kans op overlap in onderzoek werd aanzienlijk verkleind doordat het IPO zich voornamelijk richtte op spraaksynthese, en resynthese, zeker in het begin. Pols: "Ze hebben heel lang geaarzeld voordat ze in herkenning doken. Misschien is dat voor mij ook wel reden geweest dat ik dat wèl gedaan heb." Lou Boves, hoogleraar Spraaktechnologie aan de Katholieke Universiteit Nijmegen, meent de reden te kennen waarom het IPO geen spraakherkenning deed: "Cohen en Nooteboom vonden dat dat niet kon, dat het te moeilijk was. En bovendien denk ik dat op de achtergrond vanaf het begin interne Philips-politiek een rol heeft gespeeld. Philips had meerdere Europese onderzoekscentra. Een daarvan was Hamburg, met Hermann Ney, die spraakherkenning deed. Daarnaast hadden ze ook nog Christian Wellekens en Yves Kamp in Brussel, en die waren ook bezig met spraakherkenning. En het onderzoek op de verschillende plaatsen mocht natuurlijk niet overlappen." Nooteboom bevestigt dat de activiteiten elders binnen Philips een rol hebben gespeeld bij het bepalen van de focus van het onderzoek. Maar hij kan zo nog een aantal andere redenen opnoemen waarom het IPO zich minder op automatische spraakherkenning stortte: "Voor een deel was het toeval. Voor een deel had het te maken met de interesse van Cohen. En wij hadden natuurlijk die band met taal. De relatie tussen taalkundige eigenschappen en spraak ligt veel meer voor de hand wanneer je aan spraaksynthese werkt. Spraakherkenning was een heel andere wereld, daar waren geen taalkundigen bij betrokken. Denk maar aan de quote van Fred Jelinek: "Every time I fire a linguist, the performance of our speech recog-

Zink Typografie Sheet 110 of 213 - Page 102 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



nition system goes up"[4] Overigens deden we wel wat aan analyse, maar dan altijd in functie van andere zaken, niet voor automatische spraakherkenning."

Het team van de ingenieurs kende nog een tweede speler: Nijmegen. In 1972 was de Duitse ingenieur Willem Vieregge vanuit Bonn naar Nijmegen gekomen om daar de leerstoel fonetiek te bekleden. Lou Boves zat destijds als student in de benoemingscommissie. "Het was nog best lastig om kandidaten te vinden, want het moest iemand zijn die gepromoveerd was. Maar de vooraanstaande Nederlandse fonetici waren dat allemaal niet". Vieregge maakte zich sterk voor een groot nationaal programma gericht op spraakherkenning, maar dat kwam niet van de grond, onder meer doordat het IPO niet enthousiast was. De hoogleraar nam op zijn beurt Lou Boves aan, die in 1984 promoveerde, en nog in datzelfde jaar startte met een project binnen het Innovatiegerichte Onderzoekprogramma Hulpmiddelen voor gehandicapten, waarmee Nijmegen echt op de kaart kwam te staan.

Andere spelers waren er niet. Of toch? Nooteboom: "Je had natuurlijk ook nog Groningen. Maar Groningen deed nergens aan mee. Daar zat Graham Stuart, en die heeft nooit iets gepubliceerd, daar hebben we nooit iets van gehoord, dus die hebben we maar een beetje laten zitten." Doordat de hoogleraar in Groningen op geen enkele manier een bijdrage levert aan de ontwikkeling van een spraakgroep in Groningen, staat fysicus, slavist en foneticus Tjeerd de Graaf volledig geïsoleerd. De interesse is er wel. "Hij was een van eersten die spraakanalysesoftware in huis haalden", weet Louis Pols. Maar een goede voedingsbodem met technologische ondersteuning ontbrak, en De Graaf richt zich meer en meer specifiek op de Slavische talen. Er was ook nog een commercieel lab waar aan spraak gewerkt werd: het Dr. Neherlab in Leidschendam, dat later bekend werd onder de naam KPN Research, maar toen nog eigendom was van PTT. Er is weinig bekend over het onderzoek dat hier plaatsvond in deze tijd. Pols

[4] Deze uitspraak zou gedaan zijn op de *Workshop on Evaluation of NLP Systems* in Wayne, Pennsylvania in December 1988, waar hij een lezing gaf. Er zijn verschillende varianten van de uitspraak in omloop. Hijzelf herinnert zich de uitspaak als "Anytime a linguist leaves the group the recognition rate goes up" (Jurafsky & Martin 2000). Het verslag van de bijeenkomst maakt in het geheel geen melding van de uitspraak (Palmer & Finin 1990).

Zink Typografie Sheet 111 of 213 - Page 103 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



weet dat ze gespecialiseerd waren in patroonherkenning, onder meer voor het uitlezen van getallen op een girokaart. En zijn voorganger als hoogleraar fonetiek aan de Universiteit van Amsterdam, professor Hendrik Mol, werkte er ooit. Maar volgens Pols heeft hij nooit aan spraakherkenning gewerkt. Het is niet eenvoudig om te achterhalen wat er precies gebeurde op het Dr. Neherlab. Noch Louis Pols, noch Sieb Nooteboom hebben ooit samengewerkt met onderzoekers uit het Dr. Neherlab. Nooteboom: "Het Dr. Neherlab was met kranten dichtgeplakt. Dat komt doordat ze een heel gesloten bedrijfscultuur hadden. Pas veel later, bij het ASSP-programma, zijn ze betrokken geraakt. Ze hadden geen publicaties, en daardoor geen aansluiting bij onderzoek rest van de wereld. Je kunt je zelfs afvragen of het wel research was wat ze deden. Of praktische ondersteuningsactiviteiten voor de PTT? In die tijd stelde het weinig voor. Maar ongetwijfeld zullen ze dat zelf anders zien dan ik."

Vlaanderen had begin jaren zeventig nog René Collier verloren aan het IPO, omdat het er simpelweg ontbrak aan apparatuur en kennis. Wat Vlaanderen nodig had om spraaktechnologisch onderzoek op te starten, was iemand die zelf de apparatuur kon bouwen die daarvoor nodig was. Die persoon heette Jean-Pierre Martens. Net als Louis Pols had hij geen taalkundige achtergrond, en net als Pols was hij min of meer per ongeluk in het vak gerold. De elektrotechnisch ingenieur studeerde in 1974 af aan de Universiteit van Gent. Hij kon meteen aan de universiteit blijven om een project te doen voor Siemens, namelijk het oplossen van het probleem van rondzingen in telefoons. Dat probleem kon opgelost worden door ervoor te zorgen dat de beide bronnen geen overlappende frequenties hadden. Martens ging daarom op zoek naar een methode om het geluid te filteren. Wanneer nu de filters van de bronnen verschillende frequentiebanden hebben, dan is het probleem opgelost. Alleen bleek de spraak die door de filters kwam niet goed genoeg te zijn. Toen hij ging onderzoeken waarom dat zo was en probeerde de kwaliteit van de filters te voorspellen, kwam hij terecht in de psychoakoestiek en ging hij modellen maken van het menselijk gehoor. Zo werd hij de eerste spraaktechnoloog in Vlaanderen. "Ik ben een dinosaurus in spraak", zoals hij het zelf zegt.

Zink Typografie Sheet 112 of 213 - Page 104 of 205

## 3.2 De techniek

Bij de ontwikkeling van taaltechnologie zagen we dat deze enorm gedreven en beïnvloed is door de ontwikkeling in de techniek, de bredere beschikbaarheid van computers, het op de markt komen van snellere computers, het dalen van de prijs van geheugen, de ontwikkeling van programmeertalen, en de ontwikkeling van software. Voor de spraaktechnologie geldt dat zeker ook, ondanks het feit dat de spraaktechnologie ontstaan is voordat er computers beschikbaar kwamen voor spraakonderzoek. Nooteboom: "Toen Cohen bij het IPO begon in 1959, had hij helemaal niets. Hij heeft lang zitten wachten op een setje om magneetbanden mee te kunnen knippen en plakken. Dat was zijn gereedschap in die tijd." Ook jaren later nog werd in de meeste laboratoria nog onderzoek gedaan met primitieve middelen. Nooteboom herinnert zich een bezoek aan het lab van Peter Ladefoged, een vooraanstaand foneticus en pionier op het gebied van de experimentele fonetiek. "Hij deed de deur van zijn lab voor mij open, waarop een noodkreet klonk. In het lab stond een raam open, en doordat Ladefoged de deur opendeed was er een windvlaag ontstaan in de ruimte waar een promovendus bezig was met het sorteren van honderden bandlussen. Dat was zijn gereedschap voor zijn onderzoek. Ieder bandje een woord. Die tafel vol met bandlussen was nu verstoord door die windvlaag, en daarmee was wekenlang werk weg. Dat was hoe het was."

De computer deed pas halverwege de jaren zeventig haar intrede in het IPO. "Dat was een Honeywell 5.16", weet Nooteboom nog, "maar hij mocht niet zo genoemd worden. Door Philips werd hij verkocht onder de naam Philips P2000". Het was een 16-bits computer met in- en uitvoer in de vorm van ponsband. De computer ging 's nachts uit. "Omdat hij geen geheugen had, kon hij 's nachts niet werken", verklaart Nooteboom. Dan was er namelijk niemand om ponsband in te voeren. Een paar jaar later kreeg Nooteboom voor het eerst beschikking over computergeheugen, in de vorm van magnetische schijven. Ook bij het Instituut voor Zintuigfysiologie werd de signaalanalyse offline gedaan met behulp van een bandfilterset. Voor het rekenwerk met de output hiervan werd een PDP7 gebruikt. Geheugencapaciteit: 8k. In Gent bouwden de ingenieurs hun eigen hardware. Martens: "Dat was heel primitief werken, je wilt niet weten

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



hoe wij dat deden. Wij hadden een zelfgemaakte computer, dat was eigenlijk een soort kloon van een bestaande computer, die draaide dezelfde software, maar was totaal anders gemaakt. Daarin hadden wij een kaart gemaakt, tussen A3 en A4 formaat, aan weerskanten bestikt, en dat was een hardwarevermenigvuldiger. Daarmee kon je met één instructie twee getallen met elkaar vermenigvuldigen. Doordat we dat gemaakt hadden, konden we digitale filters maken, die in real time konden werken. Dat was in de jaren zeventig. Daardoor konden wij die filters bouwen, waarmee we direct konden werken. Al was dat niet heel direct, want uiteindelijk zaten daar ROMs in, read-only-memories, en dus we moesten die eruit halen, code erop programmeren, terug in die computer steken, en dan konden wij met een digitaal filter werken." De zelfgemaakte computer gebruikte een oscilloscoop als beeldscherm. "Dat had als leuk gevolg dat wij spectra konden tonen, dat we signalen voorbij konden laten komen op ons scherm in 1975. Het heeft nog geduurd tot 1995 tot dat op een pc snel genoeg kon. Wij konden dat in 1975." Voor studenten waren nog minder middelen beschikbaar. Als Boves voor zijn afstuderen met twee medestudenten intonatieanalyse wil doen op radio-interviews, kunnen die niet in Nederland gedigitaliseerd worden. "De enige computer die ik mocht gebruiken was een IBM minicomputer die in Amsterdam stond, vreselijk groot en duur, maar met niet meer dan 16kb intern geheugen en een harde schijf van 256kb", herinnert hij zich. "In Amsterdam waren ze al jaren bezig een analoog-digitaalconverter te bouwen voor die IBM-computer, maar dat lukte almaar niet." Uiteindelijk krijgt hij het toch voor elkaar om zijn bestanden te digitaliseren. "Via de contacten tussen Amsterdam en Bonn wist ik dat daar een computer stond. Dus ik ben een paar dagen met mijn audiotape naar Bonn geweest, waar geen IBM stond maar een PDP8 waar ze wel een AD-converter voor hadden, en ben teruggekomen met een koffer vol met ponsbanden."

De analyse van het spraaksignaal kan nu vrijwel automatisch gedaan worden door het softwarepakket Praat van Paul Boersma (Boersma & Weenink 2001). Een groot verschil, dat een enorme impact heeft op wat er gedaan kan worden. Pols: "Ik heb al mijn software destijds bij TNO allemaal zelf moeten schrijven. Dat blijft natuurlijk toch allemaal vrij primitief, dat kan niet anders. De echte real time component daarvan, daar hadden we dan weer een programmeur

Zink Typografie Sheet 114 of 213 - Page 106 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



voor bij TNO, die in machinetaalcode dat zat te programmeren. Want met 8k geheugen, kan je niet een Fortran of een Pascal of een Algol of een programma laten draaien. Met 8k geheugen kan je helemaal niets." Dirk van Compernolle, die in de jaren tachtig een spraakgroep in Leuven opzette, schat de invloed van rekenkracht nog hoger in: "Het is nog altijd de regel dat computing power domineert over slim zijn, in spraakherkenning. Maar het is nu wel veel minder relevant dan dertig, veertig jaar geleden."

## 3.3 Internationale netwerken

Martens begon met een zekere achterstand ten opzichte van de groepen in Soesterberg en Eindhoven. In de eerste plaats waren de groepen daar net een paar jaar eerder begonnen, en was er heel brede ondersteuning voor ontwikkeling van apparatuur. Maar daarbij waren Soesterberg en het IPO zeer goed verbonden met spraakonderzoekers in binnen- en buitenland. Pols: "Mijn mentor Reinier Plomp was heel erg internationaal georiënteerd, die nam zijn jonge medewerkers ook mee naar Amerika. In 1972 heeft hij me al meegenomen naar Amerika, en dan realiseer je je opeens wat er verder allemaal nog aan de hand is in de wereld." Na zijn promotie brengt Pols bovendien nog een heel jaar door in de Verenigde Staten. Hij krijgt een ZWO-beurs om onderzoek te doen aan het Speech Communication Research Lab in Santa Barbara, Californië. "ZWO had veel geld nog", verklaart hij de beurs. "Als je gepromoveerd was, en dat gebeurde ook nog niet zo veel, dan kon je gewoon een jaar weg." Maar na een half jaar houdt het instituut op te bestaan. Pols rijdt van de Westkust naar de Oostkust en kan de tweede helft van zijn jaar doorbrengen bij het prestigieuze Bell Labs. "Het is dus alleen maar een groot voordeel geweest. Ik heb een half jaar aan de westkust van Amerika gezeten, en een half jaar aan de oostkust. Ik heb een halfjaar gewoond op beide plekken, en heb alle congressen en alle laboratoria bezocht aan beide kanten." Ook na terugkeer in Nederland blijft Pols uitstekende internationale contacten houden. Door de nauwe banden tussen Defensie en het Instituut voor Zintuigfysiologie kwam hij terecht in een Research Study Group voor Speech Processing, waarin alle NATO-landen vertegenwoordigd waren, en

Zink Typografie Sheet 115 of 213 - Page 107 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



die hij zelfs een tijdje voorzit. Ieder jaar kwam de studiegroep bijeen in een van de deelnemende landen om de laboratoria te bezoeken en kleinschalige samenwerkingsverbanden op te zetten. In latere jaren zou Pols een actieve rol spelen binnen de *European Speech Communication Association* ESCA, dat later zijn naam veranderde in ISCA, met de I van International, en de Eurospeech (later Interspeech) congressen organiseert. Ook het IPO zat goed ingebed in de internationale spraakwereld, met dank aan de inbreng van Philips. Nooteboom: "We hadden heel goede contacten met Bell Labs. We wisselden kennis uit voor zover die niet geheim was, en jaarlijks was er een uitwisseling van octrooien met gesloten beurzen." Het dubbele karakter van het IPO, half industrieel, half universitair, was daarbij een voordeel: "Ik had vanuit Philips gemakkelijk toegang tot Bell Labs, maar anderzijds kon ik soms beter doen alsof ik van het IPO was, want dan konden de wetenschappers beter vrijuit praten. Ook al wisten zij natuurlijk ook wel hoe het zat", aldus Nooteboom. Jaarlijks bezocht Nooteboom het prestigieuze laboratorium in de Verenigde Staten, waar onder meer [Bishnu] Atal werkte, een van de grondleggers van een zeer belangrijke techniek voor spraakanalyse, lineaire predictie. "Daar hadden wij direct baat bij", geeft Nooteboom dan ook aan. "Ik reisde veel naar de laboratoria in de VS en Japan, en dat was belangrijk voor de voortgang van het onderzoek. Dat was een van de grote voordelen van werken bij het IPO. [...] Als ik destijds naar Amsterdam was gegaan, had ik nooit zo vaak naar Amerika kunnen reizen."

En inderdaad, het contrast met de spraaktechnologen aan de universiteiten is groot. Jean-Pierre Martens over de situatie in Gent: "Toen ik begon in 1974 was er geen geld om naar conferenties te gaan. Wij wisten heel weinig over wat er op conferenties gebeurde, want als je niet naar een conferentie ging, dan had je ook de proceedings niet. Dus wij waren eigenlijk aangewezen op publicaties in tijdschriften. Wij publiceerden allemaal in tijdschriften zoals de JASA [*Journal of Acoustic Society of America*] en dergelijke, maar ik ben nooit naar een conferentie geweest voordat ik mijn doctoraat haalde." Lou Boves, die begin jaren zeventig in Nijmegen spraaktechnologie op de kaart zet, herkent het beeld dat Martens schetst. "In die tijd, jaren zeventig, waren we van onderzoek op de hoogte via tijdschriften en klein aantal boeken. De rol van tijdschriften was heel groot. Het was de tijd waarin ook boeken verschenen die alleen maar bestonden uit belangrijke ar-

Zink Typografie Sheet 116 of 213 - Page 108 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



tikelen uit tijdschriften. JASA was voor spraak heel erg belangrijk." Overigens garandeerde een setting in de industrie geen internationaal netwerk. Professor Herman Tromp, die tot halverwege de jaren tachtig leiding gaf aan de spraakgroep van PTT Research, ontbrak het aan deze contacten en hij liet zich weinig zien op congressen. Volgens Boves, die in 1985 als consultant naar het Neherlab kwam, één van de redenen waarom die spraakgroep niet veel méér heeft opgeleverd dan het geval was.

Ook tussen specifieke spraakgroepen over de grens ontstaan banden. Doordat zowel Gent als de spraakgroep op het IPO zich nadrukkelijk bezighield met vocodertechnologie, ontstonden er min of meer vanzelf contacten en een vorm van samenwerking tussen de beide groepen. "Dat waren zo de eerste contacten met Nederland", aldus Martens, die bijna zelf in Eindhoven was gaan werken. "Ik had net mijn doctoraat afgemaakt, en Sieb Nooteboom zocht iemand om mensen te begeleiden", vertelt hij. "Maar omdat zij zowel met beeld als met geluid werkten, en omdat ikzelf met jonge kinderen zat, ben ik in Gent gebleven." Wel maakt hij Nooteboom attent op zijn broer, die in beeldverwerking werkzaam is. En zo kwam Jean-Bernard Martens bij het IPO terecht. De spraakgroep van Jean-Pierre Martens in Gent deed eerder ook projecten met de Universiteit van Mons, maar sinds de wetenschap in België geregionaliseerd werd, is deze samenwerking moeilijk geworden. Martens: "Vlaanderen had zijn eigen politiek op het gebied van wetenschappelijk onderzoek en Wallonië had zijn eigen politiek op het gebied van wetenschappelijk onderzoek. Een project tussen die twee, dat ging zomaar niet, want de gelden waren lokaal verdeeld. Dus voor ons is Wallonië nu even moeilijk om een project mee te doen als een buitenland."

In 1983 was Nederland even zelf het centrum van de internationale spraakwereld toen het *Xth International Congress of Phonetic Sciences* in Utrecht plaatsvond onder voorzitterschap van Toon Cohen. "Many phoneticians in the world will remember Antonie Cohen as the chairman of the Xth International Congress of Phonetic Sciences, that took place in Utrecht in 1983. I have heard it said many times by many people from all over the world that this was the best organized international congress of phonetics they had ever witnessed" (Nooteboom 1996).

Hoewel met de jaren de communicatiemiddelen beter werden,

Zink Typografie Sheet 117 of 213 - Page 109 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



bleef het hebben van goede internationale contacten bijzonder belangrijk. In 1985 startte het Amerikaanse *Defense Advanced Research Projects Agency* DARPA een spraakherkenningsprogramma. Jaren later vertelt Lou Boves hierover in het NRC Handelsblad: "DARPA, […] schreef een soort wedstrijd uit. Ze gaven geld en spullen en ieder jaar werd aan de deelnemende researchinstituten een bandje met spraak gestuurd. Binnen veertien dagen moest de automatische verwerking daarvan worden teruggestuurd. Kijken wie er wint. Een belangrijke voorwaarde voor deelname was dat ieder instituut ieder jaar in de workshop ter afsluiting van de competitie volledige opening van zaken moest geven, zodat het volgende jaar iedereen over de zelfde gegevens en technieken kon beschikken." (NRC Handelsblad 3 december 1999). Het door defensie gesubsidieerde programma stond alleen open voor Amerikaanse instellingen. En niet alle instellingen deden mee: "bijvoorbeeld IBM deed er niet aan mee", weet Van Compernolle. "Het is nog maar de laatste 10, 15 jaar dat ze dat doen. Heel lang heeft IBM in al die DARPAprogramma's niet meegedaan. Om niet te moeten delen met anderen, om de eigendomsrechten." Desondanks heeft het programma zeer veel invloed gehad op de ontwikkeling van het spraaktechnologisch onderzoek. Niet iedereen is het erover eens of die invloed positief of negatief was. Boves meent dat DARPA de vrije ontwikkeling van het vak in de weg staat: "Als er één organisatie is die het onderzoek in een richting gestuurd heeft die in zekere zin contraproductief is, dan is het niet de industrie, dan is het DARPA. Voor alles waar ik mee te maken heb gehad, en volgens mij geldt voor taaltechnologie precies hetzelfde, is DARPA de organisatie die zegt wat er gedaan gaat worden in onderzoek, al was het maar omdat de programma's die zij financieren zo genereus voorzien worden van corpora en testdata. En daarmee bepalen zij eigenlijk wat wij gaan doen." Doordat het een competitie is, zouden onderzoekers minder geneigd zijn om echt nieuwe wegen in te slaan, die wellicht niet binnen een jaar de gevestigde orde verslaan, maar op de lange termijn meer potentie hebben. Pols zet daar tegenover: "De allereerste DARPA test is gewonnen door de mensen van Carnegie Mellon, en dat waren nu juist de buitenbeentjes, zij hebben die hele Hidden Markov Modeling geïntroduceerd." HMMs werden vervolgens het dominante paradigma. Maar ook Pols geeft toe: "in die competitie waar vele miljoenen in omgaan, daar kunnen ze [de onderzoekers] zich niet veroorloven om

Zink Typografie Sheet 118 of 213 - Page 110 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



af te wijken van de mainstream." Nederlandse en Vlaamse spraaktechnologen konden niet meedoen aan de DARPA-competities, die vele jaren lang niet open stonden voor buitenlandse instellingen. Des te belangrijker was het om wel internationale netwerken te hebben, zodat de ontwikkelingen wel gevolgd konden worden. Daarnaast begonnen zowel de Nederlandse en Vlaamse als de Europese overheden langzaam maar zeker te beseffen dat zij hier iets tegenover moesten stellen.

## 3.4 1984

> The pen was an archaic instrument, seldom used even for signatures, and he had procured one, furtively and with some difficulty, simply because of a feeling that the beautiful creamy paper deserved to be written on with a real nib instead of being scratched with an ink pencil. Actually he was not used to writing by hand. Apart from very short notes, it was usual to dictate everything into the speakwrite, which was of course impossible for his present purpose.
>
> 1984, *George Orwell (1949)*

Spraaktechnologisch onderzoek in Nederland en Vlaanderen richtte zich lange tijd vooral op regelgebaseerde spraaksynthese en analoge vocoders (analyse, compressie en resynthese). Een uitzondering hierop was het intonatie-onderzoek van het IPO, dat weliswaar aan synthese gerelateerd was, maar toch echt een eigen niche vormde. Bij de regelgebaseerde synthese werd uitgegaan van vaste kenmerken van elk van de fonemen van een taal. Maar de realisatie van een foneem verandert sterk, afhankelijk van zijn context: de /k/ in 'zakdoek' klinkt meer als de /g/ in het Franse 'garçon' dan als de /k/ in 'zak'. En dat geldt voor elke klank. Al deze veranderingen ten aanzien van de 'basisvorm' van de fonemen werden vastgelegd in regels, die feitelijk generalisaties waren die gemaakt konden worden op basis van onderzoek op vele fragmenten spraak.

Het zogenaamde vocoderonderzoek kwam er simpel gezegd op neer dat op het niveau van fonemen het geluidssignaal ontleed werd met behulp van bandfilters (eerst analoog, later digitaal), die het sig-

Zink Typografie Sheet 119 of 213 - Page 111 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



naal reduceerden tot een aantal getallen dat voor iedere zoveel milliseconde weergaf wat het energieniveau was op verschillende frequenties, of het geluid stemhebbend of stemloos was, en eventueel nog andere karakteristieken. Fonetici lieten de frequenties van de bandfilters samenvallen met de formanten van de verschillende klinkers, ingenieurs maten zoveel ze konden en reduceerden later de informatie, waarbij de meest informatieve data behouden bleef, en zodoende een soortgelijke uitkomst opleverde. Het rijtje getallen was vele malen kleiner dan het oorspronkelijke signaal, wat betekende dat het een goede manier was om spraak op te slaan of te verzenden, aangezien geheugen en bandbreedte in die tijd nog schaars en duur waren. Maar deze compressie heeft alleen zin wanneer op basis van deze getallen weer goede spraak gegenereerd kan worden. En zolang op foneemniveau gewerkt werd, was dat nog best lastig.

In de jaren zeventig werd een nieuwe techniek ontwikkeld voor de analyse en resynthese van het spraaksignaal: *Linear Predictive Coding* (LPC). Volgens Dirk van Compernolle één van de twee basisalgoritmes uit de spraaktechnologie: "Spraak, dat bekijk je in het spectrale domein. Dus je moet spectra berekenen. Fourier transformatie is een manier om spectra te gaan berekenen, maar dat is iets dat rekenintensief is. Fast Fourier Transform (FFT), dat is een algoritme waardoor dat heel veel sneller gaat. Met de computers van die dagen was dat superbelangrijk. [...] En dan heb je op een paar verschillende plaatsen het LPC algoritme. Eigenlijk is het een model waarbij je spraak synthetiseert door gebruik te maken van een parametrisch model van de mondholte. [...] Dus FFT hebt je puur nodig voor analyse, LPC hebt je nodig voor codering, en voor spraaksynthese."

De ontwikkeling van LPC maakte het mogelijk om bij spraaksynthese naar andere eenheden van klank te kijken dan de traditionele fonemen. Bijzonder geschikt bleek de difoon, een stukje van het signaal dat loopt van het (stabiele) midden van een foneem tot aan het midden van de daaropvolgende foneem. Deze aanpak was erop gericht om het grote probleem van assimilatie aan te pakken: de spectrale veranderingen in een foneem door de omringende klanken was nu al ingebakken in de basiseenheid van analyse. Door telkens de overgang mee te nemen in de bouwstenen van het model, werd weliswaar de hoeveelheid basiseenheden groter, maar de variatie binnen een basissegment kleiner, en daardoor de kwaliteit van de gesynthetiseerde

Zink Typografie Sheet 120 of 213 - Page 112 of 205

spraak beter. Dit wordt gezien als een van de grote doorbraken in de geschiedenis van de spraaktechnologie. Nooteboom: "Het grote voordeel van de difoonsynthese is dat je niet over segmenteren hoeft in te zitten. Alle aandacht kan gaan naar het echte probleem." En ook Martens geeft aan: "Ik vind het wel een belangrijke doorbraak dat men ingezien heeft dat fonemen niet de eenheden waren die we nodig hadden bij spraakverwerking."

De ontwikkeling en eerste toepassingen van LPC in Nederland brengen ons naar het jaar 1984. Vanaf dat jaar neemt het spraakonderzoek een grote vlucht en gebeuren er een aantal zaken snel achter elkaar: internationale samenwerking tussen Philips en Siemens in SPICOS, het eerste grote project voor Nijmegen in de vorm van het IOP Hulpmiddelen voor Gehandicapten, de start van het Europese ESPRIT-programma, waar verschillende Nederlandse groepen projecten in hadden, gevolgd door de start van het grote Nederlandse samenwerkingsprogramma ASSP en de oprichting van de Stichting Spraaktechnologie (SST) een jaar later, en het datacentrum SPEX in 1987. In datzelfde jaar arriveerde ook Dirk van Compernolle in Leuven, waar hij een nieuwe spraakgroep opzette.

Maar eerst SPICOS (Siemens Philips IPO COntinous Speech). SPICOS was een samenwerking tussen Siemens, de Philips onderzoeksgroepen in Hamburg (later Aken) en Brussel, en het IPO, en had tot doel een Duits dialoogsysteem te ontwikkelen. Het begon allemaal in 1984. Nooteboom: "Het SPICOS-project is ontstaan op het vliegveld van München in een gesprek tussen Harry Bunt en mijzelf. We waren op weg naar huis van een vergadering bij Siemens. Daardoor hebben we ons vliegtuig nog gemist." In SPICOS kwamen enerzijds spraakherkenning en -synthese samen, en anderzijds taal en spraak. Voor de dialoogcomponent werd technologie gebruikt uit TENDUM, een soort spinoff van PHLIQA, waar Harry Bunt nog bij betrokken was. Het IPO was voornamelijk betrokken bij de synthese. "En dat was aardig om te doen, want omdat het een dialoogsysteem was, had je de betekenis tot je beschikking", vertelt Nooteboom. Dit in tegenstelling tot text-to-speech, waar wel de woorden, maar meestal niet de exacte bedoelde betekenis bekend is. En die heeft een enorme invloed op de intonatie, de specialiteit van het IPO. Spraakherkenning werd door de buitenlandse groepen gedaan, onder meer met behulp van Hidden Markov Modellen, die zeer populair werden in spraakher-

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



kenning en later ook in automatisch vertalen. Maar het IPO hield zich hier verre van. Enerzijds omdat iedere groep moest doen waar het goed in was, anderzijds omdat de insteek fundamenteel anders was. Het Philips laboratorium in Brussel, bijvoorbeeld, was een wiskundig laboratorium. "En wiskundigen zijn echt een beetje anders", aldus Nooteboom. Ook culturele verschillen speelden een rol bij de samenwerking: "De voortgang werd nog wel eens gehinderd doordat wij als we bij Siemens op bezoek gingen er altijd maar voor één dag waren. Dat was zuinig, want dan hoefde je geen hotelkosten uit te geven. Maar die ene dag werd voor een groot gedeelte doorgebracht met een uitgebreide lunch met veel wijn – dat betaalde Siemens. Siemens was rijk. Dat betekende dat de dag snel voorbij was. Dat was anders wanneer we naar Hamburg gingen, daar kregen we altijd alleen een broodlunch en geen alcohol." De samenwerking tussen Philips en Siemens heeft geen stand gehouden: Philips stapte uit de samenwerking en hoewel Siemens nog een tijdje doorgewerkt heeft aan het project en zelfstandig een vervolgproject heeft opgezet, heeft dat nooit tot een product geleid.

Ook Nijmegen start in 1984 een belangrijk syntheseproject en wel binnen het door het Nederlandse Ministerie van Economische Zaken gefinancierde Innovatiegerichte Onderzoeksprogramma (IOP) Hulpmiddelen gehandicapten. Het eerste grote project voor de groep van Lou Boves. "Dat wij ook synthese gingen doen was in wezen een uitvloeisel van twee dingen: over synthese was in Nederland kennis, en ineens was er ook behoefte aan synthesetoepassingen", aldus de dan net gepromoveerde Boves. "In dat IOP zijn we begonnen met het maken van een spraaksynthesesysteem dat je in een 2080 IBM pc kon stoppen en dat een werkstation moest opleveren voor blinden en slechtzienden. We hebben daar inderdaad een insteekkaart gemaakt voor een pc, waar aan de ene kant tekst in werd gestopt die van het scherm geplukt werd, en aan de andere kant spraak uit kwam." Het systeem bleek uiteindelijk te duur om op grote schaal geproduceerd te worden. Kennisgebaseerde synthesesystemen zouden niet lang centraal blijven staan in Nijmegen: direct na dit project verschuift de focus in Nijmegen van synthese naar herkenning. Tegelijkertijd neemt de probabilistische aanpak, zowel in synthese als in spraakherkenning, een grote vlucht.

Een derde belangrijke mijlpaal in 1984 is de start van het ESPRIT

Zink Typografie Sheet 122 of 213 - Page 114 of 205

programma. ESPRIT staat voor *European Strategic Programme for Research and Development in Information Technology*. Europese consortia konden subsidie aanvragen voor onderzoeksprojecten rondom het thema informatietechnologie. Er moesten altijd minimaal 2 verschillende landen uit de Europese Unie meedoen, en de subsidie besloeg niet meer dan 50% van het totale budget, waardoor serieuze participatie van het bedijfsleven noodzakelijk werd. Nederlandse en Vlaamse instellingen participeerden in verschillende projecten. Zo zaten TNO (zoals het Instituut voor Zintuigfysiologie inmiddels heette) en PTT Research in het project SAM, dat tot doel had een goede evaluatie en vergelijking van verschillende spraakherkenners mogelijk te maken. Pols: "We maakten ons druk om evaluatie van systemen, terwijl de systemen er nog helemaal niet waren." Hij denkt dat het een poging was om internationaal iets bij te dragen aan de ontwikkeling van spraakherkenningssystemen, zonder een zeer kostbaar programma in de stijl van DARPA neer te zetten, dat weinig kans had om betere systemen voort te brengen. "Wij konden niet tippen aan de DARPA-systemen", aldus Pols. Er waren ook taaltechnologieprojecten: Piek Vossen nam namens de Universiteit van Amsterdam deel aan Acquilex, een project dat moest leiden tot een multilinguale lexicale *knowledge base*. De Katholieke Universiteit Nijmegen deed mee aan het Polyglot-project, dat automatische vertaling van spraak tot doel had. "Het IPO zat daar trouwens ook in, net als Philips in Hamburg en later Aken", vertelt Boves. "Dat was de eerste keer dat René Collier en ik en Herman Ney moesten proberen samen te werken." Anders dan bij SPICOS zijn de Nederlanders, en met name de groep van Boves, ditmaal wèl direct betrokken bij de spraakherkenning. Boves: "Rond die tijd was het dus, dat de continue spraakherkenning opkwam, en continue spraakherkenning is echt van begin af aan gebaseerd geweest op puur probabilistische principes." In eerste instantie deed Nijmegen vooral herkenning van geïsoleerde woorden, terwijl andere groepen continue spraak deden. "Dat was voor het eerst dat wij geconfronteerd werden met de overgang van puur kennisgebaseerd naar statistisch. Maar *isolated word dictation* was nog wel een mix van die twee met een belangrijke kenniscomponent. Die systemen zijn vervolgens geeclypseerd door verbonden spraak. En daarin zijn we toen wel mee proberen te gaan." De overgang naar probabilistische methoden ging zeker niet vanzelf. Boves: "Dat was toen even slikken. Het was heel

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



veel leren. Het is een totaal andere manier van denken, het duurt toch even voordat je je die helemaal eigen gemaakt hebt." Deze nieuwe methodologie sloeg niet overal aan. "Het idee dat spraak gezien kan worden als ruizig verstoord signaal, waarbij je puur informatietechnische principes kunt gebruiken om de inhoud van het signaal, de inhoud van de boodschap te decoderen, dat heeft tijd gekost om door te dringen", beaamt ook Boves. Een vierde mijlpaal in 1984 illustreert dat.

Ruim een jaar voor het begin van ESPRIT, in april 1983, had de Nederlandse Organisatie voor Zuiver-Wetenschappelijk onderzoek (ZWO) op verzoek van het Ministerie van Onderwijs en Wetenschappen de werkgroep Taal- en Spraaktechnologie opgericht. Elf man sterk en onder leiding van Toon Cohen moest deze een voorstel doen voor een onderzoeksprogramma op het gebied van de TST. Twee van de elf werkgroepleden waren ingenieur, dan anderen hadden in grote meerderheid een taalkundige achtergrond. De werkgroep presenteerde zijn conclusies in het rapport Taal- en Spraaktechnologie in Nederland in, jawel, 1984 (Cohen et al. 1984). De taalkundige oriëntatie van de leden komt naar voren in de duidelijk uitgesproken wens om taalkunde een grotere rol te laten spelen in de spraaktechnologie. Zo pleit de werkgroep onder meer voor betere overdracht van onderzoeksresultaten vanuit de formele taalkunde en de fonetiek naar de taal- en spraaktechnologie, en verder voor stimulering van de spraakherkenning in het algemeen, en voor de ontwikkeling van een Nederlands dialoogsysteem in het bijzonder. Op dat dialoogsysteem zou Nederland nog ruim tien jaar moeten wachten. Maar de stimulering van spraaktechnologie in het algemeen zou snel gerealiseerd worden.

In 1985 start het ASSP programma, dat staat voor *Analysis and Synthesis of Speech*. Het onderzoek werd gefinancierd door het Ministerie van Economische Zaken en het Ministerie van Onderwijs en Wetenschappen en viel onder het *Stimulatie Project Informatietechnologie Nederland*. Het IPO, PTT Telecom (het dr. Neherlab), en de groepen van de universiteiten Utrecht, Leiden, Amsterdam en Nijmegen deden mee, en daarmee was het de eerste grootschalige samenwerking van alle spraakcentra in Nederland. "Dat was nieuw", beaamt Pols. "Het waren aparte koninkrijkjes, mensen die de noodzaak van samenwerking ook niet inzagen." Het ASSP-programma veranderde dat tot op zeker hoogte. Pols: "In dat ASSP was er wel samenwerking,

Zink Typografie Sheet 124 of 213 - Page 116 of 205

maar de activiteiten gebeurden wel grotendeels op het eigen instituut. Maar omdat het een geïntegreerd programma was met een heel aantal deelprojecten en een programmacommissie die daar toezicht op hield, werd het allemaal wel boven dat lokale uitgetild. Dat was een goede bodem voor samenwerking." Het project begint met slechts één commerciële partner: Philips. "Als het IPO niet gelieerd was geweest aan Philips, dan vraag ik me af of dat spraakprogramma ASSP überhaupt gefinancierd zou zijn geworden door Economische Zaken", aldus Boves. Later komt er echter een tweede partner uit de industrie bij: PTT Telecom. Voor het eerst werkten de onderzoekers van het Dr. Neherlab samen met mensen van buiten het instituut, ongetwijfeld gestimuleerd door Lou Boves, die daar inmiddels als consultant was begonnen. Het doel: de ontwikkeling van een laboratoriumsysteem voor hoogkwalitatieve text-to-speech conversie. Parallel aan elkaar werden in het project zowel de eerste Nederlandse difoonsynthesizer als een regelgebaseerde synthesemodule ontwikkeld. Naast de onderzoeksresultaten van het project (van Heuven & Pols 1993) had het ASSP ook een aantal positieve neveneffecten. Om te beginnen standaardiseerden alle partners hun hardware naar Digital VAX/VMS. Pols: "Toen ontstond ook de neiging om resultaten internationaal te toetsen, om te kijken wat je nou eigenlijk voorstelde. Dat was in het begin ook niet zo gebruikelijk." Een derde neveneffect was dat voor het ASSP project centraal spraakcorpora verzameld werden. Om ervoor te zorgen dat deze corpora beschikbaar zouden komen en blijven voor zowel kennisinstellingen als bedrijven, werd in 1987 SPEX opgericht, het *Speech Processing Expertise Centre*. SPEX was aanvankelijk gehuisvest in het Dr. Neherlab in Leidschendam. De doelstelling was "to develop and provide software, tools and databases for companies and institutes active in speech research and the development of speech technology and speech driven applications. SPEX has a special task in making available spoken language resources for research purposes in the Dutch academic environment."[5] Later is het expertisecentrum verhuisd naar Nijmegen, waar het nu nog steeds actief is, met name in de validatie van spraakcorpora.

ASSP had een duidelijke taalkundige inbreng: er was zelfs een werkgroep *Linguistic Analysis*. Dat paste in de wat afwijkende Neder-

[5] http://www.spex.nl

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



landse traditie van taalkundig geïnspireerd spraakonderzoek. Maar van samenwerking tussen taal- en spraak*technologen* was geen sprake, en zelfs de puur taalkundige inbreng binnen de spraak leek op zijn retour met de opkomst van continue spraakherkenning op basis van probabilistische modellen. Waarom trokken de beide teams niet meer met elkaar op? "Zou het niet puur historisch zijn?", vraagt Walter Daelemans zich af. "Tenminste in Vlaanderen zitten de spraakmensen bij de signaalherkenning dus bij de ingenieurs, bij de toegepaste wetenschappen. En computerlinguïstiek, taaltechnologie, zit bij de linguïsten. En dat zijn wel twee heel verschillende werelden, verschillende methodologieën, verschillende achtergronden." Hoewel het heel waarschijnlijk is dat dit verschil in achtergrond een rol gespeeld heeft, is het opvallend dat ook de Nederlandse spraakwetenschappers met een meer taalkundige achtergrond nauwelijks samenwerkingen aangingen met taaltechnologen. Boves redeneert vanuit de toepassingen: "Een van de redenen waarom taal- en spraaktechnologie lang uit elkaar gebleven zijn, is het feit dat simpele spraaktechnologieapplicaties eerder live gegaan zijn. Daardoor zijn mensen die met spraaktechnologie bezig waren eerder dan bij taaltechnologie gedwongen om zich vanuit een pure *engineering*-aanpak te richten op de problemen die je in de praktijk tegenkomt, in plaats van te bedenken wat er allemaal mooi en interessant zou zijn om uit te zoeken. In de tijd dat Eurotra alle kanten op ging en allemaal heel spannend onderzoek opleverde waar nooit iets van terechtgekomen is in termen van werkende systemen, bestond in de spraaktechnologie al heel duidelijk het besef dat je van alles op een hoog abstract niveau kunt gaan uitzoeken, maar dat de problemen die je op moet lossen om de technologie echt in te zetten andere problemen zijn, zoals onbegrip bij gebruikers, of de onvoorspelbaarheid van gebruikers." Zeker is dat spraaktechnologen, ongetwijfeld geholpen door hun meer technische achtergrond, sneller de omslag hebben gemaakt naar puur *data-driven* methodes. Boves: "De spraakwereld heeft het misschien makkelijker gehad dan de taalwereld, omdat – en dat is misschien de reden waarom het uiteindelijk alles bij elkaar redelijk snel is gegaan die omslag – mensen gedwongen waren zich te realiseren en toe te geven dat er zoveel variatie zit in spraak, tussen verschillende sprekers en verschillende condities, dat het hopeloos is om te denken dat je dat allemaal in nette regels kunt vangen. Taal heeft dat langer vol kunnen houden."

Zink Typografie Sheet 126 of 213 - Page 118 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Begin jaren negentig lijkt de tijd echter rijp voor nauwere samenwerking. Taaltechnologen beseffen dat probabilistische modellen op basis van taaldata hun ook wat te bieden hebben. Daarmee delen spraak- en taaltechnologen nu een gezamenlijk doel: het ontwikkelen, onderhouden en exploiteren van grote databestanden met taal- en spraakmateriaal. Daarnaast is nog een andere, veel verder gaande soort samenwerking denkbaar, namelijk een integratie van taalkennis in spraakherkenningssystemen. Zou het niet logisch zijn dat spraakherkenning verbetert wanneer het, net als mensen, meer kennis heeft over de taal dan alleen fonologische kennis? Als het weet welke structuren mogelijke taaluitingen zijn? Dit basisidee komt met enige regelmaat naar voren in de wereld van TST. Vaak geopperd door taaltechnologen. Maar ook spraaktechnologen hebben het geprobeerd, zelfs tot in de 21ste eeuw: Het FLaVoR-project (2002–2006) van de KU Leuven en de universiteit van Antwerpen was erop gericht om door middel van taalkundige kennis de *word error rate* van een spraakherkenningssysteem te verbeteren, door de mogelijke interpretaties van het signaal te *prunen*. "Het heeft niet echt iets opgeleverd", moet Daelemans concluderen. "Maar", voegt hij daaraan toe, "ook al helpt taalkundige kennis niet op dit moment, ze moeten het wel blijven proberen. Wat je niet wilt is dat het aparte werelden worden die elkaars problemen niet kennen. Ik denk dat het wel belangrijk is dat we pogingen blijven doen om die twee te integreren." De Nederlands/Vlaamse samenwerking tussen spraak- en taaltechnologen zal zich echter eerst richten op de ontwikkeling van resources en corpora. En daarnaast op de wens die al in 1984 werd uitgesproken, maar nog niet vervuld is: een Nederlands dialoogsysteem. Bij uitstek een toepassing waarin beide disciplines samenkomen, hoewel elk relatief onafhankelijk aan één stuk van het systeem kan werken. Meer over deze samenwerkingen in hoofdstuk 5. In het volgende hoofdstuk echter eerst meer over een factor die zowel taal- als spraaktechnologie heeft beïnvloed: de industrie rondom TST.

Zink Typografie Sheet 127 of 213 - Page 119 of 205

Jakub Zavrel (links achteraan), Walter Daelemans (tweede van rechts) en astronaut Dirk Frimout (midden achter) bij de Cam Brain Machine.

# 4 Industrie

Begin jaren zestig vertaalde Hugo Brandt Corstius op het Mathematisch Centrum zijn lettergreepprogramma in Algol. Een relatief kleine klus, want het programma had hij tijdens zijn studie al geschreven, en tijdens zijn verblijf in Berkeley al geïmplementeerd. Maar het leverde een goed werkend programma op. "Toen kwamen alle mogelijke kranten en uitgevers naar het Mathematisch Centrum om dat programma te kopen", herinnert Brandt Corstius zich. En dat plaatste het instituut voor een probleem. "Ze hadden daar nog nooit een programma verkocht, geloof ik. Niemand wist ook eigenlijk hoe dat zou moeten. Want als je een programma verkoopt, kun je dat dan ook nog aan een ander verkopen?". Een derde partij nam uiteindelijk de commerciële afhandeling over. Volgens Brandt Corstius was het Parool de eerste krant die het programma kocht en gebruikte. "Dan las ik 's avonds het Parool en dacht 'Goh, dat is verkeerd afgebroken'. Want er waren natuurlijk altijd fouten."

Het afbreekprogramma van Hugo Brandt Corstius was het eerste commerciële product dat door de Nederlandse Taal- en Spraaktechnologie (TST) gerealiseerd werd. Een primeur voor de taalkundigen, die in tegenstelling tot de ingenieurs nooit erg gericht waren geweest op de ontwikkeling van producten. En een product waar bedrijfjes nog

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



steeds geld mee verdienen, zoals het Nederlandse bedrijfje *TALŌ. Maar de mogelijkheden waren zowat onbegrensd: spellingcontrole, automatisch vertalen, vraagbeantwoording, automatische omzetting naar tekst van gedicteerde spraak en spraakgestuurde bediening van apparatuur. Toch duurde het nog jaren voor de industrie zich serieus begon te interesseren voor TST.

## 4.1 Vrijheid, blijheid

De beginjaren van de industriële taal- en spraaktechnologie in Nederland zoals we die tot nu toe gezien hebben, kenmerkten zich door onderzoek in grote laboratoria met relatief veel vrijheid. Toen Philips zich begin jaren zeventig als eerste begaf op het gebied van de taaltechnologie, deden ze dat ook meteen goed. Meeliftend op het optimisme over de digitalisering en de mogelijkheden van de computer, was niets te gek. Bunt over de begintijd in de fabriek van Philips Electrologica in Apeldoorn: "In die tijd dachten ze bij Philips nog dat de bomen tot in de hemel groeiden, en dat ze van computers ontzettend rijk zouden worden. Iedereen van wie ze dachten dat die wel slim was, kon een plaatsje krijgen." Dus toen Herman Schweigmann bedacht dat er ook iets aan Kunstmatige Intelligentie gedaan moest worden, kon hij daarvoor een groepje opzetten met onder meer het driemanschap Scha, Bunt en Landsbergen. En wie eenmaal werkte bij Philips, kreeg veel vrijheid en budget. "Meneer Schweigmann heeft dat [groepje] een beetje geleid, maar niet veel anders dan mensen enthousiast maken, en faciliteren. Er was veel geld, dus we werden met zijn drieën naar allerlei zomerscholen en conferenties gestuurd. En op ons advies werden ook wel mensen ingehuurd uit Amerika om hun kennis met ons te delen." Maar de tijden veranderden snel. "Dat heeft ongeveer anderhalf jaar geduurd", herinnert Bunt zich. "Toen heeft er een omslag plaatsgevonden bij Philips, omdat mensen zich realiseerden dat de computertak ontzettend veel geld aan het opmaken was, en dat er niet direct een perspectief was om geld te verdienen. Philips probeerde in die tijd op ieder niveau met IBM te concurreren. Ze ontwikkelden allerlei typen computers, van hoog tot laag om compatibel te zijn met alle soorten computers die IBM ontwikkelde, en ook nog een eigen

Zink Typografie Sheet 130 of 213 - Page 122 of 205

programmeertaal, een eigen alles. Ze dachten dat ze een complete tegenpool voor IBM zouden kunnen worden. Op zeker moment hebben een aantal mensen op het hoofdkantoor gedacht: dit is waanzin. Dit gaat helemaal niet goed." Bij de grote reorganisatie die volgde was de onderzoeksgroep het eerste slachtoffer: vrijwel al het onderzoek in Apeldoorn werd gestopt.

De taalgroep had geluk en werd gespaard. Wel moesten ze verhuizen naar het Natuurkundig Laboratorium in Eindhoven. Veel veranderde dat niet: de vrijheid bleef bestaan, en budget was er ook nog steeds. Maar het was onvermijdelijk dat ooit de vraag gesteld zou worden: "Kan Philips hier iets mee?" Jan Landsbergen: "Op het Natlab was het toen nog wel vrijheid blijheid, maar ooit komt natuurlijk die vraag. En het antwoord wist Philips zelf ook niet." Nu had Philips wel ervaring met onderzoek op gebieden waarvan het nog niet wist hoe het toegepast zou gaan worden. "Aanvankelijk wou Philips niet in computers", weet Landsbergen. "Daar was het niet in geïnteresseerd. Maar het Natlab was eigenwijs en vond dat we dat toch moesten doen. En op een gegeven moment sloeg het om, zijn ze die computerfabriek gaan maken. Toen was gelukkig het Natlab paraat met kennis op dat gebied. […] En dat gold ook hier wel voor. Natlab dacht: "Natuurlijke taal, dat is de toekomst, daar moeten we expertise in houden. En als dit niet de ideale toepassing is, dan is het wel iets anders." Maar geleidelijk kwam er dus druk om wat wat leukere toepassingen te vinden, en dat veel eigenlijk niet zo mee." Het lukte Landsbergen en Scha niet om de juiste toepassing voor PHLIQA te vinden, en het project werd ontbonden zonder ooit tot een product geleid te hebben. Wat restte was een filmrol met de demonstratie van het systeem.

De geschiedenis herhaalde zich op het Natlab met het Rosetta-project, dat Jan Landsbergen na afloop van PHLIQA opzette. Net als PHLIQA was het wel toepassingsgericht, maar kwam het nooit tot een product. Volgens Odijk is de reden dezelfde als bij het Europese vertaalproject Eurotra: "ook bij Rosetta waren de meeste medewerkers taalkundigen, en die waren vooral geïnteresseerd om hun eigen taalkundige theorie te implementeren en te kijken of die werkte. Ik ook hoor." Hoewel er een mooi demosysteem wordt gemaakt is Rosetta niet marktklaar. "En dat was uiteindelijk onze doodsteek, want als er geen product uitkomt, dan was Philips niet meer geïnteresseerd", aldus Odijk. Maar er waren ook andere redenen: "ze zagen dat MT onder-

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



zoek ook in Japan niet tot dramatisch goede resultaten leidde, daardoor was het concurrentiegevaar niet meer zo groot, en verdween het strategisch belang." Bovendien paste de vertaalsoftware niet goed in de productlijn van Philips. Landsbergen vond dan ook geen geschikte toepassing voor Rosetta. In totaal was er grofweg twintig jaar taaltechnologisch onderzoek gedaan bij Philips, zonder een verkoopbaar eindproduct op te leveren. En toen besloot Philips de handdoek in de ring te gooien. Rosetta werd stopgezet en daarmee het taaltechnologisch onderzoek van het Natlab. Hoewel? Sieb Nooteboom, inmiddels buitengewoon hoogleraar in Utrecht, ontdekte begin jaren negentig dat nog niet alle Rosettagelden die Philips aan de Universiteit Utrecht had betaald op waren. Nooteboom: "Dat geld moest natuurlijk verantwoord worden. Dus ik ben toen samen met Landsbergen gaan praten met de directeur van het Natlab. Ik heb daar nog een weddenschap op afgesloten met Landsbergen. Ik zei dat we het geld niet terug hoefden te geven, Landsbergen dacht dat dat niet zou kunnen. Maar Philips wist helemaal niet hoe dat moest, geld terugvorderen van zo'n regeling. Daar was geen proces voor. Dus uiteindelijk heeft die directeur ons gevraagd om een project te formuleren om dat geld op te maken."[1]

Philips had daarmee nog steeds niet helemaal gebroken met het taal- en spraakonderzoek in Nederland: aan het IPO, een samenwerkingsverband tussen Philips en de Technische Universiteit van Eindhoven, werd wel doorgewerkt aan TST-projecten. Een heel aantal van de Rosettamedewerkers verhuisden naar het IPO, dat mede door de relatie met de universiteit nog minder toepassingsgericht was dan het Natlab. In hoofdstuk 2 noemde Landsbergen het al een 'academisch clubje, dat meer gericht was op publiceren dan op bouwen'. Hier werden dan ook evenmin producten in de markt gezet. Onder meer Jan Landsbergen, Jan Odijk en Lisette Appelo werkten tussen 1993 en 1996 aan Dial-your-disc, een systeem dat de gebruiker moest helpen om via de telefoon muziek uit te zoeken. Odijk: "Het idee was dat iedereen over tien jaar zijn eigen cd zou samenstellen, en zou down-

[1] Nooteboom vermoedt dat met dat geld het boek over Rosetta geschreven is (Rosetta 1994), maar geeft aan dat niet zeker te weten. Odijk betwijfelt het. Hij geeft aan dat in ieder geval André Schenk na Rosetta een vervolgaanstelling kreeg in Utrecht, wat ook een mogelijke besteding van het overgebleven projectgeld zou kunnen zijn.

Zink Typografie Sheet 132 of 213 - Page 124 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



loaden via de telefoon. Achteraf gezien was dat een zeer vooruitziende blik. Omdat er veel muziek was, en het ging via de telefoon, en mensen niet precies zouden weten wat ze wilden, bedachten wij dat we een user interface moesten maken die ze zou helpen muziek te kiezen." Landsbergen is weinig enthousiast over deze samenwerking tussen taal- en spraaktechnologen: "Ik geloofde er niet in, maar heb dat nooit gezegd. Dat was eigenlijk niet helemaal eerlijk van mij." Misschien was had hij zijn twijfels wèl uit moeten spreken, want Landsbergen had het juist gezien. Rond 1996 werd duidelijk dat het downloaden van muziek via telefoonlijnen wel door zou zetten, maar dat het via internet zou gaan in plaats van via de telefoon. En daarmee kwam de nadruk te liggen op grafische interfaces, en boette spraakinterfaces in aan belang. Bovendien was directeur Herman Bouma, die de spraak- en taalgroepen welgezind was geweest, inmiddels vertrokken. Zijn opvolger Bösser was volgens Odijk meer gericht op user interfaces, en minder op taal. Daarbij moest er ook nog eens stevig bezuinigd worden. Verschillende taaltechnologen verdwenen al dan niet gedwongen naar banen elders, en het taaltechnologisch onderzoek aan het IPO ging als een nachtkaars uit.

Meer dan om het taalonderzoek, was IPO bekend om het spraakonderzoek. Cohen startte in 1959 de spraakgroep, waar onder meer ook Sieb Nooteboom, Rene Collier en Lei Willems aan spraaktechnologieprojecten werkten. Ook hier veel vrijheid voor de onderzoekers, zelfs als daarvoor de definitie van spraaktechnologie wat ruimer opgevat moest worden. Nooteboom, die in 1967 bij het IPO begon: "Wij deden het onderzoek waar we zin in hadden en dat was niet alleen spraaktechnologie. We hebben veel onderzoek gedaan naar woordherkenning, terwijl het niet duidelijk was dat dat ooit van belang zou zijn voor de spraaktechnologie." Sturing van het onderzoek was er feitelijk niet, De opvatting over onderzoek was in die tijd vergelijkbaar met die van het Natlab. "De filosofie was: op gebied waar je ooit iets mee wilde moest je goed onderzoek hebben, maar niet noodzakelijkerwijs productgericht onderzoek", aldus Nooteboom. "Wij weten niet op welke punten in de wereld er belangrijke technologische doorbraken gaan komen, maar als dat plaatsvindt moeten we mensen hebben die daar alles van weten, zodat we dan productgericht onderzoek kunnen gaan doen." Deze filosofie gaf Nooteboom de vrijheid om het onderzoek te doen dat hij het liefste deed, en daar heeft hij

Zink Typografie Sheet 133 of 213 - Page 125 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



optimaal van geprofiteerd, zolang het kon. Toen hij in 1988 vertrok, was de situatie voor de spraakgroep al aan het veranderen. "Het IPO was op zijn retour", zegt Nooteboom. "En aan de universiteit werd de situatie juist beter." Zelf betwijfelt hij overigens of die vrije aanpak nou de meest productieve aanpak was. "De afstand tussen onderzoek en ontwikkeling was veel te groot." Hoewel het patroon van het Natlab en het IPO erg op elkaar lijken, ging het er niet overal zo aan toen. Nooteboom: "Bij Siemens lagen research en ontwikkeling veel dichter bij elkaar." Overigens is het niet zo dat Philips geen enkel product heeft overgehouden aan het onderzoek van het IPO: in de jaren tachtig heeft het bedrijf enkele jaren de spraaksynthesechip MEA8000 en zijn opvolger de PCF8200 geproduceerd.

Het vrije onderzoek in de grote onderzoekslaboratoria liep eind jaren tachtig ten einde door tegenvallende resultaten en teruglopende financiële middelen bij Philips, de grote sponsor van dit onderzoek. Geen van de grote beloften, zoals vraagbeantwoording, automatisch vertalen en continue spraakherkenning, was ingelost. Ook een kleinere partij zoals het Bureau voor Systeemontwikkeling BSO, dat onder leiding van Toon Witkam jarenlang onderzoek had gedaan naar automatisch vertalen, slaagde er niet in om zijn onderzoeksresultaten aan de man te brengen. Net zo min als Océ van der Grinten erin slaagde om zelfstandig een systeem te ontwikkelen waarmee in gewoon Nederlands een documentenbank te bevragen was. Alice Dijkstra liep er stage in 1986/1987. "Ik heb daar een jaar lang heel leuk zitten hobbyen, maar het project was natuurlijk veel te ambitieus. Het moest alles kunnen." Later doet Océ nog mee in het SPIN-MMC programma, een gezamenlijk onderzoeksprogramma van de Katholieke Universiteit Nijmegen, de Universiteit van Tilburg, het IPO en Océ, dat 'iets' moest doen op het gebied van mens-machine communicatie. Maar Kempen, die namens Nijmegen in het project zat weet niet meer precies wàt dan. "Hoe ging dat ook al weer? Daar heb ik heel weinig aan teruggedacht." Iets met user interfaces, dat weet hij nog wel. "Maar dat moet je echt aan Harry Bunt vragen." Bunt op zijn beurt geeft aan dat hij daar "niet zulke scherpe herinneringen aan heeft." Océ was betrokken bij TRICS, een project gericht op het ontwikkelen van computerhulpmiddelen voor het opsporen van vertaalfouten, maar tot een product is het nooit gekomen. De tegenvallende resultaten van de brede programma's en het niet halen van de grote doelstellingen

Zink Typografie Sheet 134 of 213 - Page 126 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



temperden de verwachtingen over de resultaten van taaltechnologie wellicht, waardoor er ruimte ontstond om naar kleinere toepassingen te kijken. En dit leidden vaak wèl tot 'echte' producten.

## 4.2 Op zoek naar tastbaar resultaat

### Taaltechnologie

Waar Philips en BSO er niet in slaagden een product te maken op basis van het taaltechnologisch onderzoek, lukte dat Siemens wel: hun METAL was het enige vertaalsysteem met Nederlandse of Vlaamse onderzoeksinbreng dat werkelijk op de markt gekomen is. De technologie is later door het Duitse GMS en het in 1993 opgerichte Vlaamse Lant geëxploiteerd en leeft tot op de dag van vandaag voort in Xplanation, waarin Lant is opgegaan. Ook in Nederland kwam er een vertaalproduct op de markt, een product dat sterk lijkt op het systeem dat Jan Landsbergen volgens Jan Odijk met Rosetta oorspronkelijk voor ogen had: "Jan had eigenlijk met Rosetta een heel ander systeem willen maken. Hij wilde een systeem bouwen dat wat betreft het woordenboek heel rijk was, en wat regels betreft heel arm was, zodat het in eerste instantie een vertaalhulp zou zijn, waarmee je frasen kon laten vertalen." Hoewel Odijk zich destijds meer interesseerde voor het automatisch analyseren van taalkundig interessante fenomenen, noemt hij die meer beperkte focus nu een verstandig idee: "dat had veel meer kans om tot een product te leiden, dat uiteindelijk door de business gesupport zou worden." Rosetta ontwikkelde zich echter onder invloed van de vele taalkundigen in het team in een andere richting: het systeem ging zich meer en meer richten op het vertalen van lastige taalkundige constructies, zoals de verschillende gebruiken van *er*. Leo Konst, die in de groep van Kempen werkte, sprong in het gat. In 1985 richtte hij het bedrijf Linguistic Systems op en bracht hij Euroglot op de markt, een vertaalhulpmiddel dat woorden en frasen vertaalt van en naar Nederlands, Engels, Duits, Frans, Spaans en Italiaans en naar wens aangevuld kan worden met specialistische terminologie voor een aantal branches. Het bedrijf bestaat nog steeds.

Ook woordenboekmaker Van Dale zag mogelijkheden. Van Dale

Zink Typografie Sheet 135 of 213 - Page 127 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



had een aandeel in verschillende onderzoeksprojecten. Zo waren ze betrokken bij automatisch vertalen, bijvoorbeeld bij Rosetta, dat probeerde met de Van Dale woordenboeken de lexicale informatie van het systeem uit te breiden. Maar dat was lastiger dan gedacht. In hoofdstuk 2 vertelde Odijk al over de 32 verschillende coderingen van werkwoordsvervoegingen met *hebben* en/of *zijn* – een triviaal voorbeeld dat wel illustreert dat de woordenboeken niet zomaar geschikt waren voor gebruik in automatische natuurlijke-taalverwerking.. Maar taaltechnologie kon ook gebruikt worden om lexicografische producten voor Van Dale zelf te ontwikkelen. Een voorbeeld daarvan was Lexitron, dat door de groep rondom Gerard Kempen in Nijmegen ontwikkeld is. Kempen: "Van Dale had via Spectrum een heel groot kennisbestand met feitjes over allerlei begrippen. Een soort encyclopedie, maar dan niet met hele lappen tekst, maar met kleine, losse feiten. Iemand was al lang bezig geweest daar bij Spectrum, omdat op een databankachtige manier te coderen. […] Op een gegeven moment heb ik bedacht – dat moet in 1996 geweest zijn – om daar een groot elektronisch woordenboek van te maken, zoals dat in Frankrijk gebruikelijk is, waar woordenboek en encyclopedie een geheel is. Wij hebben daarvoor altijd twee aparte boeken, maar in Frankrijk voegen ze die samen. En mijn idee was om dat uit te brengen op cd. Dat heette Lexitron. Die naam heb ik bedacht." Lexitron werd onder meer verkocht aan de redactie van de omroep Veronica. Maar een echte hit is het niet geworden. Kempen denkt dat dat te maken heeft met de prijs. "Dat kostte 1200 gulden. Een belachelijk hoog bedrag. En dan moest je ook nog een cd-romlezer kopen, want die had niet iedereen." Drijvende kracht vanuit Van Dale was Bernard Al, die we eerder tegenkwamen als lid van de toezichtcommissie van BSO. "Een man met veel commercieel gevoel", zegt Kempen over Al. Lexitron was niet het enige product dat Van Dale in samenwerking met de onderzoekers uit Nijmegen ontwikkelde. De groep in Nijmegen richtte zelfs een bedrijfje op voor deze meer commerciële activiteiten: Cognitech. Binnen dit bedrijf werd ook Spelraam ontwikkeld, een hulpmiddel voor basisschoolleerlingen om te leren spellen. Ook weer vrij duur, vindt Kempen, voor 500 gulden. Naast productontwikkeling deden ze ook opdrachten direct voor Van Dale. En dan komen weer de inconsistenties in papieren woordenboeken aan het licht. "Ik weet dat wij als opdracht kregen om alle woordvormen uit te

Zink Typografie Sheet 136 of 213 - Page 128 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



rekenen van alle woorden die in Van Dale staan. Dus alle verbuigingen en vervoegingen. […]. Ik kan me herinneren dat er soms rare dingen in van Dale stonden. Dat zie je dan met die programma's. We lieten die op een toenmalige IBM pc 's nachts doordraaien. Op een ochtend hoorde ik van medewerker Edwin Bos dat hij 's nachts gestopt was. Wat bleek, het woordje 'hall' was verkeerd gegaan: de tweede 'l' in hall, dat was geen 'l' maar het cijfer '1'. Dat was nooit iemand opgevallen. Misschien heeft daar wel jarenlang een '1' gestaan in plaats van een 'l'."

De goede connecties met Van Dale legden de taaltechnologen uit Nijmegen geen windeieren. Via via hoort de groep dat een of ander Frans kaasje een prijsvraag heeft uitgeschreven. De opdracht was om met de letters van een aantal vooraf gegeven woorden zoveel mogelijk nieuwe woorden te vormen die iets maken hadden met kaas. En, zo schreef het reglement voor, alleen woorden die in Van Dale stonden, zouden geaccepteerd worden. De groep schreef snel een programmaatje om op basis van de digitale bestanden van Van Dale, waarmee de groep al zo vaak gewerkt had, anagrammen te maken. Het programma spuugde in no time een enorme lijst mogelijke woorden uit. Maar toen moest nog wel bekeken worden of ze iets met kaas te maken hadden. "Maar mijn vrouw zit in de voedingsmiddelenindustrie, die weet veel van voeding af", vertelt Kempen. "Dus onder haar leiding hebben zij en mijn oudste dochter die woorden beoordeeld op kaasachtigheid." Lange tijd hoort de groep niets meer. Maar dan komt het bericht: wilt u de hoofdprijs, een auto, komen ophalen? Dat heuglijke nieuws stelde de immer voorzichtige Kempen wel voor een dilemma. Want was het eigenlijk wel helemaal reglementair wat ze gedaan hadden? "Ik wilde niet dat ze op dat moment in de gaten zouden krijgen dat we het op deze manier elektronisch gedaan hadden." En dus deden vrouw en dochterlief net alsof ze de woorden met behulp van wat andere mensen gevonden hadden – wat strikt genomen ook zo was. "Zij zijn toen naar Delft gegaan om die auto op te halen, en die hebben ze meegekregen." Een van de medewerkers heeft de auto voor een zacht maar redelijk prijsje overgenomen. Het geld is netjes overgemaakt naar de universiteit, minus een donatie aan Van Dale, dat tenslotte de woordenboeken geleverd had en dat hiervan het jaarlijkse feest in Hotel De Dikke van Dale in Sluis wat extra opluisterde.

De universiteit Nijmegen bracht nòg een ondernemende taaltech-

Zink Typografie Sheet 137 of 213 - Page 129 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



noloog voort. Begin jaren negentig was Theo van de Heuvel onderzoeksassistent aan de universiteit, waar hij onder meer tools bouwde voor de syntactische analyse van tekstcorpora. In 1993 begon hij echter met het taaltechnologiebedrijf Polderland Language & Speech Technology, kortweg Polderland. Het bedrijf leverde als *Original Equipment Manufacturer* (OEM) taaltechnologische (deel)producten aan derden: producenten van softwarepakketten bouwden Polderlands taaltechnologie in en verkochten het onder eigen naam. In 1996 kreeg het bedrijf een enorme deal: het mocht de Nederlandse spelling- en grammaticacontrole leveren die Microsoft in Word gebruikte. Het bedrijf specialiseerde zich in de minder grote talen, en leverde zo ook bijvoorbeeld spellingcontrole voor IJslands en Catalaans aan Microsoft. In 2009 zouden de activiteiten van Polderland overgenomen worden door Knowledge Concepts, dat sinds 1998 taaltechnologie ontwikkelde, met name voor de zoekmarkt. Een van de eerste spelers in deze markt was C-content, dat zich zelfs al sinds 1987 richtte op het efficiënt ontsluiten van content.

Ongeveer even oud is het Arnhemse bedrijf Human Inference. "Iedereen noemde zijn bedrijf in die tijd compu-dit of data-dat. Wij wilden wat anders", vertelt mede-oprichter Norbert Mergen. De naam ontstond onder de douche, na een wedstrijdje squash met zakenpartner Jan Arends. Het menselijke aspect moest erin zitten, en het vermogen om conclusies te trekken uit data, zoals mensen dat ook kunnen. "Het was wel een moeilijke naam, maar daardoor onthielden mensen het ook." De gegevens waarmee geredeneerd moet worden zijn in het geval van Human Inference voornamelijk relatiegegevens. Mergen en Arends werkten midden jaren tachtig op de automatiseringsafdeling van Goudse Verzekeringen aan een nieuw motorrijtuigenverzekeringssysteem. De automatisering daarvan was voor Mergen een weinig boeiende klus, behalve één aspect ervan. "Onderdeel van dat project was de zwarte lijst. Daarin zijn de namen opgenomen van zo'n 350.000 mensen die een onaanvaardbaar risico vormen. Dat zijn wanbetalers, mensen met een veroordeling, die een rijontzegging hebben gekregen, en mensen die eerder wegens een groot ongeval gekort zijn in hun bonus/malus." Deze gegevens worden door alle verzekeraars gedeeld, en wie op de zwarte lijst staat, wordt niet zomaar geaccepteerd voor een nieuwe verzekering. "Maar wat doe je als je weet dat je op die lijst staat?", vraagt Mergen zonder antwoord

Zink Typografie Sheet 138 of 213 - Page 130 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



te verwachten. "Dan maak je 'per ongeluk' een tikfout in je naam of geboortedatum, en dan vindt de medewerker je niet in het systeem, dat nog met microfiches werkte." Dat moest slimmer kunnen. Hoewel beter zoeken geen onderdeel uitmaakte van het automatiseringsproject dat ze deden, besloten Mergen en Arends het 'ernaast' te doen. De twee ontwikkelden een systeem waarmee de waarschijnlijkheid berekend kon worden dat twee setjes persoonsgegevens (voornaam, achternaam, adresgegevens, geboortedatum etc.) van dezelfde persoon waren. Volgens Mergen leverde het eerste systeem 50% meer hits op en kostte dat de helft minder zoektijd. Een van de technieken die ze gebruikten was fonetische representatie van onder meer namen. "Ik had toevallig wel eens van Soundex gehoord. Maar het algoritme voor fonetische representatie van woorden bleek niet direct toepasbaar, aangezien het erg op het Amerikaans Engels geënt was." En zo rolde de automatiseerder in de taal- en spraaktechnologie. "Alfa en beta waren bij mij altijd al redelijk in balans", verklaart Mergen. "Ik schreef vroeger ook al voor studentenbladen."

Arends en Mergen besloten dat ze met hun systeem de markt op wilden. Hun werkgever bood een uitweg: als zij gratis gebruik mochten blijven maken van de software, dan mochten Arends en Mergen het eigendomsrecht meenemen èn gebruik blijven maken van de IBM mainframe computer. Zoals het een goede start-up betaamt, begonnen de twee op 1 april 1986 vanuit de zolderkamer van Jan Arends. Ze hadden 40.000 gulden startkapitaal. "Dat had je nodig om een BV te starten", legt Mergen uit. Maar de bank wilde geen lening geven voor een bedrijf. Dus leenden de beide oprichters elk 20.000 gulden, zogenaamd om een auto te kopen. "Als automatiseerders hadden we best een goed salaris, dus we konden gemakkelijk een persoonlijke lening krijgen." Het begin was lastiger dan gepland. "We dachten dat alle verzekeringsmaatschappijen *en bloc* voor onze technologie zouden vallen", vertelt Mergen, "Maar verzekeraars zijn bijzonder langzame beslissers." Daarbij hadden Mergen en Arends besloten dat ze hun software niet wilden verkopen, maar verhuren. "Dat zou later een gouden zet blijken, maar initieel was het doorbijten, omdat je niet alle omzet meteen kunt nemen." Het ronselen van klanten ging zó traag, dat de oprichters moesten nadenken over een plan B. "We zaten in het rokerige kamertje dat ons eerste kantoor vormde – we rookten toen beiden als ketters", herinnert Mergen zich. "En we stelden elkaar de

Zink Typografie Sheet 139 of 213 - Page 131 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



vraag: 'Ga jij weer consultancy doen of ik?'. Met onze automatiseringsachtergrond konden we onszelf natuurlijk gewoon weer als consultant verhuren om brood op de plank te krijgen." Maar volgens de overlevering belde precies op dat moment uitgeverij Wolters-Noordhoff met de mededeling dat het bedrijf de software van Human Inference wilde hebben. En vanaf dat moment gaat het bergop met het bedrijf uit Arnhem, dat later ook andere bewerkingen van relatiegegevens gaat doen: hoofdletterconversie, normalisatie van diakritische tekens, doublure-identificatie en inkortingen van complexe namen. Het bedrijf heeft anno 2010 zo'n tachtig man in dienst. De beide oprichters verlieten het bedrijf eind 2003. Mergen verkoopt nu onder meer pure chocolade via internet.

Ook de Vlamingen ontdekten eind jaren tachtig en begin jaren negentig het commerciële potentieel van relatief simpele taaltechnologische producten. Sommigen begonnen klein en zelfstandig, en bleven dat door de jaren heen, zoals Filip de Brabander en zijn Natlanco. Andere kwamen terecht in een opeenvolging van overnames en fusies. Jean-Paul Janssens-Lens, beter bekend als Paul Janssens, volgde net als Frank Van Eynde het vak computerlinguïstiek bij Dirk Geens in Leuven, en maakte deel uit van het Vlaamse Eurotrateam. In 1983 richtte hij samen met zakenpartner Rudy Montigny het bedrijfje Soft-Art op. Eerste product? Een algoritme voor het afbreken van lettergrepen. In ditzelfde jaar wees Frank Van Eynde zijn student Jan van Sas, die net zijn licentiaat afgerond heeft, op het bedrijfje van zijn oud-studiegenoot. "Ik heb een brief naar hen gestuurd, want email had ik nog niet", vertelt Van Sas, "en toen kreeg ik een telefoontje om de volgende keer dat Montigny in België was langs te komen. Dat heb ik gedaan. Hij was twee weken in het land en de eerste week ontmoetten we elkaar en hij zei: 'laat maar eens zien wat je kan. Kies zelf maar iets.' " Van Sas koos voor samenstellingen in het Nederlands. In een week schreef hij een demonstratieprogrammaatje, in de programmeertaal Pascal, waarvoor hij een illegaal gekopieerde compiler had bemachtigd. "Op die manier leerde ik ook Pascal een beetje beter kennen, met al die libraries." En zo heb ik ook een library geschreven die met samenstellingen van alles kon doen, en ik heb daar ook een user interface rond gebouwd, zodat je kon testen. Je tikte het woord in en dan begon die dat te analyseren op basis van de woordenboeken en wat simpele regeltjes. En als hij het woord niet kende, dan had je

Zink Typografie Sheet 140 of 213 - Page 132 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



de optie om het systeem wat woorden bij te leren." Een week lang werkte Van Sas onafgebroken, want hij wist dat er veel van af hing. "De vrees die bij Montigny leefde – terecht – was dat iemand die computerlinguïstiek had gedaan helemaal niet kon programmeren. Dan kom je in een bedrijf en zijn ze eerst maanden bezig om jou te leren programmeren. Dat wilde hij niet. Hij wilde iemand die meteen aan de slag kon." Van Sas slaagde voor de test en werkte eerst op het hoofdkantoor op Marco Island in Florida en later op het kantoor in Antwerpen voor Soft-Art. Het bedrijf verkocht afbreekprogramma's en spellingcorrectie voor de meeste Europese talen. Dit was een lucratieve markt, omdat er nog een heel aantal verschillende tekstverwerkingsprogramma's in omloop was. Aan elk van deze programma's konden de licenties tegen forse prijzen verkocht worden. Van Sas: "Op een goede deal kwamen miljoenen en miljoenen binnen. Daar konden ze heel wat volk van bekostigen." Er werkten ongeveer 25 man bij Soft-Art, waarvan slechts een enkeling in een commerciële rol. Een aantal mensen programmeerden, maar er waren vooral veel taalkundigen en studenten voor wat Van Sas het 'vuile werk' noemt. "Soft-Art had een Nederlandse spellchecker, maar die was heel beperkt. […] Wat ze deden was – ik weet niet hoe wettelijk dat was – ze kochten woordenboeken, Verschueren, van Dale, Koenen, allerlei woordenboeken. En dan lieten ze iemand goedkoop data intikken van het hard copy. Het hele Van Dale Nederlands. Van Dale was een heel goede bron, want die had een codering voor inflectie, voor een heel aantal talen. En dat was een heel goede basis om dan die morfologiecodering op te baseren. Het is niet zo dat ze de hele Van Dale kopieerden. Ze namen de entry-lijst. En daarna gingen ze hetzelfde doen met de Verschuren, de Koenen, hindert niet wat, zelfs de Prisma's." Toen het bedrijf tapes had van woordenboeken, schreef Van Sas programmaatjes om die informatie te parseren. Maar hij werkte vooral aan een uitbreiding van het kernsystem voor het Fins, dat met zijn meerdere lagen van infixen niet door het oude systeem verwerkt kon worden.

Montigny verruilde Soft-Art voor Reference Software, dat in 1985 van Aspen Software de allereerste grammaticachecker Grammatik had gekocht. Van Sas volgde hem. Het team in Antwerpen, waar ook David van den Akker en Peter de Bie deel van uitmaakten, verbeterde het programma van een simpele patroonherkenner tot een echte grammaticachecker in een range van talen. De activiteit rondom spel-

Zink Typografie Sheet 141 of 213 - Page 133 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



lingcorrectie en grammaticacontrole in Antwerpen paste binnen de dynamische markt van tekstverwerking. Er waren een fors aantal partijen op de markt, waarvan Wordperfect (WP) en Microsoft Word de grootste spelers waren. In Nederland had WP een zeer groot marktaandeel, in vele andere landen was Word al marktleider. Van Sas: "het succes van WP in Nederland was dat ze het gratis uitdeelden aan universiteiten, al dan niet legaal. Dat werd zo oogluikend gestimuleerd, gebruik maar aan. En wat gebeurde er: een paar jaar later gingen die mensen werken, en die wilden werken met WP. En in die bedrijven moest het legaal zijn. Dat heeft een enorme boost aan de verkoop van WP gegeven." Het was deze strategie van Nederlander Ad Rietveld, die hem groot maakte en zelfs de functie van CEO van Wordperfect opleverde. Schrijfhulpmiddelen vormden belangrijke ammunitie in de oorlog tussen WP en Word. Peter de Bie: "Dat heeft een grote boost gegeven aan de ontwikkeling van taalspecifieke software." De tekstverwerkers betaalden license fees voor de hulpmiddelen, en verkochten ze vervolgens in aparte dozen. De Bie: "In die tijd kocht je voor je computer een blauwe doos, waar WordPerfect in zat, en een gele doos, waar de taaltools in zaten." Hier kwam een einde aan toen Microsoft in 1992 een concurrerende grammaticachecker opkocht en als feature meeverkocht bij Word. Wordperfect kon toen niet achterblijven en kocht Reference Software. "Het moederbedrijf was zoveel geld aan licenties kwijt dat het goedkoper was om ons te kopen", aldus Rudy Montigny in Computable.[2] Voor de groep in Antwerpen veranderde er niet zo veel, omdat alleen het administratieve en commerciële personeel ontslagen werd. WP wilde zelfs graag verder ontwikkelen, en het team groeide tot zo'n zestig mensen, waarvan er ongeveer dertig in Antwerpen werkten. Van Sas: "Er was geld om in twee jaar tijd een Nederlandse en Spaanse versie erbij te bouwen, daarna zouden Italiaans en de Scandinavische talen volgen. Het was een goede gelegenheid om veel code weg te gooien en opnieuw te schrijven, op een manier die taalonafhankelijker was. Dat was mijn werkgebied." Ook toen WP failliet ging en in 1994 door Novell overgenomen werd, bleef de ontwikkeling aan taalhulpmiddelen doorgaan. De eerste Nederlandse

[2] http://www.computable.nl/artikel/ict_topics/loopbaan/1293278/1458016/herkenning-en-synthese-moeilijke-maar-onbegrensde-spraakmogelijkheden.html, op 26-11-1999.

Zink Typografie Sheet 142 of 213 - Page 134 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



grammaticachecker kwam op de markt, volgens Van Sas een van de slechtsten van Reference. "Gewoon omdat het een allereerste versie was." Het spellingsgedeelte was volgens Van Sas wèl goed. En toen kwam in 1995 de spellingsherziening waarin 'pannekoek' veranderde in 'pannenkoek'. WP had een overeenkomst met Van Dale waarmee het bedrijf het recht kreeg om de woordenboeken te gebruiken, in ruil voor feedback over gevonden fouten en omissies, en morfologische informatie van alle woordenboekingangen. Van Dale was niet blij met de spellingsherziening, vooral omdat de commissie die erover besliste wel samenwerkte met het Groene Boekje, maar Van Dale niet goed op de hoogte stelde van de nieuwe regels en uitzonderingen. Toen besloot Van Dale om rigoureus de regels toe te passen. Daarmee paste het woordenboek de regels dus strikter toe dan het Groene Boekje, dat enkele bijzondere uitzonderingen bevatte. "Dat was een probleem voor Novell om hun officepakket verkocht te krijgen aan Vlaamse overheidsinstellingen", weet Van Sas, "want wettelijk was vastgesteld dat ze nieuwe spelling moeten volgen. En dat was een heel grote klant." Dus klopt Novell aan bij de taaltechnologiegroep van Wordperfect om – tegen de afspraken met Van Dale in – toch een versie van de spellchecker te maken die voldeed aan het Groene boekje.

Novell is in deze tijd betrokken bij de ontwikkeling van Eurowordnet, een grote database met semantische relaties tussen concepten voor verschillende Europese talen. Dit Eurowordnet werd opgezet naar het voorbeeld van het al bestaande Amerikaanse Wordnet voor Engels, en hieraan gelinkt. Het bedrijf was geïnteresseerd in toepassing van de multilinguale semantische database voor *cross-lingual information retrieval*, het vinden van informatie in teksten die in vreemde talen geschreven zijn. Onderzoeker Piek Vossen van de Universiteit van Amsterdam initieerde en coördineerde het project.

Voorlopig kwam er nog geen einde aan de overnames voor de groep taaltechnologen in Antwerpen: met Novell ging het slecht, het bedrijf werd overgenomen door Corel en moest bezuinigen. In 1998 stapt de WP-groep uit het concern en wordt onderdeel van het snel groeiende Vlaamse taal- en spraaktechnologiebedrijf van Jo Lernout en Pol Hauspie. Lernout was zijn carrière begonnen bij Wang Laboratories, producent van tekstverwerkers en minicomputers, en een van de eerste klanten van Montigny.

Zink Typografie Sheet 143 of 213 - Page 135 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



## Spraaktechnologie

Philips had in de jaren tachtig met de op het IPO ontwikkelde spraaksynthesechips MEA8000 en PCF8200 de eerste Nederlandse spraakproducten. En vanuit het IPO gebeurde er meer. Arthur Dirksen, die van 1990 tot 1996 bij het IPO werkte, richtte in 1997 samen met collega Ludmila Menert het bedrijfje Fluency op.[3] Een jaar later bracht het bedrijf zijn eerste product op de markt: Fluent Dutch text-to-Speech, dat later Fluent TTS genoemd zou worden. Het bedrijf werd in 1999 ingelijfd bij Van Dale, maar zou vanaf 2004 weer zelfstandig doorgaan. Naast Fluent TTS bracht het ook Spika op de markt, een hulpmiddel voor mensen met dyslexie, dat geselecteerde stukken tekst voorleest.

Ook het Vlaamse Jabbla was actief op de hulpmiddelenmarkt. Het bedrijf was halverwege de jaren tachtig ontstaan vanuit de spraaktechnologiegroep van de Universiteit Gent, waar onder meer Jean-Pierre Martens werkte. Onder de naam Technology and Integration ontwikkelde het bedrijf met de Gentse spraaktechnologie toepassingen in de zorg. In 2008 is het omgedoopt tot Jabbla, en inmiddels gebruikt het bedrijf de spraaktechnologie van Nuance, maar net als toen leeft het bedrijf nog steeds van spraaktechnologie. Het bestaan van Jabbla is voor Martens reden tot tevredenheid. "Het feit dat Jabbla bestaat, dat is een rechtstreeks gevolg van de technologie, en dat ik daarin geïnteresseerd was. […] Voor mij is het genoegdoening dat Jabbla bestaat. Dat dat voortgevloeid is uit dat onderzoek. Het is niet zo dat wij direct voor dat bedrijf onderzoek gedaan hebben, maar het feit dat er activiteit was aan de universiteit op dat vlak, het feit dat er interesse was vanuit de universiteit, het feit dat er een vraag was vanuit de privesector om iets te doen, dat is een samenloop van omstandigheden die gemaakt heeft dat Jabbla is ontstaan in Vlaanderen."

In de loop van de jaren negentig kwamen er verschillende spraakherkenningspakketten op de markt, waaronder die van het Nederlandse Philips, dat echter niet in Nederland ontwikkeld was. Ook Lernout en Hauspie, Dragon en IBM brachten pakketten op de markt. In het kielzog van deze technologieleveranciers ontstonden applica-

[3] http://www.fluency.nl

Zink Typografie Sheet 144 of 213 - Page 136 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



tieontwikkelaars die toepassingen maakten met de basissoftware, en serviceleveranciers, die bedrijven hielpen met het integreren van de software in de organisatie. Soms ook begon een bedrijf als softwareontwikkelaar, maar stapte het uiteindelijk over op een pakket van een van de grote leveranciers, en ging het bedrijf verder als applicatie- of serviceprovider, zoals Speechworks[4] en Kompagne.[5]

De bekendste in deze categorie is Telecats – waarbij 'cats' staat voor Computer Automated Telephone Systems. "Maar als we bedrijven belden, dan verstonden ze vaak 'Telekids' ", vertelt mede-oprichter Martin Spitholt. "Dat lieten we maar zo, want daardoor werden we wèl meteen doorverbonden met de directeur." Martin Spitholt, Wim Luimes en Jan Willem Mulder richtten in 1992 het bedrijf op.[6] Spitholt en Luimes schreven samen aan een proefschrift, en hadden daarnaast een adviesbureau opgezet. Studentassistent Mulder werd daar ook ingezet wanneer een opdracht erg technisch werd. Via een opdracht voor een marktonderzoek naar een voicemailsysteem kwamen ze in aanraking met telefonie. De opdrachtgever ging weliswaar failliet, maar het idee om met druktoetstechnologie telefoonprocessen te gaan automatiseren was geboren. En toen een uitgewerkt plan meteen al drie prospects opleverde, hebben de drie het er maar op gewaagd. De tijden waren gunstig. Eind negentiger jaren profiteerde het bedrijf volop van de technologiehype. Amerikaanse bedrijven die de Europese telecommarkt op wilden, konden bij Telecats de infrastructuur kopen voor een prepaid-kaartensysteem voor vaste telefonie. Spitholt: "In die tijd was niets te gek. We vlogen eerste klas naar Nice, waar we met een helikopter werden opgehaald om in Monaco naar een beurs te gaan. Daar reden we in een limousine, en gingen we 's avonds in smoking naar het casino. En de aandeelhouders vonden dat normaal!" De nuchtere Tukker genoot van die gouden tijden. "Ik realiseerde me toen niet dat dat niet zo kon blijven", zegt hij nu. Maar eind jaren negentig ging het toch mis. De bedrijven waar Telecats zaken mee zou doen vielen in rap tempo om. "Dat waren rampjaren voor ons. Eigenlijk

---

4 `http://www.speechworks.nl`, een pure serviceverlener, niet te verwarren met de technologieontwikkelaar die opging in Scansoft (later Nuance).

5 `http://www.kompagne.nl`

6 Formeel is het bedrijf opgericht op 1 april 1993, maar toen waren de oprichters feitelijk al enkele maanden bezig.

Zink Typografie Sheet 145 of 213 - Page 137 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



zijn we in 1999 opnieuw begonnen", stelt de directeur. "We hadden genoeg vet op de botten om het verlies te lijden. Maar ik heb toen pas geleerd wat ondernemen is. Geleerd om hard te zijn. Als iemand de rekening niet betaalt, gaat de stekker er gewoon uit. Hoe lief en aardig hij ook is." Nog voor de crisis, zo rond 1996, benaderde de Duitse vliegtuigbouwer DASA Telecats om samen een ticketreserveringslijn te bouwen voor bioscoopketen Cinemaxx. DASA had de spraakherkenning die zij in hun vliegtuigen gebruikten overgezet naar een pc, en hiermee werd de dienst opgezet. "Later heeft Philips ons benaderd met de boodschap dat het toch echt niet kon dat een Nederlands bedrijf in Duitsland projecten deed met Duitse technologie", aldus Spitholt. Vanaf dat moment was het voornamelijk Philips' technologie wat Telecats gebruikte. Hoewel het bedrijf van Spitholt en consorten zelf geen herkenningssoftware ontwikkelde, was het wel regelmatig betrokken bij onderzoeksprojecten waarin nieuwe toepassingen van de technologie ontwikkeld werden. Zo huurde Philips Telecats in voor hulp bij het OVIS-project uit 5.1, gericht op een dialoogsysteem voor informatie over openbaar vervoer.

Een belangrijke concurrent van Telecats is Comsys, dat al in 1983 werd opgericht. Ooit spraken de twee bedrijven over een fusie, maar aangezien die niet doorging, bleven de beide partijen elkaar in de markt beconcurreren, zowel wat klanten als wat werknemers betreft. Zo regelde Telecats busreclame met de openstaande vacatures bij Telecats op de buslijn die langs het kantoor van Comsys kwam. Ondermeer spraaktechnoloog en Telecats' 'Head of imagination' Arjan van Hessen kwam vanuit Comsys bij Telecats terecht. Ook bedrijven als het Vlaamse Voxtron en Voice Insight bouwden applicaties met spraakherkenningstechnologie. Vaak specialiseerden bedrijven zich op een specifieke applicatie of bedrijfstak, zoals G2 Speech en Sensotec, die zich beide exclusief op de medische wereld richtten. En dat gold niet alleen voor spraaktechnologiebdrijven: ook taaltechnologiebedrijf Language and Computing – dat later naar de Verenigde Staten verhuisde en uiteindelijk in 2010 door Nuance overgenomen werd – specialiseerde zich in medisch taalgebruik.

Ook PTT/KPN Research richtte zich voornamelijk op het ontwikkelen van spraakapplicaties. Toen internet-entrepeneur Stan van de Burgt in 1987 bij het Dr. Neherlab in Leidschendam aan de slag ging

Zink Typografie Sheet 146 of 213 - Page 138 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



was dat echter niet op het gebied van taal- en spraaktechnologie: "Er was wel een spraakgroep, maar ik werkte zelf aan internetprotocollen", aldus Van de Burgt. Dat veranderde in ieder geval gedeeltelijk halverwege de jaren negentig, toen hij samen met een collega een voorstel indiende voor een onderzoek naar het omzetten van (geschreven) natuurlijke taal naar een databasequery, een onderwerp dat direct aansloot bij de afstudeerscriptie die hij jaren eerder bij Anton Nijholt in Enschede geschreven had. Het ging om een zogenaamd RvB-project "Het meeste onderzoek werd direct gesteund vanuit de werkmaatschappijen. Maar er was een kleine hoeveelheid geld waar de Raad van Bestuur (RvB) over besloot. Dat was wat minder toegepast." Dit project vormde het begin van een Taalgroepje, dat na een reorganisatie samengevoegd werd met de Spraakgroep. Van een echte integratie was geen sprake. "De spraakgroep had zijn handen vol aan het sprekeronafhankelijk herkennen van de getallen nul tot en met negen en 'ja' en 'nee', zij hadden geen interesse in taalproblemen." De twee blijven dan ook los van elkaar opereren, waarbij Taal alleen interne projecten deed, terwijl Spraak participeerde in grote Europese projecten, onder meer door de inspanning van projectmanager Els den Os.

Rond 1997 kwam Den Os, die eerder bij datacentrum SPEX had gewerkt als projectmanager, naar het lab in Leidschendam. De groep Taal en Spraak van KPN Research was toen een man of tien groot, en groeide daarna nog iets. Ze hadden een eigen studiootje, waar spraakopnames gemaakt konden worden voor de ontwikkeling van demo's en diensten. "We ontwikkelden zelf geen taal- en spraaktechnologie", vertelt Den Os. "Maar we probeerden diensten op te zetten met bestaande technologie. Voicedialing was daar een voorbeeld van. Daar ging echt heel veel geld in om. En we ontwikkelden een sprekerverificatiedienst voor slechtzienden. KPN bood die dienst aan. Blinden konden gratis gebruik maken van 118, omdat voor hen het telefoonboek geen optie was. Om de blinde beller te herkennen moesten ze hun eigen telefoonnummer opnemen, en dat werd dan later gematcht. We hebben dat met best veel klanten getest, en technisch werkte het wel, maar mensen snapten het trainen niet." De dienst werd geen succes. Over onderzoek op het Neherlab mocht gepubliceerd worden, maar niet voordat het artikel reviewed was om te checken dat er geen

Zink Typografie Sheet 147 of 213 - Page 139 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



geheime en ongepatenteerde informatie naar buiten ging. En de goedkeuring kwam niet altijd: Van de Burgt herinnert zich dat een artikel van zijn hand tegengehouden werd.

Een tijdlang gaat het erg goed. De groep groeide, de onderzoekers waren op ieder congres te vinden, en KPN was heel actief in Europese onderzoeksprojecten, zoals Babel, gericht op meertalige webdiensten, en MUST, een groot multimodaal project dat spraak en aanwijzen combineerde in een applicatie met toeristische informatie op mobiele apparaten. "Dat was zijn tijd vooruit", zegt Den Os nu. KPN Researh is ook betrokken bij het Taal- en Spraaktechnologie Prioriteitsprogramma, dat loopt van 1995 tot 2000, waarbij verschillende taal- en spraaktechnologiegroepen uit Nederland samenwerken aan een dialoogsysteem voor openbaar vervoer reisinformatie (zie hoofdstuk 5). Maar tegen 2000 zijn er de eerste tekenen dat het minder gaat bij KPN Research. "Je merkte dat er minder geld binnenkwam van externe projecten." De directie weet niet hoe het tij te keren. Den Os: "Achteraf was het mesjokke. Er moest een positieve wind gaan waaien op het lab. De directie heeft toen overal oranje Pipi Langkouspoppen opgehangen." Het mocht niet baten. Mensen begonnen te vertrekken, en de sfeer verslechterde. Den Os, die zelf ook vertrok in deze tijd: "Mensen klaagden en waren bang. Het was ieder voor zich." In 2000 wordt de groep opgeheven.

"Spraaktechnologie blijft een moeilijk product", stelt Dirk van Compernolle, "want voor iedere implementatie moet je werk verzetten. Dat maakt dat het niet superrendabel is, en dat het moeilijk is om producten te maken voor het Nederlands, waar afzetmarkt te klein is ten opzichte van de kosten. In Amerika zijn er veel meer applicaties op de markt. Het kleine taalgebiedje speelt ons parten." Toch was het juist in dat kleine taalgebiedje dat uiteindelijk toch dat ene bedrijf ontstond dat groot, heel groot werd met spraaktechnologie. Voor het Engels, maar met name ook gericht op de kleinere talen. Zó groot werd het bedrijf, dat het zijn Amerikaanse concurrenten kon opkopen. Lernout en Hauspie leken de grote belofte van Rekenen en Taalkunde eindelijk waar te maken. Maar het sprookje zou niet eeuwig duren.

Zink Typografie Sheet 148 of 213 - Page 140 of 205

## 4.3 De grote belofte

In 1987, publiceerde Apple een filmpje met de naam Knowledge Navigator.[7] We zien een professor die thuiskomt, zijn tablet computer openvouwt, en onmiddellijk wordt bijgepraat door zijn virtuele assistent over gemiste telefoontjes en geplande afspraken. Desgevraagd plaatst de assistent een paar telefoontjes en zoekt hij data en artikelen, die hij en passant ook even samenvat. Besturing van de computer gaat voornamelijk spraakgestuurd, aangevuld met wijzen en schuiven op het scherm. We spreken 1987, de meeste huishoudens hadden geen computer en van het internet had nog vrijwel niemand gehoord. Het HTTP-protocol moest nog uitgevonden worden.

Op 18 december van datzelfde jaar werd ten kantore van notaris Bernard Boes te Kortrijk de naamloze vennootschap Lernout opgericht, gezeteld op de Industrielaan in Ieper. Jo Lernout, voorheen verkoper en marketing directeur bij softwarebedrijf Wang, en Pol Hauspie, de boekhouder met technisch inzicht die zelf een boekhoudpakket programmeerde, zijn er heilig van overtuigd dat taal- en spraaktechnologie hèt gat in de markt is en dat Vlaanderen de aangewezen plek is om die sector op te bouwen. Pol Hauspie in een interview in Humo in december 1997: "In de Belgische politiek hoor je vaak over taalproblemen spreken; wel, wij hebben ons bedrijf op dat probleem gebouwd" (de Witte et al. 2002). Het startkapitaal van tien miljoen Belgische frank (250.000 euro) plus twee miljoen Belgische frank aan "know-how" werd gedeeltelijk ingebracht door Lernout en Hauspie zelf. Pol Hauspie verkocht zijn vrij succesvolle softwarebedrijf HPP (Hauspie Pol Poperinge) en Jo Lernout en zijn vrouw verkochten hun huis. De rest kwam van enkele bevriende particuliere investeerders. "*The three Fs*," lacht spraaktechnoloog en oud-werknemer van L&H Dirk van Compernolle, "zoals dat heet in het jargon: *fools, family and friends*. Dat zijn de mensen die in het begin van een bedrijf voor geld zorgen." Van Compernolle was net terug in Vlaanderen na aan Stanford gepromoveerd te zijn en ruim anderhalf jaar bij IBM gewerkt te hebben. Hij ontmoette Lernout op zijn 'Tour d'horizon' nog voordat hij wegging bij Wang, toen hij op bezoek ging bij spraaktechnologen aan Vlaamse universiteiten om te kijken wie en wat daar zoals te vin-

[7] http://www.youtube.com/watch?v=3WdS4TscWH8

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



den was. "Met de intentie een bedrijf te starten en ofwel mensen te rekruteren, ofwel technologie te licensiëren van de universiteit. Hij was overtuigd, dat voor de wordprocessors, waarin Wang een belangrijk bedrijf geweest was, en misschien op dat moment ook nog was, dat spraak als een rechtstreekse interface met de computer de toekomst had." De locatie op een bedrijventerrein in Ieper was weinig sexy, maar had wel andere voordelen: "ze zijn in Ieper gestart omwille van het T-zone statuut", weet Van Compernolle. "Dat was een speciaal statuut waardoor zij geen belastingen hoefden te betalen over de eerste tien jaar. Dus een financiële neus hebben ze in die zin wel gehad. Dat was een stimuleringsmaatregel voor dat industrieterrein waar ze toen zaten. Maar om het T-zone statuut te verkrijgen, moest wel het businessplan goedgekeurd worden, en daarvoor vonden ze dat ze best een 'expert' konden gebruiken om mee te gaan verdedigen" . Van Compernolle vervulde de rol van expert, en het T-zone statuur werd toegekend.

Het begin van Lernout NV, of Lernout en Hauspie Speech Products, zoals het bedrijf vanaf 1989 heette, was weinig glamoureus. Er was bijna permanent geld tekort. Wanneer er niet genoeg in kas was om het personeel te betalen, werd er gekeken wie het geld op dat moment het hardst nodig had. Het management – naast Lernout en Hauspie zelf al snel ook zakenman Nico Willaert – stonden daarbij naar eigen zeggen achteraan in de rij. Wel kregen alle werknemers opties, te verzilveren bij beursgang van het bedrijf – wat een goede regeling zou blijken te zijn voor wie op het juiste ogenblik uitstapte. Deze eerste magere jaren leverden enkele weinig hoogstaande spraakproducten op, waarvan de sprekende kerstbal de bekendste is: een chip van Texas Instruments die twee boodschappen kan bevatten, een batterijtje, interne klok en optisch oog, allemaal verpakt in een kerstbal. Het resultaat: een kerstbal die vóór Kerst zegt: 'Ho ho ho, Christmas is coming, but don't open the presents yet!' wanneer je voor de kerstboom staat (met daaronder de cadeautjes), en met Kerst zelf 'Merry Christmas, open the presents now, ho ho ho'. Het lukte de werknemers van Lernout en Hauspie (L&H) om de bal te produceren, maar commercieel werd het een flop, doordat het pas in november op de markt gebracht werd, terwijl winkels hun kerstinkopen ruim voor de zomer al doen. "En omdat ze te zwaar waren voor de kersttakken, zo heb ik mij laten vertellen", voegt Jan Odijk daaraan toe.

Zink Typografie Sheet 150 of 213 - Page 142 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Terwijl het bedrijf met dit soort spielerei zijn hoofd boven water probeert te houden, slaagt het er wel in een aantal heel goede wetenschappers aan zich te binden. Bert van Coile werkte in Gent onder professor Marc Vanwormhoudt aan spraaksynthese. L&H nam Van Coile in 1989 in dienst en nam een licentie op de Gentse technologie, die de basis zou vormen van de spraaksynthese van Lernout en Hauspie. Spraakherkenning wordt in eerste instantie gedaan door Hervé Bourlard, een Franse spraaktechnoloog van Philips. Georges Zanelatto uit Mons (Bergen) tenslotte was een expert op het gebied van compressie en decompressie van spraak. De opslag van spraak is een van de eerste wat grotere opdrachten van Lernout en Hauspie. Na een aanslag van de CCC, de Strijdende Communistische Cellen van Pierre Carette in 1985, die te wijten was aan slechte communicatie, besluit de Rijkswacht dat alle meldingen opgenomen en bewaard moeten worden. Lernout en Hauspie gaan bij de openbare aanbesteding voor deze opdracht ruim onder de prijs zitten, en beloven grotere capaciteit en betere vindbaarheid, doordat de gesprekken niet op magnetische band, maar op harde schijf worden opgenomen. Volgens Lernout in zijn geautoriseerde biografie (Joris 2005) verloren ze het contract bijna doordat ze met een malafide Israelisch bedrijf in zee gingen, maar het product komt uiteindelijk op tijd af. Dirk van Compernolle over deze technologie: "Dat is niet super high-tech. Je hebt er wel spraakcodering voor nodig, maar het was eigenlijk meer componenten bijeen doen. Eigenlijk is het heel jammer dat ze die poot niet echt goed uitgebouwd hebben, dat was een redelijk goed product op het juiste moment. Op een bepaald moment is het uit de focus geraakt, en zijn er andere bedrijven met het grote marktaandeel gaan lopen. Maar eigenlijk is dat misschien het grootste gouden ei dat ze ooit gehad hebben, maar nooit beseft hebben." Een van de mensen die met dat marktaandeel weg zijn gaan lopen is Dirks broer Geert. Lernout in zijn boek: "Met het systeem dat we voor de Rijkswacht ontwierpen heeft de Belg Geert van Compernolle in Singapore en Maleisië gouden zaken gedaan. Honderden heeft hij ervan verkocht. Die staten zijn zachte dictaturen waar veel telefoongesprekken afgeluisterd worden. In België gebeurt dat veel minder. […] Zo naïef [was ik] dat ik spionage, schaduwpraktijken en afgetapte telefoons alleen in de films van James Bond situeerde."

Een project waarmee L&H veel zichtbaarheid verwierf was het

Zink Typografie Sheet 151 of 213 - Page 143 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



quizlijnproject. Quizmasters Ben Crabbé en Chris van den Durpel presenteerden een quiz voor de BRT, waar kijkers naar Amerikaans voorbeeld via de telefoon mee konden spelen. Maar in tegenstelling tot de Amerikanen, die allemaal een druktoetstelefoon bezaten, hadden veruit de meeste mensen in België een telefoon met een draaischijf, zodat het goede antwoord niet met een druk op de knop doorgegeven kon worden. Lernout en Hauspie presenteerden de oplossing: een spraakherkenningssysteem zou de gesproken antwoorden automatisch verwerken. Dirk van Compernolle, die ondertussen een eigen spraakherkenningssysteem was gaan ontwikkelen, werd benaderd voor de klus. "Het moet ergens april of mei 1990 geweest zijn", herinnert hij zich, "dat zij mij daarover voor het eerst contacteren. "Kunt je zo'n project afkrijgen tegen het eind van het jaar?" Toen heb ik gezegd: "Zes maand, als je mij een goede ingenieur erbij geeft." Toen had ik nog geen volledig pakket dat helemaal af was, maar ik had wel studenten gehad die eraan gewerkt hadden en ik was er zelf aan blijven werken – er was al iets redelijks. De herkenner draaide. Voor woorden. Dat was geen continue herkenning, dat waren geïsoleerde woorden." Maar de overeenkomst liep wat vertraging op. "Het heeft uiteindelijk tot in juni geduurd voor het contract getekend was. En toen zeiden ze mij: "Maar eigenlijk hebben we dat in september nodig, want het project moet opgeleverd worden tegen Kerstmis, dus dan moeten we daarvoor nog alles integreren." Dus in plaats van zes maanden had ik plots maar twee maanden tijd." Met een hoop nachtwerk lukt het om iets op te leveren. "Het was verre van perfect, maar we hebben het toch min of meer up and running gekregen." Jo Lernout over de lancering, door hemzelf gedaan, live op de Belgische tv: "het zweet stond in mijn oksels toen ik de hoorn opnam en de computer belde. Die antwoordde, *so far so good*. Ik sprak mijn keuzenummer in: 'Vijf'. De computer antwoordde: 'U zei twee?' " (Joris 2005). Lernout begint opnieuw, en gelukkig gaat het vanaf dat moment vlekkeloos. Over de afloop van het verhaal verschillen de meningen. Van Compernolle noemt het programma 'een enorme afknapper' en 'een fiasco'. Lernout heeft het over 'een bescheiden succes'. Feit is dat het programma al na enkele maanden van de buis gehaald werd.

Het project was belangrijk voor de ontwikkeling van *Automatic Speech Recognition* (ASR) software binnen L&H en binnen de Katholieke Universiteit Leuven (KUL). Sporen van deze spraakherkenner

Zink Typografie Sheet 152 of 213 - Page 144 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



zijn nog terug te vinden in het huidige spraakherkenningsplatform SPRAAK,[8] dat onder meer in de spraakgroep van het ESAT (afdeling elektrotechniek) in Leuven ontwikkeld is. Toch baalt Dirk van Compernolle ervan dat er niet méér geleerd is van deze ervaring. "Tegen dat we op het einde van het project waren wist ik dat we met een paar kleine verbeteringen het serieus beter konden doen draaien. Maar toen eenmaal het project afgelopen en opgeleverd was, is het er nooit meer van gekomen. Een schande. Maar het is symptomatisch voor hoe soms ook in de toekomst projecten gerund zijn." Bovendien zijn er nooit opnames gemaakt van alles wat er binnenkwam op het systeem. Een gemiste kans om de spraakherkenning in de toekomst vooruit te helpen. Van Compernolle was op dit moment niet in dienst van Lernout en Hauspie. "Het was contractwerk, een contract tussen de universiteit en L&H." En de universiteit werd daarvoor goed betaald – uiteindelijk: "Dat was voor ons hier aan de uni een mooi lucratief contract. Daar hebben we een mooie computer van kunnen kopen. […] Maar tegen dat het betaald geweest is... " Volgens Van Compernolle heeft het vele jaren geduurd voordat L&H betaalde. "Maar", haast hij zich te zeggen, "dat is allemaal niet zo belangrijk. Het is uiteindelijk betaald. Met interest."

De betalingsachterstand van L&H aan de KUL was illustratief voor de financiële situatie van Lernout en Hauspie in die tijd: altijd krap. Verschillende keren wordt er opnieuw geld bij elkaar geharkt van grote particuliere investeerders en enkele risicofinanciers. En soms zijn er lichtpunten: in 1993 wint L&H fors aan geloofwaardigheid als het Amerikaanse AT&T een belang van 5% in het bedrijf neemt. Het uiteindelijke doel van het bedrijf is om als eerste Belgische bedrijf genoteerd te worden op de Amerikaanse technologiebeurs Nasdaq. Maar in 1994 is het onduidelijk of ze hun beursgang wel gaan halen, zo hoog is de acute geldnood. L&H besluit tot een ongebruikelijke stap: ze geven Automatisch Converteerbare Obligaties (ACO's) uit, die op vijf momenten in de toekomst omgezet zouden kunnen worden in aandelen. De intekenprijzen op elk van die momenten geven de verwachte waardestijging aan: 56% in minder dan drie jaar. Toch hebben zij die ACO's gekocht hebben – en na verzilvering niet verder belegd hebben – forse winsten gemaakt. Lernout en Hauspie benadrukken

[8] http://www.spraak.org

Zink Typografie Sheet 153 of 213 - Page 145 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



graag dat het de 'gewone man' uit West-Vlaanderen is geweest, die L&H deze moeilijke tijd doorgesleept heeft, en daar uiteindelijk goed van geprofiteerd heeft. De bekendste anekdote over de bedeltocht over het Belgische platteland:

> Hij had nog nooit een computer gezien, de varkensboer uit Poperinge met wie Pol bij het ochtendkrieken een afspraak had. Ik weet niet of hij de demonstratie echt had kunnen volgen toen hij na een half uurtje opstond en terugkwam met een kasbon van twee miljoen frank. Het papier zag er niet uit, was helemaal verfomfaaid en de randen waren zelfs afgebeten. " Voilà. Als je die kan incasseren, steek het geld dan maar in dat bedrijf van jullie", bromde hij. Hoe kwam het dat de randen afgebeten waren, vroeg Pol. "Ik bewaar mijn kasbons altijd in de meelbak van de varkens, maar de boerin heeft zich op een dag vergist en de varkens uit de verkeerde bak gevoederd."[9] (Joris 2005)

De Witte, Van Aelst en Van Peteghem laten echter zien dat het merendeel van de ACO's niet in West-, maar in Oost-Vlaanderen verkocht werd, en wel in de omgeving van Aalst, waar beursmakelaar Goethals gezeteld is (de Witte et al. 2002). Goethals organiseerde de uitgifte van ACO's. Een kwart van de ACO's, die niet op een andere manier aan de man gebracht konden worden, werden op naam van Sogespat gezet, een Zwitsers beleggingsvehikel dat Goethals had opgericht. De ACO's werden omgezet in aandelen, die later verkocht werden. Maar het geld verdween naar een Panamese vennootschap, en de beleggers van Sogespat zagen hun geld niet meer terug.

Het quizlijnproject illustreert een overgang binnen het bedrijf van applicaties naar technologie. De quizlijn zelf was nog een volledige applicatie, maar wel met vrij complexe technologie, die bovendien speciaal hiervoor ontwikkeld was. L&H neemt steeds meer mensen aan met een hoog technologisch profiel, en begint het businessmodel te veranderen. Van Compernolle: "De eerste jaren is L&H echt geen technologiebedrijf, maar vanaf eind 1991, begin 1992 beginnen ze met licenties." Het bedrijf zet al snel helemaal in op technologie: "Uit die licentiecontracten komt eigenlijk heel weinig geld binnen. Niettemin beslissen ze in 1992 – dat weet ik bijna zeker – om heel de applicatiedivisie te verkopen." Precies in die periode wordt Van Com-

[9] In iets andere vorm ook verschenen in (de Witte et al. 2002), die het weer overgenomen hebben uit het boek *De Spraakmakers* van Piet Depuydt.

Zink Typografie Sheet 154 of 213 - Page 146 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



pernolle met nog een paar hoogleraren aangesteld door een bank om het bedrijf door te lichten, met als doel in te schatten wat het bedrijf waard was. "Een van de conclusies was: ok, L&H heeft een heel mooi potentieel opgebouwd, maar daarvan is toch heel weinig realiteit in termen van getekende contracten." En de overwegend positieve evaluatie bevatte nog een kritische noot: "dat ze te veel producten hadden. Dat het niet off-the-shelve licentiëren was. Je had de basis, maar dat moest altijd nog aangepast worden. Dat is een realiteit waar we nog jaren mee geconfronteerd zouden worden." Het feit dat L&H daarmee dus geen technologiebedrijf was, maar meer een ingenieursbedrijf, zou grote gevolgen kunnen hebben voor een toekomstige beursgang. Van Compernolle: "als je een engineeringbedrijf bent, dan gaan ze je eerder waarderen op je omzet. Als je een licentiebedrijf bent, en het geld stroomt binnen, dan gaan ze zeggen je waarde is vele keren je omzet. Dus de multiple die ze hanteren in de waardering van het bedrijf ligt vele malen hoger bij een bedrijf dat leeft van licenties dan bij een engineeringclub."

De technologie binnen Lernout en Hauspie was opgedeeld in synthese, compressie en herkenning. Hervé Bourlard leidde de herkenningsgroep vanuit Brussel. Maar dan vraagt het Amerikaanse Amerigon om spraakherkenning voor zijn navigatiesysteem. Binnen een paar maanden moet er een spraakherkenner zijn die het ook doet in een lawaaierige auto. Bourlard zegt: "Dat kan niet." Van Compernolle, die zich inmiddels gespecialiseerd had op het gebied van ruisrobuuste spraakherkenning, zegt: "Dat kan wel." En krijgt de opdracht. Zo komt het dat de groep van het ESAT zo intensief samenwerkt met Lernout en Hauspie, dat er volgens Van Compernolle organogrammen in omloop moeten zijn geweest waarop twee spraakherkenningsgroepen staan ingetekend: één in Brussel, en één buiten het bedrijf, gezeteld in Leuven. Voor Van Compernolle breekt een mooie tijd aan: "Mede door het Amerigonproject en andere dingen die daaruit voortgekomen zijn, is er een deftig project opgezet met steun van de overheid, een IWT-project [van het Vlaamse agentschap voor Innovatie door Wetenschap en Technologie], voor verder onderzoek op het gebied van robuuste spraakherkenning. Dat is wel veel professioneler en beter aangepakt dan het quizlijnproject.[…] Dat was echt een joint research project tussen industrie en universiteit en zoals het zou moeten zijn. Op het juiste moment. To the point. Met een goede leve-

Zink Typografie Sheet 155 of 213 - Page 147 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



rancier van technologie – wij – en een goede klant. Daar klopte het plaatje vrij perfect. Alleen dat het aanspreekpunt binnen L&H niet het hoofd van de spraakgroep was. Dat was eigenlijk het enige gekke. Daar klopte het plaatje helemaal niet." Die imperfectie wordt midden 1994 ook weggewerkt, wanneer hij gevraagd wordt om hoofd van de spraakherkenningsgroep te worden. Van Compernolle reist naar de Verenigde Staten om onopvallend rustig naar de technologie te kijken en sollicitatiegesprekken te voeren. Hij weet dat de uitgifte van ACO's geslaagd is, wat hem vertrouwen geeft in de financiële positie van het bedrijf. "Hadden ze mij een of twee jaar tevoren gevraagd, ik zou het niet gedaan hebben. Ik wist hoe penibel de financiële situatie moest zijn, omdat we met projecten problemen genoeg hadden om het geld binnen te krijgen." De spraaktechnoloog vraagt en krijgt een jaar verlof om bij Lernout en Hauspie aan de slag te gaan. Hij blijft wel colleges geven en mensen begeleiden. Zijn projecten zijn allemaal voor minimaal twee jaar gefinancierd, het onderzoek liep goed, er was ondersteuning door een programmeur. Met heel hard werken kon het allemaal net. Van Compernolle over de combinatie Universiteit en L&H: "Toen zijn we echt in overtijd gegaan. Daarvoor was het nog mooi, maar hierna werd het niet meer zo mooi. De komende twee jaar zijn een beetje zotjes." Op 6 juni 1994 begint Van Compernolle. Bourlard wordt niet ontslagen, maar is ook geen groepsleider meer. Op die eerste dag verschijnt hij op zijn werk in gevangenisplunje. "Die had hij aangetrokken omdat hij nu aan mij moest rapporteren. Hij kwam ontslag geven, en wilde duidelijk maken dat hij het niet fair vond", aldus Van Compernolle.

De uitgifte van ACO's bracht relatieve rust in het bedrijf, dat zich nu echt kon richten op de ontwikkeling van technologie. "Ik mocht niet veel mensen werven, maar wel af en toe eens eentje. De lonen werden betaald, er kwamen geen deurwaarders, en de computers konden aanstaan", vertelt Van Compernolle. Hij was optimistisch over wat er bereikt kon worden. "Met wat opschoning van de technologie die er was zouden we stapjes vooruit kunnen zetten. We hadden een vrij uitgebreide groep van twintig tot dertig mensen, waarvan ongeveer acht echt onderzoek deden. De rest zat op klantprojecten. […] High tech is tenslotte voor 20% technologie en voor 80% implementatie, databases, opnemen, transcriberen, enzovoorts." Het was de tijd waarin neurale netwerken de belofte waren voor de toekomst. In het

Zink Typografie Sheet 156 of 213 - Page 148 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



onderzoek was er veel aandacht voor neurale netwerken. "Toen was er zo'n hele upswing van neurale netwerktechnologie. Bourlard was daar mee bezig. Hij was een van de grote advocaten daarvan en dacht dat dat veel beter ging werken dan Hidden Markov Models. Hoewel het toch ook HHMs waren waar hij mee werkte. Maar neurale netwerken zouden veel beter gaan werken." Overigens had ook Van Compernolle zelf aan de universiteit al aan dat onderwerp gewerkt. "Iedereen deed toen neurale netwerken", zegt hij daarover.

Spraaktechnologie neemt in deze tijd een grote vlucht. Binnen Europa worden grote initiatieven opgestart, zoals het Duitse megaproject Verbmobil, met 86 miljoen euro met afstand het grootste taal- en spraaktechnologieprogramma ooit. Ook de Vlaamse regering start een eigen TST-programma. Het adviesorgaan VRWB, de Vlaamse Raad voor Wetenschapsbeleid, was vrij kritisch over de plannen voor dit programma. Maar de Vlaamse minister-president Van den Brande was een groot voorstander van taal- en spraaktechnologie: hij kwam onder meer spreken op de taal en spraak-workshop op Flanders Technology, de 2-jaarlijkse technologiebeurs. Het programma kwam er dan ook, en hieruit werden onderzoeksprojecten gefinancierd waarvan de resultaten nog steeds gebruikt worden, zoals het corpusproject CoGen en het lexiconproject Fonilex. Voor Leuven was vooral het project Nerex belangrijk: "dat was Nederlandstalige spraakherkenning waarbij onze herkenner hier op ESAT de transitie heeft gemaakt van een woordgebaseerd systeem naar een large vocabulary moderne spraakherkenner", aldus Van Compernolle.

De top van het bedrijf was ondertussen druk bezig met de voorbereiding van de beursgang. Lernout en Hauspie reisden in de VS van de ene potentiële belegger naar de andere om producten te demonstreren en de prospectus onder de aandacht te brengen. Op 1 december 1995 was het dan eindelijk zo ver: Lernout en Hauspie Speech Products was het eerste Belgische bedrijf op de Nasdaq. De introductieprijs was vastgelegd op 11 dollar, maar stond aan het einde van de eerste beursdag op 16 dollar. Een stijging van 45% in één dag. Het management vierde het in de VS met vijf flessen Château Margeaux uit 1966. "We hebben de wijn tot de laatste druppel leeggedronken, iedereen heeft zijn handtekening op de etiketten gezet en de lege flessen namen we mee als souvenir." (Joris 2005). Ook in Vlaanderen was het feest, bij de werknemers van L&H: "We zijn goed zat geworden", herinnert Van

Zink Typografie Sheet 157 of 213 - Page 149 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Compernolle zich. "Maar we snapten het niet, want zoveel verkochten we niet. We zaten middenin de hype." Ook de kopers van ACO's en de geldschieters van het eerste uur zullen opgelucht adem gehaald hebben.

De beursgang luidde vijf doldwaze jaren in, gevuld met samenwerkingen, overnames, en nieuwe aan L&H gelieerde ondernemingen. Vertaalbureau Mendez was de eerste grote overname. Het bureau verzorgde vertalingen door menselijke vertalers, maar luidde ook een beweging in richting een nieuw maar gerelateerd terrein: automatisch vertalen en vertaalhulpmiddelen. Maar het is niet alleen de taalkundige kennis en de nieuwe markt waar L&H in geïnteresseerd is. Het vertaalbedrijf maakt winst. En dat in grote tegenstelling tot het moederbedrijf zelf, dat veel investeert, maar nog geen enkel kwartaal winst heeft kunnen noteren. Het daaropvolgende jaar nemen Lernout en Hauspie nog een bedrijf op het gebied van vertalen over: GMS. Gesellschaft für Multilinguale Systemen is het Duitse bedrijf dat de rechten op Siemen's vertaaltechnologie METAL voor de pc overnam. De algemeen directeur Peer van Driesten wordt het hoofd van een nieuwe afdeling Machine Translation bij L&H. Stephan Bodenkamp, die ook aan GMS gelieerd is, krijgt een sleutelrol in de later op te zetten Language Development Companies (LDC's). Een heel belangrijke in de rij van overnames is die van Kurzweil Applied Intelligence in 1997. Kurzweil AI had kennis van en ervaring met de medische wereld, een potentiële markt voor dicteeroplossingen (de Witte et al. 2002). Ook hadden ze ervaring met het ontwikkelen van eindproducten, terwijl L&H zich al een tijdje richtte op technologie. Met de aankoop van Kurzweil AI hoopte het bedrijf de slag om het eerste dicteerprogramma te winnen. Toch was Dirk van Compernolle, die met zijn groep bezig was om zelf *large vocabulary continuous speech recognition software* te bouwen, niet onverdeeld gelukkig met de gang van zaken: "Je gaat vooruit, je bouwt wat op. En dan moet je na maanden research over op het aangekochte systeem. Dan moet je je mensen vertellen dat het ophoudt. Dat is niet leuk." Bovendien was Van Compernolle er niet van overtuigd dat de technologie van Kurzweil AI zou leiden tot een snellere oplevering van continue spraakherkenning. De gang van zaken rondom de overname typeerde volgens hem hoe de top tegen eigen en andermans technologie aankeek. "De gekochte bedrijven en technologie werden door de top steeds veel hoger ingeschat dan

Zink Typografie Sheet 158 of 213 - Page 150 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



de eigen mensen en ontwikkelingen. Blijkbaar was men bang dat de mensen in de overgenomen bedrijven gedemotiveerd zouden raken. Maar men vergat wel eens de eigen mensen die zich voor hun eigen projecten hadden uitgesloofd." (de Witte et al. 2002).

De grote opmars van Lernout en Hauspie krijgt een extra stimulans wanneer Microsoft besluit het bedrijf een kapitaalinjectie van 45 miljoen dollar te geven en een partnership met L&H aan te gaan. Er gaat een grote due diligence [onderzoek naar de boekhouding en technologie] aan vooraf, zowel technisch als juridisch. Microsoft had ook een eigen spraaktechnologielijn, maar te weinig capaciteit om alle ontwikkelingen zelf te doen. Afspraak was dat MS alleen technologie voor sleuteltalen als Engels en Chinees zelf zou doen, en voor alle andere talen de technologie uit Ieper zouden gebruiken. Planning was dat die technologie voor vreemde talen op de markt zou komen met de release van Windows 2003. Microsoft was ook al een grote klant van Lernout en Hauspie, met name via contracten met vertaalbureau Mendez. Die verstrengeling baarde Van Compernolle wel wat zorgen: "Microsoft was ineens onze grootste aandeelhouder en onze grootste klant en onze grootste bedreiging. Bedrijven waar Microsoft in investeert worden leeggezogen en gaan dood." Toch overheerst het optimisme. De deal met Microsoft geeft een enorme boost aan het Belgische bedrijf. Toen Lernout en Hauspie bij varkensboeren moesten bedelen om geld, wie had toen gedacht dat die twee West-Vlamingen een paar jaar later bondgenoot zouden worden van de grootste softwareproducent van de wereld?

De stroom van overnames blijft doorgaan. In 1998 neemt L&H de taalgroep van Novell over, waarin onder meer Jan van Sas en Peter de Bie onder leiding van Rudy Montigny werken. Montigny wordt de leider van een nieuwe groep Intelligent Content Management. Binnen negen maanden veranderde de situatie opnieuw, als de groep in januari 1999 wordt ondergebracht in een nieuwe divisie, SAILlabs. Peer van Driesten, voorheen algemeen directeur van GMS, leidt de divisie van 250 taaltechnologen met een eigen Raad van Bestuur. Peter de Bie over de overgang naar Lernout en Hauspie: "Toen de overname rond was, veranderde er weinig aan de opzet of organisatie van de groep. Maar het was wel alsof er een snoeppot openging: L&H had al zo veel in huis, en met al die technologie mocht gespeeld worden. Wij konden dan proberen of we daarmee mooie nieuwe dingen konden

Zink Typografie Sheet 159 of 213 - Page 151 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



maken. Een voorbeeld daarvan is een multilingual search engine. Dat werd gemaakt door verschillende technologieën van L&H met elkaar te combineren." Piek Vossen, die eerder Eurowordnet opzette, wordt aangetrokken om hierbij te helpen. Jan van Sas: "We zeiden: "Piek, probeer eens wat met eigen wordnet te doen. Laat zien wat je moet doen om betere retrieval te krijgen." Hij heeft een analyse gedaan, die heeft geleid tot een mooi intern rapport. Maar daar is niets van naar buiten gebracht, want er was een expliciet verbod om te publiceren."

Jan Odijk was inmiddels vanuit het IPO naar Lernout en Hauspie gekomen om te werken als hoofd van de afdeling die zorgde voor taalkundige resources, die vervolgens door alle afdelingen gebruikt konden worden. Net als De Bie is ook hij enthousiast over de werksfeer: "We vonden het een fantastische tijd. Er moest ontzettend hard gewerkt worden. Dat werd ook gedaan, en mensen deden dat met plezier. Je zag dat het bedrijf groeide, en je kon heel goed aan je technologie werken. Er zijn ongelooflijke tijden gemaakt. […] Soms moest er gewoon zeven dagen op zeven gewerkt worden. Dan moest er iets geleverd worden voor Microsoft. Dat was een grote klant, dus dat moest af. Dan maar het weekend doorwerken, en iedere avond tot elf uur twaalf uur doorwerken of soms langer. "

Van Compernolle was op dat moment ook positief, maar zag wel gevaren voor de toekomst: "Het gevoel overheerste dat we goed bezig waren. De marge tussen wat gezegd wordt en wat geleverd werd was niet raar. In die periode dacht ik: "Het komt ooit goed met L&H. Misschien wordt het ooit een normaal bedrijf." Het was nog steeds een enthousiaste club, waar heel hard gewerkt werd. Maar dat kon niet eeuwig voortduren. Mensen kunnen geen jarenlang twaalf uur per dag werken. En naarmate het bedrijf groeit, is het moeilijker om die enthousiaste sfeer te behouden. We moesten *sustainable* worden, niet continu op het tandvlees hoeven te lopen. Ook verlof kunnen plannen en geen schrik hebben dat je morgen failliet gaat. Niet continu moeten bedelen om geld." Maar *sustainability* had geen eerste prioriteit. Lernout en Hauspie financierden overnames voornamelijk met aandelen, en dus was het van het allergrootste belang dat het aandeel bleef stijgen. Van Sas: "Bij L&H draaide alles erom de beursanalisten tevreden te stellen. Spectaculaire short-term resultaten. Er was een totaal gebrek aan langetermijndenken." Gaston Bastiaens lijkt de belangrijkste vertegenwoordiger van deze strategie. Van Sas: "Er waren

Zink Typografie Sheet 160 of 213 - Page 152 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



wel mensen die hun geld op lange termijn wilden zetten, maar dat kregen ze niet verkocht in de organisatie. Gaston Bastiaens deed alles om de beurscijfers omhoog te krijgen. En hij was CEO, dus hij had laatste woord." Voor Van Compernolle was de houding van Bastiaens de oorzaak van een omslag in denken over L&H: "Waar ik eerst dacht dat de ingenieur Gaston zou helpen, heeft hij de lont juist aangestoken. Van hem moest L&H elk jaar verdubbelen. Wij zeiden wel dat dat niet kon, dat daar geen markt voor was, we hebben zwaar aan de bel getrokken. Maar dat veranderde niets. In een paar maand tijd veranderde mijn geloof in L&H drastisch. Wij melken uit wat universitair gedaan is, maar we zijn niet bezig met de toekomst. We zouden zwaar moeten investeren in lange termijn research. Maar dat doen we niet." Van Compernolle komt in de clinch met de leiding van het bedrijf. Hem wordt een lange-termijn-divisie beloofd, waar hijzelf leiding aan mag geven. Wanneer deze niet van de grond komt, verlaat de spraaktechnoloog in 1999 het bedrijf.

Een breed gedeelde klacht is dat de top van het bedrijf te weinig luisterde naar de specialisten die het had aangetrokken. Van Compernolle was het niet eens met de aanhoudende stroom aankopen van Lernout en Hauspie, maar heeft anderzijds ook wel voor aankoop van een bedrijf gepleit, waarna er niets gebeurde. Ook Jan Odijk herinnert zich overnames waar de experts kritisch over waren. "Jo was altijd de boer op om nieuwe bedrijven over te nemen. Hij ging naar een bedrijf, hoorde de technologie aan, snapte die technologie maar voor de helft, en kwam dan naar ons om te vertellen wat voor fantastisch bedrijf, wat voor fantastische technologie hij had gevonden. Wij hoorden dat dan aan en begrepen niet waar hij het over had. Daarna gingen we zelf naar het bedrijf kijken en dan bleek het veel meer *down to earth* te zijn, veel realistischer, en dan snapten we weer wat ze deden. Vaak raadden we dan af om het te doen, maar ja, Jo was niet gehinderd door enige kennis van zaken. Jo ging gewoon door, ook al gaven wij negatief advies omdat het misschien wel interessante technologie was, maar niet iets waar we nou zo ontzettend veel geld mee konden verdienen en geen technologie die goed paste binnen ons pakket. Jo ging onverdroten voort." De mismatch tussen het enthousiasme van inspirator Jo Lernout en de dagelijkse praktijk van het ontwikkelen en opleveren zorgde nog wel eens voor botsingen. "Dan had hij veel te ambitieuze dingen beloofd, die wij helemaal niet waar konden maken", herinnert

Zink Typografie Sheet 161 of 213 - Page 153 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Odijk zich. "Dus dan moest er weer onderhandeld gaan worden, zodat wij toch iets leverden dat wel aardig in de richting ging." Gelukkig zat er nog een stap tussen de beloften van Lernout en een getekend contract: "Op Jo's niveau waren het allemaal mondelinge afspraken. Op contractniveau ging alles langs de juridische dienst en die consulteerden de technische mensen, die op hun beurt heel precies beschreven wat ze wel gingen leveren en wat ze niet gingen leveren. Dan was het opeens weer down to earth. Maar Jo wist zodoende dus wel heel veel mensen te enthousiasmeren, en investeerders te lokken."

Rondom het steeds verder uitdijende bedrijf ontstaan verschillende gelieerde organisaties zoals het Dictation Consortium (DC), de Brussels Translation Group, en de Language Development Companies (LDC's), allen opgericht met geld van durfinvesteerders. Het doel van deze entiteiten was om de hoge ontwikkelingskosten buiten de boeken van het beursgenoteerde bedrijf te houden. Zo werd Dictation Consortium opgericht met het doel Dragon af te troeven in de race om het eerste dicteersysteem op de markt te brengen – als niet voor het Engels, dan toch voor andere talen. En daar was veel ontwikkelingswerk voor nodig. Jan Odijk: "Dus hebben ze een nieuw bedrijf opgericht, Dictation Consortium, daar hebben ze investeerder voor gezocht en gevonden, en vervolgens heeft dat investeerdersbedrijfje L&H opdracht gegeven dictation software te maken. Zo werd je onderzoek en ontwikkeling betaald door een derde partij. Een gedeelte van dat geld was ook weer van Jo [Lernout] en Pol [Hauspie], maar er was ook een gedeelte van onafhankelijke investeerders. Dus alles wat wij aan dictation deden werd betaald door een externe partij."[10] Als de ontwikkeling zou slagen, dan zou L&H die terugkopen met een mooie winst voor de investeerders. Zou het project mislukken, dan konden de investeerders proberen de technologie door te verkopen aan een concurrent van L&H. Het succes van deze constructie was helemaal afhankelijk van de stijging van de koerswaarde van het bedrijf. Odijk: "Het is een Baron van Münchausen die zichzelf aan zijn haren omhoogtrekt. Je hebt eigenlijk niets, ook geen geld om te investeren, dus richt je een ander bedrijf op, vindt daar wat investeerders voor, en laat dat zo betalen. Je kan de ontwikkeling allemaal als revenu

[10] Odijk geeft later aan dat de bijdragen van Lernout en Hauspie zelf tijdelijke voorschotten voor de eigenlijke investeerders betreffen.

Zink Typografie Sheet 162 of 213 - Page 154 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



tellen, […] waardoor de koers omhoog gaat – het koersklimaat was ook heel gunstig in die tijd. Daardoor wordt je zo rijk, dat je het oorspronkelijke bedrijf kunt kopen. En daardoor wordt je rijker en heb je nieuwe business-producten." BTG was opgericht om de ontwikkeling van vertaalsoftware voor het internet te ontwikkelen, en werkte samen met SAILlabs. De LDC's waren opgezet om spraaktechnologie voor exotische talen als Vietnamees, Hindi en Urdu te ontwikkelen. Met de financiële en juridische constructies van durfkapitaal, ontwikkelingskosten en terugkooprecht rondom deze entiteiten zochten Lernout en Hauspie telkens de grenzen op van het toelaatbare. De rechter zou later oordelen dat hierbij ook grenzen overtreden zijn.

Een bijzondere nieuwe ontwikkeling in de schaduw van Lernout en Hauspie Speech Products was Flanders Language Valley. Het idee achter de technologievallei was dat verschillende bedrijven die met TST werken gecentreerd werden op een plek. Zo konden activiteiten gebundeld worden, en zou de branche als geheel een boost krijgen. FLV werd op 11 november 1999 officieel geopend door prins Filip van België. Op het openingsweekend zouden volgens Lernout 25.000 mensen geweest zijn. "Voor de eerste keer was er een file op de A19 vanaf Beselare tot aan de campus", aldus de topman in zijn biografie (Joris 2005). De vallei had ook een opleidingstak, FLV Education, die verantwoordelijk was voor het vinden en opleiden van geschikt personeel op het gebied van taal- en spraaktechnologie en kunstmatige intelligentie. Aan het hoofd hiervan staat Dirk Frimout, de eerste Belgische astronaut. Frank Van Eynde van de Katholieke Unievrsiteit Leuven en Walter Daelemans van de Universiteit Antwerpen coordineerden de lesmodule Computerlinguïstiek, waarin ondermeer finite-state methoden, parsing, computationele semantiek, dialoog, statistische methodes en information retrieval de revue passeerden. Middenin de hype van de onbegrensde mogelijkheden schafte Flanders Language Valley twee Cam Brain Machines aan. "Dat was een heel geavanceerde computer in de vorm van een cortex", legt Walter Daelemans uit. "Het was gebaseerd op FPGCs, *Field Programmable Gateway Chips*. Dat zijn chips die dynamisch kunnen herfigureren, gebaseerd op neurale netwerken en genetische algoritmen die neurale netwerken configureren." De machines kwamen uit het lab van Michael Korkin en Hugo de Garis. Die laatste is bekend om zijn voorspellingen. In de jaren negentig claimde hij dat met genetische algoritmen

Zink Typografie Sheet 163 of 213 - Page 155 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



neurale netwerken ontwikkeld zouden kunnen worden die menselijke intelligentieniveaus al snel zouden overschrijden. Hij voorspelt verder dat er voor het einde van de 21ste eeuw een oorlog zal uitbreken tussen voor- en tegenstanders van intelligente machines, die miljarden doden tot gevolg zal hebben. Hij leidt momenteel het Artificial Brain Lab van de Xiamen University. FLV besloot twee van die machines te kopen. "Een miljoenenbesteding", aldus Daelemans. Maar het was onduidelijk wat er met de machines moet gebeuren. "Dus toen kregen we geld van Lernout en Hauspie om interessant onderzoek te gaan doen met die machines. Taal- en spraaktechnologie met een horizon van twintig jaar. Er werd echt met geld gegooid." Jakub Zavrel werkte als student van Walter Daelemans op een project dat de computers nuttig moest maken. Hoewel hij gerelateerd onderzoek heeft gedaan, heeft hij zelf de Cam Brain Machine nooit gebruikt: "Al snel werd duidelijk dat we helemaal niets konden met die machines."

Ondertussen ging het taal- en spraaktechnologiebedrijf onverminderd door met overnames. Misschien wel de twee belangrijkste aankopen vonden plaats in maart 2000. Eerst verwierf het Dictaphone Corporation, marktleider op het gebied van medische transcripties, en twintig dagen later maakte het bekend bovendien Dragon Systems, de grote Amerikaanse concurrent van L&H gekocht te hebben. Aan beide overnames hingen forse prijskaartjes: voor Dictaphone werd 465 miljoen aan aandelen betaald, ook al had het bedrijf 395 miljoen aan schulden openstaan, die door L&H werden overgenomen. Die schulden werden door een consortium van banken met een krediet gefinancierd. Het was de bedoeling dat het geleende geld na de zomer met een nieuwe financieringsronde zou kunnen worden afbetaald. Vice-voorzitter van de Raad van Bestuur van Lernout en Hauspie en Nederlands ICT-ondernemer Roel Pieperwais degene die Dragon aankocht. De eigenaren James en Janet Baker hadden al eerder gesprekken gevoerd over overname, maar de prijs die ze vroegen was te hoog. Maar Dragon moest overgenomen worden, want het bedrijf dreigde failliet te gaan. En Lernout en Hauspie wilden Dragon graag overnemen, onder meer vanwege de 170 vakspecialisten die ze in dienst hadden. Janet Baker weigerde te onderhandelen met Gaston Bastaens. En toen schoven Lernout en Hauspie Roel Pieper naar voren. De onderhandelingen vonden plaats in een restaurant in Boston. Op een papieren servet schreef de Hollander zijn bod: het legenda-

Zink Typografie Sheet 164 of 213 - Page 156 of 205

rische bedrag van 540 miljoen dollar. Janet Baker heeft het servetje bewaard, en de 540 miljoen gekregen. Uitgekeerd in L&H aandelen.

Zo langzaam maar zeker dekte Lernout en Hauspie Speech Products het hele spectrum van taal- en spraaktechnologie af. Al deze technologie moest samenkomen in één product: Sofia, *Society of Intelligent Assistants*. Deze tabletcomputer moest zelfstandig op mobiel internet het antwoord zoeken op gesproken vragen, en deze ook weer presenteren in de vorm van gesproken antwoorden – de Belgische variant op Apple's Knowledge Navigator. Het geheime project, dat intern op de R&D afdeling bekend stond als het Charm-project, moest de grote klapper worden voor L&H. Maar de ene tegenslag volgde de andere op: Vlaamse partner De Post schendde de geheimhoudingsclausule en kondigde een samenwerking op een soortgelijk project aan met Microsoft, in Korea werden intentieverklaringen nooit omgezet in contracten, en de nieuwe CEO John Duerden zag helemaal niets in het ambitieuze project. Ondanks concrete aanwijzingen voor investeringen van buitenaf, draaide hij het project direct na aantreden de nek om.

En toen werd het 8 augustus 2000. Op de dag dat het bedrijf goede tweedekwartaalcijfers bekend maakte kwam Wall Street Journal (WSJ) met een artikel dat aantoonde dat de omzetcijfers uit Korea geflatteerd waren. De journalisten hadden Koreaanse klanten gebeld, en een aantal van hen had aangegeven niet of voor veel minder geld dan opgegeven van het Ieperse bedrijf af te nemen. Lernout en Hauspie ontkenden dat er een probleem is, en *The Register*[11] meldde later dat er inderdaad fouten gemaakt waren in het Wall Street Journal artikel, maar het kwaad was al geschied: beleggers verloren het vertrouwen in Lernout en Hauspie en de koers kelderde. WSJ kwam met een nieuw artikel over aan L&H gelieerde bedrijven in Singapore die spookfirma's zouden zijn. Zeker geen goede tijd voor een nieuwe financieringsronde. Maar de schulden van Dictaphone moesten worden afbetaald. De banken eisten hun geld. L&H restte niets anders dan uitstel van betaling aan te vragen. Op 5 december besliste de rechter echter dat uitstel van betaling niet werd verleend, en er evenmin een faillissement werd uitgesproken: er was nog een overbruggingskrediet, en daarmee moesten de bestuurders een reddingsplan opstellen en

[11] www.theregister.co.uk

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



uitwerken. De meeste mensen hadden het bedrijf toen al opgegeven. Werknemer Jan Odijk: "Ik begon bang te worden op het moment dat ik Jo Lernout's blik zag en zijn uitspraken hoorde na het befaamde artikel in The Wall Street Journal. Daarin werd beweerd dat er gesjoemeld was met de cijfers, en Jo zei: "We laten die journalisten ons bedrijf toch niet kapotmaken". Maar de manier waarop hij keek en de blik in zijn ogen, toch behoorlijk wanhopig…Toen dacht ik: 'Misschien is er wel echt iets mis'." Toch is Odijk tot het laatst blijven geloven dat het bedrijf in een of andere vorm zou blijven bestaan. Anton Nijholt van de Universiteit Twente is in deze periode nog op bezoek geweest in Ieper, en was minder optimistisch. "Ik ben daar in november geweest toen het zo'n beetje afliep. We hadden samen een conferentie georganiseerd en invited speakers konden niet meer betaald worden. Dat heeft heel wat voeten in de aarde gehad. […] Toen ik wegging dacht ik: 'Dit is de laatste keer dat ik hier ben, laat ik vragen of ze nog T-shirts verkopen bij de receptie.' Toen heb ik nog een T-shirt van L&H gekocht, voor een gulden of zeven. Dat ligt nog ergens in de kast. Een historisch attribuut inmiddels." Nijholtkreeg gelijk: op 21 oktober 2001 sprak de rechter officieel het faillissement uit van Lernout en Hauspie Speech Products.

Over de ondergang van Lernout en Hauspie doen veel verhalen de ronde. Een veelgehoorde is dat de Amerikaanse geheime dienst erachter zou zitten. De top van het bedrijf is ervan overtuigd dat dit het geval is. Motief? Het feit dat Dragon, het bedrijf dat voor de Amerikaanse inlichtingendiensten werkte, in buitenlandse handen was gekomen. Het feit dat de toppers van het spraakonderzoek, dat gefinancierd was door de Amerikaanse defensie-organisatie DARPA, in de technologische adviesraad van L&H zat: Janet Baker van Dragon, Raymond Kurzweil van Kurzweil AI, en Alex Waibel van Carnegie Mellon (de Witte et al. 2002). Dit verhaal wordt gesteund door het feit dat een spion als Stephan Bodenkamp enerzijds betrokken was bij het Europese Sensusproject, een project gericht op een Technologie voor een Europese Inlichtingendienst, en anderzijds binnen het bedrijf op strategische plekken werkte, zoals in de coördinatie van LDC's, en toch niet voorkomt in de hele rechtzaak. Tot slot zouden er aanwijzingen zijn dat de CEO in 2000, de Amerikaanse John Duerden, dubbelspel gespeeld zou hebben. Lernout claimt dat de man gegevens doorspeelde aan Mark Maremont, de journalist van The Wall

Zink Typografie Sheet 166 of 213 - Page 158 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Street Journal die met zijn artikel in augustus 2000 de val van Lernout en Hauspie inluidde, en dat hij probeerde de broncode naar Amerika over te hevelen (Joris 2005). Ook de rol van Philip Bodson, aangesteld om het bedrijf van de ondergang te redden, wordt ter discussie gesteld. Waarom had hij bijvoorbeeld het overbruggingskrediet van de Belgische bank en L&H investeerder KBC afgeslagen, en in plaats daarvan een krediet aangenomen van Cerberus, een fonds waarover bijzonder weinig bekend is, maar waarvan het lijkt alsof het onder directe controle staat van het Witte Huis (de Witte et al. 2002). Jan Odijk: "Vroeger geloofde ik nooit zo in die verhalen, maar tegenwoordig ben ik toch wel iets meer geneigd ze te geloven." Dan is er de vraag of oprichters Lernout en Hauspie bewust de boel opgelicht hebben, of verkeerd voorgelicht danwel onvoorzichtig geweest zijn. Van Compernolle gelooft niet in kwade opzet van de beide ondernemers: "Lernout en Hauspie zijn geen bewuste fraudeurs. Ik ken geen mensen die zoveel gewerkt hebben als zij. Ze hebben er hun gezondheid mee kapot gemaakt, hun gezin kapotgemaakt, en veel vrienden verloren. En ik geloof ook niet in het complotverhaal. Nooit in geloofd. L&H zou wel aan iets kapot gaan."

De oprichters en enkele sleutelfiguren van het Ieperse bedrijf zijn door het Openbaar Ministerie aangeklaagd wegens boekhoudfraude. Op 20 september 2010 werden Jo Lernout, Pol Hauspie, Nico Willaert en Gaston Bastiaens allen veroordeeld tot een gevangenisstraf van vijf jaar, waarvan twee of drie (Bastiaens) voorwaardelijk, en geldboetes. Enkele andere bestuurders kregen lagere straffen. In één van de verschillende nevendossiers over de investeringen in drie bedrijven die in de schaduw van L&H werden opgericht, was Pol Hauspie al veroordeeld tot 18 maanden voorwaardelijk. Hauspie bekende dat hij wist dat zijn financiële constructies illegaal waren, en dus dat hij bewust fraude pleegde – al zegt hij dat nooit op persoonlijk gewin uit te zijn geweest, wat klopt met de cijfers: Lernout en Hauspie hebben zichzelf tijdens de hoogtijdagen van het bedrijf nooit verrijkt. Jo Lernout blijft ontkennen dat er sprake was van fraude.

Tijdens de faillissementsprocedure, na het aanvragen van uitstel van betaling, maar voor de definitieve uitspraak van het faillissement, is Dictaphone afgesplitst en vertaalbureau Mendez snel verkocht in een poging te redden wat er te redden viel. Mendez ging van de hand voor 44 miljoen dollar – veel minder dan in 2000 voor het bedrijf

Zink Typografie Sheet 167 of 213 - Page 159 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



geboden werd (180–200 miljoen dollar). De rest van de technologie werd na het officiële faillissement verdeeld in acht groepen, die per veiling verkocht zijn. Een onzekere tijd voor de *Linguistic Resources Group* van Jan Odijk, die een dienstverlenende rol voor de andere groepen verzorgde. Hijzelf en een klein deel van zijn team werden geplaatst in de text-to-speech-groep. Maar toen werd duidelijk dat Scansoft (dat later opgaat in Nuance) drie stukken van de L&H technologie opkocht: *Text-to-Speech*, spraakherkenning en dictation. Later verwierven ze ook nog de technologie voor audiomining, die aanvankelijk was opgekocht door de oprichters van Dragon, James en Janet Baker. Odijk: "Die technologie kwam ook van Dragon. Maar ze konden het uiteindelijk toch niet financieren, en toen heeft Scansoft het alsnog overgenomen." Doordat Scansoft meerdere technologieën kocht, ontstond er weer ruimte voor een centraal data-team. Odijk: "Want al die groepen hadden data nodig. Dat was een gemeenschappelijke taak, dat moet je niet dubbel gaan doen." Odijk hoopte een groter deel van zijn team te kunnen behouden. "Maar dat viel tegen. Uiteindelijk hebben we na veel gelobby één extra persoon kunnen behouden. De rest van mijn team werd ontslagen" (p.c.). Scansoft en later Nuance kan gezien worden als de opvolger van L&H. Niet alleen heeft het bedrijf een groot gedeelte van de technologie, maar het volgt ook dezelfde filosofie van opkopen van concurrenten. Odijk werkte er jarenlang, maar het optimisme en enthousiasme van de jaren bij Lernout en Hauspie vond hij er niet terug: "Het was hetzelfde werk, en er waren ook veel acquisities, maar de manier waarop met personeel werd omgegaan bij Scansoft en Nuance werd als heel slecht ervaren. Het was een Amerikaans bedrijf en de leiding wilde het Amerikaanse personeelsbeleid toepassen. […] Als ze mensen nodig hadden dan haalden ze die binnen, en als ze die niet meer nodig hadden dan gooiden ze ze er weer uit. Je had helemaal niet het idee te werken bij een bedrijf waar je dingen kon leren, waar je kon groeien. Dat gevoel was helemaal weg." Vantage Learning kocht de taal- en zoektechnologie van Lernout en Hauspie. De groep van Jan van Sas kwam er terecht en verbeterde de essayscoringstechnologie van het Amerikaanse bedrijf. Maar ook Vantage bleek geen goede werkgever. Van Sas vertrok en begon voor zichzelf, degenen die achterbleven werden uiteindelijk ontslagen. Van Sas: "Die mensen werd wijsgemaakt dat er een nieuw gebouw gehuurd was, en dat ze dus moesten inpakken. Die dozen

Zink Typografie Sheet 168 of 213 - Page 160 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



werden vervolgens allemaal naar Amerika gestuurd. Dat was het einde van hun contract."

Het einde van TST-gigant Lernout en Hauspie heeft invloed gehad op het vakgebied in Vlaanderen. Allereerst op de aantrekkingskracht van het vakgebied. Frank Van Eynde: "Het vak kreeg heel veel belangstelling, en daardoor nam het aantal studenten voor computerlinguistiek toe. […] De val van L&H heeft geleid tot een terugval in het aantal studenten. Dat heeft een aantal jaar gespeeld." Voor de studenten die al begonnen waren was het tijdelijk lastig om stageplekken te vinden, doordat er geen plek meer voor ze was in Ieper. Van de taaltechnologen in dienst van L&H is een aantal – al dan niet met een omweg – teruggekeerd naar de universiteit, zoals Dirk Van Compernolle, Jan Odijk, en de Leuvense wetenschappers Hugo van Hamme en Vincent Vandeghinste. De meningen zijn verdeeld over de vraag of er in het onderzoek ook een Lernout en Hauspie-effect is geweest. Herman Caeyers denkt van wel: "Ik ben er zeker van dat als L&H niet te maken had gehad met malversaties, dan had de TST veel verder gestaan", aldus de Vlaming. "Vergelijk de investeringen in taaltechnologie toen en nu, dat is peanuts. Het STEVIN-programma, vergeleken met de investeringen van toen, dat is een schijntje." Jean-Pierre Martens geeft bovendien aan dat het succes van Lernout en Hauspie ook in de universitaire wereld veel ruimte creëerde om onderzoek te doen. Doordat het vakgebied erg leefde, was de kans op onderzoeksfinanciering groter dan voorheen. "In Vlaanderen heb je zoiets als de Vlaamse instituten", vertelt Martens. "Er bestaat zo'n Vlaams instituut voor biotechnologie, een voor elektronica. Die instituten zijn bedoeld als attractiepolen waar kennis over een bepaald domein gebundeld wordt om kritische massa omhoog te krijgen. Er was op zeker moment zelfs sprake van om zo'n Vlaams instituut voor taal- en spraaktechnologie te beginnen. Had L&H nog een paar jaar moeten leven, dan was het er ook gekomen. Dat om u te zeggen wat er allemaal mogelijk was." Aan de andere kant ziet Walter Daelemans helemaal geen negatieve effecten van de val van L&H. Hoogstens een positief effect: "Het was een beetje een probleem in de tijd van het hoogtepunt om te zeggen dat er nog interessante onderzoeksmogelijkheden waren in taal- en spraaktechnologie, want L&H hadden alles al opgelost. Die misvatting bestond niet meer daarna. Dus je kon weer met recht en rede zeggen dat het een onderwerp is waar je fundamenteel onderzoek naar kan

Zink Typografie Sheet 169 of 213 - Page 161 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



doen." Daarnaast ziet hij het terugkeren van toppers in het vakgebied naar de universitaire wereld als een bonus. Van Compernolle is minder positief: "Het is nu niet meer de spraakwereld van de jaren negentig toen we grote dromen hadden. Is dat erg? Ja, ik vind van wel. Spraak is *not hot anymore*. Dat is overal zo, maar door L&H in België nog wat krachtiger. De hype was groter, dus de terugval erna ook. Ik heb zelf ook niet meer hetzelfde enthousiasme als vijftien jaar geleden." Dan realiseert hij zich dat het niet alleen aan Lernout en Hauspie ligt: "Maar ja, ik ben zelf ook ouder, en heb een gezin."

## 4.4 Ondertussen in Nederland

De opkomst en ondergang van Lernout en Hauspie bepaalde jarenlang het beeld in de taal- en spraaktechnologie in Vlaanderen, maar in Nederland was er nauwelijks enig effect van de TST gigant te merken. Terwijl in Vlaanderen één bedrijf alle commerciële activiteit op het gebied van taal- en spraaktechnologie bundelde en maximaal profiteerde van de internethype, was er in Nederland ruimte voor kleine ondernemers om op de golf van optimisme nieuwe bedrijfjes op te starten. Vaak ontstonden die bedrijfjes vanuit een universitaire omgeving, zoals Carp Technologies, dat in 1996 door vier studenten van de Universiteit Twente werd opgericht, zoektechnologie en software voor automatisch samenvatten ontwikkelde, en enkele jaren opdrachten deed voor onder meer de provincie Overijssel. En in sommige gevallen kon L&H, aan de zijlijn, toch een handje bij helpen bij de start van zo'n taaltechnologiebedrijf.

In april 1999 werd professor Franciska de Jong gevraagd een bezoek te brengen aan Lernout en Hauspie. De internethype was in volle gang en met het Ieperse bedrijf ging het meer dan voorspoedig. "Volgens mij was het de Overijsselse Ontwikkelingsmaatschappij die een uitstapje naar L&H had georganiseerd, en die vroegen mij of ik meeging", herinnert De Jong zich. "Dat zal wel via de Universiteit van Twente gegaan zijn." De groep bestond uit een man of zeven. "En daar zat een investeerder tussen, die op zoek was naar kansen op nieuwe bedrijven. Hij vond het leuk om met mij te praten omdat hij in mij de verbinding zag met informaticastudenten, die soms goede ideetjes

Zink Typografie Sheet 170 of 213 - Page 162 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



hebben. Dus hij zei dat als er ooit een student was die een bedrijf wilde beginnen, dat ik ze dan naar hem toe moest sturen." Maar professor de Jong had geen studenten die bezig waren een bedrijfje op te richten. De enige die ze kende en daar wèl mee bezig was, was Stan van de Burgt, een alumnus uit Twente, die inmiddels bij KPN Research werkte. Van de Burgt vertelt hoe hij ertoe kwam om een bedrijf te starten: "Internetaanbieders dachtten toen nog dat ze meer moesten aanbieden dan connectiviteit. KPN had ons daarom gevraagd om een portalstrategie te ontwikkelen. In het kader daarvan waren collega Berend Metz en ik op reis gestuurd om relevante Amerikaanse bedrijven te bezoeken." Op die reis bezochten ze ook Ask Jeeves, een webdienst voor het beantwoorden van vragen. Het concept resoneerde bij Van de Burgt, die zelf ook al wel had gefilosofeerd over de mogelijkheden voor kostenbesparing in callcenters door online vragen te beantwoorden. "Berend en ik vonden het gaaf dat het bedrijf zich zo helemaal richtte op één dienst." In eerste instantie informeerden de twee of zij misschien de Nederlandse variant konden ontwikkelen voor Ask Jeeves, danwel via KPN, danwel zelfstandig. Maar daar werd weinig enthousiast op gereageerd. "Dan moet je maar een licentie nemen", was het antwoord van AskJeeves. "En dat was miljoenen", herinnert Van de Burgt zich. "Maar met wat ik had geleerd en de mensen die wij kenden, moesten wij dat zelf ook kunnen." Metz en Van de Burgt besluiten ervoor te gaan, en beginnen een plan uit te werken.

Franciska de Jong herinnert zich nog dat Stan van de Burgt het haar vertelde. "Hij zei: "Ik heb een plan voor een bedrijf, wil je daar niet eens naar kijken en er commentaar op leveren?" Op het station in Delft hebben ze mij toen het verhaal uit de doeken gedaan. Zij waren bezig financiers te interesseren, en verwachtten dat het snel rond zou komen." Maar kort na het bezoek aan Lernout en Hauspie hoorde De Jong dat het met de financiering van het bedrijf niet goed liep. Ze bracht de twee in contact met Ilja Bobbert, de investeerder van Prime Technology Ventures die ze in Ieper had ontmoet. "Een piepklein demootje en tien slides hebben we ze laten zien", vertelt Van de Burgt. Maar Prime was overtuigd en gaf een zogenaamde *letter of intent* af. Maar de investeringsmaatschappij was nog in oprichting, dus het geld liet nog even op zich wachten. Metz en Van de Burgt waagden het er toch op: ze verhoogden elk de hypotheek op hun huis en startten hun bedrijf: Q-go, de Nederlandse vraag-antwoord service op internet.

Zink Typografie Sheet 171 of 213 - Page 163 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Metz is de belangrijkste commerciële man in het bedrijf, maar de technologie van Q-go is van de hand van Stan van de Burgt en Theo Vosse, met wie Van de Burgt in Twente gestudeerd had. Het basisidee van Q-go is dat de informatie die ontsloten moet worden, gestructureerd wordt in de vorm van modelvragen en dat zowel de vraag van de gebruiker als de modelvraag automatisch volledig taalkundig geanalyseerd worden: spelfouten worden gecorrigeerd, samenstellingen en werkwoord-clitic-combinaties zoals in het Spaans uit elkaar getrokken, en de syntactische structuur ontleed. Zodoende wordt geabstraheerd van de formulering die de gebruiker min of meer toevallig uit vele mogelijkheden gekozen heeft om zijn vraag te formuleren, en wordt de betekenis uit de vraag gehaald. Bij het matchen van gebruikersvraag tegen de modelvragen wordt bovendien gebruik gemaakt van semantische kennis, zodat ook een modelvraag en de bijbehorende informatie gevonden kan worden wanneer die niet precies over hetzelfde onderwerp gaat, maar er wel nauw aan gerelateerd is, bijvoorbeeld "Hoe kan ik een hypotheekofferte aanvragen?" voor de gebruikersvraag "Hoe hoog is de hypotheekrente?". In eerste instantie mikt het bedrijf op een combinatie van internet- en corporate websitetoepassingen. "De internetservice zou zorgen voor naamsbekendheid en wie weet kostendekkend zijn, en met de technologie die we daar ontwikkelden, konden we dan ook de corporate applicatie bouwen", legt Van de Burgt uit. "Investeerders waren in die tijd vooral in de internettoepassing geïnteresseerd".

Eind 2000 kwam het einde van de eerste investering en het eigen geld in zicht, maar het bedrijf had een aardige naamsbekendheid gekregen, en de tijden waren nog vrij goed voor internettechnologie. In een tweede financieringsronde haalde Q-go 12 miljoen euro op en het bedrijf breidde snel uit naar Duitsland, Frankrijk, Spanje en het Verenigd Koninkrijk. De focus kwam te liggen op zakelijke websites in plaats van het hele internet. Maar toen spatte de internetbubble uiteen. En terwijl Q-go het openingfeestje vierde van het kantoor in Engeland, van waaruit de hele Europese financiële dienstverlening bediend moest gaan worden, vlogen in New York twee vliegtuigen de Twin Towers in. "Gedurende de dag kregen we steeds meer afzeggingen of gingen er mensen weg, omdat collega's van ze in de torens zaten", herinnert Van de Burgt zich. Q-go moest fors saneren: de vestigingen in Frankrijk en Engeland werden gesloten en een aanzienlijk gedeelte

Zink Typografie Sheet 172 of 213 - Page 164 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



van het personeel werd ontslagen. Beide oprichters stapten in 2003 uit de onderneming, omdat ze met de investeerders van mening verschilden over de te volgen strategie. Maar het bedrijf overleefde, groeide langzaam weer, en levert tot op de dag van vandaag web self-service oplossingen in zeven talen in heel Europa en de VS.

Ook via het Flanders Language Valley Fund heeft Lernout en Hauspie een beetje invloed gehad in het Nederlandse TST-landschap. Het fonds, dat investeerde in taal- en spraaktechnologie, ging ten onder in de nasleep van de val van L&H en als gevolg van een mislukte kapitaalsverhoging van dertig miljoen dollar in Korea, die bovendien aan de basis stond van de beleggersmanipulatie waarvoor onder meer Pol Hauspie inmiddels veroordeeld is. Maar voordat het fonds in de problemen kwam, investeerde het nog in een Nederlandse start-up, die op basis van TNO-technologie innovatieve informatieverwerking op het web wilde gaan aanbieden: Irion Technologies. Oprichter en CEO is Joop van Gent, tot 2000 manager van de afdeling informatiesystemen van TNO. In tegenstelling tot Q-go, dat zich al snel richtte op een specifiek product (web self-service) en een specifieke markt (grote tot zeer grote ondernemingen in de financiële, telecom- en verzekeringsbranches), ontwikkelde Irion een waaier aan verschillende producten. "Aanvankelijk had Irion drie basistechnologieën: automatische classificatie, crosslinguale search, en semantische netwerken. Later is daar onder andere een dialoogsysteem bijgekomen", aldus de oprichter. Het bedrijf maakt een paar grote veranderingen door, allereerst in strategie: "De eerste belangrijke wijziging was rond 2002, toen we besloten om het 'componentenmodel', waarbij Irion componenten onder OEM [*Original Equipment Manufacturer*, technologie ingebouwd in het product van de klant] levert aan IT-bedrijven, te verruilen voor een model waarbij we zelf eindproducten maakten, samen met partners. […] Een tweede beslissing was die om meer focus te ontwikkelen, ik schat rond 2005, zowel in producten als markten, en niet alles 'wat leuk is' zomaar te accepteren. Op dit moment maakt Irion een derde belangrijke strategische wijziging door: de omslag van producten naar services". Ook Irion, vernoemd naar de vrouw van de oprichter, werd beïnvloed door de tijdgeest. "Aanvankelijk is geprobeerd om het bedrijf in Silicon Valley te positioneren, en er is een Amerikaanse investeerder nauw betrokken geweest bij deze poging. Uiteindelijk is om allerlei redenen toch besloten het Ame-

Zink Typografie Sheet 173 of 213 - Page 165 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



rikaanse avontuur af te blazen. […] de internetzeepbel en het uiteenspatten ervan heeft de ambities wel een beetje bescheidener gemaakt." Overigens waren de gevolgen van het failliet van FLV voor Irion beperkt. FLV verkocht de portefeuille en daarmee hun aandeel in Irion. "Maar de banden met FLV waren uitstekend", aldus Van Gent. Verder heeft het bedrijf altijd sterke banden gehouden met de universitaire wereld. Van Gent: "Carnegie Mellon University was belangrijk omdat de wortels van het zoeksysteem er vandaan kwamen." De banden met de VU kwamen onder meer tot uiting in de aanstelling van VU-wetenschapper Piek Vossen als CTO. Overigens was volgens Alice Dijkstra die relatie met de universitaire wereld niet altijd zo goed: "Joop van Gent maakt leuke dingen. Maar hij werd vroeger verguisd door de academische wereld, omdat hij te praktisch bezig zou zijn."

Waar de meeste bedrijfjes in taal- en spraaktechnologie ontstonden rondom een bestaand product of concept rondom taal en/of spraak, volgde ICT-bedrijf Gridline een andere route. Maar wie de achtergrond van oprichter Tigran Spaan kent, zal zich niet verbazen dat hij toch geëindigd is in de taaltechnologie. Hij is gestart met een studie informatica, gevolgd door een studie wiskunde bij het ILLC van Johan van Benthem, die zich nadrukkelijk bezighield op het snijvlak van taalkunde en logica, en heeft daarna enige tijd op een puur taalkundig aio-project in Utrecht gewerkt. "Het idee was dat ik de verschillende theorieën over binding bij elkaar zou brengen, en zo een brug zou slaan tussen de verschillende taalkundige eilandjes in Utrecht, zoals de generativisten en de semantici." Maar hij maakt het project niet af. "Het lukte niet. Enerzijds denk ik dat ik te ambitieus was wat betreft mijn doelstelling, wat ik eruit wilde halen. Daar was veel te veel kennis voor nodig. Maar anderzijds was ik ook niet gemotiveerd genoeg. Ik kon er niet tegen dat je dan een tijd lang heel hard werkte, en dan eigenlijk nog steeds niets concreets gemaakt had, en eigenlijk ook niet eens wist wat dat ding dan zou moeten zijn, dat je gaat maken." En zo komt Spaan terecht bij websitebouwer Desk. "Een half uurtje programmeren, en dan had je al wat staan." Na een jaar vindt hij dat het tijd is voor iets nieuws, en bedenkt dat hij zelf wel een bedrijf zou kunnen beginnen. En zo zit hij op 1 januari 2000 op het 'kantoor' van zijn IT startup Blackeye – letterlijk een zolderkamer.

Zink Typografie Sheet 174 of 213 - Page 166 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Spaan begint zijn bedrijf zonder klanten, zonder computer, en zonder idee. "Als er een project langskomt, dan kan ik het wel", vat hij nu zijn ondernemingsplan van destijds samen. Maar aan ambitie ontbrak het hem niet: "Binnen vijf jaar wilde ik een bedrijf hebben met tussen de tien en twintig man, dat ook zonder mij gewoon door zou kunnen draaien. Een belachelijk idee." Maar er kwam inderdaad een project langs, en hij kon het inderdaad. Blackeye deed programmeeropdrachten voor onder meer Telfort. Met taaltechnologie had het allemaal niets te maken. "Ik had ook er ook geen weet van dat er een wereld van taaltechnologie bestond. Dat er taaltechnologiebedrijven waren. Ik had nog nooit van Polderland gehoord. Alleen Rosetta kende ik wel, via semanticus Theo Janssen." Pas via het jaarlijkse CLIN congres en NOTaS-magazine Dixit kwam hij erachter dat er bedrijven waren die gespecialiseerd waren in taaltechnologie. Zijn eigen bedrijf, dat ten tijde van de buzz rond The Grid en gridcomputing omgedoopt was tot Gridline, was er toevalligerwijs ingerold: "IHLIA [Internationaal Homo/Lesbisch Informatiecentrum en Archief], een van onze klanten, had een thesaurus gemaakt en wilde daar wat mee. Dat vond ik leuk!" Uit het werk aan thesauri groeide het eerste taaltechnologische product, de Gridwalker, een enterprise zoekoplossing met de mogelijkheid de zoekopdracht te verfijnen of uit te breiden en thesauri te beheren." Inmiddels telt Gridline tussen de vijftien en twintig man. Spaan: "Maar het bedrijf draait nog steeds niet door zonder mij."

Ook aan de Universiteit van Tilburg merkten ze dat er interesse was vanuit de industrie voor toepassingen van taaltechnologie. Om aan deze vraag te kunnen voldoen, werd een stichting opgericht: de Stichting Toepassingen Inductieve Leertechnieken (STIL). Jakub Zavrel, die enkele van deze opdrachten uitvoerde, vertelt hoe de opdrachten binnenkwamen: "Dat ging via via. De stichting had een eigen raad van bestuur, en daar zat iemand in die venture capitalist was bij Twinning. Via hem hebben we bijvoorbeeld een opdracht gekregen om informatie-extractie te doen uit Intermediair." Het geld dat met deze opdrachten binnenkwam, werd gebruikt om ingehuurd personeel te betalen, of stroomde terug naar de universiteit. En dat bleef niet onopgemerkt. Zavrel: "Op een goede dag werden we op het kamertje van de secretaris geroepen. De boekhouder was er ook. En de ene zei beslist: 'Wat jullie doen kan helemaal niet!', maar de ander vond het

Zink Typografie Sheet 175 of 213 - Page 167 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



juist helemaal geweldig en vond dat we vooral door moesten gaan. Dus dat laatste hebben we gedaan." Maar Zavrel realiseerde zich wel dat hij die opdrachten die hij nu voor de stichting deed, ook voor zichzelf zou kunnen doen. En toen ontstond het idee van een eigen bedrijf. In Remko Bonnema en Rob Koeling vond Zavrel geschikte partners voor de onderneming. Toen Koeling echter naar Brighton verhuisde, gingen Zavrel en Bonnema met zijn tweeën door. Met 20.000 euro startkapitaal, afkomstig uit inkomsten van STIL, en de creditcard van Zavrel, werd in 2001 Textkernel opgericht. "Dat startkapitaal hebben we geleend, en vier jaar later konden we het terugbetalen", verzekert de ondernemer. Via de bedrijfsnaam heeft Textkernel toch een klein lijntje naar Lernout en Hauspie: "In Flanders Language Valley was een bedrijf met de naam The Learning Kernel", vertelt Zavrel. "Dat vond ik mooi. Textkernel is daar een afgeleide van. Het heeft de juiste uitstraling van moeilijke dingen oplossen."

Toen Textkernel begon ziet de toekomst er zonnig uit. De starters hadden al een aantal mooie opdrachten op zak, en de internethype was op zijn hoogtepunt. Maar de twee waren nog niet goed en wel begonnen, of de bubbel barstte. Contracten werden gecanceld, en het startkapitaal slonk snel. Maar de twee konden zich altijd nog verhuren als programmeur. Eén ding heeft Textkernel gered, volgens de oprichter: "We hebben heel snel ingezien dat er geen markt is voor taaltechnologie. Er is een markt voor het oplossen van problemen, en hoe dat gebeurt maakt niet uit." Het zou jaren sappelen blijven voor Zavrel en voor Bonnema, die bovendien een gezin met drie kinderen te onderhouden had, maar uiteindelijk werd het bedrijf rendabel, onder meer door een groot contract met de PTT. Zou Zavrel het weer doen als hij van tevoren wist dat de bubbel zo snel uiteen zou spatten? "Natuurlijk niet!"

De Nederlandse taal- en spraaktechnologiebedrijven zijn sterk genoeg om het einde van de internethype te overleven, en al snel na de dip zouden al weer verschillende nieuwkomers de kop opsteken, zoals selfserviceprovider AskNow Technologies en mediareputatiespotter Trendlight. De commerciële activiteit op het gebied van taal- en spraaktechnologie wordt uiteindelijk ook verankerd in een gezamenlijk platvorm. In maart 2001 richtten de Brabantse Ontwikkelingsmaatschappij, de Katholieke Universiteit Brabant (nu Universiteit van Tilburg), Technische Universiteit Eindhoven, Polderland Language &

Zink Typografie Sheet 176 of 213 - Page 168 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Speech Technology, Van Dale Data en nog enkele kleine bedrijfjes de Stichting NOTaS (Nederlandse Organisatie voor Taal en Spraak) op. Het Brabantse accent verdwijnt vrijwel direct. In NOTaS vinden universiteit en industrie elkaar, en taal en spraak. Nederland en België ontmoeten elkaar hier echter nog niet, want ondanks dat de stichting na enige tijd expliciet opengesteld wordt voor Vlaamse instellingen, blijven deze onvertegenwoordigd.

Zink Typografie Sheet 177 of 213 - Page 169 of 205

Laboratoriumopstelling voor het Talking Heads experiment: ‘robots’ met camera’s leren woorden voor gekleurde vormen.

# 5 Taal en spraak in dialoog

De automatische verwerking van talige structuren is zó wezenlijk anders dan de automatische verwerking van het geluidssignaal dat deze taalstructuren codeert, dat het begrip *taaltechnologie* alleen gereserveerd is voor technologie die te maken heeft met de eigenschappen van een taal, maar niet voor technologie rondom de drager van deze taal: spraak. Omgekeerd verwijst de term *spraaktechnologie* alleen naar technologie die het geluidssignaal zelf produceert of analyseert, niet naar technologie die de informatie uit dat spraaksignaal analyseert en betekenis geeft. De geschiedenis van die beide vakgebieden raakt elkaar nauwelijks. Aan het IPO en bij Lernout en Hauspie werkten onderzoekers in beide vakgebieden, maar overwegend naast elkaar in plaats van met elkaar. Waarom dan toch een boek over taal- èn spraaktechnologie?

Dit boek dwingt de taal- en spraaktechnologie de bladzijden tussen voor- en achterkaft samen te delen. En daarmee schaart het zich in een bescheiden traditie van pogingen om de twee vakgebieden samen te brengen. Deze traditie krijgt vorm in de tweede helft van de jaren negentig, waar we een aantal van die pogingen voorbij zullen zien komen. Maar al eerder werden wat voorzichtige voorzetten gegeven. In 1984 bijvoorbeeld, toen ZWO de werkgroep Taal- en Spraaktechnolo-

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



gie oprichtte, die adviseerde te investeren in een Nederlands dialoogsysteem. En een beetje in het spraaktechnologieprogramma ASSP, met zijn werkgroep Linguistic Analysis. Ook CELEX kan gezien worden als een voorzichtige samenwerking tussen taal en spraak: de elektronische databanken met lexicale informatie van het in 1986 opgerichte Nijmeegse Centrum voor Lexicale Informatie bevatten naast morfologische en frequentie-informatie namelijk ook informatie over de uitspraak van de woorden. Het project werd gefinancierd door het Informatica Stimuleringsplan (INSP, 1984–1988), en de resultaten worden nog steeds gebruikt, zowel door spraaktechnologen als door taaltechnologen. Maar de flirt tussen taal en spraak werd pas wat serieuzer in de jaren negentig, toen er samen aan corpusontwikkeling gewerkt werd èn het langverwachte dialoogproject eindelijk van de grond kwam.

Van 1994 tot 1997 bestond in Vlaanderen het korte-termijnprogramma Spraak- en Taaltechnologie. Binnen dit door de Vlaamse regering met 1,2 miljoen euro gefinancierde programma zijn drie resources ontwikkeld voor taal en spraaktechnologie: FONLEX, ANNO en COGEN. FONLEX is min of meer de Vlaamse tegenhanger van CELEX. ANNO is een corpus van radio-uitzendingen van de VRT, dat niet alleen getranscribeerd is, maar ook deels automatisch voorzien is van morfologische en syntactische annotatie: een duidelijke bijdrage van taaltechnologie aan dit spraakcorpus. COGEN tenslotte bevat verscheidene opnamen van dezelfde tekst door telkens verschillende sprekers. Daarmee is het corpus duidelijk gericht op spraaktechnologische toepassingen, in het bijzonder continue spraakherkenning. Hoewel het korte-termijnprogramma positief beoordeeld werd, en voor aanvang van het programma al geanticipeerd werd op een langetermijnprogramma, is dat er nooit gekomen.

## 5.1 Samen werken aan applicaties: het Openbaar Vervoer Informatie Systeem

Ook de Nederlandse overheid had wel het idee dat er ‘iets’ gedaan moest worden aan taal- en spraaktechnologie. Al in 1984 had de

Zink Typografie Sheet 180 of 213 - Page 172 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



commissie TST in Nederland gepleit voor stimulering van de taal- en spraaktechnologie, onder meer door de ontwikkeling van een Nederlands dialoogsysteem, maar dit rapport was in een bureaulade terechtgekomen. Nederlandse initiatieven op het gebied van TST waren tot dan toe uit algemene potten voor informatietechnologie gekomen (SPIN, ESPRIT), die bovendien alleen toegepast onderzoek financierden, en geen fundamenteel onderzoek. Vijf jaar later[1] wordt er een nieuwe commissie ingesteld door de Stichting Taalwetenschap van NWO, die onder leiding van voorzitter Remko Scha de noodzaak en de haalbaarheid van een Prioriteitsprogramma voor TST moet onderzoeken. Het Algemeen Bestuur van NWO ontvangt dit rapport in 1991 en bespreekt het in de Centrale Commissie voor de Prioriteitsprogramma's. De uitkomst? De instelling van wederom een nieuwe commissie, ditmaal met de opdracht een concreet voorstel te schrijven voor een TST Prioriteitsprogramma, onder het voorzitterschap van Lou Boves.

De commissie Lou Boves komt in 1992 met een voorstel dat voorziet in een programma rondom vier kernthema's: robuust ontleden, efficiënt parseren, geavanceerde akoestische decodering en integratie van spraak- en taalrepresentaties (Boves et al. 1992). Bovendien geeft het aan dat voor de eerste twee zowel naar taalkundige benaderingen gekeken moet worden, als naar 'alternatieve aanpakken', waarmee probabilistische methoden bedoeld worden. Een derde belangrijk aspect van het voorstel was dat een concrete toepassing ontwikkeld zou worden: al het onderzoek zou samenkomen in een demonstrabel dialoogsysteem voor het beantwoorden van vragen over treinreizen. Volgens Alice Dijkstra, die speciaal voor het TST Prioriteitenprogramma bij NWO werd aangetrokken als programmasecretaris, was het cruciaal dat er expliciet gesproken werd over een demonstratiesysteem: "Bij de overheid heeft het gelieg [rondom automatisch vertalen] wel een neurose opgeleverd. Zoveel grote beloften waren niet nagekomen, dat niemand nog wat geloofde. […] Dus was het voorstel heel duidelijk: we maken een demosysteem. We leveren het bewijs dat wat we be-

[1] Volgens de brochure uit 1995 (Boves, Landsbergen, Scha & van Noord 1995). Het NWO-rapport uit 1992 noemt 1990 in plaats van 1989 als het jaar waarin de commissie werd ingesteld (Boves, Duifhuis, van Eijk, Kempen, Koster & Scha 1992).

Zink Typografie Sheet 181 of 213 - Page 173 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



loven echt kan, maar voor een eindproduct moet het netjes opnieuw geïmplementeerd worden. Dat was goed, want het vertrouwen was volledig weg, en terecht." Voor het demonstratiesysteem zou samengewerkt worden met de Openbaar Vervoer Reisinformatie (OVR), een organisatie opgericht door het Ministerie van Verkeer en Waterstaat, dat de beschikking had over de volledige dienstregeling van alle openbaar vervoersdiensten. OVR verzorgde al telefonische reisadviezen door call center agents, maar wilde deze dienst graag (ook) geautomatiseerd aanbieden.

Het NWO-bestuur was enthousiast over het plan en stelde geld ter beschikking. Er waren wel twee addertjes onder het gras: in plaats van de gevraagde tien miljoen gulden, kwam NWO over de brug met de helft: vijf miljoen gulden (2,3 miljoen euro). En dat geld was alléén beschikbaar als er aanvullende middelen vanuit het bedrijfsleven zouden komen. Maar in de loop van 1993 werd duidelijk dat het bedrijfsleven niet bereid was geld in het project te pompen. KPN en Philips 'redden' het project door aan te geven dat ze actief wilden deelnemen aan het project, en hun eigen onderzoek wilden laten aansluiten op het Prioriteitsprogramma. Concreet kwam het er op neer dat de spraakgroep de beschikking kreeg over de broncode van de spraakherkenningssoftware van de Philipsgroep in Aken. "We mochten eigenlijk alles doen met die code. We hebben de software ook gebruikt voor andere projecten, en we mochten hem ook wijzigen", vertelt Boves. Die overeenkomst was niet alleen heel prettig, maar ook noodzakelijk. "We hebben nooit overwogen om de spraakherkenning zelf te bouwen", aldus Boves. "Daarvoor hadden we de kennis en de data niet." Behalve dat Philips de spraakherkennningssoftware ter beschikking stelt, besluiten Philips en KPN bovendien om samen een commercieel onderzoeksproject te doen voor een dialoogsysteem, en dit parallel te laten lopen aan het OVIS-programma. De onderzoeksgroepen mogen ook van deze software gebruik maken. NWO was overtuigd en gaf in 1994 de subsidie vrij. De eerste activiteit in het TST Prioriteitsprogramma bestond uit het instellen van opnieuw een nieuwe commissie: de stuurgroep.

De stuurgroep kiest ervoor om geen *open call* voor onderzoeksvoorstellen te organiseren. Dijkstra: "Het idee was dat het een gesloten programma zou worden. Dus er zou een werkplan geschreven worden, en dat zou dan uitgevoerd worden. Dat zou dan eerst goedge-

Zink Typografie Sheet 182 of 213 - Page 174 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



keurd moeten worden, en dan mochten ze dat uitvoeren." Over het waarom van deze beslissing zegt de brochure over het project (Boves et al. 1995): "The most important reason for choosing this policy is that it seemed to offer the best guarantee to obtain further funding later on." Waar het voorstel uit 1992 het nog erg in het midden laat wie er zullen deelnemen aan het programma, wordt in de aanloop naar het werkplan al snel duidelijk wie de deelnemende partijen zullen zijn: Remko Scha, Lou Boves en Jan Landsbergen. Anton Nijholt uit Twente zit wel in de stuurgroep, maar doet niet mee in het programma. "Tot mijn spijt", aldus Nijholt. "Ik heb wel geprobeerd om onze groep erin te krijgen." Een mogelijke reden waarom dat niet gelukt is, is dat Nijholts specialiteit parsing al zwaar vertegenwoordigd was in het programma. Want aan het al genoemde trio, waarin Scha het probabilistisch parsen vertegenwoordigde, werd ook nog parsingspecialist Gertjan van Noord toegevoegd. Dijkstra: "Hij was wat jonger, vertegenwoordigde de nieuwe richting. De jonge generatie moest ook meegenomen worden. Daar was Lou volstrekt op tegen. 'Verspilling van het geld', noemde hij dat." Het belangrijkste bezwaar van de spraaktechnoloog was dat het Groningse parseersysteem gebaseerd was op een handgeschreven unificatiegrammatica. Een kennisgebaseerd systeem dus. En Boves was nu juist een groot aanhanger van statistische methoden op basis van data. Dat Van Noord wel een statistische disambiguatiemodule zou bouwen, was niet genoeg om Boves gunstig te stemmen. Hoogleraar Computationele Taalkunde Jan van Eijck van de Universiteit Utrecht, die bij Boves in de voorbereidingscommissie zat, had hier echter heel andere ideeën over. Hij had aan SRI Cambridge ervaring opgedaan met een kennisgebaseerde parser, en schoof zijn promovendus Gertjan van Noord juist heel nadrukkelijk naar voren om hiermee aan de slag te gaan. Alice Dijkstra probeert de betrokkenen op één lijn te krijgen: "Ik zei tegen Gertjan: 'Organiseer in godsnaam wat in Groningen zodat hij kan zien wat jullie doen'. Dus wij een hele dag naar Groningen voor een demonstratie. Op de terugweg, na een lange stilte, komen we in de buurt van Zwolle en begint Lou ineens te praten. Dan spreekt hij de beroemde woorden 'Ze doen daar toch wel leuke dingen'."

Hoewel iedereen akkoord ging met de deelname van de groep uit Groningen, laat de brochure uit 1995 wel enige bias zien ten aanzien van de te verwachten resultaten:

Zink Typografie Sheet 183 of 213 - Page 175 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



> "We want to establish a paradigm change in linguistic research in the Netherlands, by confronting the experimental, corpus-based approach to language research with the mentalistic enterprise and its results obtained up to now. […] In this respect it is interesting to call attention to the developments that have taken place in phonetics and speech technology. As recently as a decade ago there was still a widely held and strong believe that acoustic decoding of speech signals could be accomplished by means of a (possibly very large) set of essential deterministic rules. The harsh and hard practice, however, has very convincingly shown that that believe was wrong […]. A solid information theoretic, and therefore essentially probabilistic approach has closed much of the performance gap. What is more: speech scientists now understand why an integrated probabilistic approach of the recognition problem is inevitable. Thus, speech science has led the way towards a new, and more powerful research paradigm.
> It is one of the goals of this program to show a similar way on the linguistic levels of syntax, semantics, and probably also pragmatics." (Boves et al. 1995)

Het citaat illustreert meteen hoe het succes van datagedreven methoden in de spraaktechnologie de aanpak van taaltechnologisch onderzoek beïnvloedde.

Lou Boves was de leider van het project en bovendien themaleider van het kernthema spraak. "Wat onderzoek betreft probeerden we vooral om greep te krijgen op variatie in uitspraak", aldus Boves. "Daarnaast probeerden we om modellen te ontwikkelen die niet alleen op basis van fonemen werkten, maar ook op basis van grotere eenheden, zoals lettergrepen en hele woorden." Dan waren er nog de thema's dialoog (onder leiding van Landsbergen) en de beide taalverwerkingsmodules, die geleid werden door Scha en Van Noord. Elk van deze themaleiders had aan de eigen universiteit een onderzoeksgroepje dat zich bezig hield met het onderwerp, en overlegde eens in de twee maanden met de andere leden van het kernteam. Twee keer per jaar werden de resultaten gepresenteerd aan de stuurgroep onder leiding van Herman Bouma van de Technische Universiteit Eindhoven. Zoals beloofd in het voorstel werden in drie rondes prototypen van de demonstrator opgeleverd: OVIS (Openbaar Vervoer Informatie Systeem) 1–3. Er staat bovendien een tijd lang een OVIS-systeem live, waarmee data worden verzameld die weer gebruikt kunnen worden in het programma. Maar de demonstrators lieten niet al het werk uit het programma zien. Boves: "Deze tussentijdse demonstrators gebruikten wel de spraakherkenner die binnen het project ontwikkeld

Zink Typografie Sheet 184 of 213 - Page 176 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



was, maar niet de dialoogmanager of de parseertechnieken van de andere groepen." Een eindversie, waarin alle componenten van de verschillende onderzoeksgroepen aan elkaar geknoopt werden en samenwerkten, is er nooit gekomen. Ondanks het feit dat de componenten afzonderlijk het best goed deden. Boves: "het is als met slootje springen. We kwamen twee centimeter te kort. Maar voor productontwikkeling is dat is net zo erg als wanneer je een halve meter te kort komt."

Eindevaluatie vond expliciet plaats vanuit een wetenschappelijk perspectief. Boves: "Daar hadden we ook op aangedrongen." De procedure bestond uit een zelfevaluatie door de themaleiders, beoordeling door de stuurgroep, en externe evaluatie door buitenlandse specialisten. De NLP-modules werden bovendien onderworpen aan een formele evaluatie, waarbij de resultaten van de beide aanpakken van taalverwerking met elkaar vergeleken werden op 1000 niet eerder geziene testinputs. In oktober 2000 werd het eindrapport gepresenteerd. Ondanks het feit dat de uiteindelijke demonstrator niet gerealiseerd was, bleek het eindoordeel unaniem positief: het programma had volgens de stuurgroep en de externe beoordelaars de TST in het algemeen en de uitwisseling van ideeën tussen taal- en spraaktechnologen in het bijzonder gestimuleerd. Dat het eindproduct er niet gekomen was, deed daar niets aan af. Boves: "De dialooggroep heeft bijvoorbeeld een dialoogmanager gebouwd die werkte, en die mooi in elkaar zat. Gert Veldhuijzen van Zanten heeft dat gedaan. Die had het in zich om inzetbaar te zijn voor ingewikkelder dingen dan simpele treininformatie. Alleen waren complexere taaluitingen vooralsnog onmogelijk door beperkingen in de spraakherkenning." De externe beoordelingscommissie pleitte voor voortzetting van het onderzoek om toch tot een eindproduct te komen, maar dat vervolg kwam er niet. "Waar het programma dus niet in is geslaagd", concludeert Boves, "is om de industrie te overtuigen dat verder investeren in het uitontwikkelen van de onderzoeksresultaten tot commercieel bruikbare software de moeite waard is." En dat ondanks het feit dat het project voor commerciële partner KPN toch een succes mag heten: "Het is binnen KPN maar één keer gelukt om een operationele dienst te maken, en dat was binnen OVIS", aldus Boves.

Ook de opzet met twee concurrerende methoden voor NLP werd door de meerderheid van de stuurgroep geroemd. Anton Nijholt is

Zink Typografie Sheet 185 of 213 - Page 177 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



wel kritisch wanneer hij ernaar gevraagd wordt: "dat [programma] was naar mijn mening nog te veel linguïstisch georiënteerd. Met ontzettend veel aandacht voor syntactische analyse, waarbij zowel in Groningen als in Amsterdam groepen met ontleedmethoden bezig waren." Maar in de eindevaluatie prijst de stuurgroep de dialoog die het op gang bracht tussen aanhangers van probabilistische methoden en die van kennisgedreven methoden. Maar wat was nou de uitslag van de competitie tussen die twee? In de test met 1000 ongeziene testinputs deed de parser van Gertjan van Noord en Gosse Bouma het 30% beter wanneer gewerkt werd met de ruwe output van de spraakherkenner, en 40% beter op de nette zinnen (hypothetische perfecte transcriptie). Bovendien was de software van de Groningers sneller dan die van de groep uit Amsterdam (Veldhuijzen van Zanten, Bouma, Sima'an, van Noord & Bonnema 2000). Of zoals Van Noord tot op de dag van vandaag op zijn website heeft staan: Ajax–FC Groningen: 0–6.

Alice Dijkstra ziet in het OVIS-programma het nut geïllustreerd van de samenwerking tussen taal- en spraaktechnologie: "ze zijn bij elkaar gaan stelen. Mijn simpele opvatting is dat ze zich beide bezighouden met patroonherkenning en statistiek, alleen de een meer statistiek en minder patroontjes, en de ander meer patroontjes, minder statistiek. Dus kunnen ze van elkaar dingen overnemen." Ook Frank Van Eynde gelooft dat het goed is om van elkaars methodologieën te leren: "De ervaring van spraaktechnologen in de statistische verwerking van gegevens is ook nuttig voor taaltechnologen." Maar Dijkstra geeft ook meteen toe dat het 'stelen' van methodologie ook voortkwam uit de competitie tussen statistische en kennisgebaseerde methoden binnen de taalverwerking: "Gertjan [van Noord] is hier een voorbeeld van geweest. Als iemand op intelligente wijze heeft gestolen van de statistici, dan is dat Gertjan. En binnen dat programma hebben we dat ook ingebouwd, omdat we de taalmodule twee keer hebben laten implementeren, een keer door Remko [Scha], en een keer door Gertjan. Remko zat bij elke meeting te vertellen wat voor geweldige ideeën hij had, alleen implementeerde hij ze niet, en Gertjan dacht 'oh, dat kan ik er ook nog wel even naast hangen, als een heuristiekje.' " Van Noord benoemt een andere factor die heeft bijgedragen aan het succes van zijn groep in het OVIS-project: "Ik moest me bewijzen." Daarbij haalt hij

Zink Typografie Sheet 186 of 213 - Page 178 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



een quote aan van Remko Bonnema, die samen met Khalil Sima'an het Amsterdamse Data-Oriented Parsing model geïmplementeerd had. "Remko heeft wel eens gezegd: 'als Groningen het DOP-model had moeten implementeren en wij de grammatica, dan had Groningen ook gewonnen. De drive om te winnen was gewoon veel groter omdat ze in de underdog-positie zaten.'"

Terwijl Nijmegen, Eindhoven, Groningen en Amsterdam aan het OVIS-systeem werkten, werd in Twente een eigen dialoogsysteem ontwikkeld. Al ruim voor het Prioriteitsprogramma startte, bouwden Anton Nijholt en zijn informaticastudenten Schisma, een dialoogsysteem met virtueel assistent Karin voor schouwburginformatie en -boekingen. Nijholt vertelt hoe dat project ontstaan is: "Dat had te maken met het feit dat we binnen informatica een plek moesten krijgen en dat we in de richting van *human machine interaction* gingen. Dat was in het begin niet zo duidelijk, dat ontwikkelde zich langzaam, maar het idee dat we in de toekomst dialoogsystemen zouden hebben die via natuurlijke taal en eventueel via spraak aanstuurbaar moesten zijn, was duidelijk een belangrijk onderwerp. Daar hebben we toen op ingezet." Het Muziekcentrum Enschede leverde ieder jaar de volledige programmering, en Nijholts groep zorgde ervoor dat die informatie bevraagbaar was. De opzet van het systeem was gebaseerd op keyboardinput. "Maar", zo stelt Franciska de Jong, "er was wel steeds het idee dat dat ook met een gesproken interface zou moeten kunnen. En daar zijn ook wel studentprojecten voor geweest." In deze projecten werd vooral ervaring opgedaan met het inbouwen van bestaande spraakherkenningssoftware in het dialoogsysteem. Volgens de Jong een illustratie van hoe in Twente naar spraak gekeken werd: "De spraaktechnologie zelf was niet het onderzoeksterrein, maar het werd wel geïntegreerd in dingen die bij ons gemaakt en bestudeerd werden." Er is wel een versie van Schisma met geïntegreerde spraaksynthese beschikbaar geweest. "Het werkte", verzekert Nijholt. "Het heeft jarenlang gedraaid. Maar het werd niet door het grote publiek gebruikt." Belangrijkste reden hiervoor was dat het systeem wel online kaartjes kon reserveren, maar het Muziekcentrum dat niet wilde. "Dat zou problemen geven met de andere manieren van reserveren, via de telefoon of aan de balie. Ze waren niet in staat dat in hun structuur op te nemen."

Zink Typografie Sheet 187 of 213 - Page 179 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Dialoogsystemen zijn een mooi voorbeeld van samenwerking tussen taal en spraak op een toepassing. Dat is een terrein waar taal- en spraaktechnologen elkaar graag treffen. Iedereen is het erover eens dat voor veel applicaties een samenwerking nodig is tussen taal en spraak: niet alleen dialoogsystemen, maar ook bijvoorbeeld zoeken in audiobestanden of Computer Assisted Language Learning (CALL). "Toepassingen waar spraaktechnologen alleen niet tot een goed resultaat gaan komen, maar waar de taaltechnologen ook het niet waar kunnen maken zonder spraaktechnologie", vat spraaktechnoloog Jean-Pierre Martens samen. Bovendien kon voor toegepassingsgerichte onderzoekprojecten subsidie aangevraagd worden binnen de Europese Kaderprogramma's. Taaltechnoloog Franciska de Jong komt op die manier bijna vanzelfsprekend bij spraaktechnologie terecht als een volgende stap na het Pop-Eyeproject, dat zich richtte op crosslinguaal zoeken in videobeelden op basis van geïndexeerde ondertitels. De Jong: "Toen we zagen dat we retrieval deden met een afgeleide waarvoor handwerk nodig was, en waarbij we te maken hadden met IPR op ondertitels èn we merkten dat er ook heel veel materiaal was waarvoor dat niet beschikbaar was, zagen we dat dat dus aandacht voor spraakherkenning vergde." Het erop volgende project in Europees verband, Olive genaamd, richtte zich dan ook op crosslinguaal zoeken in de geïndexeerde spraak van televisieuitzendingen. En dat gaf op zijn beurt weer een impuls aan het Nederlandse spraakonderzoek, omdat de Franse partij die voor de spraakherkenning zou zorgen, geen zin had in het ontwikkelen van een Nederlandse spraakherkenner. In Nederland was zo'n herkenner ook niet voorhanden, omdat de de spraakherkenning van OVIS geen *broad-coverage* spraakherkenner was. Dat leverde dus een probleem op, dat echter snel omgezet werd in een mooie kans: het Telematica-instituut beloonde het voorstel van TNO, het Amsterdamse CWI en de Universiteit Twente voor onderzoeksproject Druid met een forse subsidie om een open domein spraakherkenner voor het Nederlands te bouwen en in te bedden in het multimedia retrieval systeem dat binnen Olive ontwikkeld werd.

Zink Typografie Sheet 188 of 213 - Page 180 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



## 5.2 Samen werken aan corpora: het Corpus Gesproken Nederlands

Een ander terrein waarop taal- en spraaktechnologie elkaar gemakkelijk vonden, was de ontwikkeling van corpora. Spraaktechnologen zaten al enige tijd te springen om trainingsdata voor hun herkenners op basis van statistische modellen zoals Hidden Markov Models. "Er was een duidelijk tekort aan materiaal. 'No better speech than more speech' ", parafraseert Louis Pols de beroemde woorden 'no data like more data' van Bob Mercer[2]. Maar ook taaltechnologen en taalkundigen hadden baat bij een corpus van gesproken Nederlands.

In de eerste plaats vormen corpora een heel goede bron van taalkundige voorbeelden. Lange tijd bestond het testmateriaal voor een taalkundige theorie vooral uit handmatig geconstrueerde voorbeelden op basis van de eigen intuïtie over wat wel of niet grammatikaal is en voorbeelden uit de vakliteratuur. Hoewel die voorbeeldzinnen vaak heel duidelijk zijn en de mogelijkheid bieden om alles weg te laten wat niet relevant is, hebben ze ook hun beperkingen. Met name voor infrequente constructies is het lastig om goede, natuurlijk klinkende voorbeelden te verzinnen. Soms wordt daaruit onterecht geconcludeerd dat een bepaalde constructie ongrammaticaal is. Zo is geclaimd dat reflexieve werkwoorden niet met 'zichzelf' kunnen combineren. Maar Van Noord en Bouma laten zien dat in grote corpora wel degelijk grammaticale voorbeelden hiervan gevonden kunnen worden ('Nederland moet stoppen zichzelf op de borst te slaan') (van Noord & Bouma 2009). In hetzelfde artikel ontkrachtten ze ook de claim dat rechtsextrapositie uit een topic niet mogelijk is – eveneens met voorbeelden die in een groot, automatisch geparseerd corpus gevonden werden.

Corpusmateriaal is helemaal onmisbaar als voor wie erkent dat grammaticale regels niet per definitie categoriaal zijn. Zo is de algemene regel dat in het Engels meervouden niet kunnen voorkomen als modifiers in samenstellingen (*'jobs market'). En toch zijn er een heel aantal voorbeelden waarin wel meervouden voorkomen ('financial services industry', 'bonds market'). Zodra de taalkundige toestaat

[2] Tijdens de Arden House Workshop on Speech Recognition in 1985.

Zink Typografie Sheet 189 of 213 - Page 181 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



dat iets infrequent of onwaarschijnlijk is, heeft hij kwantitatieve data nodig om de hypothese te onderbouwen. Ook hiervoor is corpusmateriaal onontbeerlijk.

Daarnaast begonnen ook de taaltechnologen geannoteerde corpora te gebruiken als trainingsdata. Voor sommige op grammatica's gebaseerde parsers waren modules gebouwd op basis van probabilistische modellen, bijvoorbeeld disambiguatiemodules. Parsers genereren vaak grote hoeveelheden syntactische analyses van inputzinnen op basis van de (handgeschreven) grammaticaregels. De overgrote meerderheid van deze parses is correct in de zin dat alle deelregels goed zijn toegepast, maar veel van de combinaties van deelanalyses zijn heel onwaarschijnlijk. Welke combinaties wel en niet waarschijnlijk zijn, kan geleerd worden op basis van syntactisch geannoteerde corpusdata, feitelijk niet meer dan een grote verzameling expliciete voorbeelden van goede analyses. En natuurlijk waren er inmiddels ook taaltechnologen die hun systemen volledig *data-driven* maakten, zoals het DOP-systeem. Voor deze doeleinden hadden ook taaltechnologen corpusdata nodig. En hoewel er al wel wat corpusmateriaal van geschreven Nederlands was, ontbrak het nog aan een verzameling teksten gesproken taal, die duidelijk verschilt van de geschreven variant.

Tenslotte wilden taaltechnologen corpusdata gebruiken om hun eigen werk te evalueren. Zij hadden geleerd dat het een goed idee was om niet alleen naar zelfverzonnen of op grammaticaboeken gebaseerde voorbeeldzinnen te kijken, maar de kwaliteit van een parser te testen op echte taaluitingen. En een belangrijk deel van die taaluitingen bestaat nu eenmaal uit gesproken taal. Op het terrein van de corpusontwikkeling vonden beide taal- en spraaktechnologen elkaar dan ook als vanzelf.

Het verzamelen van taaldata is niet alleen een taak waar taal- en spraaktechnologen elkaar ontmoetten, maar ook een waar Nederlandse en Vlaamse wetenschappers elkaar vonden. Weliswaar verschilt het Nederlands zoals dat gesproken wordt in Nederland wel van dat in Vlaanderen, maar de twee regio's vormen toch duidelijk één taalgebied, en het leek dan ook verstandig om de inspanningen te bundelen. Dat gebeurde dan ook in 1998, in het project Corpus Gesproken Nederlands (CGN). Dit project werd gefinancierd door de Vlaamse en Nederlandse overheden, die hier in totaal bijna vijf

Zink Typografie Sheet 190 of 213 - Page 182 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



miljoen euro in investeerden, waarvan tweederde uit Nederland, en eenderde uit Vlaanderen. Het doel was een corpus op te bouwen van tien miljoen woorden, waarvan eveneens tweederde uit Nederland, en eenderde uit Vlaanderen zou moeten komen. Verschillende instellingen uit beide gebieden werkten mee, en het resultaat zou publiekelijk toegankelijk gemaakt worden. In eerste instantie distribueerde ELDA (Evaluations and Language resources Distribution Agency) de data, maar in 2004 werd het corpus overgedragen aan de Vlaams Nederlandse organisatie de Taalunie.

De grootste driver voor CGN was het succes dat in het buitenland bij spraakherkenning behaald werd met de combinatie van statistische methoden en grote hoeveelheden data. Louis Pols: "DARPA had ieder jaar weer geweldige collecties van nieuw spraakmateriaal, waarmee je de herkenners trainde en testte, en telkens werden weer nieuwe variabelen, meervoudige microfoons, of telefoonkanaal, of ruis, of sprekers in een vreemde taal, en dat ging maar door, steeds meer materiaal. Je zag dat met die HMM-techniek, dat die *brute force* benadering van data buitengewoon succesvol was. Maar wij hadden niks in het Nederlands. Je kon wel zo'n HTK-pakket [Hidden Markov Model Speech Recognition Toolkit] kopen of lenen of huren, maar er was geen materiaal om het te trainen. De applicaties lagen allemaal in de telefoonsfeer, maar we hadden niet één opname van telefoonspraak. Daar was buitengewone behoefte aan."[3] Volgens Pols is het uiteindelijk Pim Levelt van het Max Planck Instituut geweest die via goede contacten bij de overheid de doorslag heeft gegeven." En toen ontbrandde de discussie over wat er dan precies in dat corpus moest komen. Iedereen had zijn eigen wensen, voor spraakherkenning, synthese en dialectonderzoek. Uiteindelijk werd het een diverse verzameling opnames van monologen en dialogen, voorgelezen teksten, spontane spraak, telefoongesprekken en vergaderingen. Sommigen zouden zeggen: mooi gebalanceerd. Maar Pols noemt het 'een schaap met vijf poten': "[Door de opzet] kon geen van de applicaties goed

[3] Jan Odijk nuanceert deze uitspraak: "Binnen Europa is in Nederland het telefoonspraakcorpus gemaakt dat vervolgens als model heeft gediend voor de SpeechDat telefoondatabanken, verkrijgbaar via ELDA. Er was dus wel data voor het Vlaams en het Nederlands, maar het kostte wel enkele tienduizenden euro's." (persoonlijke communicatie)

Zink Typografie Sheet 191 of 213 - Page 183 of 205

gerealiseerd worden. Bijvoorbeeld een synthesizer ontwikkelen met het CGN is bijna niet mogelijk, want je hebt niet genoeg materiaal van één spreker. En een echte herkenner trainen met genoeg materiaal kan ook net niet." Ondanks die kritische kanttekening werd het CGN direct na release een zeer veel gebruikt corpus. Wel was duidelijk dat er nog winst te halen was met méér data. Martens: "Als je nu ziet waarom conversationele spraak niet goed werkt, dan zou dat wel eens kunnen zijn dat één van de redenen daarvoor is dat we eigenlijk onvoldoende data hebben om taalmodellen voor spontane spraak te maken. Neem nu CGN, dat zijn maar tien miljoen woorden, en daarvan is maar drie of vier miljoen spontaan. Op drie miljoen kun je geen taalmodellen genereren die dezelfde kracht hebben als een taalmodel dat je op vijf biljoen woorden traint."

## 5.3 Samen werken aan technologie?

De samenwerking tussen taal en spraak binnen CGN bestond er niet alleen uit dat beide partijen het corpus nodig hadden, maar zeker ook uit het feit dat besloten was dat de spraakcorpora geannoteerd moesten worden met allerlei taalkundige informatie. Louis Pols: "Ook de spraakmensen hadden ondertussen wel door dat je met talige informatie je herkenner kon helpen. Dus dat materiaal moest niet alleen fonetisch getranscribeerd worden op woordniveau, maar dat moest geparseerd worden, daar moest syntactische analyse op losgelaten worden. Zodat je ook die kennis kon gebruiken. Dat hadden we wel door, maar we wisten zelf niet hoe dat moest, dus dan haalde je de taaltechnologen erbij."

Dat doet de vraag rijzen of er nog een derde terrein is waarop spraak- en taaltechnologie kunnen samenwerken: taaltechnologie als component voor betere spraaktechnologie. Wordt spraaktechnologie beter door kennis van taal aan het model toe te voegen? Hierover zijn de meningen verdeeld. Dat een bepaalde hoeveelheid taalkundige informatie nodig is om spraaksynthese te realiseren, daar is men het wel over eens. "Allicht moet je taalkundige kennis hebben", stelt ook Jean-Pierre Martens. "Maar het punt is een beetje dat dat alleen maar heel algemene kennis hoeft te zijn. Als je die eenmaal hebt, is het ge-

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



daan. Dus als ik kijk naar de spraaksynthese, er zijn misschien al vier generaties synthesizers gemaakt, maar die bevatten nog altijd de taalkundige kennis van de eerste generatie, in essentie." Frank Van Eynde legt uit: "Bij spraaksynthese is belangrijk te weten welke woordsoort een bepaald woord is, bijvoorbeeld om te kiezen tussen *voornáám* en *vóórnaam*. Het leggen van de juiste klemtoon en intonatie wordt vergemakkelijkt als je syntactische analyse van de zin hebt."

Maar hoe zit dat met spraakherkenning? Voegt taalkundige kennis daar iets toe? Martens: "Allicht wel. Maar er zijn al heel veel mensen die zich daarop stukgebeten hebben. Het is zeer moeilijk om louter statistische methoden zoals N-grammen te kloppen met taalkennis." Walter Daelemans heeft later nog, in het FLAVOR-project (2002–2006) geprobeerd om met taalkundige kennis de spraakherkenning te verbeteren. In hoofdstuk 3 zagen we al dat hij heeft moeten constateren dat het niet echt iets opgeleverde. Maar Daelemans wil het wel blijven proberen. Boves heeft echter een heel andere richting in gedachten voor onderzoek naar automatische spraakherkenning. Boves is – ondanks de overwinning van FC Groningen – nog steeds overtuigd dat de formele beschrijving van taalkundige kennis in een grammatica een metaniveau van modelleren is dat voor spraakherkenning geen toegevoegde waarde heeft. "Ik denk echt dat als je automatische spraakherkenning helemaal goed wilt doen, je dichter bij het soort representaties moet komen dat kinderen opbouwen als ze hun taal leren, dan bij een formele beschrijving zoals we die nu hebben. Het is lastig om over spraak te praten anders dan dat je woorden hebt en dat woorden bestaan uit rijtjes van klanken. Maar als je gaat kijken hoe het in werkelijkheid is, dan is dat helemaal niet zo. Woorden zitten aan elkaar, en als je al gaten ziet in een signaal, dan zijn het even vaak gaten binnen wat normaliter een woord heet als daartussen. Klanken zijn geen dingen die na elkaar komen, dat overlapt elkaar allemaal. Al onze bestaande spraakherkenningsystemen zijn gebaseerd op het idee dat je spraak kunt beschrijven als een sequentie van klanken en van woorden. En daar moeten we op een of andere manier vanaf. […] En misschien moet je je ook van gewone kinderen afvragen of de representaties die ze bouwen wel fonemen worden voordat ze leren lezen."

In de taaltechnologie is de rol van statistiek (ondanks de uitkomsten van OVIS) zó hard gegroeid, dat zelfs de vraag gesteld kan worden

Zink Typografie Sheet 193 of 213 - Page 185 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



of voor taaltechnologie eigenlijk nog taalkundige kennis nodig is. Ook hierop zijn verschillende antwoorden mogelijk. Van Eynde denkt dat de rol van taal juist weer zal gaan toenemen: "De puur statistische aanpak is aan een plafond gekomen. Bij automatisch vertalen zie je ook dat weer meer gebruik wordt gemaakt van features die taalkundig zijn. Taaltechnologie heeft wel heel sterke transformatie ondergaan, toen Eurotra, Rosetta, BSO en METAL nog liep. De transformatie is dat veel van de huidige systemen vaak hybride zijn. Een parser met een grammatica, maar ook met een statische disambigueringscomponent. […] Alpino bijvoorbeeld: had eerst alleen regels, maar na Rob Malouf ook veel statistiek. Dat creeërt een grote rol voor treebank mining. Ik zie dat uitgebreid worden met semantische representaties, dat is nu aan de gang bij SONAR. […] Het Paco-project voor *corpus based machine translation* is ook een voorbeeld. Monolinguale analyse met kennisgebaseerde parsers, en dan de bilinguale stap op een manier die gebaseerd is op statistiek. Hybride modellen. Dat is toch zo'n beetje de richting waarin ik het zie evalueren. Weinigen die puur regels of puur statistiek doen. Iedereen is op zoek naar de juiste mix."

De grote vraag bij hybride systemen is dan welk van de twee componenten de leiding neemt. Daelemans: "Ga je eerst proberen om de statistiek goed te hebben en dan taalkundige kennis toevoegen, of ga je eerst proberen de taalkundige kennis goed te hebben en dan statistiek eraan plakken. Ik denk dat de meeste hybride systemen het tweede type onderzoek doen." Zelf is hij meer voorstander van de omgekeerde werkwijze: eerst de statistiek goed hebben, zorgen dat je goede modellen hebt, en daarna eventueel je representaties verrijken met taalkundige kennis. "En dan ga je zien dat dat toch niet zo heel veel toevoegt aan wat je kan met oppervlakkige representaties. Volgens mij heb je aan oppervlakkige representaties genoeg, moet je gewoon harder werken aan de statistische modellen. Meer data, andere data." Het grote struikelblok voor het combineren van kennisgebaseerde en probabilistische technieken is volgens Daelemans dat de tools om taalkundige kennis te genereren niet goed genoeg zijn. "Als je parser maar 80% accuraat is, dan gaat het meer kwaad doen dan goed." Hoewel de accuratesse van parsers inmiddels wel een stuk hoger ligt, komt Daelemans dit probleem nog steeds tegen in zijn huidige onderzoek: "Ik doe nu onderzoek naar auteursherkenning op basis van tekst. Daar hadden we zulke mooie linguïstische features,

Zink Typografie Sheet 194 of 213 - Page 186 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



patronen van woordsoorten, dependenties, heel mooie taalkundige features die toch iets zouden moeten zeggen over stijl. Maar wat werkt goed? Bigrammen van karakters. Dat werkt significant beter. Omdat het robuust is: je kan geen fouten maken bij het genereren van de representaties. Terwijl één cruciale *part of speech* die verkeerd gelabeld is, zeg een werkwoord in plaats van een zelfstandig naamwoord, kennis kan vernietigen wanneer je het voor verdere verwerking gebruikt. Dus het grote probleem voor hybride systemen is dat je voor het linguïstische gedeelte je corpus of je data niet accuraat genoeg kan annoteren. Zolang je niet boven de 97% zit, genereer je meer ruis dan kennis."

De Jong benadrukt dat het niet alleen gaat om wat er de beste cijfers oplevert, maar ook hoe de resultaten geïnterpreteerd kunnen worden: "Natuurlijk zie je dat naarmate de statistische systemen geperfectioneerd raken dat de kans op verbeteringen daar afneemt. Op een gegeven moment zit je aan je maximum. De vraag is of er dan nog verbetering te bereiken is door alsnog iets te doen aan die lagen waar je met expliciete regels iets kunt bereiken, en of dat in performancepercentages imponerende cijfers zijn. Neem de part-of-speech taggers. Dat kun je op basis van statistiek doen, en daarmee kun je performancepercentages krijgen van 98% correct ofzo. Stel dan dat je er dan meer regels instopt, dan kan je misschien 99% bereiken. Imponerend wordt het langs die as nooit. Maar bevredigender misschien wel. Misschien is het van belang dat er mensen zijn die benadrukken dat niet alleen de performance belangrijk is, maar ook het inzicht in de onderliggende modellen. Maar dat verhaal komt natuurlijk alleen maar aan bij mensen die ook een belang hebben in taalkunde onafhankelijk van de toepassing. De gemiddelde ingenieur zit daar niet op te wachten."

Er is ook nog een heel andere richting in het onderzoek naar taal en spraak en die richting wordt aangevoerd door Luc Steels. Steels keerde rond 1995 na een studie computerwetenschappen en loopbanen in expertsystemen, genetische algoritmen en robotica terug bij zijn stiel: taal. En wel de structuur van taal. "Het is nog steeds *deep language processing* dat mij interesseert. Niet de statistische benadering van taal." Tegelijkertijd is hij pessimistisch over de resultaten en het potentieel van de kennisgebaseerde aanpak: "We gaan nooit taal kunnen programmeren. Taal is te ingewikkeld om te programmeren.

Zink Typografie Sheet 195 of 213 - Page 187 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Ik denk dat dat niet kan. Dus wat is dan het alternatief? We gaan een systeem bouwen dat zelf taal leert." De Vlaming heeft van Sony *carte blanche* gekregen om een onderzoekslab in Parijs op te zetten, waar hij met de robots van Sony het beroemde Talking Heads experiment uitvoerde (Steels 1999). In dit project modelleerde Steels de ontwikkeling van een taal door de interactie van twee 'agents'. In de eerste experimenten waren de agents nog vrij statisch, computers met camera's, die naar gekleurde figuren kijken in een plat vlak. Later werden ze vervangen door meer geavanceerde exemplaren, echte robots, uitgerust met een camera en microfoon. De ene wijst naar een object in een driedimensionale ruimte, waarbij hij een woord zegt. De andere robot ziet waarnaar gewezen wordt, en gaat ervan uit dat het woord naar dat object verwijst. Maar het weet nog niet of het woord verwijst naar de kleur, maat, vorm of positie van het woord. Pas door heel veel interacties kan het die beslissing maken, en zo ontwikkelen de agents een manier om met elkaar over de objecten te praten.

Een laatste, pragmatische reden voor samenwerking van taal- en spraaktechnologie is het verwerven van onderzoeksgelden. Met de sterk toenemende digitalisering leeft de angst dat een kleine taal als het Nederlands niet mee zou kunnen komen. Voor bedrijven is het namelijk weinig lucratief om tijd en geld te steken in de ontwikkeling van technologie en applicaties voor het Nederlands, omdat de afzetmarkt relatief klein is in vergelijking met talen als het Engels. Deze angst raakt zowel de spraak- als de taaltechnologie, en aangezien de vereiste ontwikkelingen voor een deel overlappen, in ieder geval op het gebied van de ontwikkeling van goede datacollecties, zijn de twee samen opgetrokken bij het aanvragen van subsidiegelden. In 1998 schreven Gosse Bouma en Ineke Schuurman het rapport 'De positie van het Nederlands in Taal- en Spraaktechnologie',[4] waarin de auteurs precies op basis van deze zorg pleiten voor gezamenlijk onderzoek en onderwijs, en een Vlaams-Nederlands platvorm voor TST. Dat platvorm komt er (tijdelijk), en geeft op haar beurt opdracht tot de studie 'Het Nederlands in taal- en spraaktechnologie: prioriteiten voor basisvoorzieningen' van Walter Daelemans en Helmer Strik, waarin gepleit wordt voor de ontwikkeling van de ontbrekende componenten

[4] http://www.ccl.kuleuven.be/Papers/rapport_NTU_BoumaSchuurman.pdf

Zink Typografie Sheet 196 of 213 - Page 188 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



van de BAsisTAalVOorzieningen (BATAVO) voor TST. Het rapport opent: "Om de status van het Nederlands te vrijwaren in een snel evoluerende multilinguale informatiemaatschappij moeten ICT toepassingen snel en adequaat in een Nederlandstalige versie ontwikkeld worden." In de jaren die hierop volgden zou dit een sterk argument blijken, waarmee aanzienlijke subsidieprogramma's gefinancierd werden. Terecht, vindt Louis Pols: "Taal- en spraaktechnologie is een noodzakelijk vakgebied. De technologische ontwikkeling van je eigen taal moet je in eigen hand houden. Dat zou best een boodschap mogen zijn voor zo'n historisch overzicht."

Zink Typografie Sheet 197 of 213 - Page 189 of 205

Vlaams wis- en natuurkundige Simon Stevin (1548–1620).

# 6 Epiloog

"De technologische ontwikkeling van je eigen taal moet je in eigen hand houden." De woorden van professor Louis Pols toen ik hem op 16 oktober 2009 sprak. Onder meer om deze reden ging op 15 september 2004 al het STEVIN-programma van start. STEVIN staat voor Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands. Het programma maakt bovendien graag gebruik van de associatie met Simon Stevin, een belangrijk Vlaams wetenschapper in – hoe kan het ook anders – de natuurkunde en de wiskunde. STEVIN is een onderzoeksprogramma op het gebied van taal- en spraaktechnologie, waarin Vlaanderen en Nederland nauw samenwerken, en ook de samenwerking tussen industriële en academische spelers sterk benadrukt wordt. De doelen van het programma zijn het stimuleren van de taal- en spraaktechnologische sector en de vraag producten uit deze sector en het ontwikkelen van een digitale taalinfrastructuur voor het Nederlands. Veiligstellen van de technologische ontwikkeling van het Nederlands dus. Het programma heeft met een budget van 11,4 miljoen euro bijna twintig onderzoeksprojecten gefinancierd en ook nog eens bijna twintig flankerende activiteiten, zoals demonstratie- of educatieve projecten. Ook het belangrijkste doel,

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



de ontwikkeling van een zeer groot corpus geschreven Nederlands, wordt binnen dit programma gerealiseerd. Het programma loopt in 2011 af. Hoewel er andere programma's lopen waarin TST een component kan zijn, is het nog onduidelijk of er een vervolgprogramma zal komen voor fundamenteel onderzoek in TST.

Zink Typografie Sheet 200 of 213 - Page 192 of 205

# Bibliografie

Alinei, M. (1962). *Dizionario Inverse Italiano*, Mouton, Den Haag.

Alinei, M. (1963). Lexical, grammatical, and statistical indexing of Italian texts with the help of punched card machines at the University of Utrecht, *Levende Talen* **219**: 166–173.

Alinei, M. (1965). La lista di frequenze della divina commedia, *Miscellanea Dantesca*, Utrecht/Antwerpen, pp. 138–270.

Alinei, M. (1968/1977). *Spogli Elettronici dell'Italiano delle Otigini e del Duecento, II, Forme (A Linguistic Inventory of Old Italian) 20 Volumes*, Il Mulino, Bologna.

ALPAC (1966). *Language and Machines. Computers in Translation and Linguistics*, National Academy of Sciences.

van Bakel, J. (1970). Automatische Syntactische Analyse van Nederlandse Teksten, *Technical report*, Universitair Rekencentrum KU Nijmegen.

van Bakel, J. (1975). Automatische Zinsontleding met de Computer, *Technical report*, KU Nijmegen.

Battus, H. (1983). *Rekenen op taal*, Querido, Amsterdam.

van der Beek, L. (2001). Van Beth tot Van Benthem: de opkomst van de Nederlandse semantiek, *Tabu* **1–2**: 1–60.

van der Beek, L., Bouma, G., Malouf, R. & van Noord, G. (2002). The Alpino Dependency Treebank, *in* M. Theune, A. Nijholt & H. Hondorp (eds), *Computational Linguistics in the Netherlands CLIN 2001; Selected Papers from the Twelfth CLIN Meeting*, Rodopi.

Bennett, W. S. & Slocum, J. (1985). The LRC machine translation system, *Computational Linguistuistics* **11**(2–3): 111–121.

Bennett, W. S., Slocum, J., Whiffin, L. & Norcross, E. (1985). An evaluation of METAL: the LRC machine translation system, *Second conference of the European Chapter of the Association for Computational Linguistics*, University of Geneva, Geneva, Switzerland, pp. 62–69. `http://acl.ldc.upenn.edu/E/E85/E85-1009.pdf`.

van Berckel, J., Brandt Corstius, H., Mokken, R. & van Wijngaarden, A. (1965). *Formal properties of newspaper Dutch*, Mathematisch Centrum, Amsterdam.

Bernhard, U. (1991). Evaluation of Machine Translation Systems from a User's Viewpoint. Some critical comments, *Proceedings of the Evaluators' Forum*.

Boersma, P. & Weenink, D. (2001). Praat: doing phonetics by computer, `http://www.fon.hum.uva.nl/praat`. computer program.

Boves, L., Duifhuis, H., van Eijk, D., Kempen, G., Koster, C. & Scha, R. (1992). Prioriteitsprogramma taal- en spraaktechnologie. Voorstel voor de inrichting van een prioriteitsprogramma Taal- en Spraaktechnologie door NWO, intern rapport.

Boves, L., Landsbergen, J., Scha, R. & van Noord, G. (1995). NWO Priority Programme Language and Speech Technology.

Brandt Corstius, H. (1970). *Exercises in computational linguistics*, PhD thesis, Universiteit van Amsterdam.

Brandt Corstius, H. (1974). *Algebraïsche taalkunde*, Oosthoek, Utrecht.

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Brandt Corstius, H. (2003). De desillusie van mijn leven of Remember November, *in* T. Gaustad (ed.), *Computational Linguistics in the Netherlands 2002: Selected Papers from the Thirteenth CLIN Meeting*, Rodopi, Amsterdam/New York.

Bronnenberg, W., Bunt, H., Landsbergen, S., Scha, R., Schoenmakers, W. & van Utteren, E. (1980). The question answering system PHLIQA1, *in* L. Bolc (ed.), *Natural Language Question Answering*, Hanser, pp. 217–305.

Brown, P., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Mercer, R. & Roossin, P. (1988). A statistical approach to language translation, *Proceedings of the 12th conference on Computational linguistics*, Association for Computational Linguistics, Morristown, NJ, USA, pp. 71–76.

Bunt, H. (1980). Reisverslag 114. IPO, Eindhoven.

Bunt, H. (1981). *The formal semantics of mass terms*, PhD thesis, Universiteit van Amsterdam.

Burton, D. M. (1981). Automated concordances and word indexes: the early sixties and the early centers, *Computers and the Humanities* **15**: 83–100.

Chomsky, N. (1957). *Syntactic Structures*, Mouton, Den Haag.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*, MIT Press.

Chomsky, N. (1975). *The logical structure of linguistic theory*, Plenum Press, New York. In 1955 verschenen op microfiche.

Cohen, A. & 't Hart, J. (1963). Speech synthesis of steady-state segments, *Proceedings of the Speech Communication Seminar, Stockholm 1962, F1*.

Cohen, A. et al. (1984). Taal- en Spraaktechnologie in Nederland, Persoonlijk archief van Alice Dijkstra.

Coppen, P.-A. (2001). Een mooie manier om taal te beschrijven; interview met drie hoogleraren computerlinguistiek, *Onze Taal* **11**.

de la Court, J. (1937). *De meest voorkomende woorden en woordcombinaties in het Nederlandsch; Verslag van een onderzoek in opdracht van het Departement van Onderwijs en Eeredienst*, Batavia.

Zink Typografie Sheet 203 of 213 - Page 195 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Debille, L. (1986). Het basismodel van het EUROTRA-vertaalsysteem, *Automatische vertaling aan de K.U. Leuven*.

Dijkstra, E. W. (1980). A programmer's early memories, *in* N. Metropolis, J. Howlett & G.-C. Rota (eds), *A History of Computing in the Twentieth Century: a Collection of Essays*, Academic Press, New York, p. 563–573.

Droste, F. (1969). *Vertalen met de computer; mogelijkheden en moeilijkheden*, Wolters-Noordhoff.

Gazdar, G., Klein, E. H., Pullum, G. K. & Sag, I. A. (1985). *Generalized Phrase Structure Grammar*, Basil Blackwell, Oxford.

Gebruers, R. (1986). Het vertaalsysteem METAL, *Automatische vertaling aan de K.U. Leuven*, pp. 29–78.

van Heuven, V. J. & Pols, L. C. W. (eds) (1993). *Analysis and Synthesis of Speech*, Mouton de Gruyter.

Hutchins, W. J. (1988). Recent developments in machine translation: a review of the last five years, *in* D. Maxwell, K. Schubert & T. Witkam (eds), *New directions in machine translation: conference proceedings, Budapest 18–19 August 1988*, Foris Publications.

IJlander, G. (1990). *Een fabelachtig uitzicht*, Veen uitgevers.

Joris, F. (2005). *Jo Lernout; mijn verhaal*, Houtekiet.

Jurafsky, D. & Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (Prentice Hall Series in Artificial Intelligence)*, Prentice Hall.

Kaplan, R. & Bresnan, J. (1982). Lexical-Functional Grammar: A formal system for grammatical representation, *in* J. Bresnan (ed.), *The mental representation of grammatical relations*, The MIT Press.

Kučera, H. & Nelson Francis, W. (1967). *Computational Analysis of Present-Day American English*, Brown University Press, Providence.

Zink Typografie Sheet 204 of 213 - Page 196 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Landsbergen, J. (1981). Adaptation of Montague grammar to the requirements of parsing, *in* J. Groenendijk, T. Janssen & M. Stokhof (eds), *Formal methods in the Study of Formal Language Part 2, MC tract 136*, Mathematical Centre, Amsterdam, pp. 399–420.

Landsbergen, S. (1977). Wat is computerlinguistiek?, *in* B. T. Tervoort (ed.), *Wetenschap en Taal*, Countinho, pp. 47–65.

Landsbergen, S. (2000). Schoonheid en troost van de taaltechnologie. Afscheidsrede uitgesproken op 29 juni 2000.

Linschoten, J. (1963). *De la Court's frekwentietelling van Nederlandse woorden*, Psychologisch Laboratorium, Rijksuniversiteit Utrecht.

Martin, W. (1967). *De inhoud van krant en roman*, Uitgeverij Plantyn, Antwerpen.

Martin, W. (1970). *Analyse van een vocabularium met behulp van een computer*, AIMAV, Brussel.

Mineur, A.-M. (1992). "Wij waren misschien wat minder slim dan Gazdar en Kaplan en Bresnan, maar wij dachten toch wel op dezelfde manier", *Ta!* **1**(4). interview met Louis des Tombe.

Mineur, A.-M. (1993). "Even rustig van het leven genieten", *Ta!* **2**(1). interview met Gertjan van Noord.

Mols, B. (2010). De computer kletst maar wat, *NRC Handelsblad* (25 september, Wetenschapsbijlage).

Montague, R. (1973). The Proper Treatment of Quantification in Ordinary English, *in* J. Hintikka, J. Moravcsik & P. Suppes (eds), *Approaches to Natural Language*, p. 221–242. Reprinted in Thomason, 1974.

Newman, M. H. A., Turing, A. M., Jefferson, G. & Braithwaite, R. B. (1952). Can automatic calculating machines be said to think?, From the 1952 BBC Broadcast. Available from `http://philosophy.tamucc.edu/readings/minds-and-machines/can-automatic-calculating-machines-be-said-to-think` and reprinted in "The Turing Test: Verbal

Zink Typografie Sheet 205 of 213 - Page 197 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Behavior as the Hallmark of Intelligence", ed. S. Shieber, Cambridge, Massachusetts: The MIT Press.

van Noord, G. & Bouma, G. (2009). Parsed Corpora for Linguistics, *Proceedings of EACL Workshop The Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, Athens, pp. 33–39.

van Noord, G., Dorrepaal, J., van der Eijk, P., Florenza, M. & des Tombe, L. (1990). The MiMo2 Research System, *MT conference 1990 Austin Texas*. Available from `http://www.let.rug.nl/~vannoord/papers`.

Nooteboom, S. G. (1996). Antonie Cohen: obituary, *Phonetica* **53**: 230–232.

Palmer, M. & Finin, T. (1990). Workshop on the Evaluation of Natural Language Processing Systems, **16**(3): 175–181. `http://acl.ldc.upenn.edu/J/J90/J90-3005.pdf`.

Pedersen, T. (2008). Empiricism Is Not a Matter of Faith, *Computational Linguistics* **34**(3): 465–470.

Plomp, R., Pols, L. & van der Geer, J. (1967). Dimensional analysis of vowel spectra, *Journal of the Acoustical Society of America* **41**: 707–712.

Pollard, C. & Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*, University of Chicago Press.

Reichling, A. (1935). *Het woord: een studie omtrent de grondslag van taal en taalgebruik*, Berkhout, Nijmegen.

Rosetta, M. (1994). *Compositional translation*, Kluwer Academic Publishers.

Scha, R. (1983). *Logical Foundations for Question Answering*, PhD thesis, Rijksuniversiteit Groningen.

Sciarone, A. (1980). Over automatisch vertalen, Inaugurele rede.

Smeiers, F. (2007). De Eerste Nederlandse Computers, `http://www.student.tue.nl/p/f.a.m.smeijers/2r930/index.htm`.

Zink Typografie Sheet 206 of 213 - Page 198 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Steels, L. (1999). *The Talking Heads Experiment. Volume 1. Words and Meanings*, Antwerpen. Special pre-edition for LABORATORIUM, Antwerpen 1999.

Stoop, A. M. (1995). *TRANSIT: a linguistically motivated Dutch-Turkish MT-system*, PhD thesis, Katholieke Universiteit Nijmegen.

Swift, J. (1726). *Gulliver's reizen naar verschillende verre landen*, J.H. Gottmer & Co., Haarlem. Bewerking van G. Blom, 1940.

Takeda, K., Nasukawa, T., Uramoto, N. & Tsutsumi, T. (1992). Shalt2: a symmetric machine translation system with conceptual transfer, *Proceedings of the 14th conference on Computational linguistics*, Association for Computational Linguistics, Morristown, NJ, USA, pp. 1034–1038.

Uit den Boogaart, P. (ed.) (1975). *Woordfrequenties in geschreven en gesproken Nederlands*, Oosthoek, Scheltema & Holkema, Utrecht.

Veldhuijzen van Zanten, G., Bouma, G., Sima'an, K., van Noord, G. & Bonnema, R. (2000). Evaluation of the NLP Components of the OVIS2 Spoken Dialogue System, *in* F. van Eynde, I. Schuurman & N. Schelkens (eds), *Computational Linguistics in the Netherlands 1998*.

Weaver, W. (1949). Translation, Repr. in: Locke, W.N. and Booth, A.D. (eds.) *Machine translation of languages: fourteen essays* (Cambridge, Mass.: Technology Press of the Massachusetts Institute of Technology, 1955), pp. 15–23.

van Wijngaarden, A. (1952). *Rekenen en vertalen*, Uitgeverij Waltman, Delft.

Willems, L. (1966). The intonator, IPO Annual Progress Report 1, 123–125, beschikbaar via `http://www.let.uu.nl/~Sieb.Nooteboom/personal/HistoryIPOspeechsynthesis/Willems66a.pdf`.

Witkam, T. (1988). DLT – An Industrial R&D Project for Multilingual MT, *Proceedings of the 12th conference on Computational linguistics*, Association for Computational Linguistics, Morristown,

Zink Typografie Sheet 207 of 213 - Page 199 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



NJ, USA, pp. 71–76. http://www.aclweb.org/anthology/C/C88/C88-2154.pdf.

Witkam, T. (2005). A new road to Automatic Translation, http://www.mt-archive.info/systems-1.htm.

Witkam, T. (2006). History and Heritage of the DLT (Distributed Language Translation) project, http://www.mt-archive.info/Witkam-2006.pdf.

de Witte, R., van Aelst, L. & van Peteghem, L. (2002). *De L&H files; Lernout en Hauspie, het verhaal en de geheimen*, Lanno.

Zink Typografie Sheet 208 of 213 - Page 200 of 205

# Index

Aarts, Jan, 84, 87
Adriaens, Geert, 61, 63
Aelst, Luc van, 64, 146
Akker, David van den, 133
Al, Bernard, 68, 128
Alinei, Mario, 18
Appelo, Lisette, 74, 76, 81, 82, 87, 89, 124
Arends, Jan, 130, 131
Armstrong, Susan, 93
Arnold, Doug, 53
Atal, Bishnu, 108

Bösser, 125
Bakel, Jan van, 15, 20, 21, 37, 74, 83, 84, 87, 88
Baker, James, 156, 160
Baker, Janet, 156–158, 160
Bakker, D.M., 19
Bakker, J., 18
Bakker, J.J.M., 19
Bar-Hillel, Yehoshua, 6, 43
Bastiaans, Gaston, 81, 152, 153, 156, 159
Bennett, Winfield, 60
Benthem, Johan van, 28, 29, 166
Berckel, J.A.Th.M. van, 18
Berg, B. van den, 19
Beth, Evert Willem, 7–9, 28, 29, 42, 43, 46, 93
Bie, Peter de, 133, 134, 151, 152
Bobbert, Ilja, 163
Bodenkamp, Stephan, 64, 150, 158
Bodson, Philip, 159
Boersma, Paul, 106
Boes, Bernard, 141
Boitet, Christian, 47, 51, 52

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Bonnema, Remko, 168, 179
Boot, Martin, 12, 37, 38, 55
Booth, Andrew, 6
Bos, Edwin, 129
Bosch, Linde van den, 92
Bouma, Gosse, 90, 91, 178, 188
Bouma, Herman, 125, 176
Bourlard, Hervé, 143, 147–149
Boves, Lou, 99, 102, 103, 106, 108–110, 114–118, 173–177, 185
Brandt Corstius, Hugo, 8–14, 16–23, 30, 36–39, 42, 48, 57, 65, 71, 77, 84, 85, 88, 121
Bresnan, Joan, 37, 52
Bronnenberg, Wim, 30
Bunt, Harry, 12, 22–25, 27, 29, 30, 33, 38, 68, 84, 85, 113, 122, 126
Burgt, Stan van de, 88, 138–140, 163, 164

Caeyers, Herman, 60–64, 161
Cambell, H. Wilfred, 89
Chomsky, Noam, 23, 36, 37
Cohen, Toon, 98–100, 102, 105, 109, 116, 125
Coile, Bert van, 143
Collier, René, 100, 104, 115, 125
Compernolle, Dirk van, 107, 110, 112, 113, 140–153, 159, 161, 162
Compernolle, Geert van, 143
Coppen, Peter-Arno, 87

Daan, J., 19
Daelemans, Walter, 87–89, 92, 95, 118, 119, 155, 156, 161, 185, 186, 188
De la Court, J.F.H.A., 17–19
De Smedt, Koenraad, 39, 89
De Witte, René, 64, 146
Deemter, Kees van, 89
Dijkstra, Alice, 38, 55, 82, 88, 89, 126, 166, 173–175, 178
Dik, Simon, 85
Dirksen, Arthur, 136
Dorrepaal, Joke, 93
Dotson-Smith, Bill, 14
Driesten, Peer van, 150, 151
Droste, Flip, 10, 12
Dudley, Homer, 97, 98
Duerden, John, 157, 158

Eggermont, J.P.M., 19
Eijck, Jan van, 175
Emde Boas, Peter van, 29
Engels, L.K., 18, 38, 47, 100

Feitsma, A., 19
Fellinger, Carel, 76
Fillmore, Charles J., 16
Frimout, Dirk, 155

Garis, Hugo de, 155
Gebruers, Rudi, 61
Geens, Dirk, 46–48, 52, 132
Geer, J.P. van der, 101
Gent, Joop van, 165, 166
Gerhard Daalenoord, 23, 25
Goethals, 146

Zink Typografie Sheet 210 of 213 - Page 202 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Graaf, Tjeerd de, 103
Groenendijk, Jeroen, 29
Guiraud, P., 18

Hacken, Pius ten, 59
Hamme, Hugo van, 161
Hart, Johan 't, 100
Hauspie, Pol, 135, 141, 142, 145, 149, 154, 155, 159, 165
Hessen, Arjan van, 138
Heuvel, Theo van de, 130
Heylen, Dirk, 87
Hoekstra, Heleen, 87
Hoenkamp, Ed, 11
Hutchins, John, 44, 65, 73

IJlander, Gijs, 66

Jaap Calis, 23, 25
Janssen, Theo, 29, 63, 82, 93, 167
Janssens, Paul, 132
Jaspaerts, Lieven, 53
Jelinek, Fred, 102
Johnson, Mark, 87
Jong, E.D. de, 19
Jong, Franciska de, 14, 32, 76, 82, 89, 90, 162, 163, 179, 180, 187

Kaiser, L., 18
Kamp, Hans, 28, 42, 43, 57
Kamp, Yves, 102
Kaplan, Ron, 37, 52
Kay, Martin, 16, 90
Kempelen, Wolfgang von, 97, 98
Kempen, Gerard, 68, 70, 87–89, 126–129
King, Maghi, 46–49, 52
Klooster, Wim, 36, 42
Koeling, Rob, 168
Konst, Leo, 127
Korkin, Michael, 155
Kraak, Albert, 23, 36, 42
Krauwer, Steven, 16, 37, 38, 48, 49, 52–59, 74, 75, 84, 87, 93, 94
Kurzweil, Raymond, 158
Kwee Tjoe Liong, 37

Ladefoged, Peter, 105
Landsbergen, Jan, 8, 10, 12, 15, 22–27, 29–31, 33, 35, 36, 38, 47–49, 63, 66–68, 71–84, 89, 122–125, 127, 175, 176
Landsbergen,Jan, 76
Leermakers, René, 76
Lernout, Jo, 135, 141–145, 149, 153–155, 158, 159
Lernout, Jol, 135
Levelt, Pim, 183
Lisman, J.H.C., 18
Louvezijn, Ellen van, 74
Luimes, Wim, 137

Maas, Dieter, 47, 52
Malouf, Rob, 186
Maremont, Mark, 158
Martens, Jean-Bernard, 109
Martens, Jean-Pierre, 104, 105, 107–109, 113, 136, 161, 180, 184, 185

Zink Typografie Sheet 211 of 213 - Page 203 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Martin, Willy, 18, 21, 38, 92
Medema, Jeroen, 76
Medema, Piet, 25, 26, 30
Meinsma, G.L., 18
Menert, Ludmilla, 136
Mercer, Bob, 181
Mergen, Norbert, 130–132
Metz, Berend, 163, 164
Meulen, Alice ter, 29
Mierlo, A.M. van, 19
Mijnhout, Agnes, 76
Mineur, Anne-Marie, 93
Mokken, R.J., 18
Mol, Hendrik, 104
Monigny, Rudy, 132–135, 151
Montague, Richard, 8, 28, 29, 72, 89
Moortgat, Michael, 93
Mulder, Jan Willem, 137
Munster, Elly van, 76

Nerbonne, John, 91
Ney, Hermann, 102, 115
Nijholt, Anton, 89, 90, 139, 158, 175, 177, 179
Noël, Jacques, 39
Noord, Gertjan van, 54–57, 87, 91, 175, 176, 178
Nooteboom, Sieb, 98, 100, 102–105, 108, 109, 113, 114, 124–126

Odijk, Jan, 58, 59, 65, 73, 76–82, 89, 123–125, 127, 128, 142, 152–154, 158–161, 183
Olde Kalter, Alex, 67
Ollongren, Alexander, 87
Os, Els den, 139, 140

Pedersen, Ted, 12
Perschke, Sergei, 46, 47, 56, 59
Peteghem, Luc van, 64, 146
Pieper, Roel, 156
Pinillos Bartolomé, Elena, 76
Plomp, Reinier, 101, 107
Pols, Louis, 101–104, 106–108, 110, 115–117, 181, 183, 184, 189, 191

Rayfeld, Burt, 28
Reichling, Anton, 36, 98
Rietveld, Ad, 134
Rijk, Rudolf de, 42
Rolf, Piet, 37
Rooij, J. de, 19
Rous, Joep, 71, 73, 76

Sanders, Margreet, 76
Sas, Jan van, 132–135, 151, 152, 160
Scha, Remko, 22–33, 38, 85, 86, 94, 95, 122, 123, 173, 175, 176, 178
Schank, Roger, 27
Schenk, André, 74, 76, 81, 124
Schoenmakers, Wijnand, 25
Schouten, Jan, 98
Schultink, Henk, 36, 48
Schuurman, Ineke, 92, 188
Schweigmann, Herman, 23, 24, 122
Sciarone, Bondi, 39, 44, 45, 48, 54, 57, 68, 75, 83, 84
Seuren, Pieter, 36, 42
Sikkel, Klaas, 89
Sima'an, Khalil, 179
Slocum, Jonathan, 60
Smit, Harm, 76, 77

Zink Typografie Sheet 212 of 213 - Page 204 of 205

29 december 2010, 15:14 Leonoor van der Beek - Van rekenmachine tot taalautomaat



Smits, Harm, 42
Soest, J.L. van, 18
Spaan, Tigran, 166, 167
Spitholt, Martin, 137, 138
Staal, Frits, 36
Steels, Luc, 16, 39, 92, 187, 188
Steen, Gert van der, 37
Stevens, Jan, 76
Stokhof, Martin, 29
Stoop, Albert, 83
Strik, Helmer, 188
Stuart, Don Graham, 91, 103
Stumpers, F.L., 18

Tarski, Alfred, 9, 10
Tervoort, B.Th., 19
Tesnière, Lucien, 51, 52
Tombe, Louis des, 37, 48, 49, 52–58, 74, 81, 84, 87, 93
Tomita, Masaru, 90
Tromp, Herman, 109
Troost, Albert, 84
Tsuji, Junichi, 69

Uit den Boogaart, P.C., 19
Uittenbogaard, Frank, 76
Utteren, Eric van, 25, 30

Van den Brande, 149
Van Eynde, Frank, 46, 47, 51, 53, 55, 58, 63, 92, 132, 155, 161, 178, 185, 186
Vandeghinste, Vincent, 161
Vanwormhoudt, Marc, 143
Vauquois, Bernard, 47, 51
Veldhuijzen van Zanten, Gert, 177
Veltman, Frank, 29
Verkuyl, Henk, 29
Vieregge, Willem, 103
Visser, Henk, 42
Vosse, Theo, 70, 88, 89, 164
Vossen, Piek, 115, 135, 152, 166
Vries, M. de, 18
Vuijst, Jan de, 86, 90
Vuysje, D, 18

Waibel, Alex, 158
Walker, Don, 28
Weaver, Warren, 6, 43, 94
Wellekens, Christian, 102
Welling, George, 91
Wijngaarden, Adriaan van, 5–10, 13, 18–20, 22, 35, 39, 42, 43
Wilks, Yorick, 16, 47
Willaert, Nico, 142, 159
Willems, Lei, 99, 125
Winograd, Terry, 11, 16, 27, 28, 39
Wintzen, Eckart, 65, 67, 68, 70, 75
Wit, Petra de, 76
Witkam, Toon, 63, 65–70, 75, 80, 84, 126
Woods, Bill, 16

Zampolli, Antonio, 16, 47
Zanelatto, Georges, 143
Zavrel, Jakub, 156, 167, 168
Zuijlen, Job van, 70
Zwaneveld, A.M., 19
Zwarts, Frans, 11, 86, 90, 91

Zink Typografie Sheet 213 of 213 - Page 205 of 205