

DIXIT

TIJDSCHRIFT OVER TOEGEPASTE TAAL- EN SPRAAKTECHNOLOGIE

Eurospeech 2005

Ubiquitous Speech Processing

Intelligent Design

Over wetenschapsdefinities en de zin van het leven



De stem van
WFH ontrafeld
*Audiovisuele archieven
geïndexeerd*

VAN DE MAKERS VAN VAN DALE XIII

van Dale

XIV



NU IN DE
BOEKHANDEL

"Opnieuw een meesterwerk ...
een literaire standaard voor jong en oud."
Thomas Rosenboom

"Van Dale, de enige vorm
van vandalisme die ik toejuich."
Harry Mulisch

"Van Dale is voor mij
de moeder van alle woordenboeken."
Daphne Deckers

IN DRIE DELEN MET CD-ROM

DOOR DRS. TON DEN BOON EN PROF. DR. DIRK GEERAERTS
ETYMOLOGIE DOOR DR. NICOLINE VAN DER SIJS

www.VanDaleXIV.nl

VAN DALE LEXICOGRAFIE UTRECHT . ANTWERPEN

van Dale

INHOUD

Het Groene Boekje en de Hulk

Op 15 oktober is de nieuwe editie van de Wordenlijst Nederlandse Taal verschenen. De Wordenlijst is op verschillende punten aangepast volgens de herziene spellingregels. Een gevolg van die herziening is dat een zo groot mogelijke spellingseenheid kon worden gecreëerd: alle toonaangevende woordenboeken, spellinghandboeken en spellingcheckers zullen vanaf nu één en dezelfde spelling hanteren.

4

Eurospeech 2005

Congressen hebben de naam verkapt plezierisjes te zijn. Niet altijd, zo ontdekten de 1300 taal- en spraaktechnologen die zich hadden verzameld in Lissabon. Twee dagen praten over dialogen (SigDial), één dag besteed aan de huidige stand van zaken (Tutorials) en vier hele sessiedagen. Ronduit slopend, maar de resultaten mogen er wezen. Arjan van Hessen doet verslag.

12

De stem van WFH ontrafeld

De WFH-portal biedt een grote hoeveelheid informatie over Hermans en zijn schrijven én, uniek voor Nederland, ook de mogelijkheid om gedetailleerd te zoeken in Herman's gesproken woord: in interviews, lezingen en voorgelezen verhalen. Dat er nog een hoop te verbeteren valt, is duidelijk. Het verbeteren van de mogelijkheden om snel en doeltreffend een spraakherkenningsysteem aan te passen aan de eisen van een willekeurig taakdomein behoort hiertoe. Ook wat betreft zoekopties en de presentatie van de zoekresultaten kan nog het nodige werk worden verzet. Niettemin is duidelijk dat manieren om ook audiovisueel materiaal te 'googelen', steeds dichterbij komen.

15

En verder...

TST als hulpmiddel voor mensen met communicatieve beperkingen
Intelligent Design
Taal in Bedrijf
TST-centrale actief in Antwerpen
NOTaS Nieuws

6

9

10

14

18

Evolutie van een vakgebied

Het knusse Team voor Taal heeft plaatsgemaakt voor een professioneler Taal in Bedrijf. Een totaalprogramma met workshops, plenaire zittingen, een informatiemarkt, en een locatie die technologie en onderzoek door zijn betonnen fundamenten heeft stromen (het Evoluon). Waar Team voor Taal een eerste kennismaking bleek van de activiteiten van de diverse NOTaS-leden, daar wil Taal in Bedrijf de vele toepassingen van taal- en spraaktechnologie voor het voetlicht werpen. Niet meer voor elkaar, maar voor de buitenwereld. In zekere zin belichaamt de overgang naar Taal in Bedrijf de gestage volwassenwording van het vakgebied TST. Louder onderzoek volstaat niet meer, het gaat meer en meer om praktische en commerciële toepassingen.

Zonder NOTaS geen Team voor Taal, en zonder Team voor Taal geen Taal in Bedrijf. Eén van de doelstellingen van NOTaS was en is het vervullen van een brugfunctie tussen onderzoek en toepassing, en tussen toepassing en commercialisering. Tijdens Team voor Taal bleek al snel hoe nodig een dergelijke brug was. Bedrijven en instellingen kenden elkaar wel, maar dan van horen zeggen. Wat vooral in het oog sprong was dat sommige instellingen bezig waren met onderzoek waar de inbreng van commerciële partijen node gemist werd, terwijl bedrijven worstelden met problemen waar de onderzoekswereld al geruime tijd een oplossing voor had.

De tijd leek rijp voor een grote doorbraak: van NOTaS maar zeker ook voor het gehele TST-veld. Leek rijp, want hoewel er veel ten goede is veranderd, is er nog geen sprake van een kentering. De rituele paringsdans die werd uitgevoerd tijdens Team

voor Taal heeft nog niet geleid tot een bijzonder vruchtbaar huwelijk. NOTaS is wel gegroeid, maar het is een teken aan de wand dat de grote spelers zich nog altijd niet hebben aangesloten. De marktrijen zijn kennelijk nog niet gesloten, en dat betekent dat de gehele branche nog altijd niet met één mond spreekt. Of, op termijn veel gevaarlijker, dat de belangen te ver uiteen liggen.

Even zorgelijk is de trend dat onderzoeksinstellingen en academische semi-commerciële spin-offs zich te makkelijk laten laven door allerlei subsidies. Een gezonde kenniseconomie krijg je niet door geld te fourneren, en echte innovaties zijn ook al zelden het gevolg van forse subsidie-injecties. Innovatie komt eigenlijk altijd tot stand onder druk van noodzaak, bijvoorbeeld de noodzaak om investeringen in onderzoek en ontwikkeling binnen een bepaalde termijn terug te verdienen (in het volgende nummer van DIXIT zal Norbert Mergen zijn licht over deze zaak laten schijnen; onder de provocerende titel 'Innovatiesubsidies werken averechts' gaat hij even op de ondernemersstoel zitten).

Genoeg gemijmerd, en zeker 'gemopperd'. TST is een vakgebied dat meer dan toegevoegde waarde heeft. Zo kunt u verderop lezen hoe TST mensen met communicatieve beperkingen het leven aangenamer maakt. En hoe we met behulp van TST audiovisuele archieven kunnen ontsluiten. Als het congres Taal in Bedrijf de opmaat is voor Taal én Bedrijf, is een terugkeer naar Team voor Taal, maar dan brancheoverkoepelend, voor de hand liggend.

Pierre Pieterse
Hoofdredacteur DIXIT

Het Groene Boekje en de

Op 15 oktober verschijnt de nieuwe editie van de Woordenlijst Nederlandse Taal. De Woordenlijst is op verschillende punten aangepast volgens de herziene spellingregels. Een gevolg van die herziening is dat een zo groot mogelijke spellingseenheid kon worden gecreëerd: alle toonaangevende woordenboeken, spellinghandboeken en spellingcheckers zullen vanaf nu één en dezelfde spelling hanteren.

Om die eenheid te bevorderen krijgen uitgevers en producenten van taalmaterialen van de NTU een keurmerk waaruit blijkt dat het product conform de herziene spellingsregels is. Het INL heeft daaraan voorafgaand wél het product door de mangel van de spellingscontrole gehaald: het Hulpmiddel Keurmerk. Over een vruchtbare samenwerking tussen de Nederlandse Taalunie, Polderland, het INL, én de belangrijkste woordenboekproducenten.

MACHIENTJE

Binnen drie maanden de trefwoorden van de vijf meest toonaangevende woordenboeken- en lexiconproducenten en uitge-

vers van spellinghandboeken van het Nederlandse taalgebied verifiëren én corrigeren volgens de herziene spellingsregels. Hoe doe je dat? Hadden we nu maar een machientje dat ons daarbij zou kunnen helpen, verzuchtte men bij de Taalunie. Welnu, die *deus ex machina* kwam er ook.

In opdracht van de Taalunie maakten Polderland en het INL een online service die een groot deel van het correctiewerk uit handen nam. Polderland zorgde voor de technologie; het INL voor de benodigde lexica en, last but not least, menskracht. Het resultaat is het **Hulpmiddel Keurmerk** (alias 'HulK').

PROCES

De spellingcorrectie van trefwoorden met de HulK gaat als volgt in zijn werk: eerst gaat de woordenboekproducent of producent van spellinghandboeken (de 'aanbieder') naar een speciaal voor de HulK ingerichte internetsite, gehost door de TST-centrale (de Centrale voor Taal- en Spraaktechnologie) aan het INL. Nadat de voorwaarden en bepalingen verwoord in een gebruikerslicentie door middel van een vinkje zijn geaccepteerd, wordt de aanvrager geïnstrueerd over de uploadprocedure en over de resultaten die automatisch naar hem worden gemaïld.

Eenmaal ingelogd uploadt de aanvrager in één keer alle trefwoorden van alle producten waarvoor hij een keurmerk heeft aangevraagd. Elke aanbieder kreeg, dat spreekt voor zich, ruim van tevoren een gebruikersnaam en wachtwoord om in te loggen op deze HulK-site.

Vervolgens begint de HulK aan zijn analyseslag. Is het woord onbekend, dan wordt er geprobeerd toch een toelichting bij te geven. Hierbij wordt gebruik gemaakt van de linguïstisch intelligente componenten die Polderland in de HulK heeft aangebracht. Toelichtingen variëren van: 'dit woord is een samenstelling', tot: 'het is niet *cottage cheese*' maar '*cottagecheese*'. Of: 'dit woord ken ik niet, maar moet waarschijnlijk met een tussen-n geschreven

| MICHEL BOEKESTEIN |



HulK: spelling is groen

worden'. De woorden waar de HulK een toelichting bij heeft gevonden, stuurt hij terug naar de aanbieder, zodat deze zijn trefwoordenlijst meteen kan corrigeren. Woorden die de HulK niet kent, en waarvoor hij bovendien geen suggestie kan vinden, moeten handmatig doorlopen worden. Hier komen de speciaal voor dit project opgeleide controleurs in actie, waarover later meer.

Met de resultaten van hun manuele werk worden de lexica van de HulK vervolgens geüpdatet. De aanbieder uploadt zijn aangepaste trefwoordenlijst opnieuw, en de cyclus (uploaden – corrigeren – updaten) herhaalt zich totdat de HulK alle woorden kent, en bovendien goed gespeld vindt. De HulK maakt na iedere upload een logbestand aan, waarin de aanbieder onder andere het aantal onbekende woorden en het aantal woorden met toelichting kan terugvinden. Als beide aantallen nul zijn, weet de aanbieder dus dat zijn geüploade trefwoorden allemaal goedgekeurd zijn door de HulK.

DE CONTROLEURS

De controleurs, zeven linguïsten uit Vlaanderen en Nederland, hebben zich in zeer beperkte tijd de herziene spellingsregels eigen gemaakt. Samen hebben zij in twee maanden tijd ruim 300.000 woorden bekeken en, zo nodig, gecorrigeerd. In totaal werden ruim 1,8 miljoen woorden geüpload.

Het overgrote deel van die woorden bestonden uit de hiervoor beschreven onbekende woorden: woorden die bij de HulK onbekend waren, en waar bovendien geen toelichting (van het type 'veilige samenstelling' of 'bekende fout') bij gevonden werd.

De woorden met toelichting ontsnapten echter ook niet aan de rode pen van de controleurs. Deze woorden geven namelijk, samen met hun toelichtingen, waardevolle informatie over de kwaliteit van de door HulK gebruikte lexica. Zo worden niet alleen de aangeboden trefwoorden gecontroleerd, maar konden ook de HulK-

lexica constant worden gemonitord en up-to-date gehouden.

Bij twijfelgevallen communiceerden de controleurs via een schoolbordsysteem.

Het HulK-team

Aan de HulK hebben gewerkt Tineke Koster, Wilko Apperloo, Marijke Koster (Polderland); de keurmerkers waren Griet Depoorter, Dirk Glandorf, Dirk Kinable, Kenny Louwen, Katrien Van Pellicom, Bart Roelandts, Veronique Verreycken en stonden onder leiding van Jeannine Beeken en Michel Boekestein.



Vragen en antwoorden werden naar de hele groep gemaild. Bovendien was er een centrale plaats voor de referentielexica die constant door de projectleider bijgewerkt werden aan de hand van de genomen beslissingen.

KEURMERK

Het resultaat van deze samenwerking mag er zijn en zal worden gezien. Van Dale Lexicografie, Het Spectrum, Polderland, SDU Uitgevers, Lannoo en Wolters Plantyn mogen het keurmerk op verschillende van hun producten voeren; de spellingeenheid is er en zij is mede het gevolg van de uiterst efficiënte en zeer gewaardeerde samenwerking tussen de verschillende ontwikkelaars, de aanvragers, de controleurs en de HulK.

Michel Boekestein is computerlinguïst verbonden aan de TST-centrale INL

Wat heeft TST te bieden voor met communicatieve beperkin

| THEO VAN DEN HEUVEL & LISANNE TEUNISSEN |

Gevraagd naar de zin van taal- en spraaktechnologie verwijst menig ontwikkelaar naar de oplossingen die het vakgebied oplevert voor blinde of dove gebruikers en voor mensen met andere communicatieve beperkingen. Denk aan spraaksynthese om tekst voor te lezen aan visueel gehandicapten, of spraakherkenning om doven juist toegang te bieden tot gesproken informatie. Toegang tot talige informatie en de mogelijkheid om taal te

gemaakt van de beschikbare technische mogelijkheden. De Taalunie gaf Polderland Language & Speech Technology de opdracht deze vraag te beantwoorden. Polderland schakelde de hulp in van prof. dr. Toni Rietveld (Radbouduniversiteit Nijmegen), dr. Jetske Klatter-Folmer (Radbouduniversiteit Nijmegen en ViaTaal) en een aantal andere experts op het gebied van hulpmiddelen en talige handicaps. Er zijn

opmerkingen te maken. Het maakt een groot verschil of iemand een aangeboren stoornis heeft, of dat die op later leeftijd ontstaat.

Blindgeborenen kunnen bijvoorbeeld veel beter en sneller braille lezen; op latere leeftijd is deze vaardigheid erg moeilijk om aan te leren.

Doofgeborenen blijken vaak problemen te hebben met lezen en schrijven (een grote groep is zelfs functioneel analfabeet), terwijl schrift juist heel belangrijk is voor wie op latere leeftijd gehoorproblemen krijgt.

Beperkingen gaan vaak samen ('comorbiditeit'), hetgeen de mogelijkheden om TST-toepassingen te gebruiken weer beperkt. Er zijn allerlei graden waarin iemands gehoor, gezichtsvermogen en spraakorganen kunnen zijn aangetast. Leeftijdsverschillen zijn erg bepalend voor de behoeften. Voor ouderen speelt de leerdrempel vaak een grotere rol en is er vooral een behoefte aan eenvoudig aan te leren en te gebruiken oplossingen.

Synthetisch fluisteren

Mensen die van kunstmatige spraak afhankelijk zijn om te communiceren, zijn gedwongen te kiezen uit het arsenaal van bestaande synthetische stemmen. Het klinkt natuurlijk raar als een jong meisje spreekt met een kunstmatige oudeman-nestem. Het blijkt dat de gebruikers ook behoefte hebben aan een fluisterstem. Soms wil je immers iets zeggen wat niet voor ieders oren bestemd is ('ik vind jou ook heel leuk', 'moet je hem daar zien' of 'ze is morgen jarig'). De hoorder zal een gefluisterd bericht ook anders behandelen. Niet kunnen fluisteren is een behoorlijke beperking.

kunnen produceren zijn sterk bepalend voor de relatie tussen de mens en zijn omgeving. Elke hindernis die wordt afgebroken is een bevrijding. Je zou dan ook verwachten dat deze groep enthousiast van allerlei nieuwe technologieën gebruik maakt en dat hun wensen een belangrijke inspiratiebron zijn voor genoemde ontwikkelaars.

ONDERZOEKEN...

De Nederlandse Taalunie wilde wel eens weten of die aansluiting in Nederland en Vlaanderen inderdaad bestaat en of er gebruik wordt

tientallen gesprekken gevoerd met vertegenwoordigers van belangengroepen, instellingen, onderzoekers en ervaringsdeskundigen. Het resultaat is een rapport, dat de Taalunie u graag aanbiedt (http://taalunieversum.org/taal/technologie/communicatieve_beperkingen/).

Een samenvatting is ook beschikbaar. In dit artikel proberen we daarom niet om nog een samenvatting te geven, maar lichten we er wat zaken uit.

De wensen en behoeften van gebruikers zijn zo verschillend, dat het moeilijk is om algemeen geldende

...EN CONCLUDEREN

Het rapport benoemt tientallen manieren waarop taal- en spraaktechnologie effectief gebruikt kan worden om stoornissen te diagnosticeren, te helpen herstellen of beperkingen te omzeilen en de gevolgen ervan te verminderen. Het is duidelijk dat technologie al intensief gebruikt wordt in een breed scala van gebieden. Sommige gebruikersgemeenschappen zijn groot, bijvoorbeeld de groep van slechtzienden. De grootte van de groep (naar schatting zijn er 685.000 blinden en slechtzienden in Nederland en Vlaanderen) en het niveau van organisatie maken een goed gebruik van beschikbare

gebruikers gen

Toegang tot websites

De mode om websites steeds meer te voorzien van flash animaties en andere dynamische versiering, heeft de toegankelijkheid van het internet voor visueel gehandicapten behoorlijk begrensd. Het rapport ondersteunt dan ook het beroep dat verschillende organisaties (zoals 'Drempels weg' en 'Blindsurfer') doen op website-bouwers om de toegankelijkheid voor blinden te garanderen.

technologieën mogelijk. Andere toepassingen zijn heel specifiek, bijvoorbeeld die die betrekking hebben op combinaties van stoornissen.

De betrokkenen slagen er goed in om aan te geven wat ze zouden wensen aan verbeteringen en verfijningen. Zo wordt spraakherkenning op allerlei plaatsen toegepast, bijvoorbeeld door RSI-patiënten. Het biedt hen de mogelijkheid teksten te schrijven zonder dat ze hun handen hoeven te belasten. In tegenstelling tot wat je misschien zou verwachten, zijn dictateersystemen voor blinden juist geen oplossing, omdat blinde gebruikers geen feedback krijgen of kunnen geven over de ingevoerde tekst. Vanwege het zelflerend karakter van de systemen leidt dit snel tot een verslechtering van de herkenning. In situaties waar een beperkt bereik aan invoerwoorden voldoet, zoals bij aansturing van een kopieerapparaat, lijkt spraakherkenning wél nuttig voor blinden en slechtzienden.

De criteria voor bruikbaarheid van de technologie zijn:

- gebruiksvriendelijkheid (de wensen van gebruikers voor eenzelfde instrument kunnen heel verschillend zijn, ook bijvoorbeeld afhankelijk van hun leeftijd);
- bekendheid met de mogelijkheden (er zijn vele duizenden instrumenten en het is niet eenvoudig om een goede keuze te maken kijkend naar de specificaties alleen);
- kwaliteit (bijvoorbeeld, de betrouwbaarheid van spraakherkenners bepaalt hun bruikbaarheid; kwaliteit kan alleen beschouwd worden in relatie tot het beoogd gebruik);

- toepasbaarheid (denk bijvoorbeeld aan de mensen rond de gebruiker: soms moet een therapeut of verzorger een instrument instellen op de specifieke gebruikssituatie).

ROL VAN HET BEDRIJFSLEVEN

Het rapport gaat uitdrukkelijk in op de rol van het bedrijfsleven. Er blijkt een gat te gapen tussen de wetenschappelijke en technologische mogelijkheden en de beschikbaarheid van producten. Dat komt onder andere omdat het Nederlandse taalgebied relatief klein is en de producten waaraan behoefte bestaat, elk afzonderlijk maar een kleine doelgroep kunnen bedienen. De gebruikers zijn veelal afhankelijk van bijdragen van overheden en verzekeraars om het product te kunnen betalen. Omgekeerd wordt daardoor de prijs van het product en de nazorg sterk bepaald door budgetten die hiervoor ter beschikking worden gesteld. Daarbij is er een groot verschil tussen de Vlaamse en de Nederlandse situatie. Ook is er een groot verschil tussen producten die het leven aangenaam maken en producten die een gebruiker toegang geven tot de arbeidsmarkt, omdat de laatste zichzelf terugverdienen. Het blijft wenselijk om producten die voor de grotere markten zijn ontwikkeld, ook geschikt te maken voor het Nederlands (lokaliseren). Daarvoor is het van belang dat bij het ontwerp al met lokalisatie rekening wordt gehouden, wat helaas lang niet altijd gebeurt. Daarnaast is er behoefte aan meer ondersteuning bij toepassingen (door de leverancier of door een intermediair), en aan meer aandacht voor flexibiliteit van producten. Zo

zou er rekening gehouden moeten worden met verslechterende of verbeterende beperkingen. Samenvattend, ondanks dat er een behoorlijk aantal producten beschikbaar is voor gebruikers met communicatieve beperkingen, is er ruimte voor en behoefte aan nieuwe producten en lokalisatie van bestaande buitenlandse producten.

Doelgroepen

Het onderzoek onderscheidt de volgende stoornissen:

Afasie (taalstoornis ten gevolge van hersenletsel)

Dyslexie (problemen met lezen en spellen)

Verstandelijke beperkingen

Blind- en slechtziendheid

Doof- en slechthorendheid

Doofblindheid

Dysartrie en anartrie (problemen met articulatie)

Mutisme (niet kunnen spreken)

Stotteren

RSI/ABBE ('muisarm')

Dyspraxie en apraxie (problemen met motoriek)

Theo van den Heuvel is managing director van Polderland Language & Speech Technology BV en Lisanne Teunissen is verbonden aan de Nederlandse Taalunie



Een prikkelende vraag aan alle verzekeraars:

'U doet er alles aan om het uw klant zo gemakkelijk mogelijk te maken... maar waarom lukt dat niet via uw website?'

Uw website is een uitstekend middel voor het verkopen van verzekeringsproducten maar dan moeten uw klanten het verhaal wél begrijpen. Veel verzekeraars maken het hun klanten niet makkelijk een keuze te maken en direct online een verzekering af te sluiten. Een gemiste kans.

Met de oplossingen van Q-go kunt u online de dialoog met uw klant aangaan. Het resultaat is dat u online meer verkoopt en betere service verleent tegen lagere kosten. In tegenstelling tot zoekmachines begrijpt Q-go de betekenis van vragen die klanten stellen en geeft meteen gericht antwoord. U genereert gedetailleerde klantinformatie waarmee u de interactie met uw klanten al binnen enkele weken volledig optimaliseert.

Q-go is dé toonaangevende Europese aanbieder van oplossingen voor Online Marketing & Selfservice. Met onze oplossingen helpen wij bedrijven meer rendement te halen uit hun online dienstverlening.

Op dit moment verwerken wij bijna vier miljoen vragen per maand voor o.a:



Wilt u weten wat Q-go voor u kan doen? Kijk op www.q-go.nl of bel 020 531 3800.



Online Marketing & Self Service

Intelligent Design

Misschien moet u inmiddels al weer even graven in uw geheugen, maar in juni van dit jaar woedde er een verhit debat over *Intelligent Design*. Dat soort onderwerpen wordt naar mijn idee altijd als aftrainoefening aan kamerleden aangeboden, ter afsluiting van het seizoen, en als appetizer voor de traditionele barbecue. Wat was er gebeurd?

Minister Maria van der Hoeven had op haar weblog opgemerkt dat ze zo'n boeiend gesprek had gehad met Cees Dekker, hoogleraar moleculaire biofysica in Delft, over zijn overtuiging dat de wereld om ons heen, inclusief wijzelf, niet op toeval berust maar het product is van een Intelligent Ontwerp. Dekker heeft hierover met een aantal collega's een boek geschreven, getiteld *Schitterend ongeluk of sporen van ontwerp*. De minister zag in die opvatting mogelijkheden om verzoenende initiatieven op te starten tussen religie en wetenschap. De Kamerleden kweten zich met verve van de hun toegedachte taak. Zij rolden bijkans over elkaar heen achter de interruptiemicrofoon fulminerend dat de evolutietheorie niet uit het lesprogramma van scholen mocht verdwijnen, dat straks de religie nog de inhoud van de biologieles ging bepalen, en dat die hele ID-theorie onzin was omdat je dan in God moet geloven. Geen van deze dingen had de minister beweerd of beoogd, maar daar ging het de geachte afgevaardigden ook helemaal niet om; de goesting voor de barbecue was in het geding!

ID is intuïtief zeker een aantrekkelijk idee. In mijn vakgebied proberen we automatische spraakherkenners en spraaksynthesatoren te maken. Nou, daar komt heel wat intelligent design bij kijken! Als je dan beweert dat het ding toevallig ineens op je bureau stond, wordt je ook niet meer serieus genomen. Nee, van onze eigen spulletjes geven we hoog op als producten van intelligent ontwerp. Toch doen die spraakherkenners over het algemeen een stuk minder goed dan wij, maar desalniettemin valt het ons een stuk moeilijker om aan te nemen dat wijzelf aan een Intelligent Design(er) ontspringen.

ID is dus een interessant idee, maar vooralsnog geen wetenschap. Althans, het hangt er vanaf hoe je wetenschap definieert. Naar mijn idee gaat wetenschap over de vraag *hoe* de wereld

in elkaar zit, niet over de vraag wat de *zin*, de *reden*, of het *doel* van de wereld is. Over dat laatste gaat de religie en daar komt God in het spel. Toch is de opvatting van Dekker wel een intrigerende, omdat naar zijn mening de vraag naar het *hoe* ook niet meer zonder een Intelligent Ontwerp uitkomt. Tja, is dat zo? Geen van Dekkers medeauteurs is bioloog, op de gereformeerde hoogleraar Lever na, en juist die laatste geeft aan dat Intelligent Design voor hem geen wetenschappelijke relevantie heeft. Natuurlijk zit de natuur verbazend geraffineerd in elkaar maar het bestaan van een Ontwerper hoeft je niet te bewijzen of te veronderstellen omdat dat geen vraag naar het *hoe* is. Verder zijn alle auteurs (net als uw columnist) christen. Dus vermoedt men al snel een bias in het antwoord dat men bij de vraag wil vinden (maar ja, als het tegendeel door niet-gelovigen wordt beweerd, zegt dat net zoveel of weinig – geloof ik).

Een andere kwestie is welke onderzoeksvraag het bestaan van de Intelligent Designer dan wel zou kunnen bewijzen. Er bestaat het wat belegen idee van de wiskundige Dembski dat je een externe intelligent ontwerper moet aannemen als de complexiteit van het object zodanig is dat de kans op toevallig ontstaan op minder dan 10 tot de macht min 100 wordt geschat. Probeer daar voor het ontstaan van deze wereld maar eens een vraag en methode bij te verzinnen waar iedereen het over eens is. Het wonderbaarlijke is dat je dan het toeval over het bestaan van het toeval laat beslissen, wat mij ook wat hersenpijn bezorgt.

In dit verband wil ik u een vakantie-ervaring niet onthouden. Wij waren van de zomer in het Zwarte Woud. Mijn kinderen verdrongen zich op ons logeeraadres rondom de schommel zoals Kamerleden rondom de interruptiemicrofoon. Toen een van die schommels zich (eindelijk) in beweging zette, klonk dat karakteristieke gepiep van een slecht geolied exemplaar. Een uiterst naar geluid en ik wilde mij al afwenden, toen ik ineens heel duidelijk een thema uit de negende symfonie van Dvořák hoorde en verbijsterd bleef staan. Onmiskenbaar die twee maten, steeds weer opnieuw. De bijnaam van die symfonie luidt 'From the New World'. Onze verblijfplaats daar in het Zwarte Woud heette, jawel, 'Neuwelt'. Intelligent Design? U zegt het maar.

Wetenschap gaat over de vraag hoe de wereld in elkaar zit, niet over wat de zin is

Henk van den Heuvel is directeur van het Centre for Language and Speech Technology, CLST, Radboud Universiteit Nijmegen

Taal in Bedrijf

De succesvolle NOTaS-bijeenkomst 'Team voor Taal' op 21 januari 2004 krijgt een prominent vervolg op *dinsdag 22 november 2005* in het Evoluon te Eindhoven: **Taal in Bedrijf**. De focus van deze bijeenkomst ligt op de brede toepassingsmogelijkheden van de combinatie van taal- en spraaktechnologie in het bedrijfsleven en binnen de overheid.

PcM nabezorgingslijn



TeleCats heeft voor het Algemeen Dagblad en andere uitgaven van PcM, zoals De Volkskrant, het NRC Handelsblad en Trouw een spraakherkenningssysteem ontwikkeld, dat wordt ingezet voor de geautomatiseerde afhandeling van telefonische klachten over de bezorging. Een abonnee die belt vanwege een bezorgklacht krijgt een spraakherkenningssysteem aan de lijn. Eerst vindt de identificatie van de beller plaats aan de hand van de postcode (inclusief de letters), huisnummer en toevoeging. Vervolgens vraagt het systeem over welke datum de bezorgklacht gaat. Door een koppeling met de klantendatabase kan er worden nagegaan of dat er recent vaker bezorgklachten zijn geweest. In dat geval wordt de klant natuurlijk direct doorgeschakeld naar een medewerker. Door het inzetten van spraaktechnologie bij PcM wordt piekbelasting van de medewerkers gedurende de ochtenduren afgevlakt en wordt de wachttijd voor bellers verkort.

NOC*NSF Nederland Sportland Digitaal



Nederland Sportland Digitaal is een initiatief van Nederlands Olympisch Comité*Nederlandse Sport Federatie (NOC*NSF). Het biedt een platform waarop verschillende ICT-toepassingen worden aangeboden aan de aangesloten sportbonden van de koepelorganisatie. Het onderdeel Person Handling zorgt voor kwalitatief goede namen en adressen en tellingen die meer inzicht geven in sportend Nederland. Human Inference tekende voor dat deel van de infrastructuur, dat op maat is ontwikkeld. Person handling is een belangrijk onderdeel van het platform. De ledenadministratie blijkt voor veel verenigingen een grote klus. Maar ook voor NOC*NSF zelf is het van groot belang over actuele en correcte ledenaantallen te beschikken. HIquality Identify van Human Inference ziet bijvoorbeeld dat de gegevens van T. Putter (lid van de Nederlandse Ski Vereniging) en A.T.W. Putter (lid van het Nederlands Watersport Verbond) in zo'n mate overeenkomen dat gesteld kan worden dat het om dezelfde persoon gaat. De gegevens van deze sporter worden dan ook één keer opgeslagen in de person handling module. De bij de koepel aangesloten bonden worden afgerekend op het aantal leden; een belangrijke reden om goed te weten hoeveel leden zij hebben.

Hoewel taaltechnologie en spraaktechnologie al een aantal jaren worden gebruikt, wordt het steeds duidelijker dat beide technologieën – onder meer door verbeterde hardware en software en een toenemend inzicht in de achterliggende processen – meer en meer naar elkaar toe groeien. Ook aan de vraagzijde zien we dit: contacten tussen bedrijven en hun klanten lopen steeds vaker via meerdere media (telefoon, internet, e-mail). Het (semi-) automatisch afhandelen van deze dialogen met een combinatie van taal- en spraaktechnologie levert dan ook een significante bijdrage aan de effectievere en efficiëntere informatie-uitwisseling tussen organisaties en hun cliënten. Zo kunnen deze technologieën gebruikt worden voor segmentatie van de markt, identificatie van de klant, classificatie van het 'gespreksonderwerp', het automatisch genereren van adequate antwoorden, of het stellen van een logische vervolgvraag.

Taal in Bedrijf is het Vlaams-Nederlandse ontmoetingsplatform op het gebied van taal- en spraaktechnologie voor investeerders, managers, bestuurders, wetenschappers én vooral voor gebruikers. De organisatoren van Taal in Bedrijf streven ernaar om de mensen en organisaties die taal- en spraaktechnologie toepassen of willen toepassen bij elkaar te brengen, zodat nieuwe kansen worden gecreëerd. Tijdens het congres zullen recente succesvolle toepassingen van taal- en spraaktechnologie worden gepresenteerd, geanalyseerd en bediscussieerd.

Op het congres kan onder andere via de informatiemarkt kennis gemaakt worden met de instellingen en bedrijven die al suc-

Programma

- 08.30 - 09.45 uur Registratie en ontvangst
- 09.50 - 10.55 uur Plenaire ochtendprogramma met prominente sprekers over de toepassingen van taal- en spraaktechnologie
- 10.55 - 11.20 uur Pauze en informatiemarkt
- 11.20 - 12.40 uur Vervolg plenaire programma
- 12.40 - 14.00 uur Informatiemarkt
- Allerlei bedrijven en instellingen presenteren succesvolle voorbeelden van toepassingen van taal- en spraaktechnologie op de informatiemarkt
- 14.00 - 15.00 uur Workshops - sessie 1
- 15.00 - 16.00 uur Workshops - sessie 2
- 16.00 - 17.00 uur Workshops - sessie 3
- Tijdens de workshops worden recente succesvolle toepassingen van taal- en spraaktechnologie gepresenteerd, geanalyseerd en bediscussieerd. Meer dan twintig verschillende business cases uit verschillende sectoren in Nederland en Vlaanderen passeren de revue
- Twee voorbeelden staan beschreven
- 17.00 - 18.00 uur Borrel

Het definitieve programma met alle sprekers en workshops is op www.taalinbedrijf.org te vinden.

Locatie

Ooit werd het Evoluon in Eindhoven gebouwd om jong en oud te interesseren voor de toekomst. Nu, ruim 35 jaar later, spreekt deze vliegende schotel van beton nog steeds tot de verbeelding. De magische veelzijdigheid van het Evoluon is een ervaring die u zelf moet beleven. Een locatie die prima past bij Taal in Bedrijf!



Noord Brabantlaan 1a - 5652 LA Eindhoven - Tel.: +31(0)40 250 46 66 - www.evoluon.nl

Informatie en aanmelden

Alle informatie over Taal in Bedrijf is te vinden op www.taalinbedrijf.org. Hier kunt u zich ook aanmelden.

Organisatie

De organisatie van Taal in Bedrijf ligt in handen van een team van enthousiaste mensen bestaande uit een aantal NOTaS-leden, Nederlandse Taalunie, Programmabureau STEVIN, Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) en SenterNovem.

Medewerking wordt verleend door het Ministerie van Vlaamse Gemeenschap (AWI) en het Instituut voor de Aanmoediging van Innovatie, Wetenschap en Technologie in Vlaanderen (IWT).

Mocht u vragen hebben dan kunt u contact opnemen met Karianne Derksema (SenterNovem), telefoon: 070 373 55 58 of k.derksema@senternovem.nl.

cesvol taal- en spraaktechnologie toepassen. Tijdens de parallelsessies in de middag wordt nader ingegaan op de kansen van taal- en spraaktechnologie en op de laatste ontwikkelingen op het gebied van commerciële toepassingen. Meer dan twintig verschillende business cases uit verschillende sectoren in Nederland en Vlaanderen passeren de revue.

Taal in Bedrijf wordt *het* netwerkevenement om kennis te maken met de spelers in de markt en de wetenschap. Reserveer nu alvast 22 november in uw agenda. In oktober wordt het definitieve programma gepubliceerd en zullen persoonlijke uitnodigingen worden verzonden.

STEVIN

Taal in Bedrijf wordt georganiseerd in het kader van het Vlaams-Nederlandse STEVIN-programma. **STEVIN** (Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands) is een meerjarig onderzoeks- en stimuleringsprogramma voor Nederlandstalige taal- en spraaktechnologie dat gezamenlijk door de Vlaamse en Nederlandse overheid wordt gefinancierd. Bedoeling van het STEVIN-programma is het stimuleren van de taal- en spraaktechnologische sector in Vlaanderen en Nederland, zodat de innovatiecapaciteit van deze sector kan worden vergroot en tegelijkertijd ook de positie van het Nederlands in de moderne informatie- en communicatiewereld kan worden versterkt. Het STEVIN-programma is van start gegaan op 15 september 2004 en heeft een looptijd van vijf jaar. STEVIN wordt gecoördineerd en financieel beheerd door de Nederlandse Taalunie.

Eurospeech 2005: ubiquitous speech processing

Je hoort dikwijls dat congressen voor medici (georganiseerd door de farmaceutische industrie) eigenlijk verkapte plezierreisjes zijn. Wie denkt dat dat voor taal en spraaktechnologen ook zo is, komt echter bedrogen uit. Twee dagen praten over dialogen (SigDial), een dag 'n algemeen overzicht over de stand van zaken (Tutorials) en vier dagen van 8:30 tot 18:00 met steeds gelijktijdig vier gesproken en drie postersessies, is ronduit slopend. Dit alles gebeurde met ongeveer 1288 deelnemers in de eerste week van september aan de oevers van de Taag in het culturele centrum van Belém (Lissabon).

| ARJAN VAN HESSEN |

Uiteraard is het bij zo'n aanbod volstrekt onmogelijk om alles te zien en horen en de meeste deelnemers besluiten dan om een geselecteerd aantal presentaties te bekijken/beluisteren. Dit jaar was er voor het eerst de mogelijkheid om dat vooraf via Internet te doen. Je gaf de voor jou interessante praatjes op en kreeg een geprinte route beschrijving (wat, waar, hoe laat). Erg handig, maar bijna onmogelijk om vier dagen vol te houden.

Deze keer was er behalve een ruime aandacht aan de traditionele onderwerpen (spraakherkenning en synthese, ruisonderdrukking enzovoort) ook aandacht voor de menselijke kanten van spraakperceptie: wat horen wij eigenlijk, hoe gaan wij mensen met onvolledige spraak om, waarvoor gebruiken we het ritme in de taal enzovoort). Het is de meeste deelnemers wel duidelijk dat, wil taal en spraaktechnologie echt succesvol kunnen worden gebruikt, de menselijke en de gecomputeriseerde spraak en taaltechnologie goed naar elkaar moeten luisteren hoewel dit naar elkaar niet altijd even duidelijk is gegeven de zeer felle reacties in de sessie 'bridging the gap between ASR and HSR' (automatische versus menselijke spraakherkenning).

TENDENSEN

Het 'Leitmotiv' van de conferentie was 'Ubiquitous Speech Processing'. Dit betekent zoiets als 'overall spraakprocessing' en houdt in dat men zich afvraagt of over niet al te lange tijd spraakprocessing (zowel synthese, herkenning als spreker identificatie/verificatie) zo goed zal werken en zo algemeen geaccepteerd zal zijn dat het overall zal worden toegepast. We praten dan tegen het broodrooster, computers bellen ons op en vragen naar onze gezondheid en alle afspraken en zoekopdrachten verlopen via een gesproken interface (zie hiervoor de erg leuke clip van Apple: the Knowledge Navigator op www.billzarchy.com/clips/clips_apple_nav_navigator.htm). Of dit zal

gebeuren en of het überhaupt wenselijk is weet niemand, maar zo'n conferentie is een uitstekende gelegenheid om er uitgebreid over te filosoferen. Gaan we, ook als het zeer goed kan, echt allemaal met onze apparaten praten of prefereren we toetsen en stylus. Wat moet er nog gedaan worden om zulke applicaties mogelijk te maken waarin we op een bijna menselijke manier met apparaten omgaan? Wel: erg veel.

Hieronder een aantal 'hot topics' van de afgelopen conferentie. Erg veel interessante zaken zullen wegens ruimte gebrek ongenoemd moeten blijven maar we zullen proberen de belangrijkste tendensen (onderzoek dat door veel mensen op veel verschillende plekken op de wereld gedaan wordt) te noemen.

SPREKERHERKENNING

Of het nu door de angst voor terrorisme komt of niet, duidelijk is dat door de enorme inspanningen op dit gebied er steeds meer succes gehaald wordt op het gebied van sprekerherkenning (wie spreekt er) en sprekerverificatie (is dat echt de spreker die hij/zij zegt te zijn). Het is niet waarschijnlijk dat deze technologie ooit dezelfde resultaten (>>99% betrouwbaarheid) zal halen als bijvoorbeeld vingerafdrukken of irisscans. Het feit dat identificatie/verificatie echter zonder extra apparatuur en zonder dat je je er van bewust bent, over de telefoon gedaan kan worden, is er de oorzaak van dat er toch grote belangstelling voor is (was dat bandje echt door Bin Laden ingesproken of niet?).

NIUWSUITZENDINGHERKENNING

Een succesvolle ontwikkeling is die van het herkennen en ontsluiten van nieuwsuitzendingen. De spraak in nieuwsuitzendingen bestaan in de regel, doordat de sprekers meestal professionals zijn, uit duidelijk en meestal zonder

aarzelingen gesproken spraak. Afhankelijk van de taal en de universiteit/instelling die het uitvoert, worden hier verbluffende herkenningresultaten bereikt. Zo wordt in het herkennen van het 8-uur journaal door de Italianen (IRST) en Fransen (LIMSI) meer dan 90% van de tekst goed herkend (in Nederland zitten we pas op 72%). Redelijke herkenning (>66%) van dit soort materiaal maakt het echter al mogelijk om te 'audio googlen': je typt een aantal woorden in en krijgt een link naar de uitzendingen terug waarin deze woorden ook gebruikt worden. Op die manier kan er op eenvoudige en goedkope wijze gezocht worden in gesproken nieuwsuitzendingen: iets dat door veel programmamakers zeer gewaardeerd wordt omdat ze, ondanks de fouten die er gemaakt worden, de dag erna al in de uitzending kunnen zoeken.

ORAL HISTORY

Omdat er op het gebied van nieuwsuitzendingen niet zo heel veel winst meer valt te halen (het is duidelijk waarom we de laatste procenten nooit zullen halen), zie je dat de aandacht verschuift naar de spraak van speciale groepen zoals ouderen, kinderen en non-natives en naar andere domeinen. Met name de toepassingen op het gebied van 'oral history' komen erg op. Het gaat hier om opnamen in historische archieven waarvan slechts vaag bekend is waar het over gaat. Voor onderzoekers zou het geweldig zijn wanneer het mogelijk zou worden naar bepaalde woorden of fragmenten te kunnen zoeken. Een aansprekend voorbeeld dat op Eurospeech gedemonstreerd werd, is het Malach-project (www.clsp.jhu.edu/research/malach) van Steven Spielberg: ±116000 uur opname (ca. 13,2 jaar) van spraak in 32 talen van 52000 overlevenden de Holocaust. Dit is zoveel, dat zonder automatische spraakherkenning, het niet binnen afzienbare tijd ontsloten kan worden!

AARZELINGEN

Een meer fundamenteel onderzoeksthema is het herkennen van niet-vloeiende spraak zoals aarzelingen, het halverwege de zin veranderen van onderwerp, het herhalen van woorden of fragmenten ('ik wil.. eh, ik wil.. eh.. doe die blauwe koffer maar')? Mensen hebben met het herkennen van dit soort uitingen volstrekt geen moeite, maar spraakherkenners kunnen er (nog) bijna niets mee. Dit komt waarschijnlijk omdat wij de herkende spraak direct interpreteren en niet alleen maar naar de eigenlijke woorden luisteren. In het voorbeeld hierboven is het duidelijk dat de spreker iets wil want dat wordt zelfs 3x gezegd. Uiteindelijk wordt aangegeven wat hij wil en dus is het simpel: de spreker wil een blauwe koffer. Hoe dit met de computer na te doen?

EMOTIE

Hierbij aansluitend is het onderzoek naar het gebruik van emotie in spraak (zowel door machines als mensen). Emotie speelt een belangrijke rol in de interpretatie van spraak en moet dus worden meegenomen in de interpre-

tatie van hetgeen er gezegd is. Maar ook zonder dat we gaan interpreteren kan emotie-detectie al zeer nuttig zijn. Zo zou het bijvoorbeeld toegepast kunnen worden in een call center. Boze klanten kunnen dan anders te woord worden gestaan of direct worden doorgeschakeld naar een begripvolle telefoniste. Ook zou de manier waarop computers antwoorden (zowel woordgebruik als stem) aan de gedetecteerde emotionele staat van de beller kunnen worden aangepast. Het is een lastig onderwerp, zeker daar de betekenis van bijvoorbeeld volume stijging of toonhoogte verandering redelijk cultuurgebonden zijn. Toch worden ook op dit vlak vorderingen gemaakt en zijn er zelfs computers die een grap weten te waarderen: ze kunnen al met een redelijke betrouwbaarheid een lach detecteren.

VAN HERKENNING NAAR BETEKENIS

Het achterhalen van de betekenis van zowel geschreven als gesproken taal is waarschijnlijk de grootste uitdaging voor de komende jaren. Wat wordt er bedoeld met 'doe die maar', of 'print dat eerste document van Klaas'. Voor echt semi-menselijke mens-machine conversatie is dit soort interpretatie noodzakelijk. Wanneer we goed kijken naar de bovengenoemde Knowledge Navigator waarin de spreker iets zegt in de trant van 'iets van Dr Flemson of zoiets' en de computer terugkomt met 'u bedoelt Dr. Flemming': wel dat is echt interpretatie! Wanneer de computer in staat is om betekenis uit eventueel foutief herkende spraak te halen, kunnen we spreken van kunstmatige intelligentie. Zo makkelijk is dit echter niet en als het lukt dan zal het langzaam gaan en met behulp van grote, kostbare databases waarin de kennis over onderwerpen ligt opgeslagen.

VERGADERINGEN

Een toepassing waarbij de vorige twee onderwerpen gebruikt zullen worden is automatische verslaglegging van vergaderingen. Accurate hiervan is lastig, tijdrovend en daarom bij bijna niemand populair. In de meeste gevallen beperkt men zich daarom tot het produceren van korte aantekeningen en een besluitenlijstje. Toch zou het mooi zijn wanneer we wel een veel uitgebreider verslag zouden hebben waarin we bovendien naar uitspraken van bepaalde sprekers zouden kunnen zoeken: wanneer zei de burgemeesters iets over de nieuwe autoweg? In tegenstelling tot de bovengenoemde voorbeelden, wordt er bij vergaderingen door meerdere sprekers al dan niet door elkaar heen gesproken. Voor een succesvolle toepassing moet dus de spreker geïdentificeerd kunnen worden en moeten de door elkaar sprekenden uit elkaar gehaald kunnen worden. Ook hier is interpretatie cruciaal: hoe wordt iets gezegd (tempo, toonhoogte, volume), wat bedoelt de spreker, enzovoort.

De volgende Eurospeech-conferentie - september 2007 - wordt georganiseerd door taal en spraakonderzoekers uit Nederland en België en zal plaatsvinden in Antwerpen.

TST-centrale actief in Antwerpen

Op 4 juli 2005 organiseerde de TST-centrale voor de derde keer een taal- en spraaktechnologiedag. Locatie was kasteel Middelheim in Antwerpen. Een bijzondere editie omdat minister van OCW Maria van der Hoeven bij deze gelegenheid de Antwerpse vestiging van het INL officieel opende.

De TST-dagen zijn bedoeld om bedrijven en kennisinstellingen op de hoogte te brengen van de activiteiten van de TST-centrale, de behoeften van bedrijven en kennisinstellingen ten opzichte van de TST-centrale te monitoren en waar mogelijk kennis over te dragen. Daarnaast wil de TST-centrale bruggen slaan tussen bedrijfsleven, kennisinstellingen en vertegenwoordigers van de overheid.

| SIMO GODIJN |

TST-PRODUCTEN

Een van de kerntaken van de TST-centrale is het onderhoud en beheer van digitale taalmaterialen. De belangrijkste vorderingen op dat gebied kwamen in de eerste twee presentaties aan bod: het Referentiebestand Nederlands (RBN) is sinds kort ook online toegankelijk via de website van de TST-centrale en Nederlands grootste spraakdatabank, het Corpus gesproken Nederlands, wordt momenteel grondig gebugd, zowel qua data als qua exploitatie-software. In het najaar verschijnt versie 2.0.

INTELLECTUELE EIGENDOMSRECHTEN

Vanwege de vele vragen die de TST-centrale de afgelopen maanden heeft gekre-

gen over intellectuele eigendomsrechten hadden we 'onze' jurist Adonna Alkema van advocatenkantoor Klos, Morel, Vos & Schaap uitgenodigd dit onderwerp te komen toelichten. Mevrouw Alkema plaatste begrippen als 'auteursrecht' en 'databankrecht' in TST- (en STEVIN-) context, hetgeen duidelijk in een behoefte van het publiek voorzag.

FORUMDISCUSSIE

De TST-medewerkers willen graag op de hoogte zijn van wat er speelt in het veld. Daarom was er veel tijd ingeruimd voor een forumdiscussie, die de gelegenheid bood tot het uitwisselen van meningen tussen TST-centrale, Nederlandse Taalunie, bedrijfsleven en kennisinstellingen.

Discussieleider Steven Krauwer gaf het publiek de gelegenheid door middel van stemming te reageren op een aantal stellingen. Forumleden uit bovenstaande gelederen mochten desbetreffende stelling verdedigen of afbreken, waarna het publiek nog eens stemde. Drie thema's kwamen aan de orde: informatievoorziening, prijzen van TST-materialen en industriële samenwerking met de TST-centrale.

Opvallend was dat industriële spelers vaak niet op de hoogte zijn van de takenpakketten van NWO, de Nederlandse Taalunie en de TST-centrale. Daardoor missen ze inzicht in de mogelijkheden tot samenwerking en financiering. Omgekeerd is het voor de TST-centrale lastig industriële geïnteresseerden te bereiken en van informatie te voorzien: kennelijk zijn daar andere kanalen voor nodig dan die we tot nu toe gebruiken. Daarnaast blijkt de ontwikkeling van een goede en transparante prijspolitiek met betrekking tot de producten van de TST-centrale noodzakelijk.

Zowel kennisinstellingen als bedrijven willen duidelijkheid over de prijzen van de producten en over de totstandkoming van die prijzen.

Tot slot bleek dat de industrie wel materialen wil onderbrengen bij de TST-centrale, mits ze nauw wordt betrokken bij het vaststellen van de voorwaarden waaronder door haar geleverd materiaal door anderen gebruikt mag worden.

Simo Godijn is medewerkster aan de TST-centrale

Opening Antwerpse dependance

Vanaf oktober 2004 zetelt de TST-centrale deels in de Universiteit van Antwerpen. Hoog tijd om deze INL-dependance feestelijk te openen: niemand minder dan minister Van der Hoeven van OCW werd bereid gevonden de plechtigheden te verrichten. Onder toezicht oog van de minister werd de overeenkomst tussen de Universiteit Antwerpen en het INL getekend. Uit handen van ALVV-bestuursvoorzitter Willy Martin ontving de minister het Rapport woordenboekenbeleid in Nederland en Vlaanderen en van CGN-bestuursvoorzitter Sieb Nooteboom kreeg ze het 33 DVD's tellende CGN cadeau. Met haar openingstoespraak en het starten van een filmpje bekrachtigde Van der Hoeven de opening. Vanaf nu staat de Antwerpse vestiging op de kaart. Meer informatie: <http://www.tst.inl.nl/tstdago40705.html>.

Automatische indexering van audiovisuele archieven

De stem van Willem Frederik Hermans ontrafeld

‘Willem Frederik Hermans (1921-1995) wordt algemeen beschouwd als de belangrijkste Nederlandstalige schrijver van de twintigste eeuw. Zijn oeuvre is van een onovertroffen veelzijdigheid, geeft tot op de dag van vandaag aanleiding tot discussie en dwingt de nieuwe generaties lezers en auteurs steeds weer tot standpuntbepaling.’ Zo staat te lezen op de Willem Frederik Hermans web-portal van het gelijknamige instituut (WFHi) dat dit jaar, in het tiende sterfjaar van de schrijver, werd gelanceerd. De portal biedt een grote hoeveelheid informatie over Hermans en zijn schrijven en, uniek voor Nederland, ook de mogelijkheid om gedetailleerd te zoeken in Herman’s gesproken woord: in interviews, lezingen en voorgelezen verhalen.

| ROELAND ORDELMAN |

Audiovisuele opnameapparatuur en opslagruimte wordt steeds goedkoper en zowel het aantal als de omvang van audiovisuele archieven in Nederland neemt daardoor snel toe. Dankzij digitaliseringsinitiatieven bij instellingen en bedrijven die van oudsher beschikken over audiovisuele collecties (retrospectieve digitalisering) wordt het aanbod nog eens drastisch verbreed. Een aantal voorbeelden. Een aantal gemeenten in Nederland zijn gestart met het online brengen (‘webcasten’) van raadsvergaderingen (zie bijvoorbeeld <http://www.bestuuronline.nl>) in het kader van de ‘openbaarheid van bestuur’. Onderwijsinstellingen zijn aan het experi-

menteren met het audiovisueel archiveren van lessen, colleges, lezingen en conferentiepresentaties. Ook bedrijven nemen in toenemende mate vergaderingen en presentaties op ten behoeve van de interne informatievoorziening en natuurlijk kent iedereen die mensen die overal een videocamera naar toe zeulen om alles op te nemen wat ze niet willen vergeten of graag willen delen met anderen via weblogs.

STIEFKINDEREN VAN HET ARCHIEF

Een speciale categorie vormen instellingen zoals Beeld&Geluid (<http://www.beelden-geluid.nl>) die een van de grootste audiovisuele archieven in Europa aan het digitali-



seren is, het gemeentearchief van Rotterdam waar binnenkort een project start om haar gesproken-woord-archieven te ontsluiten, en instellingen zoals het eerdergenoemde WFHi. Deze categorie heeft in toenemende mate de belangstelling van historici omdat de betreffende archieven een belangrijk onderdeel vormen van het Nederlands cultureel erfgoed.

Het zijn allemaal prachtige en interessante collecties natuurlijk, maar zonder mogelijkheden om erin te zoeken verliezen ze veel van hun waarde, vooral wanneer de collecties omvangrijk zijn. Het is soms verbazingwekkend dat soms zelfs een minimale beschrijving van het materiaal ontbreekt. Archieven verworden daardoor tot een 'stiefkind van een archief', zoals Franciska de Jong, professor multimedia

weten te komen of een bestand misschien iets bevat waarnaar je op zoek bent, moet alsnog het hele bestand afgeluisterd of bekeken worden. Vooral als het bestand groot is, is dat vervelend. Bij voorbaat is de kans om op een potentieel interessant document uit te komen al niet zo hoog omdat de (specifieke) zoekvraag moet passen bij de in algemenere termen gestelde beschrijvingen.

AUTOMATISCHE INDEXERING

Om de geschetste problemen het hoofd te bieden, wordt er al een aantal jaren gekeken naar mogelijkheden om automatisch beschrijvingen toe te kennen aan collectiebestanden door gebruik te maken van audio- en videoanalysesoftware.

Het zoeken binnen een audiovisueel bestand zou al een stuk makkelijker gemaakt kunnen worden door het aanbrengen en zichtbaar maken van wat meer structuur. Het is goed mogelijk om automatisch te detecteren waar er bijvoorbeeld spraak zit in een bestand en waar muziek. Of waar een scène-overgang zit. De visualisatie hiervan in een zoekomgeving zou het zoeken al vereenvoudigen.

Een stap verder is dat individuele sprekers of zelfs het gesproken woord automatisch worden herkend en voorzien van een tijdelabel, zodat naar het voorkomen van sprekers of woorden binnen een bestand gezocht kan worden. Nog een stap verder is dat op basis van het herkende gesproken woord onderwerpen kunnen worden gedetecteerd. In het ideale geval ontstaat een goed gestructureerd audiovisueel document dat vanuit verschillende invalshoeken bevraagd kan worden.

Het gebruik van spraakherkenning voor het automatisch generen van meta-data, ofwel automatische indexering, ligt voor de hand. Geleid door professor Franciska de Jong doet de Human Media Interaction (HMI) groep van de Universiteit Twente al sinds de jaren '90 onderzoek naar het gebruik van spraakherkenning voor multimedia-ontsluiting. Een belangrijke stap was het ontwikkelen van een Nederlands spraakherkenningsstelsel dat spraak één op één kan omzetten naar tekst. Een bestaand commercieel spraakherkenningsstelsel van de plank halen was geen optie omdat het toepassingsgebied een flexibi-

Het is goed mogelijk om automatisch te detecteren waar er bijvoorbeeld spraak zit in een bestand en waar muziek



technologie aan de Universiteit Twente het onlangs uitdrukte, waar je weinig mee kunt.

Op zichzelf is het niet zo vreemd dat collectiebeschrijvingen minimaal zijn, dan wel ontbreken. Het moet allemaal handmatig gebeuren en een hogere mate van detail vereist dat de betreffende audiovisuele bron van voor naar achter moet worden beluisterd of bekeken. Doorgaans zijn alleen instellingen met een professionele archivariissen hiertoe in staat. Maar dan nog, de zogenaamde 'meta-data' die voor met zorg aangelegde collecties beschikbaar is, beperkt zich meestal tot een titel, een datum en een korte beschrijving van de inhoud. Dit soort beschrijvingen kan gebruikt worden om *binnen een collectie* naar een specifiek bestand te zoeken, maar niet om binnen een ongestructureerd audiovisueel bestand zelf te zoeken. Om te

liteit van de software vereist (aangepassen van akoestische modellen, vocabulaire, taalmodellen) die in commerciële systemen doorgaans niet of nauwelijks kunnen geven. Aanvankelijk richtte het onderzoek zich op het ontsluiten van Nederlandse nieuwsuitzendingen. De laatste jaren wordt steeds meer naar andere domeinen gekeken, zoals vergaderingen, lezingen, raadsvergaderingen en historische archieven.

Wanneer spraakherkenning wordt ingezet voor ontsluitingsdoeleinden is het niet zozeer van belang dat alle woorden in de spraak correct herkend worden, in tegenstelling tot wat bij dicteertaken het geval is. Internationaal onderzoek suggereert als minimale vereiste voor bruikbare herkenning voor ontsluitingsdoeleinden dat er zo'n 50% van de woorden correct moet zijn. Belangrijk is dat vooral inhoudswoorden en namen goed herkend worden. Dat zijn immers de woorden waarnaar gezocht gaat worden. De algemeen gangbare maat om de kwaliteit van een spraakherkenningssystemen mee uit te drukken, de 'word error rate' (WER), is dan ook niet erg veelzeggend als het er om gaat hoe bruikbaar een systeem is voor automatische indexering. Deze maat telt ook fouterkende functiewoorden of bijna-goedherkende woorden (enkelvoud in plaats van meervoud, fout in vervoeging, of een deel van een samenstelling fout) mee die voor het zoeken toch goed bruikbaar kunnen zijn.

HISTORISCH MATERIAAL

Met name bij historisch materiaal kan de 'word error rate' hoog zijn. De kwaliteit van de audio zelf kan daar debet aan zijn maar ook de manier van spreken ('gezwollen' taal) en het taalgebruik (ouderwetse woorden) dragen ertoe bij dat de spraakherkenning zich nogal eens verslijkt. In spraakherkenningstermen is 'adaptatie' het sleutelwoord. De herkenner moet worden aangepast aan de akoestische karakteristieken en taal van het taakdomein met behulp van voorbeelddata: spraakdata waarbij handmatig is aangegeven welke woorden werden gezegd, en tekstdata die zoveel mogelijk lijken qua woord- en taalgebruik op de spraak in het taakdomein.

Voor het genereren van spraaktranscrip-

ties voor ontsluiten van de interviews en lezingen van Willem Frederik Hermans werd in eerste instantie een standaard spraakherkenningsconfiguratie gebruikt die normaal gesproken wordt toegepast voor het herkennen van nieuwsdata. Het was dan ook niet verwonderlijk dat zo'n 80% van de woorden verkeerd werd herkend. Dat is een resultaat dat zelfs voor zoekdoeleinden ongeschikt wordt geacht. Door de zeer kleine hoeveelheid voorbeelddata die voor dit domein beschikbaar was te gebruiken voor de adaptatie van de herkenner, werd uiteindelijk een 'word error rate' van rond de 65% gehaald. Hoewel dit nog ruim boven de eerdergenoemde kwaliteitseis voor zoekdoeleinden van 50% ligt, is het resultaat al best bevredigend. Wanneer je 'tranen der acacia's' intypt als zoekvraag, levert het zoekstelsel ook daadwerkelijk audiofragmenten op waarin de term voorkomt. Het is ook goed om te bedenken dat zonder deze vorm van automatische generatie van meta-data, zoeken in de bestanden al helemaal niet mogelijk was.

AUDIOVISUEEL MATERIAAL GOOGELEN

De Willem Frederik Hermans audio-zoekdemo laat zien dat door gebruik te maken van beschikbare Nederlandse spraaktechnologie het mogelijk is om in audiobestanden te zoeken. Dat er nog een hoop te verbeteren valt is duidelijk. Het verbeteren van de mogelijkheden om snel en doeltreffend een spraakherkenningssysteem aan te passen aan de eisen van een willekeurig taakdomein behoort hier toe. Ook wat betreft zoekopties en de presentatie van de zoekresultaten kan nog het nodige werk worden verzet, bijvoorbeeld door automatisch gegenereerde segmentgrenzen (sprekerwisseling, pauzes, eventueel onderwerpen) aan te bieden. Maar in ieder geval laat het zien dat de mogelijkheid om audiovisueel materiaal ook eenvoudigweg te googelen, langzaam dichterbij komt.

De Willem Frederik Hermans demo is te zien via <http://www.willemfrederikhermans.nl> (onder 'stem/beeld') of direct via <http://www.home.cs.utwente.nl/~huijbreg/demopages/hermans/hermans.php>. De demo is gemaakt door Marijn Huijbregts en Roeland Ordelman van de Human Media Interaction groep (<http://hmi.ewi.utwente.nl>) van de Universiteit Twente, in het kader van het Multimediana project (<http://www.multimediana.nl>). door Contact: ordelman@ewi.utwente.nl of fdejong@ewi.utwente.nl



De zomermaanden vormen ook voor NOTaS een wat rustigere periode. De komkommertijd is echter nuttig besteed voor het in gang zetten van nieuwe projecten en het doorwerken aan lopende projecten. De website van 'Taal voor Afrika' is in de lucht, er zijn Casimir-voorstellen ingediend en een nieuwe Masteropleiding op het gebied van taal en spraak is van start gegaan. Hieronder volgt een update van de zomer.

Taal voor Afrika

In de vorige DIXIT heeft u kunnen lezen over 'Taal voor Afrika'. Dit door NOTaS-leden geïnitieerde project is opgestart om de voorzieningen te verbeteren voor Ghanese docenten in opleiding die zich richten op onderwijs voor doven en slechthorenden. Het bestuur van Stichting Taal voor Afrika is druk met het werven van fondsen en het ontwikkelen en beschikbaar maken van informatie en lesmateriaal.

Het project Wilde Ganzen kan een belangrijke financiële steun leveren aan Taal voor Afrika. Wilde Ganzen steunt kleinschalige, concrete projecten financieel door ingezamelde bedragen bij wervingsacties te verdubbelen of verdrievoudigen. Stichting Taal voor Afrika heeft een aanvraag ingediend.

Met de University of Education in Winneba (Ghana), waar het project in eerste instantie wordt opgestart, worden de mogelijkheden en behoeften met betrekking tot de invulling van het lesmateriaal doorgenomen. De website waarop de studenten in de toekomst informatie en lesmateriaal kunnen vinden voor doven en slechthorenden is inmiddels in de lucht. De komende tijd zal deze voorziening verder worden uitgebreid. De site is te vinden via www.e-viataal.org. Op de website vindt u tevens meer informatie over het project Taal voor Afrika.

Casimir

Een van de doelstellingen van NOTaS is het stimuleren van kennisuitwisseling en samenwerking tussen kennisinstellingen en bedrijven. Het Casimir-programma van NWO (Nederlandse Organisatie voor Wetenschappelijk Onderzoek) sluit daar volledig bij aan door het stimuleren van mobiliteit tussen onderzoekers van kennisinstellingen en R&D-ers van bedrijven. Doel is om onderzoekers voor een langere tijd bij een bedrijf te laten werken en R&D-ers bij kennisinstellingen. Er is dit jaar drie miljoen euro beschikbaar voor nieuwe projecten.

Verschillende NOTaS-leden hebben een Casimir-voorstel ingediend. NOTaS wenst hen veel succes!

Nieuwe input voor taal en spraak - Research Master

De universiteiten van Nijmegen en Tilburg hebben hun krachten gebundeld voor een nieuwe Master-opleiding op het gebied van taal en spraak. Het betreft een tweejarige

onderzoeksmaster genaamd 'Language & Communication'. De Master is opgebouwd uit vier onderdelen, te weten Communicatiewetenschap, Taalgebruik en Taalverwerving, Taalanalyse en Taal- en spraaktechnologie.

Er is een toelatingscommissie die aan de hand van strenge criteria beslist of studenten aan de eisen voldoen om de Master te starten. Inmiddels zijn 8 studenten gestart met een algemene inleiding op de vier vakgebieden, waarna ze een hoofd- en een bijrichting kiezen. Onderdeel van de Master is een stage die zowel op de universiteiten als in het bedrijfsleven kan worden gedaan. We hopen dat NOTaS-leden een aantal stages kunnen verzorgen. De colleges vinden plaats op beide universiteiten en worden gegeven door docenten van zowel de Radboud Universiteit als de Universiteit van Tilburg. Voor meer informatie over de researchmaster kunt u kijken op: http://www.ru.nl/letteren/onderwijs/overzicht/research_master_/language_and

Demonstratieprojecten

In het vorige NOTaSnieuws heeft u kunnen lezen over de projectaanvragen voor Stevin. 1 September is de deadline voor de uitgewerkte voorstellen.

Bij de flankerende activiteiten van het STEVIN-programma horen de oproepen voor demonstratieprojecten. Doel daarvan is het stimuleren van de vraag naar Nederlandstalige taal- en spraaktechnologie door middel van kortlopende voorbeeld- of demonstratieprojecten (maximaal 15 maanden), waarbij gebruik wordt gemaakt van bewezen technologie. De nadruk ligt op het ontsluiten van nieuwe markten en het toepassen van bestaande taal- en spraaktechnologische applicaties in andere domeinen. In het STEVIN-programma wordt een totaalbudget van maximaal één miljoen euro beschikbaar gesteld voor de uitvoering van demonstratieprojecten. De sluitingsdatum voor het indienen van demonstratieprojecten is 15 oktober 2005. Voor meer informatie kunt u terecht op de website: http://taaluniversum.org/taal/technologie/stevin/projectoproepen/#oproepen_demo

NOTaS

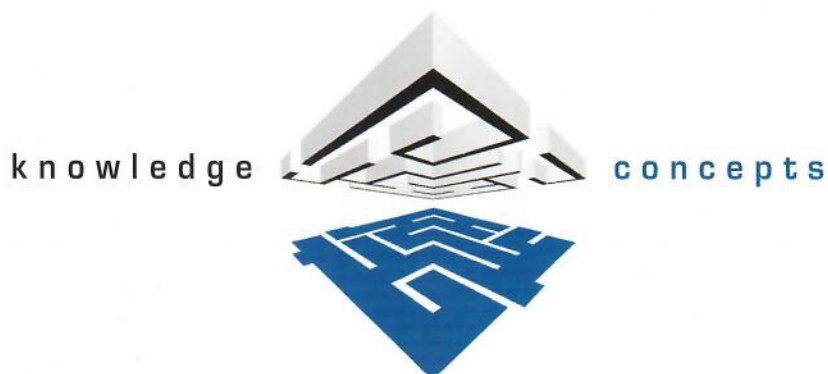
Deelnemers van de Stichting NOTaS zijn:

Human Inference Enterprise B.V. • Universiteit van Tilburg
Faculteit Letteren • TNO Management Consultants
• Knowledge Concepts • Polderland Language & Speech
Technology B.V. • TeleCats • TNO-TPD • Universiteit Twente,
HMI • Van Dale Lexicografie • Dutcheer • Q-go B.V.
• Asknow Solutions B.V. • Comsys B.V. • Radboud Universiteit,
CLST • FNB • Stichting Instituut voor Nederlandse Lexicologie,
TST-Centrale • Em@ilco

Sponsor: Nederlandse Taalunie

(Meer informatie: www.stichtingnotas.nl)

Wij werken aan de derde generatie Intranet

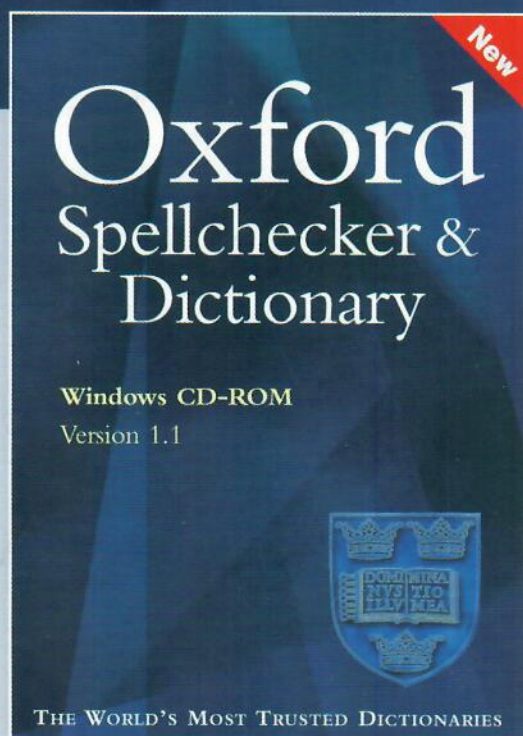


De Handboog 9, 5283 WR Boxtel, Nederland
Tel: +31(0)411 610802 | Fax: +31(0)411 611027
E-mail: info@knowledge-concepts.com | Web: www.knowledge-concepts.com

The first ever British English electronic
spellchecker with the unrivalled
authority of Oxford fully integrated in
Microsoft® Word

Oxford Spellchecker & Dictionary

Now also with American spelling



New in version 1.1:

- UK and US English
- -ise versus -ize spelling
- Includes complete
Oxford Dictionary of English

OXFORD
UNIVERSITY PRESS

 **Polder-
land** LANGUAGE AND SPEECH
TECHNOLOGY

Kerkenbos 11-03a
6546 BC Nijmegen
www.polderland.nl