

DIXIT

TIJDSCHRIFT OVER TOEGEPASTE TAAL- EN SPRAAKTECHNOLOGIE

Spraaktechnologie de puberale jaren voorbij:
van herkenning naar interpretatie

Dictionary: zie Spraaksynthese; zie Spraakherkenning; zie ook Fonetische Transcriptie



Schrijfhulp op
maat, ofwel:
*beste software is
onzichtbaar*

INHOUD

Schrijfhulp op maat, ofwel: beste software is onzichtbaar

Bij veel van wat we doen, laten we ons helpen door computers en software. Maar daarbij zijn we wel behoorlijk kieskeurig. Software mag natuurlijk intelligent zijn, zolang die software het met ons eens is. Want anders wordt die intelligentie als hinderlijk ervaren. Goede hulp valt eigenlijk niet op. Vandaar dat de beste software 'onzichtbaar' is.

4

Eurospeech 2003: overdadige dis aan praktische informatie

Begin september werd in het international conference center in Geneve het 8ste Eurospeech congres gehouden. Ongeveer duizend bezoekers hielden zich vier dagen bezig met het luisteren naar en lezen van elkaars werk. Een van de dingen die opvallend duidelijk werd gedurende dit congres, is dat spraaktechnologie de puberale jaren definitief voorbij is. Er valt duidelijk een wenkend perspectief te signaleren, namelijk van herkenning naar interpretatie.

8

Lexicografische producten op maat gesneden

Wat zou Johan Hendrik van Dale schrikken als hij vandaag het bedrijf kon bezoeken dat zijn naam draagt! In plaats van een arbeidsverslaafde onderwijzer die woordverklaringen noteert op kaartjes, zou hij een team van meer dan vijftig professionals ontmoeten die achter computers zitten en bepaald niet allemaal bezig zijn met woordjes verzamelen. Lexicografie in de eenentwintigste eeuw is meer dan woordenboeken schrijven. Het is taaltechnologie!

10

Rubrieken

Column	7
Uit de boeken	13
TST Dictionary	15
Uitgelicht!	16
NOTaS-nieuws	17

DIXIT: omdat het moet

Voor u ligt DIXIT. Een nieuw vakblad dat volledig is gewijd aan taal- en spraaktechnologie (TST). In dit eerste nummer maak ik graag gebruik van de 'hoofdreductie ruimte' om DIXIT kort te introduceren en uit te leggen waarom we hebben gemeend dit magazine het licht te doen laten zien.

DIXIT is geïnitieerd door Stichting NOTaS (Nederlandse Organisatie voor Taal- en Spraaktechnologie) en behandelt vooral *toegepaste* taal- en spraaktechnologie, dat wil zeggen: zoals TST zou kunnen worden toegepast om uw bedrijfsprocessen sneller, efficiënter en bovenal klantvriendelijker te maken. In klare taal en dus wars van 'laboratoriumpraat'.

Taal- en spraaktechnologie staat aan de vooravond van een enorme vlucht voorwaarts. Waarbij taaltechnologie vooralsnog een kleine voorsprong heeft, omdat het de laatste tien jaar zijn weg definitief heeft gevonden. Dagelijks worden we geconfronteerd met toepassingsgerichte oplossingen: de spellingchecker in MS Word, het doorzoeken van omvangrijke databases met ongestructureerde data, het fouttolerant vergelijken en corrigeren van namen en adressen, noem maar op.

Spraaktechnologie daarentegen maakt een wat langzamere, maar niettemin zeer gestage en welhaast onstuitbare opmars. Met name spraakherkenning binnen een beperkt (begrippen)domein wordt meer en meer toegepast. En vandaag de dag wordt hard gewerkt aan de integratie van taal- en spraaktechnologie.

Ongekende mogelijkheden dienen zich aan. En dan is er nog Europa. De Europese integratie voegt een extra

boeiend element toe aan TST: de taligheid over de grenzen. Bovendien wordt in diverse landen onderzoek en ontwikkeling gedaan naar het semantisch web: gebruik maken van internet op basis van de betekenis van de gegevens in plaats van het één-op-één vergelijk van een reeks letters.

De 'affaire Lernout & Hauspie' heeft de sector taal- en spraaktechnologie in De Lage Landen geen goed gedaan. De negatieve berichtgeving gaf voeding aan het idee dat TST ofwel te vaak verpakte lucht is ofwel nog slechts in een embryonaal stadium verkeert. Maar niets is minder waar! In NOTaS hebben de meeste Nederlandse bedrijven en onderzoeksinstituten (universiteiten) zich verenigd om de specifieke belangen van deze sector te behartigen. Waarom? Het Nederlands vormt een relatief klein taalgebied en investeringen in TST zijn fors. Samenwerking en synergie liggen daarom voor de hand. Maar Nederland is ook weer veel meer internationaal georiënteerd dan andere EU-landen. En dat biedt ongekende mogelijkheden.

Met DIXIT willen wij TST dichterbij brengen. En wel op zo'n manier dat u ziet wat het voor uw organisatie kan betekenen. Dat doen we vooral door praktijkverhalen, cases, relevante signaleringen van baanbrekend onderzoek, en natuurlijk achtergrondartikelen waar de stem van de klant de hoogste toon voert.

Enfin, ik heb genoeg gezegd. DIXIT zagezegd.

Geert Kobus
Voorzitter NOTaS

Hulp op maat bij het schrijven

Gemak dient de mens

Bij veel van wat we doen, laten we ons helpen door computers en software. Maar daarbij zijn we wel behoorlijk kieskeurig. Software mag natuurlijk intelligent zijn, maar moet het (desondanks?) wel met ons eens zijn. Want anders wordt de intelligentie als hinderlijk ervaren. Hulp die heel goed is, valt ons minder op. De beste software is eigenlijk onzichtbaar.

| THEO VAN DEN HEUVEL |

Er zijn dus vele en verschillende situaties denkbaar waarbij we de hulp van programmatuur goed kunnen gebruiken. Bijvoorbeeld bij het produceren van een tekst. Voor veel talen zijn de regels zo complex dat schrijvers onzeker zijn over hun beheersing van spelling en grammatica. Dat geldt al voor de moedertaal van de schrijver, laat staan voor vreemde talen. Onzekerheid over spelling speelt ook op andere terreinen een rol, bijvoorbeeld bij het intypen van een zoekstring, zoals een familie- of straatnaam in een elektronisch telefoonregister. En zelfs waar de gebruiker helemaal zeker is van de spelling, heb je bij zoeken in vrije tekst nog het risico dat de documenten spellingvarianten bevatten, zoals in Nederlandstalige documenten van vóór de spellingsherziening van 1996.

Applicaties waarin spelling op één van deze manieren een rol speelt, kunnen gebruik maken van modules voor tolerant zoeken. De term *fuzzy matching* wordt in dit verband wel genoemd, maar die term heeft betrekking op het vergelijken van

een rijtje tekens, niet noodzakelijk letters. Bij taal spelen specifieke soorten tolerantie een rol. Iemand die sneller typt dan zijn vingers kunnen bijhouden, maakt andere fouten dan iemand die de precieze spelling van het woord niet kent. Bij familienamen is het heel vaak onmogelijk op basis van de klank de juiste spelling te raden. OCR (*Optical Character Recognition*) levert weer heel andere misers op. Een goede automatische spellingcontrole houdt rekening met al deze foutbronnen.

COMPROMIS

De veelgebruikte thuis- en kantoorapplicaties voor het typen van documenten hebben allemaal wel een vorm van spellingcontrole. Deze spellingcontrole is algemeen en daarmee een compromis. De taal van bijvoorbeeld een jurist is nu eenmaal heel anders dan die van een basisschoolleerling, een medicus of een romanschrijver.



ver. Termen die in een bepaalde organisatie worden gebruikt, zijn vaak geen onderdeel van de standaardtaal en worden door de spellingcontrole vaak gemarkeerd als mogelijke fouten. De meeste pakketten hebben voor dat soort situaties de mogelijkheid van een zogenaamde *custom dictionary*: een eigen lijst van termen die de spellingcontrole als correct moet accepteren. Het nadeel daarvan is dat elke gebruiker zo met een uitzonderingenlijst wordt opgezadeld. Deze lijst is vogelvrij: als er een fout in staat, wordt die daarna nooit meer gezien, laat staan verbeterd.

Sjoklaa

Het Groot Dictee der Nederlandse taal bewijst jaarlijks hoe lastig het is om het Nederlands foutloos te spellen. Voor kinderen is spellen extra lastig. Uw kind zoekt voor een spreekbeurt misschien naar informatie over 'sjoklaa'. De meest gebruikte speller van Nederland weet hier nog steeds 'chocola' van te maken.

Dat doet natuurlijk afbreuk aan het nut van spellingcontrole. Nieuwe termen in een organisatie komen en gaan (namen van producten, afdelingen, medewerkers, enzovoort), maar er is geen structuur voor het beheer van deze uitzonderingen. Wij stellen dan ook een alternatief model voor. Dat model helpt om de terminologie centraal te beheren. Daardoor worden de medewerkers ontheven van de taak om zelf woordenlijsten bij te houden. Een goede integratie met de applicaties die binnen de organisatie worden gebruikt, is wezenlijk voor het succes ervan.

ADAPTIVITEIT

Een spellingcontrole is veel meer dan een simpele woordenlijst. De meeste talen zijn zo flexibel dat een vaste woordenlijst nooit meer kan zijn dan een eerste indruk van wat wel of niet kan in die taal. Het heeft geen zin te proberen een dergelijke woordenlijst steeds maar groter te maken, want daarmee maak je spelfouten dan weer onzichtbaar. Er is namelijk altijd wel een context te verzinnen waarin die rare reeks letters toch wel iets betekent. Zo is 'wort' een term uit de bierbrouwerij en 'aardbij' volgens Van Dale een insectensoort. Een zoekactie op het web naar het woord 'aardbij' levert vooral verkeerd gespelde vruchten op.

Voor jonge gebruikers – maar ook voor andere taalleerders – gelden heel andere wensen voor een spellingcontrole dan voor auteurs van bijvoorbeeld zakelijke brieven. De leerders zijn vooral gediend met terugkoppeling waarbij de spellingregels worden uitgelegd. Een adaptief systeem dus. Daarvoor moet er dan ook een boek-

Medische spelling

Een voorbeeld van een vakgebied met veel specifieke termen is de medische en paramedische wereld. Een medisch rapport zal bijvoorbeeld binnen Word veel woorden rood onderkringelen, simpelweg omdat ze niet tot de standaardtaal behoren. Er bestaat een volledige geïntegreerde voorziening die de spellingcontrole uitbreidt met medische terminologie. Niet alleen verdwijnen daarmee veel van de kringeltjes maar spelfouten op dergelijke termen leiden daarmee ook tot goede suggesties. De medische speller laat desgevraagd ook alle termen zien die aan een bepaald patroon voldoen, zodat je hem zelfs kunt gebruiken als je niet op het woord kunt komen. Voor andere gebieden is een soortgelijke voorziening te realiseren.

houding zijn van de eerder gemaakte fouten en hun verbeteringen. De software zou bij de derde ontbrekende 'tussen-n' een pop-up kunnen genereren die de regel voor de 'tussen-n' nog eens toelicht.

Ten slotte nog even dit. Het blijkt dat taalvaardigheid een belangrijke factor is bij de beoordeling van mensen. Bijvoorbeeld in

Software mag intelligent zijn, maar moet het wel met ons eens zijn

een sollicitatieprocedure – waarbij een kandidaat in eerste instantie alleen op een brief geselecteerd moet worden voor een mogelijk gesprek – kan spellingsvermogen een onderscheidende rol spelen. Onze wereld hangt van tekstuele informatie aan elkaar. Het blijft voorlopig de moeite waard om zorg te besteden aan spelling en daarbij blijven goede hulpmiddelen onontbeerlijk. |

Spelling en smaak

Terwijl spelling eigenlijk niets anders is dan een afspraak – een conventie – kan het bij lezers diepe gevoelens van allerlei aard oproepen. Het voorstel van de spellingcommissie die door de Vlaamse en Nederlandse ministers van Onderwijs daartoe was ingesteld, leidde indertijd tot een verbale veldslag. Veel van de boze briefschrijvers zagen de verloedering van de taal versneld op zich afkomen. Ten onrechte. Spelling blijft een spel van mode en gewinning. Hoe het ook zij, de meeste lezers reageren al geërgerd als hun naam verkeerd gespeld wordt.

Theo van den Heuvel is managing director van Polderland Language & Speech Technology

Databaas

De spraaktechnoloog is de grootgrondbezitter onder de fonetici, zoals de taaltechnoloog dat is onder de taalwetenschappers. Waar de traditionele taalwetenschapper elk taalfeit koestert als een kostbare plant, daar zwaait de technoloog met een handvol cd's in de veronderstelling dat het taaluniversum aan zijn voeten ligt. Maar aan beider voeten gaapt echter de afgrond. Het is niet anders. Ik zal u vertellen waarom.

Met de traditionele taalwetenschapper zijn we gauw klaar. Met drie datapunten beschrijf je de wereld niet. Wie meent zo de wereld in kaart te brengen, duwt zichzelf wel de afgrond in. En anders zal de gemiddelde DIXIT-lezer hem graag het laatste zetje geven. Nee, interessanter is de taal- en spraaktechnoloog. Met zijn databases is hij toch de databaas. De waarheid over de taal ligt binnen zijn handbereik, want de wet der grote getallen is op zijn hand. Wat zal de DIXIT-lezer hem in de weg leggen?

En de waarheid zelf. Zij ligt als gewoonlijk in het midden. Maar eigenlijk zou ze daar moeten *staan*, en wel fier en trots.

Want zeg nu zelf, wie is de echte databaas? Als technoloog kun je het idee hebben dat je van de data zelf niets hoeft te weten om ze de baas te zijn. Het is dezelfde instelling die sommige managers hebben. Kijkend naar de door het 'grondpersoneel' opgeleverde cijfers nemen zij hun beslissingen zonder ook maar iets over de aard van hun product af te weten. Zo kun je erover denken. En eerlijk is eerlijk, deze 'grondhouding' is uitermate vruchtbaar bij het verzamelen van commissariaten.

Maar voor mij is de lol er wel af als de band met het product uit het zicht verdwijnt. Kijkend naar mijn vak – de spraaktechnologie – meen ik dat het zelfs kwalijk is. De spraaktechnoloog die de onbeduidendheid van elk afzonderlijk feit benadrukt, en bovendien als aanleiding gebruikt om niet eens goed naar een aantal feiten te kijken, kan lelijk op zijn snufferd gaan. Hij zou iets meer van de verwondering van de taalwetenschapper moeten hebben. De verwondering over de

data in de database kenmerkt de ware databaas. Althans, dat is mijn filosofie.

Nu schreef de filosoof Seneca al dat de weg via het voorbeeld doeltreffender is dan die via het opgedrongen voorschrift. Dus volgen een paar voorbeelden uit de praktijk. In talloze spraakdatabases worden de woorden 'ja' en 'nee' opgenomen door sprekers een spontane vraag te stellen. Door de grote hoeveelheid voorbeelden die zo worden verzameld kun je een spraakherkenner trainen nieuwe voorbeelden van die woorden te herkennen. Tot ik die ene database in handen kreeg, waar de herkenning toch wel erg slecht verliep. Een nader onderzoek van wat opnamen leerde dat de woorden waren verkregen door de vraag: 'bent u een man?'. Een aldus getrainde spraakherkenner maakt er een rommeltje van als hij een 'nee' uit een mannenmond en een 'ja' uit een vrouwenmond voor zijn digitale kiezen krijgt.

Een ander voorbeeld. In het kader van de automatisering van de reisinformatie door de Nederlandse Spoorwegen is er een automatische spraakherkenner gebouwd die u ook zelf kunt raadplegen via 0900-1475. Om deze herkenner te trainen zijn grote hoeveelheden dialogen opgenomen. En hoewel het om veel meer dan 100.000 uitingen gaat, is de database beslist niet algemeen bruikbaar voor ander onderzoek. Zelfs de regels voor uitspraakvariatie die zich uit de database lieten destilleren, bleken niet representatief voor andere databases. Wel ben ik regelmatig onder de indruk geweest van het inlevingsvermogen dat een nu verouderde versie van deze spraakherkenner aan de dag legde. Zo was er eens een grapjas die vroeg om de weg naar de hemel (het mocht ook de hel zijn, zo theologisch correct was de applicant nog wel). Hoewel er akoestisch geen enkele verklaring voor te geven valt, probeerde de spraakherkenner de beller binnen het beperkte bereik van zijn mogelijkheden zo goed mogelijk te helpen door hem de weg naar station 'Den Haag Mariahoeve' aan te bieden. Zo'n schat vindt een databaas niet vaak.

'Vragen om de weg naar de hemel?'

Henk van den Heuvel is verbonden aan het Nijmeegs Instituut voor Spraak- en Taal-Technologie

Spraaktechnologie de puberale jaren voorbij: van herkenning naar interpretatie

Eurospeech 2003: overdadige

Begin september werd in het international conference center in Geneve het 8ste Eurospeech congres gehouden. Ongeveer duizend bezoekers hielden zich vier dagen bezig met het luisteren naar en lezen van elkaars werk. Een van de dingen die opvallend duidelijk werd gedurende dit congres, is dat spraaktechnologie de puberale jaren definitief voorbij is.

| ARJAN VAN HESSEN |

Nieuw – tenminste voor mij – was de mogelijkheid om op het congres via een draadloos netwerk de godgansedag ‘connected’ te zijn, iets dat vooral door Aziaten massaal werd gedaan. In alle zitjes, hoeken en gaten van het redelijk grote congresgebouw vond je van ’s morgens vroeg tot na de laatste lezing mensen met een laptop op schoot. Aanvankelijke ergernis (ik ben boven de veertig) veranderde toen bleek dat je er ook nuttig gebruik van kon maken. Samen met de cd-rom met de volledige versie van voordrachten en de daarop aanwezige links, kon je gedurende de toespraken je eigen multimediashow creëren. Werd er gesproken over een slim antwoordapparaat op het MIT dat zelf de boodschappen sorteert (<http://web.media.mit.edu/~vidya/talk-back.html>), dan klikte je lustig mee en bekeek de webpagina’s terwijl de auteur doorging met zijn betoog. Wel vermoeiend

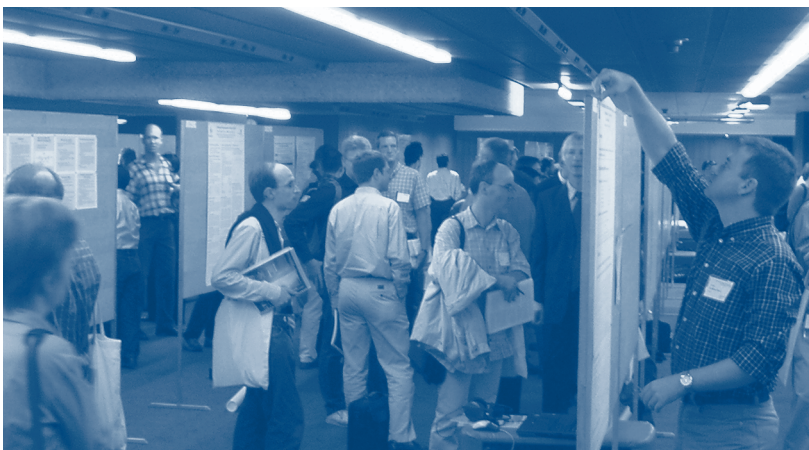
maar ook weer eens iets anders. Helaas had mijn batterij een levensduur van circa vijf minuten zodat het uiteindelijk niet effectief bleek, maar als concept is het zeker de moeite waard. Men kan zich voorstellen dat in de nabije toekomst – wanneer iedereen een laptop met power-batterij bij zich heeft, of er een kan lenen bij de balie – toespraken niet meer via een beamer worden geprojecteerd maar rechtstreeks op de laptop te volgen zijn.

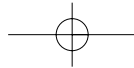
SPRAAKTECHNOLOGIE IN GROEISTUIP

Er waren opvallend veel presentaties over ‘Adaptation’, ‘Sprekerverificatie’, ‘Large Vocabulary Speaker Independent Speech Recognition’, ‘Spoken Document Retrieval’, ‘Subtitling by Recognition’ en ‘Speech Classification’. Spraaktechnologie begint duidelijk volwassen te worden, resulterend in opvallend veel toepassingsgerichte presentaties. ‘Akoestische Adaptatie’ is zo’n toverwoord: spraakherkenning heeft veel baat bij het zich zo snel mogelijk aanpassen aan de spraakkenmerken van de spreker. Dit fenomeen dat we al langer kennen van dicteersoftware (‘het voorlezen van een aantal zinnen zodat de computer leert hoe jij spreekt’) zorgt voor een flinke verbetering in de herkenning. Zeker na elf september 2001 is het niet verbazingwekkend dat veel 007-clubs zich bezig houden met het identificeren en/of verifiëren van sprekers (met een voorkeur voor niet-westerse talen). Het was fascinerend om een posterpresentatie te zien van de FBI: ‘Automated Speaker Recognition In Real World Conditions’. De beide mannen bij de poster zagen er inderdaad uit zoals in de X-files: groot, kort geknipt, wit overhemd, rode stropdas, jasje uit en je indringend aankijkend. Niks mis mee (de Nederlandse Geheime Dienst was er ook), maar soms denk je dat het leven de film imiteert!

VAN HERKENNING NAAR INTERPRETATIE

Een andere zeer interessante ontwikkeling is de verschuiving van herkenning naar





naar interpretatie

dis aan praktische informatie

interpretatie. Nu spraakherkenningapplicaties veel meer kunnen dan vooraf gedefinieerde zinnen herkennen, is het mogelijk 'How Can I Help You'-diensten te ontwikkelen. Bellers (het gaat bijna altijd om op telefonie gebaseerde applicaties) kunnen gewoon zeggen wat ze willen (zolang ze maar iets zeggen dat binnen de context van de dienst ligt). De herkenning op woordniveau is beslist niet perfect (circa 70 procent) maar met hulp van allerlei statistiek kan de gesproken vraag toch goed geclassificeerd worden. Dit soort applicaties is zeer geschikt voor het intelligent doorsturen van binnenkomende telefoontjes naar de meest geschikte callcenter-medewerker (skill based routing). AT&T en BT lieten indrukwekkende resultaten zien op dit gebied (het wordt in de praktijk al gebruikt voor allerlei storingsdiensten) waarbij ze twee miljoen telefoontjes per maand afhandelen.

RETRIEVAL KOPPELEN AAN SPRAAKHERKENNING

Een andere wereld waar spraaktechnologie aan een heuse opmars bezig is, is bij omroepen en multi-modale archieven zoals Beeld & Geluid in Hilversum. Die hebben erg veel materiaal dat allicht in aanmerking komt voor hergebruik, maar niemand weet precies waar wat staat (de uitzending van NOVA ging over goudhamsters, maar waar dat precies staat is onbekend). Ook hier wordt Information Retrieval gecombineerd met spraakherkenning. Men is in de regel niet op zoek naar één specifiek woord, maar juist naar items waar bepaalde woorden in combinatie gebruikt worden. Dit soort technieken (Spoken Document Retrieval) zijn snel en goedkoop, maar minder goed dan een documentalist van vlees en bloed. Hierdoor zijn deze technieken vooral geschikt om snel ingezet te worden zodat men in ieder geval iets heeft. Uiteindelijk zal een mens het nog wel beter kunnen doen, maar snelheid wordt door het grote informatieaanbod steeds belangrijker.

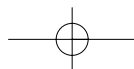


ONDERTITELING AUTOMATISCH GENEREREN

Een van de leukste toepassingen bleek het ondertitelen van televisieprogramma's. Engelstalige programma's zijn in Canada dikwijls ondertiteld, maar Franstalige niet (te duur voor de kleine groep Franstaligen). Uiteraard kan dat niet. En dus wordt tegenwoordig spraakherkenning gebruikt om de ondertiteling te genereren. Directe herkenning van het gesproken materiaal gaat echter nog niet goed genoeg en dus spreekt iemand met een 'goede' stem de TV-teksten na. Deze teksten worden dan door een tot uit den treure getrainde 'dictation engine' herkend en als ondertiteling op het scherm gezet. De vertraging is minimaal en blijkbaar acceptabel. Waar de angst dat de Franse taal niet meer mee zal tellen al niet goed voor is!

Valt er ook iets negatiefs te zeggen over Eurospeech? Ja, er is eigenlijk veel te veel informatie. Zeker in je eentje valt de enorme dis aan informatie niet te consumeren. Je loopt van Poster naar Voordracht en hoopt constant dat iets uitloopt omdat het anders helemaal niet meer gaat. Middag en avond worden besteed aan het uitvlooien wat je morgen allemaal moet horen en zien zodat je na vier dagen volkomen gaar (maar wel voldaan) terug bent. Daar heb je dan minstens drie weken nodig om alle informatie te filteren en aan de juiste personen door te geven. Vervolgens besluit je: volgende keer weer! |

Arjan van Hessen is verbonden aan de Universiteit van Twente



Digitale producten op maat van de gebruiker gesneden

Van Dale is meer dan een

Wat zou Johan Hendrik van Dale schrikken als hij vandaag het bedrijf kon bezoeken dat zijn naam draagt! In plaats van een arbeidsverslaafde onderwijzer die woordverklaringen noteert op kaartjes, zou hij een team van meer dan vijftig professionals ontmoeten die achter computers zitten en bepaald niet allemaal bezig zijn met woordjes verzamelen.

Lexicografie in de eenentwintigste eeuw is meer dan woordenboeken schrijven. Het is taaltechnologie!



Natuurlijk blijft de *Grote Van Dale*, tegenwoordig in drie kloeke delen, het vlaggenschip. Maar het is slechts één van de zeventig titels van de uitgeverij. Die is namelijk polyglot geworden en publiceert ook woordenboeken Engels, Frans, Duits, Spaans, Zweeds, Italiaans en Pools. En dat voor een divers publiek. Van Dale heeft woordenboeken voor kinderen, voor studenten, voor gezinnen, voor kantoorgebruikers, voor taalprofessionals, voor buitenlanders die Nederlands leren.

DE FICHES VOORBIJ

Al die woordenboeken op elkaar afstemmen, is al een opdracht op zich voor technologen. Stel dat een onderwijsminister een nieuwe term lanceert voor een schoolvak, dan moet die in de volgende uitgave van elk verklarend woordenboek worden toegelicht (in het ene boek voor kinderen, in het andere voor taalkenners) en dan moet dat woord ook in de vertaalwoordenboeken terecht komen. Daarom volstaat één bijgewerkt fiche in een kartonnen doos lang niet. Zo werkte Johan Hendrik aan het eind van de negentiende eeuw. Vandaag heeft Van Dale daar een omvangrijke database voor, die constant wordt bijgewerkt.

“Een database is natuurlijk op zich al een handig hulpmiddel, maar voor de redactie is het veel meer”, zegt Pyter Wagenaar, een van de redacteurs van de Grote Van

Dale. “Het is ontzettend praktisch dat álle informatie nu bij elkaar op één plek voor elke redacteur toegankelijk is, dus trefwoorden, spelling, vertalingen, betekenissen, gebruiksdomeinen en woordkenmerken als geslacht, afbreking, uitspraak, enzovoort, enzovoort.” Maar de digitale informatieverwerking maakt het ook mogelijk elk woord te koppelen aan andere woorden: aan synoniemen (woorden met dezelfde betekenis), antoniemen (tegengestelde betekenis), hyperoniemen (ruimere betekenis) en hyponiemen (engere betekenis). Daardoor ontstaan wat Wagenaar noemt ‘semantische netwerken’. Het betekenisprofiel van een woord wordt daardoor gedetailleerder. “Voor de redactie betekent dit dat zij scherp moet formuleren om überhaupt relaties te kunnen leggen, en dat zij erdoor op zoek gaat naar betekenisaspecten die een woord onderscheiden van zijn betekenisverwante soortgenoten.”

GRONDSTOF VOOR BEDRIJVEN

Uit die database, het geheugen van Van Dale, wordt meer inhoud gehaald dan alleen voor de woordenboeken.

Technologiebedrijven die toepassingen maken met taal, kloppen vaak bij Van Dale aan om ‘content’ aan te schaffen: woordenlijsten met een precieze beschrijving van de betekenis en het gebruik. Zo is er lexicografische grondstof verwerkt in intelligente zoekmachines voor het internet, in automatische vertaalsystemen, in software die teksten samenvat, in pennen die gedrukte teksten kunnen lezen.

Daarnaast zijn er in het buitenland woordenboeken op de markt gebracht met ‘Van Dale inside’: Wizzcom en Babylon in Israël, Wordfinder in Zweden en Paragon in Rusland en Duitsland.

Van Dale heeft ook zelf toepassingen die gebruik maken van taaltechnologie. Zo is in samenwerking met TNO de *Spelling-corrector* ontwikkeld. Dit is een intelligent controlesysteem dat met een grote precisie alternatieven voorstelt voor verdachte

| MARJOLIEN VAN DEN BERGH |

woordenboek

woordvormen. Dat de *Spellingcorrector* van Van Dale inderdaad beter werkt dan minder geavanceerde systemen, bewijst zijn gebruik door de redacties van grote kranten in Nederland en België. Ook in de spraaktechnologie is Van Dale actief. Het dochterbedrijf Fluency heeft samen met Vocalibur Language and Speech Technology onder meer *Belme* ontwikkeld. Dat is een 'digitale telefoniste' die voor een bedrijf de telefoon opneemt en de oproep doorschakelt naar de gevraagde medewerker. De Fluency e-maillizer zet binnenkomende elektronische post om in gesproken berichten. Daarbij wordt dezelfde technologie gebruikt die ook de elektronische Grote Van Dale in staat stelt om van alle trefwoorden de uitspraak te laten horen.

PAPIER IS (BIJNA) NIET MEER

Ook op andere punten is het woordenboek van vandaag niet te vergelijken met dat van Johan Hendrik. Toegegeven, de papieren versie (de folioversie, zeggen vaklui) bestaat nog en zal blijven bestaan. Maar van de verkoop van de *Grote Van Dale* maakt ze minder dan zeventig procent uit. Bij de Grote vertaalwoordenboeken is deze tendens nog beter zichtbaar: minder dan de helft van de verkochte woordenboeken is van papier. Dat komt doordat de professionele markt (kantoren, vertalers, schrijvers, journalisten) definitief gewonnen is voor het digitale woordenboek.

"De tijd die ik uitspaar door de Van Dale op mijn computer, is niet te berekenen", zegt Ludo Permentier, taalbewaker van de Vlaamse krant *De Standaard*. Hij vlooit de krant elke dag na op formuleringen die beter hadden gekund en krijgt tussendoor taalvragen van collegajournalisten en ook van lezers. "Vroeger zat ik soms met de hoorn van mijn telefoon geklemd tussen schouder en oor in de drie boekdelen te bladeren, op zoek naar een juiste spelling, een precieze betekenis of een woordcombinatie. Dat was vaak frustrerend, want je

moest soms lange kolommen doorlezen voor je vond wat je zocht. Vandaag tik ik een deel van een woord in, en ik zie het antwoord al verschijnen."

MEER DAN ÉÉN VOORDEUR

Het digitale woordenboek heeft trouwens meer dan één voordeur. Je kunt inderdaad, zoals Permentier dat doet, een trefwoord intikken. Je krijgt dan het antwoord

Nieuwe generatie woordenboeken is afgestemd op de context van de gebruiker

dat ook het papieren woordenboek geeft. Maar je kunt ook heel andere vragen stellen. Bijvoorbeeld omgekeerd: je hebt een idee van een betekenis, en je zoekt er een woord voor. Hoe noem je ook al weer een dokter die alles afweet van de lever? (een hepatoloog.) De Grote Van Dale op cd-rom helpt je dat woord te vinden. Of je kunt door het gehele woordenboek op zoek gaan naar uitdrukkingen. Tik een of meer woorden in, en je krijgt een lijstje.

Voor de ene gebruiker een *gadget*, voor de andere een dagelijks instrument. Dat is typisch voor het nieuwe woordenboek.

"De nieuwe generatie van woordenboeken – ik zou beter zeggen: lexicografische producten, want er komt onder meer spellingcontrole en taalcorrectie bij – is afgestemd op de context waarin de gebruiker werkt", zegt Onno Makor, die belast is met de ontwikkeling van die nieuwe toepassingen.

"Het is goed mogelijk dat een zakelijke gebruiker een andere woordenschat tot zijn beschikking wil hebben dan een student. En dat geldt niet alleen voor de inhoud, maar ook voor de functionaliteit. Die kan steeds beter geïntegreerd worden in de digitale werkzaamheden van de gebruiker, maar dat doen we alleen als hij dat als nuttig ervaart." |

Marjolien van den Bergh is
hoofd communicatie en PR bij
Van Dale Lexicografie

Uit de boeken

L E E S W I J Z E R

De praktijk is altijd een goede leermeester, maar zonder gedegen achtergrondkennis – trends, ontwikkelingen, toegepast onderzoek, noem maar op – kan die zelfde praktijk al snel een zeer beperkte leermeester worden. Het praktische blikveld heeft al gauw de neiging zich te vernauwen tot de waan van de dag. Vandaar dat Antal van den Bosch – verbonden aan de Universiteit van Tilburg, faculteit ILK/ Computational Linguistics – telkens een aantal relevante boeken of artikelen kort zal bespreken. Te beginnen met twee standaardwerken die de TST-dekens in uw werkkamer even flink opschudden.

De literatuur bijhouden in het vakgebied taaltechnologie is – zoals in zo veel vakgebieden – geen eenvoudige opgave. Er verschijnen talloze boeken (de een rijper, de ander tamelijk groen), er zijn vele al dan niet gerenommeerde tijdschriften (bijvoorbeeld *Computational Linguistics* en *Natural Language Engineering*), en we kunnen op de hoogte blijven via een aantal gevestigde *proceedings* van internationale conferenties (bijvoorbeeld *ACL* en *COLING*). En daarbuiten heerst weer een vrolijke wanorde van workshops en soortgelijke evenementen die een enorme stroom aan publicaties veroorzaken. Daarbij is taaltechnologie, naast een domein op zich, ook een toepassingsgebied geworden, en een bron van gereedschappen voor weer andere domeinen en toepassingsgebieden, met hun eigen publicatiemedia.

Gelukkig is met internet de beschikbaarheid van dit soort media geen flessenhals meer: je bestijgt je favoriete browser, en met behulp van zoekmachines en diensten als CiteSeer vind je doorgaans alles wat je wilt. Eigenlijk houdt iedere literatuurwijzer hier op – bepaal je eigen zoektermen en daar ga je.

Boeken vormen hierop een uitzondering. Je vindt de meeste niet op internet, maar tegelijkertijd bieden ze voor de geïnteresseerde maar niet gespecialiseerde lezer de breedste informatie, breder dan het gemiddelde artikel waar de achtergrondkennis-lat vaak (te) hoog ligt. Vandaar in deze rubriek een scala aan boekentips, bedoeld voor diegene die op de hoogte wil blijven over de laatste ontwikkelingen, de *what's in and what's out*, en de bijgestelde toekomstperspectieven. Te beginnen met twee standaardwerken die we even kort de revue laten passeren.

Speech and Language Processing

Een lezenswaardig, maar wel lijvig boekwerk dat op veel universiteiten gebruikt wordt als standaardcursusboek, is *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* van de hand van Dan Jurafsky en James Martin. Deze twee Amerikaanse computerlinguïsten geven een behoorlijk compleet en eerlijk beeld van de huidige stand van het vakgebied. Het boek wordt door kenners geprezen

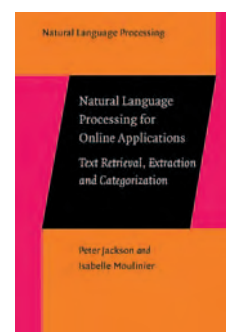
om zijn encyclopedische compleetheid, het coherente en consequente beeld dat van het vakgebied en alle deeldisciplines wordt geschilderd, gelardeerd met veel voorbeelden en humor. Letterlijk een dikke aanrader.

Dan Jurafsky en James Martin (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall.
<http://www.cs.colorado.edu/~martin/slp.html>



Natural Language Processing

Het tweede boek dat ik van harte aanbeveel – en als docent als verplichte stof opgeef voor een vak dat gaat over de toepassing van taaltechnologie in 'internet'-technologie – is *Natural Language Processing for Online Applications*. Dit boek behandelt uiteenlopende onderwerpen als information retrieval voor search engines, informatie-extractie van het web, tekstclassificatie, automatisch samenvatten, vraag/antwoord-systemen, en text- en webmining.



Zoveel hips bij elkaar zou je niet verwachten in een goed boek – boeken doen er doorgaans een jaar of wat over om uitgebracht te worden en verspreid te raken, en wat is er over van de technologie die in het boek wordt beschreven? In het geval van dit praktische handboek uit 2002 van auteurs Peter Jackson en Isabelle Moulinier is de inhoud nog altijd bijzonder actueel – de houdbaarheidsdatum ligt nog wel even in het verschiet. Een goed moment dus om dit relatief goedkope en aardige boek aan te schaffen en direct te lezen. Het is minder diepgaand dan het boek van Jurafsky en Martin, maar het onderwerp leent zich dan ook meer voor treffende voorbeelden en sprekende anekdotes. Als de verjaardagsdiscussie gaat over *search engines*, dan kan de lezer van dit boek de show stelen met een complete recitatie van het gepatenteerde PageRank-algoritme van Google.

Peter Jackson and Isabel Moulinier (2002). *Natural Language Processing for Online Applications*. John Benjamins.
<http://members.aol.com/JacksonPE/music1/nlp4olap.htm>

TST Dictionary

Dictionary: zie **Spraaksynthese** en **Spraakherkenning**; zie ook **Fonetische Transcriptie**, zie ook...

Zoals elk vakgebied heeft ook TST (taal- en spraaktechnologie) een compleet eigen vocabulaire, inclusief volstrekt duistere afkortingen. Handig natuurlijk om een zeker aureool rond het vakgebied te garanderen en in stand te houden, maar bepaald niet bevorderlijk voor de onderlinge communicatie tussen TST-expert en leek (lees: klant). 'Multi-modaal'? Hij bedoelt toch zeker 'multi-mediaal'! En IVR kunnen we inmiddels wel duiden, en ook TTS zal een belletje doen rinkelen, maar ASR, VXML, NLDS of DTMF?

Arjan van Hessen (Universiteit Twente) zal de mythische woorden- en begrippen-sluier, en daarmee het TST-vakgebied, voor u oplichten, resulterend in een heuse TST-dictionary (= de basis van iedere spraaksynthese en spraakherkenningapplicatie is het dictionary: een lijst met correct geschreven woorden en hun fonetische transcriptie(s). Alleen woorden die in het dictionary staan, kunnen herkend worden).

begrijpen *Begrijpen* en *verstaan* zijn twee verschillende begrippen die in de spraakherkenning dikwijls door elkaar gehaald worden. In TST wordt *begrijpen* gezien als het (juist) interpreteren van een (met spraakherkenning verkregen) tekst. Hoewel niet altijd, leidt foute *herkenning* op woordniveau meestal wel tot onjuiste interpretatie. Op zinsniveau is dit niet noodzakelijk het geval. Wanneer voldoende woorden juist herkend zijn, kan het gebruik van allerlei technieken toch tot juiste interpretatie leiden. Het programma heeft het dan toch begrepen.

compounding *Compounding* is het aan elkaar plakken van twee of meer woorden tot een nieuw woord. *Compound*-talen zoals het Nederlands en Duits hebben daarvoor veel meer woorden dan bijvoorbeeld Engels of Italiaans. Om toch met hetzelfde aantal woorden evenveel tekst te kunnen verklaren, moeten *compounds* eerst *gedecomposeerd* worden. Vaak gaat dit automatisch goed, maar niet altijd: zeeroverschatten – zeerover + schatten of zeeroverters + chatten

herkennen Het juist omzetten van *gesproken-spraak* in *geschreven-spraak*. Het vormt de basis van alle spraakherkenningssystemen. In tegenstelling tot dialoogsystemen waarbij

het om de interpretatie van het gesprokene gaat, is bij dicteren slechts de herkenning van belang.

judgements Een van de lastigste problemen is de *judgement* van informatie. Hoe relevant is een artikel eigenlijk gegeven de zoekvraag? Hoe logisch is het dat de spraakherkenner hier geen chocola van kan maken, want ik versta het zelf ook nauwelijks.

kennis van de wereld Steeds duidelijker wordt de nadruk gelegd op het gebruik van '*kennis van de wereld*' in TST-applicaties. Herkenning zonder enige kennis van de context waarbinnen het gesprek zich afspeelde, is uiterst lastig. Maar hoe modelleren ik die kennis van de wereld? Iedere Nederlander weet ongeveer over wie het gaat als je het over 'ze wilde gewoon prinses worden joh!' hebt, maar een computer?

lexical coverage De mate waarin unieke woorden een tekst kunnen verklaren. Stel dat we een tekst hebben met 10 keer het woordje 'koffie', 5 keer 'bier' en éénmaal thee, dan is de *lexical coverage* van de twee meest frequente woorden circa 94 procent (=15/16). De *lexical coverage* is hoger naarmate minder verschillende woorden gebruikt worden. Voor het Nederlands zijn ongeveer 3 keer gro-

dictatorship [dik'tetəʃɪp] 0.1 **dictation** [ˈdɪkʃən] 0.1 **dictie** ⇒ *voordwoordkeus, dictie*.
dictionary [ˈdɪkʃənəri] <mv.: -ies> 0.1 **con.**
dictum [ˈdɪktəm] <mv.: ook dicta> 0.1

tere woordenboeken (65K unieke woorden) nodig dan in het Engels (20K) om 97.5 procent van de tekst te kunnen verklaren.

multi-modaal Het op meerdere manieren voeren van een dialoog of representeren van informatie. Veel TST-onderzoekers hebben zich *mono-modaal* verkeken op hun eigen specialisatie. Mensen geven in dialogen met een systeem de voorkeur aan de snelste manier van communiceren: spreken of typen wanneer één keuze uit velen gemaakt moet worden, klikken, toetsen of wijzen wanneer het aantal opties zeer klein is. De stationsnaam spreekt of typt men daarom in, de keuze tussen vandaag of morgen klikt of toetst men liever. De resultaten ziet men liever op een beeldscherm dan dat het voorgelezen wordt. Foutcorrectie gaat makkelijker door het niet goede item aan te klikken (of wijzen) en het te verbeteren. De beschikbaarheid van PDA's en andere kleine mobiele apparaten met schermen maakt de vervanging van *singel-modale* door *multi-modale* dialogen waarschijnlijk. Een probleem blijft de gelijktijdige invoer van verschillende modaliteiten (bijvoorbeeld iets zeggen en ergens op klikken): wie wint?

waarschijnlijkheid Taal en spraaktechnologie zijn net als mens/mens-communicatie gebaseerd op waarschijnlijkheden. Hoe beter de kans geschat kan worden welk woord waarschijnlijk uitgesproken zal worden (gegeven reeds herkende woorden en/of het onderwerp van de conversatie), hoe beter de spraakherkenning zal werken. De waarschijnlijkheid van het woord 'sluier' zal hoger zijn dan dat van 'luier' in een conversatie over minderheden, terwijl het juist lager zal zijn in een conversatie over kraamzorg.

Uitgelicht!

Startpagina krijgt TST-dochter

Op initiatief van Rob Boeyink (AskNow) wordt de startpagina-familie uitgebreid met een nieuwe dochter: taaltechnologie.pagina.nl. 'Everything you wanted to know about TST, but were unable to find'. Bedrijven, instellingen, brancheverenigingen, publicaties, enzovoort, het is er allemaal te vinden, zij het nog in embryonaal stadium. Vandaar dat Boeyink iedereen oproept suggesties aan de hand te doen. En dat geldt niet alleen de doorgewinterde professional maar zeker ook de dagelijkse gebruiker. Want de bedoeling van deze pagina is nadrukkelijk om taal- en spraaktechnologie toegankelijker te maken voor een breed publiek. In NOTaS-terminologie: de digitale pendant van markteducatie. (u kunt uw suggesties rechtstreeks mailen aan: rob@asknow.nl)

Klantherkenning op basis van spraak

Hoe zorg je ervoor dat het IVR-systeem 'Kreuningen' herkent als 'Groningen'. Of dat de ingesproken naam 'Jan Konings, Ede' wordt omgezet in de correcte en unieke identificatie 'Han Cooningsh, Eelde'? Door de IVR-applicatie op fouttolerante wijze te vergelijken met klankcoderingen uit de database. Newtel Essence en Human Inference hebben hiertoe gezamenlijk IRON Hlquality ontwikkeld. Hierdoor wordt 95 procent van de binnenkomende gesprekken in een call center

direct herkend. Uit het persbericht: 'Het resultaat dat tot nu toe met IVR-klantidentificatiemethoden wordt bereikt, is dat slechts 25 procent van de klanten uniek wordt geïdentificeerd. Drie van de vier klanten wordt dus gevraagd om hun gegevens nogmaals in te spreken of aanvullende informatie te verschaffen. Een groot deel heeft daar geen zin in en zal ophangen.' Met als resultaat een verbetering van de volautomatische klantherkenning met 300 procent.

Voor meer informatie...

...toets een 1. Om vervolgens in een doolhof aan keuzen te geraken waar zelfs een ervaren keuzegids de weg kwijtraakt. Kortom: bij het zoeken naar manieren om kosten te besparen, is men ergens onderweg de klant kwijtgeraakt. Volgens een aantal TNO-onderzoekers is de gemiddelde menustructuur van een geautomatiseerde telefoondienst veel te complex. Met als gevolg de hoorn subiet op de haak. Ook wordt de beller nogal eens verkeerd toegesproken. Niet imperatief (toets uw code in), maar vriendelijk, dus te uitgebreid (wilt u nu uw code intoetsen). Bij spraaksystemen is het vaak de techniek die tekort schiet. Als je de filedienst '9395' belt, doe je er goed aan het snelwegnummer te noemen, en dus niet – wat op zich logisch is – vertrekpunt en eindbestemming, omdat het systeem daar niet op reageert. De richtlijnen die worden gegeven om veel ergernissen te voorkomen zijn blijkbaar zo obligaat dat ze zelden in acht worden genomen: denk vanuit de gebruiker, maak korte dialogen, gebruik gangbare taal en test met 'echte' gebruikers op duidelijkheid en gemak.

(bron: Computable, 21-11-03)

~~Vragen is een vorm van intelligentie...~~
~~...nu het antwoord nog!~~

~~The right answer at once!~~
 Met de AskNow! Intelligente Web Selfservice Software

~~Alles weten?~~
www.asknow.nl

Ask NOW SOLUTIONS

Customer Selfservice
 make your customers satisfied

Call Center
 A call for the right answer

E-government
 De bron voor een succesvolle dienstverlening

Employee Selfservice
 Kennis krijgt een waarde als deze gedeeld wordt

NOTaSnieuws

Kenniseconomie?

Kenniseconomie heet het in het Hoofdlijnenakkoord van het kabinet Balkenende II. Vurig bepleit door D'66, want Nederland is achterop geraakt als het gaat om fundamenteel onderzoek en innovatieve ontwikkeling. En als we niet oppassen, worden we nog een derdewereldland op dat gebied. Vanzelfsprekend omarmt NOTaS dit aspect van het kabinetsbeleid, maar constateert dat tussen het met de mond belijden en het in de praktijk doen, grote verschillen bestaan. Als er verandering nodig is, is het niet alleen een kwestie van geld. De regering en het ambtelijke apparaat moet zich ook de vraag stellen hoe het komt dat Nederland achterop is geraakt. NOTaS is van mening dat er vanuit de betreffende ministeries durf moet worden getoond. Durf om dingen anders aan te pakken dan in het verleden, waarbij lange tenen en heilige huisjes niet worden ontzien. Want 'if you do what you did, you get what you got'.

TST-tijdperk

Na de automatisering van de financiën, productie, logistiek en CRM, breekt nu het tijdperk van de taal- en spraaktechnologie aan. Nederlands is geen wereldtaal en in Europa neemt het de zesde plaats in. Veel softwareleveranciers respecteren het Nederlands al in die mate, dat ze de moeite nemen schermteksten en foutboodschappen te vertalen. Daarnaast zijn woordenboeken en spellingcheckers inmiddels gemeengoed. Om de doelgroep optimaal te bereiken is het dus commercieel aantrekkelijk om software te 'nationaliseren'. Voor wat betreft taal- en spraaktechnologie ligt dit iets anders: de investering in onderzoek en ontwikkeling is erg groot in relatie tot de marktomvang voor het Nederlands. Dit lijkt in contrast te staan met hetgeen iedereen wel degelijk beseft: in de komende vijf tot tien jaar is TST maatschappelijk gemeengoed. Maar het ontwikkelen van een markt kost tijd en daarvoor zijn pioniers nodig: zowel producenten en leveranciers met lef, als de witte raven onder de beoogde groep afnemers. Doch als bedrijven net iets te terughoudend zijn, zal Amerikaanse technologie op een gegeven moment de boventoon voeren: of we spreken binnen enkele jaren allemaal Engels met systemen of ergeren ons dood aan de krakemikkigheid en de 'onlogica' van de spraakherkenning en taalverwerking. Daarom biedt de keus van de regering voor kennisontwikkeling en kenniseconomie een uitgelezen kans om Nederlandse ondernemingen en universiteiten door middel van gerichte stimulering nu net het duwtje in de goede richting te geven. En dat zal, gezien de positie van Nederland in Europa, haar bredere uitwerking zeker niet missen.

Wie heeft de regie?

Als het gaat om taal- en spraaktechnologie, dan telt het veld de volgende spelers: de ministeries van EZ en OC&W, NWO, Nederlandse Taalunie, bedrijfsleven en kennisinstellingen (universiteiten). Ieder met eigen doelstelling, belangen, gevestigde posities en bestaande regelingen. In NOTaS zijn kennisinstellingen en bedrijfsleven verenigd.

Uit het Hoofdlijnenakkoord Balkenende II

Uit de paragraaf 'Onderwijs en kennis' halen we de volgende citaten:
'Nederland moet tot de Europese voorhoede behoren op het terrein van hoger onderwijs, onderzoek en innovatie' en 'Ondanks de moeilijke financieel-economische situatie wordt er niet bezuinigd op onderwijs en kennis. Integendeel, het kabinet trekt juist fors extra middelen uit voor deze prioriteit in het beleid.'
In de laatste alinea staat ten slotte: 'Het behoud van de Nederlandse taal en cultuur in een steeds kleiner wordende wereld is van groot belang.'

De Taalunie is een Nederlands-Vlaams intergouvernementeel orgaan, met als doel gemeenschappelijk beleid ten aanzien van het Nederlands te voeren. Dit doet ze door zich onder andere te richten op het tot stand brengen van 'een elektronische 'taalinfrastructuur' om de positie van het Nederlands in de internationale informatiemaatschappij te versterken' en zo behoort ook 'de meertaligheid in de instellingen van Europese Unie' tot haar aandachtsgebied. De Taalunie heeft het zogenaamde TST-Platform opgericht dat een onderzoek initieerde naar de benodigde taal- en spraakbouwstenen om het Nederlands in de automatisering te waarborgen. Dit leidde tot het 'BATAVO-rapport' (basis taal- en spraakvoorziening). In principe hebben de Nederlandse en Vlaamse overheid zich bereid getoond de BATAVO te ondersteunen en daartoe geld vrij te maken. Voor Nederland zijn dit de ministeries van EZ en OC&W. Het ministerie van EZ huldigt het standpunt dat de wijze waarop aan het plan uitvoering wordt gegeven, de steun dient te hebben van het bedrijfsleven. NOTaS is daarom uitgenodigd mee te praten en heeft op basis van de gesprekken een plan opgesteld. Dat plan richt zich op twee sporen: synergie tussen bedrijfsleven en universiteiten, en beheer. Dat laatste is nodig om een eind te maken aan de kapitaalvernietiging. In het verleden is namelijk maar al te vaak gebleken dat door het ontbreken van structuur met overheidsgeld gesubsidieerde ontwikkelingen bij universiteiten op de plank blijven liggen en in de vergetelheid raken. Het NOTaS-plan mag op algemene steun rekenen.

IOP-studie

Omdat middelen kennelijk uit bestaande regelingen moeten komen heeft EZ onlangs op eigen initiatief een zogenaamde IOP-studie gestart (Innovatie Onderzoeksprogramma; het enige bestaande ambtelijke middel waarover EZ beschikt en dat gericht is op subsidies aan universiteiten met daarbij de hoop/wens van een spin-off naar het bedrijfsleven). En parallel hieraan blijkt de Nederlandse Taalunie, die het eigendom ➔

NOTaSnieuws

heeft gekregen van vijf majeure onderzoeksresultaten uit het verleden, ten aanzien van het beheer hiervan een eigen koers te varen.

Na talloze sessies, gesprekken en vergaderingen heeft bij het NOTaS-bestuur de indruk postgevat dat door het bijna dwangmatig willen toepassen van bestaande regelingen en het niet tegen het licht willen houden van de doelmatigheid van de gevestigde orde, niet het optimale resultaat wordt gehaald. Terwijl dit wel binnen het bereik ligt. Kenniseconomie, zoals in het regeerakkoord is bedoeld, komt op deze wijze niet van de grond. Net als bij de laatste kwalificatiewedstrijd van het Nederlands elftal tegen Schotland, moet bij de overheid een kleine generaal opstaan die de touwtjes in handen neemt en op het allerlaatste moment het roer toch radicaal durft om te gooien. Met bestaande spelers die het spel breed houden en de bal terugspelen, wordt het doel niet bereikt.

Plan

NOTaS heeft een plan voor structurele samenwerking in de TST-sector geschreven. Hierin leest u meer over de visie van NOTaS en de posities van de diverse belanghebbenden. U kunt dit plan downloaden op de website www.stichtingnotas.nl

Met dank aan...

DIXIT wordt u toegestuurd, omdat u behoort tot het relatiebestand van een van de NOTaS-deelnemers. Omdat de kans groot is dat u in meerdere relatiebestanden voorkomt, is de grootste zorg besteed aan het perfect ont-dubbelen van alle bestanden.

DIXIT is ongetwijfeld interessant, maar een hele stapel is bepaald 'overdone'. De redactie hecht eraan Human Inference te danken voor het geheel belangeloos ontdubbelen van de vele relatiebestanden.

NOTaS

Deelnemers van Stichting NOTaS zijn: AskNow Solutions - Comsys - EliTech - GIOS Voice Professionals - Human Inference - Institute for Logic Language and Computation (ILCC) - Knowledge Concepts - UvT Faculteit der Letteren - Het Nijmeegs Instituut voor Spraak- en Taaltechnologie (NISTT) - ORCAvoice - PAT Document Systemen - Polderland Language and Speech Technology - Q-go - Technische Universiteit Eindhoven - TeleCats - TNO TPD - Universiteit Twente Faculteit Informatica - Van Dale Data - Vocalibur Language & Speechtechnology

Meer informatie: www.stichtingnotas.nl

NOTaS groeit met dank aan de NV BOM

In maart 2001 is Stichting NOTaS (Nederlandse Organisatie voor Taal- en Spraaktechnologie) opgericht. Het was een initiatief van de N.V. BOM (Brabantse Ontwikkelingsmaatschappij) en enkele (bijna-) Brabantse bedrijven en de universiteiten van Tilburg en Eindhoven. Inmiddels zijn er bijna twintig deelnemers en heeft de BOM haar taak als aanjager van het proces volbracht. Onlangs heeft de BOM dan ook formeel afscheid genomen als adviseur van het bestuur van NOTaS.

NOTaS is belangenbehartiger en ontmoetingsplatform tegelijk. Veel van de bedrijven die actief zijn op het terrein van taal- en spraaktechnologie in Nederland hebben zich inmiddels aangesloten. Als belangenbehartiger van de sector taal- en spraaktechnologie is NOTaS nadrukkelijk geaccepteerd. Er wordt op regelmatige wijze overlegd met de direct betrokken ministeries (EZ en OC&W), met de Nederlandse Taalunie en met NWO.

Doordat TNO/TPD alsook de universiteiten van Twente, Nijmegen en Amsterdam zich hebben aangesloten, is NOTaS niet alleen spreekbuis van het bedrijfsleven maar een serieus ontmoetingsplatform voor betrokkenen uit de hele branche. Bedrijven en instellingen die willen deelnemen aan de stichting kunnen zich via www.stichtingnotas.nl aanmelden.

Activiteiten

De doelstellingen en het activiteitspectrum van NOTaS zijn ruim. Op dit moment wordt gefocust op:

- **Belangenbehartiging** NOTaS vertegenwoordigt de sector in gesprekken met de Nederlandse Taalunie, EZ, NWO enz. De Nederlandse en Vlaamse regering hebben de bereidheid getoond middelen vrij te willen maken om de taal- en spraaktechnologie een duw in de rug te geven.
- **Stimuleren samenwerking** Al door de deelnemersvergaderingen brengt NOTaS bedrijven onderling en met onderzoeksinstituten c.q. universiteiten dicht bij elkaar. Op .. januari 2004 wordt de eerste Team voor Taal gehouden. Een besloten NOTaS-bijeenkomst waar de deelnemers zich uitgebreid presenteren.
- **Markteducatie** Het is een zwaar woord voor het dichter bij brengen van taal- en spraaktechnologie. DIXIT is daar een voorbeeld van, net zoals de website. Maar NOTaS brengt dit onderwerp ook naar voren in overleg met de (semi-) overheidsinstanties.
- **Gezamenlijke voordelen** Onderling wordt gewezen op informatiebronnen, regelingen, publieke inschrijvingen enz. Maar ook deelname aan congressen en beurzen.