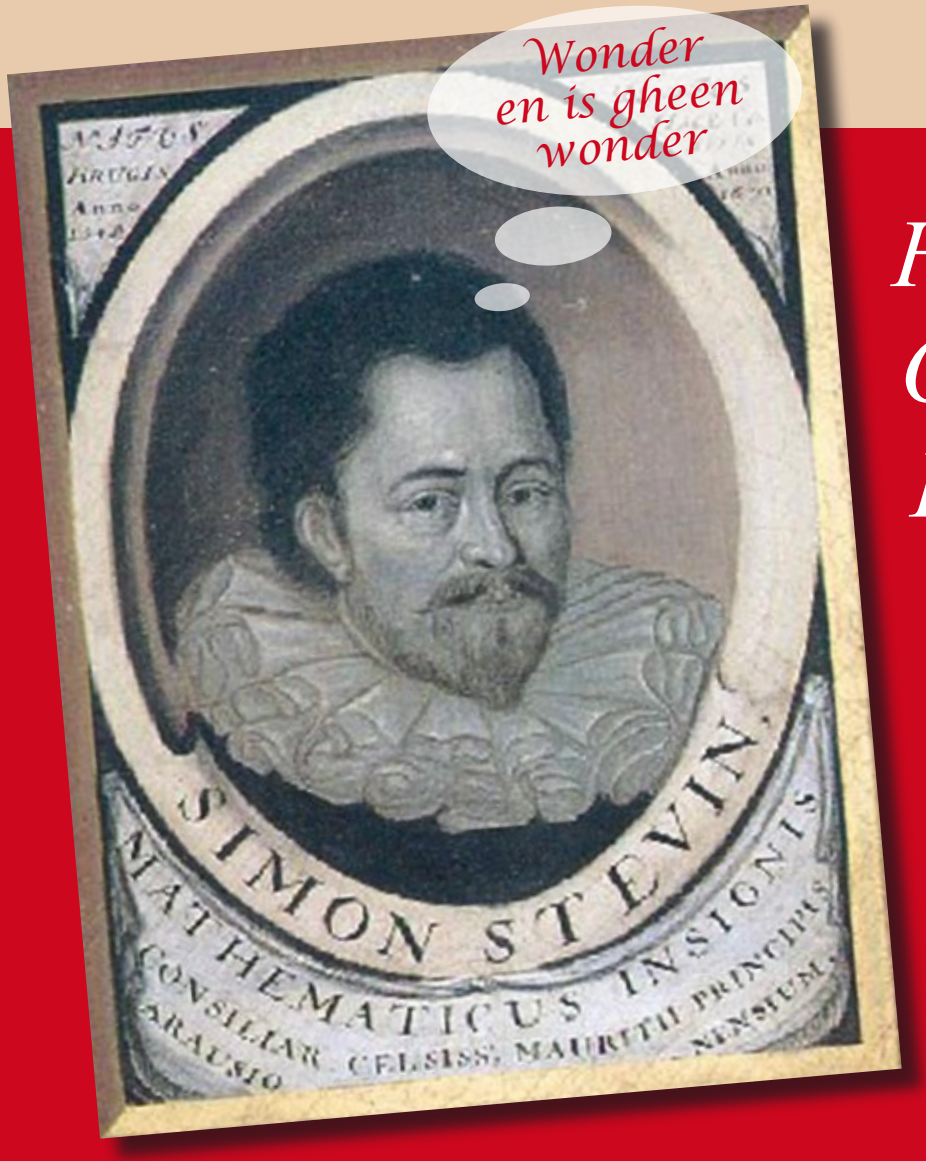
DIXIT

TIJDSCHRIFT OVER TOEGEPASTE TAAL- EN SPRAAKTECHNOLOGIE

JAAR BOEK

Het STEVIN onderzoeks- en stimuleringsprogramma

Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands



Het Radio Oranje Project:

Googlen met Hare Majesteit Koningin Wilhelmina.

JAARGANG 4 NUMMER 2 DECEMBER 2006

geschreven en gesproken Nederlands

in digitale vorm

TST-CENTRALE
voor taal- en spraaktechnologie

Bent u op zoek naar digitale taalkundige data? Bij ons bent u aan het juiste adres!

Beschikbare producten:
- Corpus Gesproken Nederlands
- INL-tekstcorpora
- Eindhoven Corpus
- ParoleCorpus en ParoleLexicon
- Bilinguale lexica
- Woordenlijst Nederlandse taal
- e-Lex
- Referentiebestand (Belgisch-)Nederlands
- Frequentielijsten van het Nederlands

De centrale voor taal- en spraaktechnologie is hét Nederlands-Vlaamse loket voor digitale taalkundige bronnen. De TST-centrale beheert en distribueert digitale taalmaterialen die met overheidsgeld zijn gefinancierd. De materialen worden onderhouden en beschikbaar gesteld voor onderwijs, onderzoek en ontwikkeling van taaltechnologische producten. Daarnaast ondersteunt de TST-centrale het gebruik van de materialen door uw vragen over bijvoorbeeld IPR te beantwoorden en themadagen en workshops te organiseren.

Wilt u weten wat de TST-centrale voor u kan betekenen of wilt u meer informatie over de producten? Surf naar *www.tst.inl.nl* of stuur een e-mail naar *tst@inl.nl*.

TST-centrale Leiden
p/a Instituut voor Nederlandse Lexicologie
Postbus 9515, 2300 RA Leiden, Nederland
Tel. +31 (0)71-527 2482, Fax +31 (0)71-527 2115

TST-centrale Antwerpen
p/a Instituut voor Nederlandse Lexicologie
Universiteitsplein 1, kamer A1.24, 2610 Wilrijk, België
Tel. +32 (0)3- 820 2784, Fax +32 (0)3-820 2784

INHOUD

Wiskunde en Taal 4

Het Radio Oranje Project 6
'Googlen' met Hare Majesteit Koningin Wilhelmina

Basismateriaal: Taal 8

Het Stevin-onderzoeksprogramma 11
Voorwoord van gastredacteur Peter Speyns

De IPR-regeling van STEVIN 12

STEVIN projecten o.a.:

MIDAS 16
Hugo van Hamme verteld hoe de spraaktechnologie door middel van het MIDAS project om kan gaan gaan met de robuustheid van taal.

Autonomata 24
Zou het niet fantastisch zijn als een autobestuurder aan zijn navigatiesysteem kon zeggen waar hij of zij naartoe wil? Waarom nog aan wieltjes draaien of in menu's navigeren om letter per letter de gewenste bestemming in te voeren? Jean Pierre Martens verteld hoe men dit droomscenario kan realiseren.

Spraakgestuurde Nummerbord Retrieval Tool 30
Dutchear realiseerde voor Politie Utrecht de Nummerbord Retrieval Tool (NRT). Agenten kunnen lopend, op de mountainbike, in de auto en op de motor met hun gsm bellen met de NRT. De NRT zorgt ervoor dat agenten van Politie Utrecht altijd op een snelle, gemakkelijke en veilige manier voertuiginformatie kunnen krijgen. Lees meer over de NRT in dit artikel.

Stevin Programmadag

De kleur van spelling 32
Theo van den Heuvel (Polderland) legt nog eens haarfijn uit waarom er maar één officiële spelling is.

IMIX-Midterm 34

Automatisch beantwoorden van e-mail 36

En verder
Speeddating 10
NOTaS Nieuws 39
Directory 38

Yesterday, today, tomorrow...

Yesterday

Met taal- en spraaktechnologie is het nu makkelijk te achterhalen van wie Yesterday is. Dit schrijft prof. Lou Boves van het CLST in Nijmegen naar aanleiding van de tussentijdse resultaten van het IMIX-programma. In IMIX zien we diverse taal- en spraakprojecten die nauw samenwerken en de uitdaging aan zijn gegaan om kennis en technologie te ontwikkelen die nodig zijn om relevante antwoorden op specifieke vragen in Nederlandstalige documenten te vinden.

Today

Als we in de speelgoedfolders voor de komende feestdagen kijken wordt het duidelijk dat de kinderen van vandaag taal- en spraaktechnologie (TST) letterlijk met de paplepel ingegoten krijgen. Van de sprekende pop, FIFI, tot de leukste educatieve taalspelletjes en de nieuwste LEGO®robot, die zelfs een "sound sensor" heeft om op de commando's van zijn jonge eigenaar te kunnen reageren: alles kan. Maar een insider zou u snel kunnen vertellen dat deze spelletjes alleen het topje van de ijsberg bevatten als het om TST gaat want voor de kinderen is dit nog maar het begin van hun reis met TST. Straks als het om hun spreekbeurt of huiswerk gaat hoeven ze de vragen maar in te typen om een compleet verhaal bij elkaar te krijgen – o.a. met behulp van wat uit de IMIX- en sommige van de STEVIN-projecten gaat komen. Dan begint TST pas echt interessant te worden...

Tomorrow

Zoals u in deze speciale editie van DIXIT zult lezen, bieden de indrukwekkende STEVIN-projecten unieke kansen voor onderzoekers en ontwikkelaars op het gebied van Nederlandstalige TST. De eerste resultaten beginnen nu vrij te komen en bieden veel voor de toekomst van TST. STEVIN loopt tot 2009 en u kunt zeker ervan zijn dat hieruit zeer boeiende nieuwe dingen zullen komen. DIXIT houdt u hiervan regelmatig op de hoogte.

Voor de sponsoring van dit eerste jaarboek van NOTaS zijn we STEVIN dankbaar en in het bijzonder onze gastredacteur, Peter Spyns van de Nederlandse Taalunie. Zijn frisse kijk op de STEVIN-projecten zorgt ervoor dat de inhoud en de relevantie van de projecten helder worden.

Verder is er ruimte in deze extra dikke editie van de DIXIT om een kijkje achter de schermen te geven van het reilen en zeilen binnen TST in Vlaanderen en Nederland en in het bijzonder bij de thema-bijeenkomsten, die georganiseerd worden door de groeiende brancheorganisatie NOTaS, (de Nederlandse Organisatie van Taal en Spraaktechnologie).

...En wat uw verlanglijstje betreft, is er gelukkig nog steeds de oude, vertrouwde Scrabble en ook nog Monopoly – ook al kent deze al het fenomeen betalen met je PIN!

Namens STEVIN, NOTaS en bijdragers aan dit DIXIT-jaarboek wens ik u veel leesplezier en vooral veel TST-inspiratie in 2007.

Debbie Kenyon-Jackson
voorzitter Stichting NoTaS

COLOFON

DIXIT: Tijdschrift over toegepaste taal- en spraaktechnologie - 4e jaargang, speciale editie: STEVIN DIXIT. DIXIT is een uitgave van Stichting NOTaS, Postbus 31070, 6503 CB NIJMEGEN. Tel. 024 - 352 88 88 - Fax 024 -354 00 90 - www.notas.nl **Redactie-adres:** Stichting NOTaS, Postbus 31070,6503 CB NIJMEGEN **Redactie:** Arjan van Hessen -hessen@cs.utwente.nl - Sylvia Hendriks - sylvia@malta-online.nl - Henk van den Heuvel - h.vandenheuvel@let.ru.nl - Lisanne Teunissen - lteunissen@taalunie.org - René Ouëndag - rene.ouendag@q-go.com **Gastredacteur:** Peter Spyns - pspyns@taalunie.org **Advertenties:** Stichting NOTaS - Sylvia Hendriks - sylvia@malta-online.nl - 024 352 88 88 **Abonnementen:** Voor een gratis abonnement dient u zich te wenden tot een van de NOTaS-deelnemers (zie www.notas.nl) **Opmaak en druk:** DesignPrint, Nijmegen **Verantwoording:** DIXIT is een uitgave van Stichting NOTaS.

Wiskunde en Taal

Soms is er een directe relatie tussen de studie die je volgt en de baan waar je uiteindelijk in terecht komt, maar dat verband hoeft niet erg direct te zijn. Veel van de oud-taalstudenten van mijn generatie (taaltechnologie bestond nog niet als studie) zijn in vakgebieden terechtgekomen waarin taalwetenschap noch literatuur een rol spelen. Ik ken wel wat wiskundigen die in de taaltechnologie zijn beland; daar hoor ik zelf bij. Iets heel anders, hoor ik u denken. Nee, hoor. Er zijn veel raakvlakken tussen taaltechnologie en wiskunde.

| THEO VAN DEN HEUVEL |

De eerste naïeve omschrijving van wiskunde die ik hoorde, toen ik nog op de basisschool zat, was: “wiskunde is net als rekenen, maar dan met lettertjes”. Daar lijkt de parallel met taaltechnologie al aardig in te zitten, ware het niet dat dit een uiterst beroerde definitie is. In feite gaat wiskunde over het bouwen van abstracte modellen en van de taal om die modellen in te beschrijven. Taaltechnologen doen dat ook, maar dan op het gebied van natuurlijke taal.

> “Kennelijk zijn er best veel wiskundige kanten aan taal.”

We vereenvoudigen de werkelijkheid nauwelijks als we zeggen dat er twee hoofdstromingen te onderscheiden zijn in de taaltechnologie: de ene werkt met regels en de andere met statistiek. Van de laatste kun je je wel voorstellen dat er wiskunde bij komt kijken, maar dat geldt ook voor de regelgestuurde benadering.
De taaltechnologen die bijvoorbeeld de grammatica van het Nederlands met formele regels beschrijven, zodat een computer die regels vervolgens kan toepassen op echte zinnen, zijn eigenlijk bezig de taal in een stelsel van vergelijkingen te vangen. De wiskundige gebruikt symbolen die aanduidingen zijn van (nog onbekende) getalswaarden, of van lijnstukken, matrices of andere wiskundige objecten. De taalkundige gebruikt symbolen als ‘NP’ en bedoelt daarmee een niet nader bepaalde zelfstandig-naamwoordgroep. En net zoals wiskundige vergelijkingen in de computer kunnen worden gevangen om oplossingen te genereren, kan een formele grammaticale beschrijving zinnen genereren of ten minste toetsen of die zinnen door de grammatica worden toegestaan.
Sommige taalkundigen werken met featurestructuren waarbij een feature een taalkundig kenmerk is. In de featurestructuren zijn eigenschappen als enkel- en meervoud, verleden en tegenwoordige tijd, maar ook semantische kenmerken zoals ‘menselijk’ en ‘vragend’ opgenomen met de relaties die daartussen kunnen gelden. Neem het zinnetje: “ze zingt”. Elk van de twee woorden heeft een aantal kenmerken. Zo is ‘ze’ een persoonlijk voornaamwoord, maar het kan op zichzelf genomen zowel enkelvoud (‘ze loopt’) als meervoud (‘ze lopen’) zijn. “Zingt” kan alleen enkelvoud zijn en daarom wordt de hele zin enkelvoud. De informatie per woord kan behoorlijk ingewikkeld worden, zodat het combineren verre van triviaal is. Het mechanisme dat we hier beschrijven is hetzelfde als in de wiskunde wordt gebruikt voor het vereenvoudigen van logische expressies.
Kennelijk zijn er best veel wiskundige kanten aan taal.
Eerder noemden we als alternatief voor de regelgestuurde aanpak in taaltechnologie de statistiek. De statistiek is de studie die helpt om best passende modellen te vinden bij experimentele data. Daarvoor bouwt de statistiek voort op kanstheoretische basis. Een mooi voorbeeld van taalstatistiek is de “wet van Zipf”. Dit is een experimentele waarneming (geen wiskundige ‘stelling’) over taal: als je woorden sorteert op frequentie, dan is hun rangnummer in die sortering ongeveer omgekeerd evenredig met de frequentie. Deze wet wordt ook vaak in de context van kennismanagement

toegepast. (Overigens blijkt bij nadere bestudering dat de wet van Zipf niet alleen tot natuurlijke taal beperkt is).

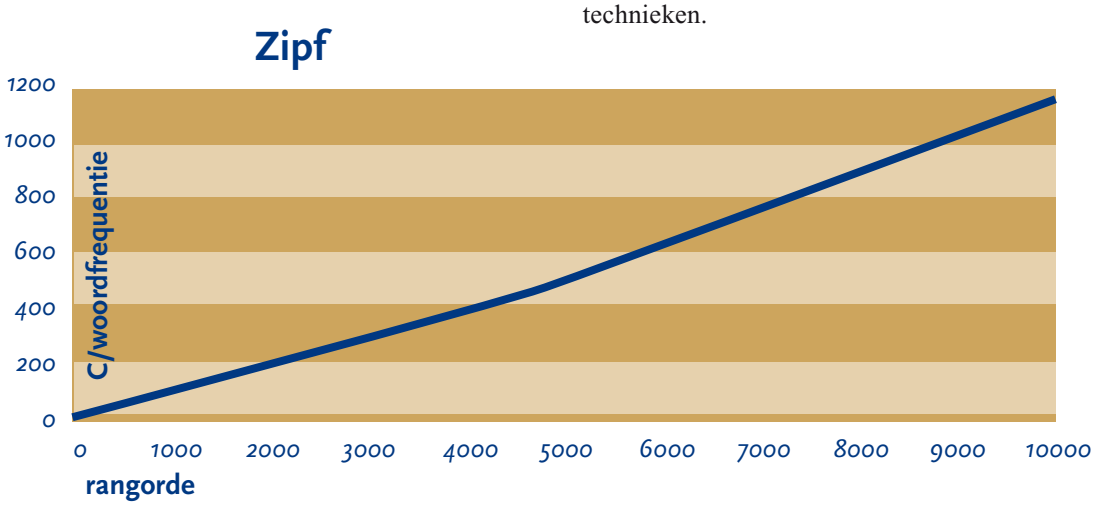
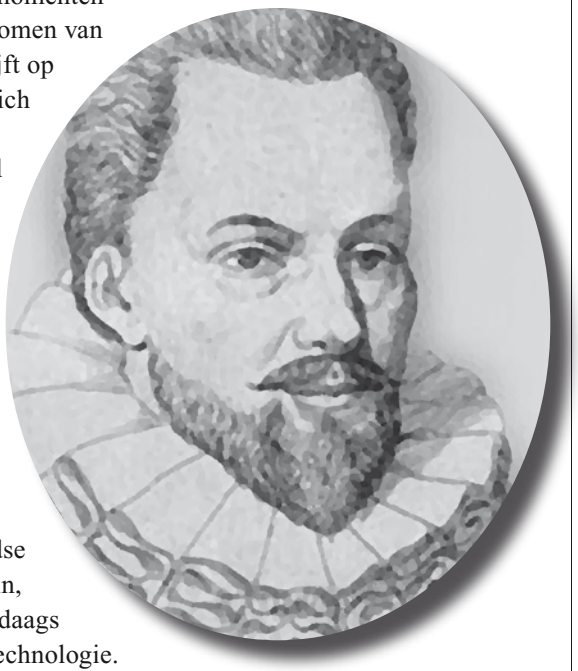


Fig. 1: wet van Zipf gebaseerd op de 10000 meest frequente woorden uit het Twente Nieuws Corpus. Op de horizontale as staat het rangnummer (meeste frequente woord (7389084) op 1, minst frequente woord op 5000. Verticaal staat een constante (=106) gedeeld door de woordfrequentie.

Laten we nog een eenvoudig voorbeeld geven van het gebruik van statistiek in taal. Neem een stuk Nederlandse tekst, bijvoorbeeld een bladzijde uit een boek. Tel nu alle a's, b's, en zo verder. Als de bladzijde representatief is voor onze taal dan levert dit experiment getallen op die je zou kunnen beschouwen als de kans op een 'a' in het Nederlands (namelijk het aantal a's gedeeld door het totaal aantal letters). Het woord 'kans' suggereert dat de auteur bij schrijven voortdurend met een dobbelsteen aan het gooien was. Iedereen snapt dat dat niet zo is, maar toch levert dit telwerk een rudimentair model voor het Nederlands op. De verhoudingen zullen voor de meeste Nederlandse teksten ongeveer hetzelfde zijn en anders zijn dan voor teksten uit het Duits of Engels. Ze vormen een soort vingerafdruk van de taal. We kunnen natuurlijk ook woordjes tellen in plaats van letters, maar dan kunnen we pas over verhoudingen gaan spreken als we een enorme hoeveelheid taalmateriaal tot onze beschikking hebben. Anders zullen veelgebruikte woorden als 'banaan' een grote kans hebben om niet in het overzicht voor te komen. Dit verklaart de voortdurende honger van taaltechnologen naar steeds grotere taalcorpora. De statistieken vormen een model van onze taal. Als de woorden in ons materiaal zijn gemarkeerd met woordklassen als 'zelfstandig naamwoord' of 'werkwoord', dan kunnen we nog meer statistieken uit onze teksten knijpen. Die modellen kunnen we gebruiken om voor nieuwe teksten de bestpassende analyses te berekenen. Programmatuur voor kennismanagement gebruikt dit soort technieken.

Dezelfde techniek wordt ook voor spraakherkenning gebruikt, maar bij het vertalen van spraaksignalen naar tekenreeksen of omgekeerd komt veel kijken dat op het terrein van techniek ligt. Cruciaal daarin zijn zogenaamde Fouriertransformaties: wiskundige bewerkingen die een signaal vertalen naar een spectrale weergave en omgekeerd. Anders gezegd, je wisselt tussen een weergave die de luchttrilling beschrijft voor een reeks van momenten en een weergave die het voorkomen van elke geluidsfrequentie beschrijft op diezelfde momenten. Kunt u zich de gekwelde discussies herinneren op de middelbare school over het nut van integreren, differentiëren, en complexe getallen? Hier heb je ze allemaal.
Alles bij elkaar blijkt de wiskunde een belangrijke bijdrage aan de taal- en spraaktechnologie te leveren. Een aardig historisch dwarsverband tussen wiskunde en taal is dat de laatste Nederlandse homo universalis, Simon Stevin, zijn naam geeft aan een hedendaags subsidieprogramma voor taaltechnologie. Dezelfde Stevin was een groot voorvechter van de Nederlandse taal en heeft ons allerlei Nederlandse vertalingen van Latijnse wiskundige termen nagelaten, waaronder het woord 'wiskunde' zelf.

THEO VAN DEN HEUVEL
IS DIRECTEUR VAN
POLDERLAND LANGUAGE &
SPEECH TECHNOLOGY
TE NIJMEGEN



Het Radio Oranje Project

"Het was op zondagavond 28 juli 1940 dat voor het eerst in de aether de woorden weerklonken: ´Hier Radio-Oranje´, zulks ter inleiding van een uitzending, waarin Hare Majesteit Koningin Wilhelmina haar eerste grote toespraak zou houden tot haar verdrukte volk."

| PETER VD MAAS, WILLEMIJN HEEREN EN ARJAN VAN HESSEN |

Tijdens de oorlog werden er vanuit Londen radio-uitzendingen verzorgd voor de Nederlanders in het door Duitsers bezette Nederland onder de naam Radio Oranje. Naast allerlei informatie over de voortgang van de oorlog werden ook toespraken van

hoogwaardigheidsbeklederes uitgezonden. Voorbeelden hiervan zijn toespraken van Minister-president Gerbrandy, Prins Bernard en natuurlijk Koningin Wilhelmina. Omdat de toespraken van deze hoogwaardigheidsbeklederes niet live voor de microfoon in de studio werden uitgesproken, maar vooraf werden opgenomen, is een deel van deze toespraken bewaard gebleven. De toespraken werden door de Engelse regering als "strategisch belangrijk" gezien en moesten vooraf aan de Engelse inlichtingendienst worden voorgelegd. Daardoor is ook de schriftelijke versie van de toespraken bewaard gebleven.
Het Radio Oranje Project, een samenwerking tussen de Universiteit Twente (oplijning), het NIOD (hosting en tekstdata) en Beeld en Geluid (geluidsbestanden), heeft als doel de schriftelijke en gesproken versie van de toespraken van Koningin Wilhelmina aan elkaar te koppelen zodat de geschreven tekst zo nauwkeurig mogelijk opgelijnd is met de gesproken tekst. Als deze oplijning eenmaal gedaan is, is het mogelijk om de toespraken van Wilhelmina te doorzoeken op woordniveau en de gevonden fragmenten direct af te luisteren. Bovendien is het mogelijk de tekst als ondertiteling te laten meelopen met het geluid. Hiervoor zal een web-interface worden ontwikkeld.

DATA

In totaal werden 32 toespraken opgelijnd, elk bestaande uit een gesproken en een geschreven document. De teksten, die in de oorlog in Londen op een typemachine waren uitgeschreven en waarvan soms hele regels met een rode stift door de Engelsen waren gecensureerd, werden na de oorlog in boekvorm gepubliceerd. Met behulp van OCR (Optical Character Recognition) werden deze pagina's omgezet in een digitaal bestand . Hoewel ook OCR fouten kan introduceren, bleek dat in minder dan 1% van de woorden het geval te zijn. De geluidsopnamen van de radiotoespraken zijn voor een groot gedeelte bewaard gebleven. Deze bestanden worden in het audioarchief van Nederlands instituut voor Beeld en Geluid gearchiveerd en werden recentelijk gedigitaliseerd.

OPLIJNEN

De tekst en het geluidsbestand van elke toespraak werden door de Twentse spraakherkenner opgelijnd. Aan de hand van de tekstuele versie bepaalt de computer hoe de zin moet klinken; de uitspraak van de woorden in de zin kan namelijk worden afgeleid met behulp van een "grafeem-naar-foneem" conversie. Daarin worden geschreven woorden

omgezet in de klanken waaruit ze zijn opgebouwd. Voor een aantal woorden dat in de oude spelling geschreven werd, zoals "mensch" of "landgenooten", klopte de grafeem-naar-foneem conversie niet en is een vertaling van oud naar nieuw Nederlands geïntroduceerd. Tijdens de oplijning werd de serie klanken op basis van de tekst gekoppeld aan de audio. Hiervoor is gebruik gemaakt van akoestische modellen die waren getraind op de stem van Koningin Wilhelmina.
Bij de oplijning van tekst met geluidssignaal geldt voor mooie, ruisvrije spraak dat de overeenkomst tussen de audio en de klankrepresentatie van de geschreven tekst zo groot is, dat het relatief eenvoudig is de woorden er precies onder te krijgen. Bij de spraak van Wilhelmina is dat anders. De opname-apparatuur van die tijd was primitief hetgeen resulteert in sterke ruis en tikken in de opname. Bovendien zijn de wasplaten waarop de spraak werd opgenomen in de loop der jaren verhard, waardoor de kwaliteit sterk achteruit gegaan is.
Het bleek niet mogelijk om een gemiddeld 10 minuten durende toespraak van Wilhelmina in één keer op te lijnen. De tekst werd daarom handmatig in fragmenten opgedeeld, en per fragment opgelijnd. Op deze manier was het begin van ieder fragment in ieder geval juist en nam de kans op fouten enorm af.

INFORMATIEBRONNEN TOEVOEGEN

Nu voor elke toespraak de audio en tekst zijn opgelijnd, willen we er automatisch extra informatie aan toevoegen. Deze extra informatie zal bestaan uit relevante foto´s die tijdens het afspelen van de audio getoond zullen worden en zo de ervaring van de luisteraar kunnen vergroten. De tekstuele inhoud van de zinnen wordt gebruikt om in het fotoarchief van het NIOD foto's te zoeken die qua omschrijving sterk op de zinnen lijken. Daarnaast wordt de uitzenddatum van de toespraken gebruikt om precies díe foto´s te vinden die uit dezelfde periode komen als de geluidsbestanden. Op deze manier worden er bij elke zin één of meerdere foto's gezocht die tijdens het afspelen van die zin zullen worden getoond.

CONCLUSIE

Hoewel oplijnen van oude en zeer ruizige teksten nog niet zó eenvoudig is, dat in één keer lange bestanden succesvol gedaan kunnen worden, is het resultaat na handmatige segmentatie toch zeer veelbelovend. Doordat er nu direct naar woorden en woordcombinaties in de audio gezocht kan worden, is het eenvoudig om dit soort audio-archieven via Internet voor iedereen toegankelijk te maken. In plaats van het afluisteren van talloze bestanden om net dat ene onderwerp te horen, kan er nu "op z'n Googles" gezocht worden en kan het resultaat eenvoudig via Internet worden afgespeeld.
De combinatie van beeld (de foto's), geluid (de radio-opnamen) en tekst (de transcripties) wordt op deze manier geheel automatisch gebundeld tot een multimedia-presentatie. Het NIOD is hier zo enthousiast over, dat ze besloten hebben de gehele presentatie binnen het thema "Wilhelmina in de oorlog" te hosten: vanaf volgend jaar is de presentatie dan via de website van het NIOD (www.niod.nl) te bekijken.

> "In plaats van het afluisteren van talloze bestanden, kan er nu 'op z'n Googles' gezocht worden."

PETER VD MAAS IS VERBONDEN AAN DE NIOD, WILLEMIJN HEEREN IS VERBONDEN AAN DE UNIVERSITEIT TWENTE / CHORAL PROJECT, ARJAN VAN HESSEN IS VERBONDEN AAN DE UNIVERSITEIT VAN TWENTE EN TELECATS

Benut de waarde van woorden

Basismateriaal: taal

Geschreven taal is een van de krachtigste uitingsvormen van bedrijfscommunicatie. Zwart op wit vormen woorden een blijvende representatie van een onderneming. Hoe benut een organisatie de kracht van tekst?

| DOOR ANNEBETH LASSEUR |

Lezen is een reflex. Ogen die letters en woorden zien, lezen onwillekeurig. Hierdoor vergeten we soms hoeveel tekst we tegenkomen op een dag. Op straat en thuis, maar vooral in zakelijke omgeving. Bedrijven zijn een continue stroom van geschreven communicatie: van e-mail aan een leverancier of een briefje over de personeelsfuif tot en met jaarverslag en website. Die laatste teksten krijgen vaak veel aandacht: ze vormen een visitekaartje van de organisatie en ondersteunen de uitstraling, het imago van het bedrijf. Aan zulke merkimago's wordt zorgvuldig gebouwd, door middel van (marketing)communicatie. Maar veel communicatie onttrekt zich aan het 'toezicht' van de afdeling marketing, communicatie of PR. Correspondentie, offertes, documentatie en nota's schrijven individuele medewerkers zelf. Vaak wel *professionals*, maar niet op het gebied van tekst en communicatie. Toch vormen deze schriftelijke uitingen evengoed visitekaartjes. Het gevaar is dat ze de zorgvuldig opgebouwde merkbeleving tenietdoen, door een heel eigen koers te varen in toon en stijl. Of, nog erger, doordat ze onbegrijpelijk zijn.

"~~Veel~~ **zinnen kunnen** ~~over het algemeen aanzienlijk~~ **korter** ~~worden opgeschreven~~."

TAAL ALS INSTRUMENT

Sabel Communicatie is een bureau voor geschreven communicatie. We maken onderscheidende communicatie-uitingen en adviseren over taal als imago-instrument. Bij dat advies kijken we ook naar teksten die een bedrijf niet doelbewust inzet voor het eigen imago. We helpen organisaties en individuele medewerkers om de juiste woorden te kiezen. We vertalen hun kernwaarden naar kernwoorden, zodat alle communicatie-uitingen het gewenste imago versterken. We doen dat voor organisaties die veel te vertellen hebben. Letterlijk: ze hebben veel boodschappen te communiceren. En figuurlijk: ze spelen een prominente rol in de samenleving. Sabel Communicatie ondersteunt hen daarbij en ontwikkelt middelen om taal en tekst zo effectief mogelijk in te zetten.

IMPACT

Taal speelt een zeer belangrijke rol in het ordenen van ervaringen. Dat hangt soms af van kleine woordjes. De manager die zijn collega's wil aanmoedigen met 'we moeten er tegenaan', bouwt onbewust een blokkade in met 'moeten'. Waarom niet: 'we gaan er tegenaan'? En soms zit het niet in woordkeuze, maar juist in het weglaten ervan. ~~Veel~~ **zinnen kunnen** ~~over het algemeen aanzienlijk~~ **korter** ~~worden opgeschreven~~. Hoe beknopter geformuleerd, hoe groter de impact. Minder is meer. Dat geldt ook voor jargon. Vakspecialisten ontlenen er graag status aan. Onderzoek toont echter aan dat ook onder vakgenoten klare taal de voorkeur verdient. In 'gewonemensentaal' komt de boodschap beter over en blijkt het aanzien van de zender juist toe te nemen.

KRACHT VAN TAAL

Taaltechnologie kan helpen bij het benutten van de kracht van taal. Om te beginnen valt er al veel te winnen met consistentie: als een organisatie consequent is in taalgebruik, versterkt dat het beeld van één afzender. En met die duidelijkheid en herkenbaarheid zijn niet alleen commerciële merken gebaat, maar ook non-profit en overheidsinstellingen. Sabel Communicatie ondersteunt daarom organisaties bij het schrijven volgens één afgesproken stijl. Hiervoor ontwikkelen we een 'teksthuisstijl' voor het bedrijf. Dit is een vertaling van de communicatiedoelstellingen, het imago en de gewenste merkbeleving van het bedrijf in de tekstkenmerken woordgebruik, stijl en spelling. Afhankelijk van de organisatie kiest de teksthuisstijl niet *geachte* maar *beste*, niet voor het *aanwenden van financiële middelen* maar voor *geld uitgeven*. Deze afspraken kunnen goed worden ondersteund met taaltechnologie. Naast het bewaken van consequente communicatie, helpt Sabel Communicatie medewerkers om toegankelijk, communicatief Nederlands te schrijven. Ook daar kan taaltechnologie een rol spelen. Bijvoorbeeld door omslachtige (passieve) constructies en omhaal van woorden te signaleren. Of alternatieven aan te dragen voor jargon, holle kreten en archaïsche formuleringen.

TEKSTANALYSE

In samenwerking met Polderland Language & Speech Technology ontwikkelde Sabel Communicatie een tool om teksten te analyseren. Veel tekstkenmerken en afspraken uit de teksthuisstijl komen daarin al aan bod. Met de taalwetenschappelijke kennis en journalistieke ervaring die Sabel Communicatie in huis heeft, ontwikkelen we steeds meer mogelijkheden om kenmerken van teksten te analyseren en te verbeteren. Nog niet alle tekstkenmerken zijn eenvoudig te definiëren. Woordvariatie bijvoorbeeld: schrijvers zijn geholpen met een goede synoniemenlijst als ze één woord vaak herhalen. Maar variatie kan ook verwarring oproepen. De kunst is dus om parameters voor woordvariatie te vinden die de leesbaarheid daadwerkelijk verbeteren, en niet gebruik van synoniemen tot doel op zich verheffen. De relatie tussen zinslengte en leesbaarheid is ook al goed gedocumenteerd. Korte zinnen lezen makkelijker. Maar een tekst is niet gebaat bij enkel korte zinnen of een 'correcte' gemiddelde zinslengte. Wel is het handig als de schrijver gewezen wordt op één zin van vijftig woorden, zodat hij die kan opbreken.

EIGEN WERK

Schrijven is heel persoonlijk. Het gaat om het verwoorden van gedachten van de schrijver. Naast de doelen die een organisatie wil bereiken met (geschreven) communicatie, hebben medewerkers vaak ook een uitgesproken mening over wat mooi is of duidelijk. Het doel van de 'elektronische eindredacteur' is dan ook niet om iedereen precies dezelfde teksten te laten schrijven. Mensen houden hun eigen voorkeuren voor bepaalde woorden of constructies en niet elke suggestie wordt op prijs gesteld. Wel kan taaltechnologie helpen om een globale schrijfstijl af te bakenen, om teksten snel te analyseren en om de leesbaarheid te verbeteren. En juist omdat een tekst zo persoonlijk is, leggen de meeste mensen hun eigen werk niet graag voor aan een collega. Terwijl ze wel behoefte hebben aan feedback. Dan voegt taaltechnologie meer toe dan een uitgebreide spellingcontrole: iedere schrijver wordt over de drempel geholpen om eigen schrijfwerk kritisch te bekijken en te verbeteren.

> "Het is de kunst om parameters te vinden voor woordvariatie te vinden die de leesbaarheid daadwerkelijk verbeteren."

ANNEBETH LASSEUR IS TEKSTSCHRIJVER EN PROJECTMANAGER BIJ SABEL COMMUNICATIE IN UTRECHT.

Speed dating: NOTaS STEVIN

Op 20 juni 2006 organiseerde NOTaS voor alle geïnteresseerden een speed-date event in de Volksuniversiteit in Utrecht.

| HANS JONGEBLOED |

Het doel van dit evenement was om aanvragen voor STEVIN-demonstratieprojecten te stimuleren. Ook konden taal- en spraak-technologiepartijen kennismaken met partijen die van deze technologie gebruik willen maken en omgekeerd. Partijen konden ideeën aandragen voor mogelijke samenwerkings-projecten en aangeven met welke andere partijen ze in contact wilden komen. Voor de STEVIN-demonstratieprojecten is een van de voorwaarden dat minimaal een taal- en/of spraaktechnologiebedrijf in het consortium zit.

> "NOTaS heeft weer nieuwe deelnemers mogen verwelkomen."

Binnen NOTaS zijn al deze bedrijven vertegenwoordigd, zodat men makkelijk contact kon leggen met mogelijke partners! Om te zorgen dat er tijdens alle rondes genetwerkt kon worden, werden er ook "blind dates" gepland, zodat er in totaal tegen de 100 gesprekjes plaatsvonden.

In totaal waren er ongeveer twintig partijen aanwezig, waaronder ook een aantal bedrijven uit Vlaanderen, om zo ook de Vlaamse commerciële inbreng in STEVIN te stimuleren.

HANS JONGEBLOED IS BESTUURSLID STICHTING NOTAS

Hier volgen een aantal quotes van de deelnemers:

Alice Dijkstra (NWO) – "Een opwindende sessie, zou ik zeggen, met oude en nieuwe mensen, spannende plannen, stimulerende interactiemogelijkheden en daadwerkelijke interacties."

Sander de Graaf (Dutchear) – "Goed om kennis te maken met partijen die op zoek zijn naar wat met de huidige technologie allemaal kan. Bovendien is het gelukt om een projectidee verder op te harden. De vervolgafspraak is al gepland!"

PETER PYNS

Remco van Veenendaal (TST centrale) – "Leuke projectideeën. Ik ben heel tevreden met wat ik heb bereikt. Ook voor mijn collega was het nuttig, samen hebben we nog dubbel zo veel gehoord."

Lisanne Teunissen (Nederlandse Taalunie) – "Een hele nuttige bijeenkomst! Ook voor mij weer nieuwe kennismakingen en veel contacten. Dit was makelen en schakelen pur sang."

Uiteindelijk heeft de dag veel nieuwe kennismakingen en minimaal twee ingediende voorstellen voor demonstratieprojecten en twee nieuwe samenwerkingsprojecten opgeleverd. Ook heeft NOTaS weer een aantal nieuwe deelnemers mogen verwelkomen.

Spraak- en **T**aaltechnologische **E**ssentiële **V**oorzieningen **I**n het **N**ederlands

Het STEVIN-onderzoeks- en stimuleringsprogramma

STEVIN: weer zo'n vergezocht letterwoord, dat een vermoeide onderzoeker op een nachtelijk uur bedenkt om zijn projectvoorstel de volgende dag wat sexyer te laten klinken bij de beoordeling? Of zit er meer achter? Zestiende-eeuwse zeilkarren gebouwd door Simon Stevin lijken niet direct verband te houden met moderne taal- en spraak-technologie (TST) voor het Nederlands. Of toch?

| PETER SPYNS |

Simon Stevin heeft meer verwezenlijkt dan zijn beroemde zeilwagentocht op het strand van Scheveningen. Hij was een Vlaming met een grote kennis van bouwkunde en toegepaste wiskunde. Toen hij uitweek naar Nederland, werd hij de leermeester en later vertrouweling van prins Maurits van Nassau. Het liefst loste hij praktische problemen op met behulp van de wetenschap.
Daarbij gebruikte hij zo veel mogelijk Nederlandse termen. Zo hebben we bijvoorbeeld de woorden 'evenwijdig', 'omtrek' en 'stelling' aan hem te danken.
STEVIN is een gezamenlijk onderzoeks- en ontwikkelingsprogramma van Vlaanderen en Nederland. Wetenschappers leveren basismaterialen die nodig zijn voor TST-toepassingen voor het Nederlands. Het programma draagt de naam STEVIN dus met recht en rede.

Simon Stevin (Brugge, 1548 - Den Haag of Leiden, 1620)

Het programma moet de versnippering tegengaan van TST-initiatieven die voorheen in Vlaanderen en Nederland los van elkaar opgezet werden. Het is aldus een voorloper van de Europese Onderzoeksruimte (European Research Area, ERA). Ook binnen ERA wordt onderzoek in grensover-schrijdende programma's gezamenlijk gefinancierd en beheerd.

Er zijn grote investeringen nodig om voor een taal het basismateriaal voor TST te ontwikkelen. Vanuit een marktgedreven benadering zijn die investeringen voor een kleine taal als het Nederlands niet haalbaar. In cultuur-politiek opzicht speelt het "elektronische Gütenbergeffect": zonder moderne TST zouden talen op termijn verdwijnen in het tijdperk van high tech ICT.

Voldoende redenen voor de Vlaamse en Nederlandse overheid om TST-onderzoek voor het Nederlands financieel te steunen. De resultaten komen laagdrempelig beschikbaar voor onderzoekers én industrie, zodat ze tot innovatieve producten en diensten kunnen leiden en zo de economie stimuleren. Het laatste aspect is trouwens een algemeen pijnpunt. De Europese Unie lijdt namelijk aan een zogeheten innovatiekloof: de industrie pikt de resultaten van hoog-staand academisch onderzoek niet voldoende op.

Dit themanummer van DIXIT geeft een overzicht van de projecten die STEVIN momenteel financiert. De redactie hoopt dat het veel lezers op ideeën brengt voor voorstellen en (transnationale) samenwerking tussen industrie en kennisinstellingen. Want in 2007 komen er nog twee oproepen voor projecten.

Het is nu wachten op een Vlaams-Nederlands team dat onderzoek wil verrrichten naar robuuste spraakherkenning in een winderige omgeving, en de resultaten toepast in een moderne zeilwagen die met de stem bediend wordt. Als zulke zeilwagens in 2020 over het Scheveningse strand zoeven, krijgt het STEVIN-programma nog meer raakvlakken met zijn naamgever.

Ik wens u veel leesplezier en vruchtbare inspiratie met dit themanummer.

PETER SPYNS IS STEVIN-PROGRAMMACOÖRDINATOR BIJ DE NEDERLANDSE TAALUNIE VANUIT DE VLAAMSE OVERHEID

Het STEVIN-pogramma loopt van 2004 tot 2009. Zie: http://www.taalunieversum.org/stevin.

Beschikbaarheid van resultaten via de TST-centrale

De IPR-regeling van STEVIN

Het STEVIN-programma beoogt een stimulans te geven aan de Nederlandstalige taal- en spraaktechnologie (TST). Bedoeling is om de innovatiecapaciteit van de sector te vergroten en tegelijk de positie van het Nederlands in de moderne informatie- en communicatiewereld te behouden. *Maximale beschikbaarheid van de resultaten van onderzoek en ontwikkeling* op het gebied van TST is een absolute vereiste.

| LINDE VAN DEN BOSCH |

Het TST-bestuur[i] streeft ernaar de STEVIN-projectresultaten
1. *maximaal*,
2. zo *goedkoop* mogelijk,
3. en waar onvermijdelijk, tegen *marktconforme* prijzen beschikbaar te stellen, waarbij verder
4. een *speciale regeling* wordt getroffen voor geprivilegieerde gebruikers, te weten de STEVIN-*projectpartners en leveranciers van materialen*.

Punt (1) wordt het best bereikt door de resultaten op te nemen in een centrale verzamelplaats, onderhouden door de TST-centrale. Daarmee is er één plaats voor alle resultaten, wat het vinden en verkrijgen van beschikbare data voor iedere potentiële gebruiker makkelijker maakt. In speciale gevallen kunnen, na overleg met de TST-centrale, de resultaten fysiek elders worden ondergebracht. Wel moeten ze dan bereikbaar blijven via de TST-centrale.

Punt (2) is noodzakelijk om de barrières voor het gebruik van de resultaten zo beperkt mogelijk te houden. De huidige prijslijst van de bij de TST-centrale beschikbare data en tools reflecteert dit streven. In veel gevallen zijn de kosten voor het verkrijgen van de data beperkt tot de kosten van de datadragers en de verzending ervan.

Punt (3) is noodzakelijk om verstoringen van de markt te voorkomen. Het betreft vooral data en tools die al eerder commercieel verhandeld werden (vóór een STEVIN-project er gebruik van maakt), of voor afgeleiden hiervan.

De STEVIN-projectpartners die de projectresultaten gecreëerd hebben, krijgen hierop een ruime licentie **(punt 4)**, waarbij echter verstoring van de markt vermeden moet worden. Zij hebben het recht “hun” resultaten waar mogelijk kosteloos te gebruiken voor verder onderzoek, en kunnen er ook onbeperkt verdere ontwikkeling op doen.

De STEVIN-IPR-werkgroep heeft voor ieder soort contract dat nodig is tussen de verschillende spelers, een model opgesteld om de IPR-kwesties adequaat te regelen. Deze modelcontracten zijn te vinden op de websites van STEVIN en van de TST-centrale (http://taalunieversum.org/stevin/projectuitvoerders/ en http://ww2.tst.inl.nl/index.php?option=content&task=view&id=389).

Het TST-bestuur hoopt dat hierdoor zoveel mogelijk barrières weggenomen worden voor een optimaal functioneren van de STEVIN-projecten. De Nederlandstalige TST moet de beoogde impuls kunnen krijgen, zodat de resultaten leiden tot nieuwe en innovatieve projecten, producten en diensten.

LINDE VAN DEN BOSCH IS ALGEMEEN SECRETARIS VAN DE NEDERLANDSE TAALUNIE

nu | STEVIN

[i] Het TST-bestuur bestaat uit vertegenwoordigers van de financierende organisaties: de Vlaamse Overheid - departement Economie, Wetenschap en Innovatie, het IWT-Vlaanderen, het FWO-Vlaanderen, het Ministerie van Economische Zaken, het Ministerie van Onderwijs, Cultuur en Wetenschap, NWO en de Nederlandse Taalunie, aangevuld met enkele experts met raadgevende stem. Voorzitter is Linde van den Bosch (Taalunie). Tevens nemen de voorzitter van de STEVIN-programmacommissie en de programmacoördinator deel aan de bestuursvergaderingen, evenals het programma-bureau (een samenwerking tussen NWO en Senter-Novem).

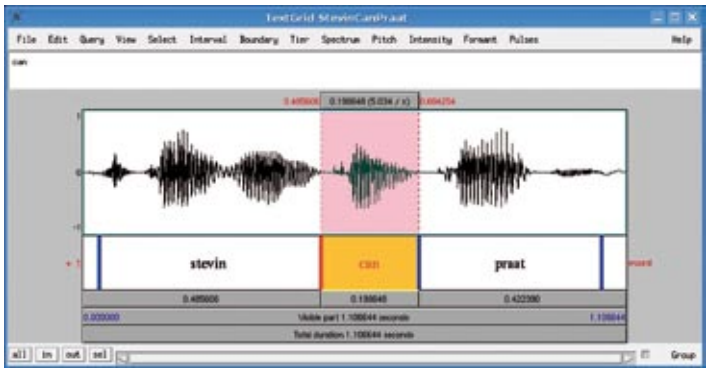
STEVIN can PRAAT

Voor wetenschappers die zich met gesproken taal bezig houden is een hulpmiddel om het spraaksignaal te visualiseren en te analyseren essentieel.

| DAVID WEENINK |

PRAAT NU …

Het computerprogramma PRAAT dat ontwikkeld is aan de Universiteit van Amsterdam door Paul Boersma en David Weenink, wordt wereldwijd door meer dan 10.000 wetenschappers gebruikt om geluiden te analyseren die door mensen én dieren kunnen zijn gemaakt. Deze geluiden kunnen variëren van babygeluidjes tot walvisgeluiden.



de werkomgeving van PRAAT

De populariteit van het programma PRAAT is onder andere verklaarbaar door zijn vele analysemogelijkheden, door zijn mogelijkheid om analyses te automatiseren via scripts, door zijn gebruikersvriendelijkheid, door de snelheid waarmee de ontwikkelaars reageren op wensen, door zijn actieve gebruikersgemeenchap die onderling kennis uitwisselt via het internet, en doordat het programma gratis is. Het programma (zie figuur) werkt op Windows-, Macintosh- en Linux-computers.

> “Deze geluiden kunnen variëren van babygeluidjes tot walvisgeluiden.”

Ondanks de vele mogelijkheden die al in PRAAT zitten ontbreekt nog gewenste functionaliteit. Het project “STEVIN can PRAAT” beoogt deze nieuwe functionaliteit toe te voegen.

DAVID WEENINK
COÖRDINEERT HET
STEVINcanPRAAT-PROJECT

… PRAAT DAN

In het project worden twee direct te gebruiken uitbreidingen aan PRAAT gemaakt.

Grafische manipulatie van formant-frequenties en bandbreedtes.
Voor het testen van spraakperceptiemodellen is een nauwkeurige specificatie van signalen nodig. Met behulp van een grafische editor kunnen we van een te maken geluid voor elke formant de frequentie en bandbreedte als functie van de tijd nauwkeurig specificeren. Met behulp van de editor kunnen we dan óf een helemaal nieuw signaal maken óf de klanken in een reeds bestaand signaal veranderen. Het uiteindelijke maken (synthetiseren) van de geluiden kan dan bijvoorbeeld gebeuren door een referentiesynthesizer, de Klatt-synthesizer. Deze synthesizer zal ook door ons gemaakt worden.

PAUL BOERSMA

Klinkerachtige signalen maken via muisbewegingen.
Door de horizontale en verticale positie van de muiswijzer op het beeldscherm te laten corresponderen met de eerste en tweede formant-frequentie kunnen we het geluidje maken dat hierbij hoort. Door muisbewegingen kunnen we dan veranderende klinkerachtige geluidjes maken. Dit kan voor onderwijsdoeleinden erg handig zijn.

Verdere uitbreidingen behelzen het zoeken en vervangen met behulp van reguliere expressies, het toevoegen van meer wiskundige functies binnen de scriptomgeving, en het implementeren van een nauwkeuriger algoritme voor formant-frequentiemetingen.

Al met al genoeg nieuwe dingen waar de gebruiker van PRAAT zich op kan verheugen.

STEVINcanPRAAT is een project uit de tweede ronde. Het project loopt tot eind april 2008. Zie: http://www.praat.org/.

N-Best

(Nederlandse Benchmark Evaluatie van Spraakherkenner-Technologie)

Bij de ontwikkeling van een automatisch systeem voor het herkennen van spraak is het belangrijk te weten hoe goed zo'n systeem presteert.

In een typische ontwikkelslag wordt iets aan de herkenner veranderd, bijvoorbeeld een algoritme of de gebruikte modellen, en vervolgens wordt bepaald of de wijziging een positief effect heeft op de prestatie.Om de prestatie te kunnen meten is een evaluatie nodig. Zo'n evaluatie bestaat uit het laten herkennen van een verzameling spraak, het evaluatie-materiaal, waarna het resultaat vergeleken wordt met wat er werkelijk gezegd is, de referentie.

Bij de ontwikkeling van een spraakherkenner is nog een andere verzameling spraakmateriaal nodig, voor de training van de akoestische modellen. Ook dit materiaal moet vergezeld zijn van een letterlijke beschrijving van wat er gezegd wordt.

De hoeveelheid spraakmateriaal voor een training is vaak veel groter dan wat nodig is voor evaluatie. Daarom wordt het evaluatiemateriaal meestal 'geknipt' uit het trainingsmateriaal. Dit is een goede methode voor de ontwikkeling van een systeem, maar *het is gevaarlijk om prestaties op een deel van het trainingsmateriaal te interpreteren als representatief voor het werkelijke gebruik.*
Om echt te bepalen wat de prestaties zullen zijn tijdens het gebruik van een herkenner, moet met nieuw materiaal getest worden. Het is daarom gebruikelijk om regelmatig nieuw evaluatiemateriaal te verzamelen en dit ter beschikking te stellen aan onderzoekers. Omdat het samenstellen van een complete evaluatie relatief arbeidsintensief is, wordt deze vaak centraal geleid.

Een grote kracht achter de evaluaties is het NIST in de VS, dat evaluaties organiseert in talen als het Engels, Chinees en Arabisch. Deze NIST-evaluaties vinden regelmatig plaats en onderzoekers zijn er erg vertrouwd mee.

Het Nederlands ontbeert tot nog toe deze evaluaties. Het project N-Best heeft als doel deze lacune op te vullen. Het project heeft partijen in verschillende rollen. TNO (Soesterberg) is als evaluator verantwoordelijk voor het houden van de evaluatie. SPEX (Nijmegen) verzamelt het audio-materiaal en voorziet deze van de referentie-transcripties. Van vijf universitaire onderzoeksgroepen worden de spraakherkennersystemen geëvalueerd: ESAT (Leuven) , ELIS (Gent), CLST (Nijmegen), HMI (Twente) en EWI (Delft). De evaluatie is ook open voor onderzoeksgroepen buiten het Nederlandse taalgebied.

N-Best legt protocollen vast en definieert de taak, evaluatiematen en procedures, zodat de resultaten van toekomstige evaluaties vergeleken kunnen worden met deze. In N-Best zijn de spraakherkenningsdomeinen nieuwsuitzendingen (radio/tv) en conversationele telefoonspraak.

DAVID VAN LEEUWEN
COÖRDINEERT HET
N-BEST-PROJECT

N-Best is een project uit de tweede ronde. Het project loopt tot eind september 2008. Zie: http://speech.tm.tno.nl/n-best/.

SPRAAK
(Speech Processing, Recognition and Automatic Annotation Kit)

Zowel voor onderzoekers in het domein van de taal- en spraaktechnologie, als voor de gebruikers van deze technologie is de beschikbaarheid van een spraakherkenningssysteem voor het Nederlands essentieel. Momenteel beschikken heel wat onderzoekers en ontwikkelaars niet over goede software voor spraakherkenning, of gebruiken zij systemen die tekort schieten op het vlak van functionaliteit of flexibiliteit.

| PATRICK WAMBACQ |

In het project SPRAAK worden in die context twee doelstellingen nagestreefd, die in eenzelfde software-raamwerk worden aangepakt.

Een eerste doelstelling is de ontwikkeling van een zeer modulaire toolkit, bedoeld voor onderzoek naar algoritmes voor spraakherkenning. Met de toolkit kunnen onderzoekers zich concentreren op bepaalde aspecten van spraakherkenning, zonder dat zij zich moeten bezighouden met andere componenten.
Daarom is modulariteit en een plug-and-play architectuur uiteraard zeer belangrijk.

> “Gebruikers kunnen onmiddelijk van start gaan met het systeem.”

De tweede doelstelling is het ter beschikking stellen van een state-of-the-art herkenner voor het Nederlands met een eenvoudige interface, die ook niet-specialisten kunnen gebruiken.

PATRICK WAMBACQ
COÖRDINEERT HET
SPRAAK-PROJECT

Het systeem is bedoeld voor de herkenning van continue spraak met een grote vocabulaire, maar is ook bruikbaar voor gerelateerde taken. Zo kan het gebruikt worden om segmentaties te maken van grote spraakdatabanken of om transcripties op te stellen van hoge kwaliteit. Andere mogelijke toepassingen zijn keyword spotting voor indexering van archieven of voor call centers, het aanleren en trainen van juiste uitspraak, enzovoort.

De basis voor SPRAAK is het bestaande herkenningssyteem van ESAT (K.U.Leuven), met toegevoegde kennis en software van de andere project-partners (Radboud Universiteit Nijmegen, Universiteit van Twente en TNO).
Er worden gebruikersinterfaces gemaakt om de flexibiliteit en gebruiksvriendelijkheid te verbeteren. De interfaces geven onderzoekers toegang tot de volledige interne werking van de herkenner op een laag niveau. Applicatiebouwers krijgen toegang tot de meest gangbare taken via een eenvoudige interface.

Goede documentatie is uiteraard van groot belang en zal worden geschreven voor zowel applicatie-ontwerpers als onderzoekers. In het Engels, zodat het systeem ook bruikbaar is voor de internationale onderzoeksgemeenschap.

Daarnaast worden ook twee voorbeeldherkenners gemaakt, voor Noord- en Zuid-Nederlands, en voor breedband- en telefoonspraak. Daardoor kunnen gebruikers onmiddellijk van start gaan met het afgewerkte systeem. Zij hebben uiteraard ook de mogelijkheid om met de software zelf een herkenner te bouwen, door de juiste modellen te trainen met eigen audio- en tekstmateriaal.

SPRAAK is een project uit de eerste tenderoproep. Het project loopt tot eind februari 2008. Zie: http://www.esat.kuleuven.be/psi/spraak/projects/index.php?proj=SPRAAK.

MIDAS
(MIssing DAta Solutions)

Spraakherkenning is een technologie die steeds meer toegang vindt tot de markt en waarmee verbluffende producten gemaakt kunnen worden, bijvoorbeeld directory assistance, dicteersoftware, audio mining enzovoort. Een voorname reden waarom de technologie moeilijk “out-of-the-box” inzetbaar is, is het gebrek aan robuustheid.

| HUGO VAN HAMME |

Dit gebrek aan robuustheid wordt door heel veel verschillende factoren veroorzaakt: dialectische spraak, spreeksnelheid, slordige uitspraak, sprekervariatie, al dan niet pathologische stemafwijkingen, de eindigheid van de woordenschat, haperingen, hernemingen, versprekingen, en “out-of-domain”-spraak.
Een andere heel belangrijke factor is achtergrondlawaai, in het bijzonder wanneer de storing snel verandert in de tijd.

> “Een belangrijke factor is achtergrondlawaai, bijvoorbeeld het geroezemoes in een restaurant.”

Bijvoorbeeld achtergrondmuziek, het geroezemoes in een restaurant, of het lawaai in een rijdende trein. Het is met name de gevoeligheid van automatische spraakherkenning voor achtergrondlawaai waarvoor in MIDAS een oplossing gezocht zal worden .

HUGO VAN HAMME
COÖRDINEERT HET
MIDAS-PROJECT

De aanpak die we in MIDAS voorstaan is gebaseerd op missing data theory. Wanneer spraak op een bepaald tijdstip en bij een bepaalde frequentie wordt overstemd door achtergrondlawaai, dan erkennen we dat die informatie onherroepelijk verloren is gegaan. Omdat de spraak overstemd is, weet je enkel dat de spraak in dat tijd-frequentiegebiedje een lagere energie moet hebben gehad dan de observatie. Om precies aan te geven wanneer en bij welke frequenties de spraakinformatie ontbreekt en wanneer ze betrouwbaar is, stellen we een “masker” op.

Spraakhypothesen die compatibel zijn met dit masker (in die zin dat ze uitgaan van een spraakenergie die lager is dan de observatie wanneer de data ontbreken) worden door een ingreep in het akoestische model van de herkenner niet langer afgestraft. Net als mensen kan de herkenner nu omgaan met ontbrekende data, die als het ware worden ingevuld door “hogere” informatiebronnen, zoals lexicon en grammatica.

Zolang je niet weet wat het spraaksignaal geweest is, is het verre van triviaal hoe je kunt vaststellen welk deel van de geluidenergie aan het spraaksignaal moet worden toegeschreven en welk deel aan achtergrondlawaai. Toch is dat precies wat er gevraagd wordt wanneer je een masker moet opstellen.

Het begin van een oplossing kan gevonden worden door de hele gehoorscène te analyseren. Door kenmerken als harmoniciteit, gelijktijdigheid en regelmaat te gebruiken wordt het in principe mogelijk bepaalde tijd-frequentiegebiedjes aan voor- of achtergrond toe te kennen. In de literatuur vind je echter geen eensluidend antwoord over hoe dergelijke kennis in de schatting van het masker moet verwerkt worden. Voor de backend hebben de voornaamste onderzoeksvragen te maken met de manier waarop het akoestische model moet gemodificeerd worden om het masker accuraat en snel in rekening te brengen.

MIDAS is een samenwerking tussen Nuance, Radboud Universiteit Nijmegen en de K.U.Leuven. Nuance staat voornamelijk in voor de definitie van test- en trainingsmateriaal en voor een vergelijking met een state-of-the-art herkenner. Nijmegen spitst zich toe op de schatting van het masker, terwijl Leuven zich voornamelijk over de back-end ontfermt. De softwareresultaten van MIDAS worden in de herkenner van het SPRAAK-project geïntegreerd.

MIDAS is een project uit de tweede ronde. Het project loopt tot eind september 2009. Zie: http://www.esat.kuleuven.be/psi/spraak/projects/index.php?proj=MIDAS.

Hetzelfde, maar dan anders

DAESO (Detecting And Exploiting Semantic Overlap)

Er zijn veel manieren om hetzelfde te zeggen. Vergelijk bijvoorbeeld de volgende twee openingszinnen, uit het NRC Handelsblad en de Telegraaf van 11 september 2006:

| EMIEL KRAHMER |

De 44-jarige Steve Irwin - bekend door zijn tv-programma's over dieren - stierf maandagmiddag (plaatselijke tijd) nadat hij tijdens het duiken voor de Australische noordoostkust bij Port Douglas in zijn borstkas werd gestoken door een giftige pijlstaartrog.

wijlen Steve Irwin

Steve Irwin, de Australische televisiepresentator die bekend is als The Crocodile Hunter, is maandag overleden nadat hij tijdens een duikexpeditie was gestoken door een pijlstaartrog.

Hoewel deze twee zinnen dezelfde gebeurtenis beschrijven, doen ze dit in heel verschillende bewoordingen. Dit fenomeen wordt wel **semantische overlap** genoemd. Vanuit een taaltechnologisch perspectief vormt het automatisch detecteren van semantische overlap een hele uitdaging.

> "De resultaten zijn potentieel interessant voor organisaties die met grote hoeveelheden tekstuele data werken."

Neem een **information-retrieval**-toepassing (IR): een gebruiker die zoekt naar informatie over *de dood van de Crocodile Hunter* wil waarschijnlijk zowel het NRC- als het Telegraaf-artikel lezen, hoewel de term *Crocodile Hunter* in het eerste tekstfragment niet eens voorkomt. Ook voor een **automatische multi-document samenvatter**, een systeem dat meerdere teksten over hetzelfde onderwerp kan samenvatten, zou het nuttig zijn om te weten dat de beide zinnen grotendeels hetzelfde uitdrukken. Helemaal ideaal zou het zijn wanneer de samenvatter tevens in staat zou zijn om de inhoud van de gerelateerde zinnen samen te voegen tot één nieuwe zin die als het ware beide zinnen combineert. Dit is een vorm van taalgenerering die bekend staat als **zinsfusie**.

Voor deze en andere taaltechnologische toepassingen zou het dus heel nuttig zijn wanneer automatisch bepaald kon wor-den in hoeverre twee zinnen semantisch overlappen. Hoe dit gedaan kan worden is de centrale onderzoeksvraag van het DAESO-project.

DAESO bestaat uit drie fases. Allereerst zal een **monolinguaal parallel corpus** worden ontwikkeld (1 miljoen woorden) bestaande uit Nederlandse tekstparen die vergelijkbare informatie bevatten. In eerste instantie zullen overlappende woorden en frases handmatig worden opgelijnd. Mede op basis van dit corpus zullen **software tools** voor het detecteren van semantische overlap worden ontwikkeld. Belangrijk hierbij is dat niet alleen gekeken zal worden naar welke frases uit de parallelle teksten gerelateerd zijn, maar ook naar de wijze waarop ze samenhangen. In de derde en laatste fase van het project, de **externe evaluatie**, zal gekeken worden of de tools daadwerkelijk bijdragen tot een verbetering van taaltechnologische applicaties (IR, question answering en multi-document samenvatters).

De resultaten zijn potentieel interessant voor organisaties die met grote hoeveelheden tekstuele data werken (uitgeverijen, nieuws- en persbureaus en taaltechno-logiebedrijven). Wanneer u op de hoogte wilt blijven van het verloop en de resultaten van DAESO kunt u zich aanmelden voor de gebruikergroep (e.j.krahmer@uvt.nl). Het project wordt uitgevoerd door de universiteiten van Tilburg, Antwerpen en Amsterdam, en het bedrijf Textkernel.

DAESO is een project uit de tweede ronde. Het project loopt tot eind september 2009.

EMIEL KRAHMER
COÖRDINEERT HET
DAESO-PROJECT

Syntactische annotatie van hele grote corpora

LASSY (LArge Scale SYntactic annotation)

In het LASSY-project beogen we de beschikbaarheid van syntactisch geannoteerde corpora flink uit te breiden. Naast een uitbreiding van handmatig gecorrigeerde syntactische annotaties, gaan we op grote schaal automatisch toegekende syntactische annotaties opleveren. In totaal zal een syntactisch geannoteerd corpus van 500 miljoen woorden ontstaan.

| GERTJAN VAN NOORD |

De beschikbaarheid van handmatig gecorrigeerd syntactisch geannoteerd materiaal is cruciaal voor het ontwikkelen, afstemmen en evalueren van een grote verscheidenheid aan toepassingen binnen de natuurlijke-taalverwerking.

Binnen LASSY breiden we het in D-Coi syntactisch geannoteerde corpus uit tot een corpus van één miljoen woorden. Deze syntactische annotaties zijn volledig handmatig gecorrigeerd, en bevatten informatie over de woordsoort van elk woord, het lemma van elk woord, en een dependentiestructuur zoals die binnen het Corpus Gesproken Nederlands is voorgesteld, en verder werd ontwikkeld binnen D-Coi.

> “Recentelijk is er flinke vooruitgang geboekt met de automatische syntactische analyse van het Nederlands.”

Het gebruik van automatisch toegekende syntactische annotaties is tot dusver minder verbreid. Recentelijk is er flinke vooruitgang geboekt op het gebied van de automatische syntactische analyse van het Nederlands. Binnen LASSY gebruiken we de vrij beschikbare Alpino-parser om een corpus van zo’n 500 miljoen woorden syntactisch te annoteren.

GERTJAN VAN NOORD
COÖRDINEERT HET
LASSY-PROJECT

Wij verwachten hierbij dat het nadeel van de lagere kwaliteit (de kwaliteit loopt natuurlijk terug, omdat de handmatige correctie achterwege blijft) voor een groot aantal toepassingen ruimschoots zal worden gecompenseerd door de veel grotere hoeveelheid materiaal.
Dit geldt met name voor toepassingen in informatie-extractie, lexicale acquisitie, het afleiden van ontologische informatie en dergelijke. Het is bijvoorbeeld de ervaring van gerelateerde projecten op het gebied van vraag-antwoordsystemen, dat het goed mogelijk is betrouwbare feitelijke informatie uit automatisch syntactisch geannoteerde corpora te extraheren (wat betekent welke afkorting, wat is de hoofdstad van welk land, wie bekleedt welke publieke functie, enzovoort). Deze betrouwbaarheid is groter dan bij systemen die de syntactische analyse veronachtzamen en bijvoorbeeld met reguliere expressies direct in de tekst vergelijkbare informatie proberen te achterhalen.

Andere recente onderzoeken laten zien hoe hele grote syntactisch geannoteerde corpora kunnen worden gebruikt om automatisch af te leiden welke woorden semantisch vergelijkbaar zijn, of om beter inzicht in de woordvolgorderegels van het Nederlands te krijgen.
De in het project op te leveren zeer grote hoeveeldheid syntactisch geannoteerd materiaal zal, om echt bruikbaar te zijn, op een goede wijze te doorzoeken en te bewerken moeten zijn. Per zin zal de syntactische analyse als een XML-bestand beschikbaar zijn. In het project zal veel aandacht besteed worden aan efficiënt zoeken binnen enorme collecties gecomprimeerde XML-bestanden, waarbij minimaal gestreefd wordt naar expressiviteit zoals die binnen XPATH en Xquery beschikbaar is.
De projectpartners zijn de Rijksuniversiteit Groningen en de Katholieke Universiteit Leuven.

LASSY is een project uit de tweede ronde. Het project loopt tot eind oktober 2009. Zie: http://www.let.rug.nl/~vannoord/Lassy/

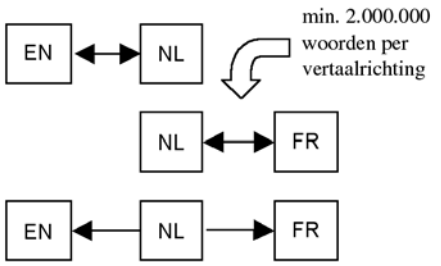
Een multifunctioneel en meertalig corpus

DPC (Dutch Parallel Corpus)

Gealigneerde parallelle corpora zijn van cruciaal belang binnen een brede waaier van meertalige applicaties, gaande van automatische vertaling en computer-ondersteunde vertaaltools over meertalige informatie-extractie tot computer-ondersteund talenonderwijs, alsook binnen (ver)taalkundig onderzoek.

| PIET DESMET |

Het doel van het DPC-project is het opbouwen van een kwalitatief hoogstaand gealigneerd corpus van 10 miljoen woorden voor de talenparen Nederlands-Engels en Nederlands-Frans.



Het corpus zal een zo breed mogelijke waaier aan vertaalde Nederlandse teksten bevatten. Het gaat hierbij over gepubliceerd materiaal van kwaliteitsvolle tekstproducenten. Om naar een goed evenwicht te streven, worden de teksten in verschillende domeinen geselecteerd, in functie van de specifieke noden van de diverse gebruikersgroepen.

> “Een initiële taak is het oppoetsen en normaliseren van de teksten en karakter-codering.”

PIET DESMET
COÖRDINEERT HET
LASSY-PROJECT

Type tekst	*Sector*
journalistiek	kranten & tijdschriften
essayistisch & literair	literaire uitgevers
zakelijk	bank en verzekering
technisch	software-bedrijven, medische sector
administratief	overheid

NORMALIZATION EN TOKENIZATION

De teksten worden door de verschillende tekstleveranciers onder allerlei formaten en karaktertypes aangeleverd. Een initiële taak is dan ook het oppoetsen en normaliseren van de teksten en karaktercodering, alsook de tokenization en het indelen in zinnen.

ALIGNERING

De tekstsamples worden vervolgens automatisch gealigneerd op zinsniveau, waarbij een aanzienlijk deel van de gealigneerde output ook handmatig zal worden geverifieerd. Om de kwaliteit van de alignering te garanderen en om handmatige controle te vergemakkelijken, zal de output van twee aligneringsprogramma’s worden samengevoegd en zullen de probleemgevallen manueel worden verbeterd.

ANNOTATIE POS + LEMMA

Het corpus wordt POS-gemarkeerd (part of speech, woordsoort) en gelemmatiseerd. Om annotaties consequent te houden ten opzichte van gelijkaardige monolinguale Nederlandse corpora zal het POS-annotatieschema van D-coi voor het Nederlandse deel van DPC worden overgenomen.

Een webzoekfunctie wordt ontwikkeld, die zal toelaten om zowel eenvoudige als meer verfijnde zoekopdrachten uit te voeren, gebaseerd op pattern-matching van woorden of annotaties.

De K.U.Leuven Campus Kortrijk en de Hogeschool Gent vormen samen het DPC-kernteam. Andere partners zijn de Rijksuniversiteit Groningen, de Radboud Universiteit Nijmegen, de Universiteit van Tilburg, de Katholieke Universiteit Leuven, de Universiteit Antwerpen en de Universiteit Gent. De gebruikersgroepen bevatten zowel industriële als academische partners. De externe validatie is toevertrouwd aan CST (Copenhagen) en aan Xplanation, die een casestudie rond terminologie-extractie zal uitvoeren.

DPC is een project uit de tweede ronde.

Verslag

STEVIN-programmadag

Op maandag 11 september werd door de de Nederlandse Taalunie een interne STEVIN-programmadag georganiseerd in het Hof van Liere te Antwerpen. Deze programmadag was speciaal bedoeld voor informatie-uitwisseling tussen alle betrokkenen bij de STEVIN-projecten. Dat wil zeggen de projectuitvoerders en coördinatoren, leden van de programmacommissie en het TST-bestuur, en de contactpersonen van de projecten bij de TST-Centrale.

| HENK VAN DEN HEUVEL |

De presentaties hadden als voornaamste doelstelling om elkaar als experts te informeren over de voortgang van de projecten. Het Hof van Liere bood daartoe een uitstekende accommodatie. De locatie is per auto niet echt eenvoudig te bereiken, maar als je er eenmaal bent, is het statige gebouw met zijn prachtige binnentuin een aangename plek voor overpeinzing en overleg.

Een kort verslag van wat er op deze dag zoal gepresenteerd werd, is ook voor DIXIT-lezers zeker de moeite waard.

Dagvoorzitter was Frank Van Eynde. Op een prettige en informele wijze verbond hij de verschillende presentaties aan elkaar en zag erop toe dat er voldoende vragen gesteld konden worden.

De voorzitter van de programma-commissie, Jan Odijk, opende de rij van presentaties met een overzicht over de verschillende oproepen in STEVIN, en wat deze tot dusver hebben opgeleverd aan projecten. Hij wees op de aanstaande Call for Tender voor een groot corpus geschreven Nederlands en op de derde oproep in 2007. Ook blikte hij vooruit op het leven na STEVIN en, meer concreet, op de mogelijkheid om het thema Taal & Veiligheid binnen een Smart Mixprogramma vorm te geven.

De rest van het ochtendprogramma werd gevuld met presentaties van de vijf onderzoeksprojecten die vanuit de eerste ronde gefinancierd worden: AUTONO-MATA, COREA, D-Coi, IRME en JASMIN-CGN. Deze projecten stellen zichzelf aan u voor in deze DIXIT, dus ik volsta graag met een verwijzing naar deze bijdragen. De programma-commissie was de taak toebedeeld om de discussie met gerichte vragen op gang te brengen. Gegeven de mondigheid van het overige publiek was dat een wat overbodige opdracht. Dat bleek temeer toen Catia Cucchiarini (JASMIN-CGN) na haar presentatie zélf een aantal pregnante vragen voor de luisteraars in petto had.

De soep-met-broodjeslunch (maar dan op zijn Belgisch) werd in een belendende kantine gebruikt. Daarna was er een presentatie van de drie lopende demonstratieprojecten: de Nummerbord Retrieval Tool, C-Content en Gemeente-Connect. De demo's toonden dat je met bestaande technologie en een gezonde dosis inventiviteit in korte tijd heel interessante en nuttige toepassingen kunt maken, die de potentie van taal- en spraaktechnologie op aansprekende wijze laten zien.

De dag werd voortgezet met een korte presentatie van Lou Boves over het aan STEVIN gerelateerde IMIX-project. Ook hiervoor verwijs ik graag naar de IMIX-bijdrage in deze DIXIT.

Voor het volgende onderdeel van het programma ging iedereen eens goed zitten. Tijdens de voorafgaande presentaties hadden verschillende sprekers al gewezen op kwesties met betrekking tot intellectual property rights (IPR). Het was daarom een goede zet van de organisatie dat men Jeannine Beeken van de TST-centrale gevraagd had om op de IPR-regelingen in STEVIN in te gaan. Zij presenteerde een model waarin de IPR-verantwoordelijkheden van de projectpartners richting Taalunie (en

vice versa) worden uiteengezet, en welke licenties hierbij horen. Uit de reacties uit de zaal bleek wel dat hiermee nog niet alle vragen de wereld uit zijn.Er zijn met name vragen over de achtergrondkennis die moet worden overgedragen. Stel je voor dat jouw bedrijf of afdeling jaren aan iets gewerkt heeft en dat deze kennis nu in zo'n STEVIN-project wordt ingebracht. Je wilt de rechten op deze inspanning niet zo maar "weggeven", hoewel je het gebruik van de kennis best met anderen wilt delen. Hoe doe je dat precies?

In elk geval ligt er nu een basis die aan individuele behoeften kan worden aangepast. Verder belichtte Jeannine Beeken de modellen voor licentie-overeenkomsten die zijn opgesteld tussen de Nederlandse Taalunie en eindgebruikers en de rol van de TST-centrale daarbij. De presentatie van Jeannine Beeken is terug te vinden onder http://taalunieversum.org/stevin/projectuitvoerders/#ipr.

Een vliegensvlugge performance werd hierna gevraagd van de presentatoren van de onderzoeksprojecten die in de tweede ronde zijn goedgekeurd (DAESO, DPC, LASSY, MIDAS, N-BEST, STEVINcan-PRAAT). Zij moesten ieder in vijf minuten hun project voor het voetlicht brengen. Gert-Jan van Noord kweet zich glansrijk van zijn taak. Hij slaagde erin LASSY helder te presenteren en ook nog tijd over te houden.

Na de thee werden de twee lopende tenderprojecten gepresenteerd: SPRAAK en CORNETTO. Hierna volgde nog een korte algemene discussie.

Na de borrel werd informeel nagepraat tijdens een voortreffelijk diner in Horta Gand Café in een bijzondere Art Nouveau entourage. Ik kijk terug op een geslaagde dag waarop het goed mogelijk was om relevante informatie te verkrijgen en contacten te leggen. Duidelijk werd dat alle projecten op schema liggen en een belangrijke bijdrage aan de ontwikkeling van de taal- en spraaktechnologie voor het Nederlands vormen; iets dat als een groot compliment voor het Stevin-programma gezien mag worden.

Enkele markante reacties

Ik vond het boeiend om te horen waar iedereen mee bezig was en te constateren dat er al zoveel STEVIN-projecten lopen of bijna lopen, en dat er gewerkt wordt aan een waaier van onderwerpen.

Het was mooi om een als het ware vanzelfsprekende samenwerking tussen Nederland en Vlaanderen enerzijds en tussen de onderzoekswereld en het bedrijfsleven anderzijds te kunnen vaststellen.

De presentaties toonden dat samenwerking en gebruik maken van elkaars resultaten effectief ook gebeurt, waar-door overlap vermeden wordt. Dit is toch een van de doelstellingen van STEVIN. Hierdoor en ook omdat een en ander getrapt wordt uitgevoerd, kan men versneld tot resultaten komen.

Het was bemoedigend om te zien dat er inderdaad een STEVIN-netwerk ontstaat waarbij mensen spontaan samenwerken of gebruik maken van elkaars resultaten zonder dat dit "vereist" wordt in het kader van het eigen project waarbinnen men actief is.

De STEVIN-dag zou jaarlijks georganiseerd moeten worden, want het is een zeer snelle manier om uit te vissen waar andere onderzoekers uit het veld mee bezig zijn. Door hier aanwezig te zijn, hoef ik de publicaties over die projecten niet meer te lezen om te weten waarover het gaat. Zo spaar ik tijd uit.

Spijtig dat de lunch elders opgediend werd, zodat er te weinig tijd overbleef om de demonstratieprojecten goed te bekijken.

Een dergelijk moment biedt een zeer goede gelegenheid om andere TST-onderzoekers en collega's zowel uit Vlaanderen als Nederland te ontmoeten en te leren kennen, zeker als men niet in de "binnenste cirkel" van het wereldje actief is.

HENK VAN DEN HEUVEL IS DIRECTEUR VAN CLST, RADBOUD UNIVERSITEIT NIJMEGEN

Een uitbreiding van het Corpus Gesproken Nederlands

JASMIN-CGN

Automatische spraakherkenning, oftewel herkenning van spraak door computers, wordt onder andere gebruikt in spraakgestuurde toepassingen zoals navigatiesystemen, educatieve computer-programma's, en hulpmiddelen ter ondersteuning van mensen met beperkingen (bijvoorbeeld voor het bedienen van apparatuur met de stem).

| CATIA CUCCHIARINI |

J ongeren
A nderstaligen
S enioren
en
M achine
I nteractie
voor het
N ederlands

Computers kunnen spraak redelijk goed herkennen als ze "getraind" zijn met grote hoeveelheden spraak, denk bijvoorbeeld aan het informatiesysteem van de NS. Hiervoor zijn spraak-databanken nodig. Voor het Nederlands bestaat bijvoorbeeld het Corpus Gesproken Nederlands (CGN), een verzameling van ongeveer 900 uur gesproken Standaardnederlands, afkomstig van volwassen Nederlandstalige sprekers uit Vlaanderen en Nederland.

Op dit moment kunnen computers de spraak van volwassenen redelijk goed verstaan, maar ze hebben nog steeds veel moeite met spraak van jongeren, anderstaligen en senioren. Dit komt doordat de spraak van deze groepen sprekers heel anders is dan die van volwassenen. Er zijn op dit moment geen databanken met spraak van jongeren, anderstaligen en senioren, althans niet voor het Nederlands.

> "Computers hebben nog steeds veel moeite met spraak van jongeren, anderstaligen en senioren."

Om aan deze behoefte tegemoet te komen is het project JASMIN-CGN gestart, dat als doel heeft het samenstellen van een databank met spraak van jongeren, anderstaligen en senioren.

Een ander bekend probleem bij het inzetten van automatische spraakherkenning om mensen met computers te laten spreken is dat wanneer mensen tegen een computer praten, hun spraak kenmerken vertoont die problematisch zijn voor computers. Bijvoorbeeld aarzelingen, onderbrekingen, versprekingen, schreeuwen en overdreven articuleren.

Beter inzicht in de aard en het voorkomen van deze verschijnselen zou kunnen helpen om te bepalen hoe de technologie aangepast moet worden zodat deze natuurlijke kenmerken van menselijke spraak geen obstakel meer vormen voor automatische spraakherkenning. In het project JASMIN-CGN zullen dan ook opnames gemaakt worden van situaties waarin jongeren, anderstaligen en senioren een dialoog met een computer voeren.

JASMIN-CGN is een samenwerkingsproject tussen de Radboud Universiteit Nijmegen (CLST), de K.U. Leuven (ESAT) en het bedrijf Talking Home.

De resultaten van dit project kunnen gebruikt worden voor onderzoek en voor het ontwikkelen van spraakgestuurde toepassingen voor deze specifieke groepen sprekers.

Talking HOME

KATHOLIEKE UNIVERSITEIT LEUVEN

Radboud Universiteit Nijmegen

JASMIN-CGN is een project uit de eerste ronde. Het project loopt tot eind maart 2003. Zie: http://homes.esat.kuleuven.be/~spch/projects/JASMIN/.

CATIA CUCCHIARINI
COÖRDINEERT HET
JASMIN-CGN-PROJECT

IRME
(Identificatie en lexicale Representatie van Multiwoord-Expressies)

In het IRME-project werken UiL-OTS (Utrecht), Alfa-Informatica (Groningen) en Van Dale Lexicografie (Utrecht) samen op het gebied van de zogenaamde multiwoord expressies (MWE's).

| JAN ODIJK |

MWE's zijn woordcombinaties met eigenschappen die niet voorspelbaar zijn uit de eigenschappen van de individuele woorden of uit de normale grammaticaregels, en die daarom opgenomen moeten worden in een woordenboek. Een combinatie van woorden kan bijvoorbeeld een onvoorspelbare betekenis hebben door een idiomatische interpretatie *(de boeken neerleggen)*, of een specifieke technische betekenis *(aandelen aan toonder)*, beperkte gebruiksmogelijkheden *(met vriendelijke groet als afsluiting van een brief)*, of een onvoorspelbare vertaling *(nuclear plant-kerncentrale)*.

> "Onderzoek is in het algemeen te veel gericht op individuele woorden en op generieke taal."

De huidige state-of-the-art natuurlijke-taalverwerkende systemen, zoals auteursystemen, vertaalsystemen en intelligente zoeksystemen, werken nog steeds het best als ze afgestemd zijn op een specifiek domein. Daarom is het noodzakelijk dat een generiek taalverwerkend systeem snel aangepast kan worden aan zo'n specifiek domein. Ieder domein brengt echter zijn eigen vocabulaire mee, en met name ook veel MWE's die alleen in dat domein voorkomen.

Daarom is technologie die snel en zo automatisch mogelijk MWE's en hun eigenschappen kan identificeren in bestaande documenten, van groot belang om taaltechnologie nuttig in te kunnen zetten.
En dit is precies waar één deel van het IRME-project zich mee bezig houdt: het onderzoeken en ontwikkelen van innovatieve methodes en bijbehorende tools om MWE's en hun eigenschappen automatisch of semi-automatisch te identificeren in tekstcorpora.

Om de geïdentificeerde MWE's te kunnen gebruiken is het van belang dat dergelijke expressies lexicaal gerepresenteerd worden op een manier die toelaat ze efficiënt te integreren in een willekeurig taalverwerkend systeem. En dat is waar het andere deel van het IRME-project zich mee bezighoudt: het onderzoeken en ontwikkelen van een methode die MWE's zo theorie- en systeemonafhankelijk mogelijk representeert en, onlosmakelijk daarmee verbonden, het ontwikkelen van methodes om aldus gerepresenteerde expressies te integreren in een willekeurig taalverwerkend systeem.

Het probleem van de MWE's wordt door de taaltechnologische industrie als een belangrijk probleem gezien. De *Technologieverkenning Nederlandstalige taal- en spraaktechnologie* van M&I/Partners en Montemore, die de laatste barrière voor het opstarten van het STEVIN-programma weggenomen heeft, vermeldt dat NOTaS erop wijst dat het onderzoek in het algemeen te veel gericht is op individuele woorden en op generieke taal, terwijl de aandacht meer zou moeten liggen op multiwoord-expressies en domeinspecifieke taal. Ook Gregor Thurmair, van Linguatec uit Duitsland, dat onder andere automatische vertaalsystemen maakt, heeft hier bij verschillende gelegenheden op gewezen. Met het IRME-project trachten we tegemoet te komen aan deze wens.

JAN ODIJK
COÖRDINEERT
HET IRME-PROJECT

IRME is een project uit de eerste ronde. Het project loopt tot eind mei 2007. Zie: http://www-uilots.let.uu.nl/irme/.

Naam maken dankzij

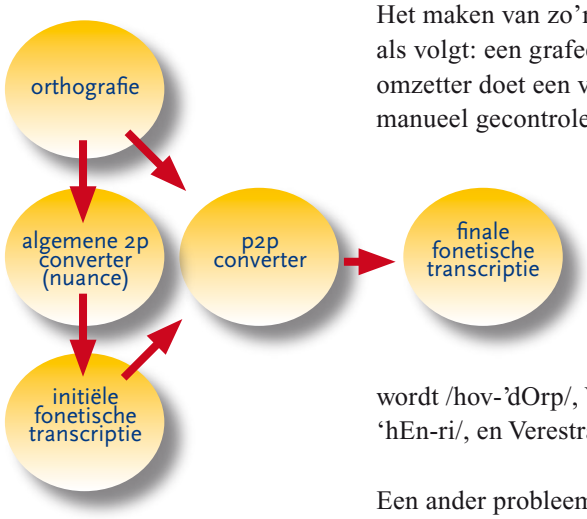
Autonomata

De uitdaging. Zou het niet fantastisch zijn als een autobestuurder aan zijn navigatiesysteem kon *zeggen* waar hij of zij naartoe wil? Waarom nog aan wieltjes draaien of in menu's navigeren om letter per letter de gewenste bestemming in te voeren?

| JEAN-PIERRE MARTENS |

Om dit droomscenario op een economisch verantwoorde wijze te kunnen realiseren, moet men snel de correcte uitspraken van adresgegevens kunnen genereren en in het lexicon van het systeem kunnen opslaan.

> "Na telling van de namen met betere en slechtere transcripties blijkt dat netto 40% van de foutieve namen een betere transcriptie krijgt."



Het maken van zo'n lexicon gebeurt nu als volgt: een grafeem-naar-foneem-(g2p)-omzetter doet een voorstel, dat vervolgens manueel gecontroleerd wordt. De dure manuele controle is nodig omdat een algemene g2p-omzetter bij namen nog te vaak incorrecte transcripties oplevert. Bijvoorbeeld: Hoofddorp wordt /hov-'dOrp/, Victor Henri /'vIk-tOr 'hEn-ri/, en Verestraat /v@-'rEs-trat/.

Een ander probleem dat optreedt bij de herkenning van namen is dat er verschillende gangbare uitspraken van dezelfde naam kunnen bestaan. Indien een bepaalde uitspraak niet voorkomt in het lexicon van de herkenner, dan zal deze ook verkeerd herkend worden.

JEAN-PIERRE MARTENS
COÖRDINEERT HET
AUTONOMATA-PROJECT

Naar een oplossing. Het project Autonomata beoogt de volgende doelstellingen:

- het ontwikkelen van nauwkeurige g2p-omzetters voor persoonsnamen (voornamen, familienamen) en toponiemen (straatnamen, plaatsnamen);
- het ontwikkelen van een transcriptietoolbox waarmee men op basis van eigen trainingslexica nauwkeurige g2p-omzetters voor andere naamsoorten (bijvoorbeeld merknamen) kan construeren;
- het opnemen en manueel transcriberen van 60.000 uitgesproken namen om zo een groot deel van de te verwachten uitspraakvariaties te kunnen blootleggen.

Methodiek. Bij de bouw van de transcriptietoolbox werd vertrokken van algemeen beschikbare state-of-the-art technologie (de RealSpeak-g2p-omzetter van Nuance). Vervolgens werden zogenaamde foneem-naar-foneem(p2p)-omzetters ontwikkeld die een aantal fouten in de initiële transcriptie kunnen verbeteren en verschillende plausibele transcripties per naam kunnen genereren (zie figuur).

Voorlopige resultaten. De transcriptietoolbox is klaar. De p2p-omzetters werden getraind op grote trainingslexica (circa 100.000 namen) met vaak meerdere manuele transcripties per naam.

Voor persoonsnamen realiseert de p2p-omzetter een daling van de word error rate (WER, percentage namen met een fout) van 43,4 naar 34,5%, en na telling van de namen met betere en slechtere transcripties blijkt dat netto 40% van de foutieve namen een betere transcriptie krijgt. Voor toponiemen daalt de WER van 51,2 naar 32,9% en krijgt 61% van de foutieve namen een betere transcriptie.

Projectuitvoerders zijn de universiteiten van Gent (ELIS), Nijmegen (CLST) en Utrecht (UiL-OTS).

Autonomata is een project uit de eerste ronde. Het project loopt tot eind mei 2007. Zie: http://speech.elis.ugent.be/autonomata/

Aanzet tot een Nederlandstalig tekstcorpus

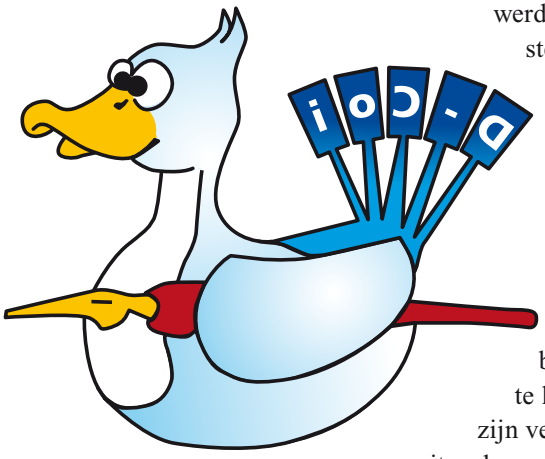
D-Coi (Dutch language Corpus Initiative)

Een van de speerpunten van het STEVIN-programma is het versterken van de infra-structuur voor de Nederlandstalige taal- en spraaktechnologie. Het project D-Coi bereidt voor op de aanleg van een zeer groot corpus hedendaags geschreven Nederlands. Daarbij wordt gedacht aan een omvang van minimaal 500 miljoen woorden.

| NELLEKE OOSTDIJK |

In het corpus worden Nederlandstalige teksten opgenomen die verschillende genres en teksttypen vertegenwoordigen. Naast de meer traditionele tekstsoorten die men gewoonlijk in corpora aantreft (zoals krantenteksten, boeken en tijdschriftartikelen), is er speciale aandacht voor teksten die verschijnen in nieuwe media: sms, e-mail, blogs en wikipedia. Het corpus omvat zowel teksten die in het Nederlands werden geproduceerd, als teksten die vanuit een vreemde taal in het Nederlands werden vertaald.

> “Er is speciale aandacht voor teksten die verschijnen in nieuwe media: sms, e-mail, blogs en wikipedia.”



Eenmaal voorzien van linguïstische annotaties zoals POS-tagging en lemmatisering, kan het corpus worden gebruikt om taalmodellen af te leiden. Deze taalmodellen zijn vervolgens te gebruiken in uiteenlopende toepassingen.
In D-Coi wordt momenteel gewerkt aan het corpusontwerp en de ontwikkeling van de benodigde protocollen, procedures en tools. Uitgangspunt daarbij is dat de verschillende annotaties zoveel mogelijk automatisch moeten kunnen worden aangebracht, dus zonder menselijke tussenkomst. Ook wordt een pilotcorpus van zo’n 50 miljoen woorden aangelegd, om de voorgestelde procollen op hun bruikbaarheid te toetsen.

Waar mogelijk wordt aansluiting gezocht bij de praktijk van het Corpus Gesproken Nederlands (CGN). Zo is in het geval van POS-tagging de CGN-tagset als uitgangspunt gekozen. Deze is uitgebreid met een aantal tags die voor geschreven tekst nodig waren (onder andere voor het taggen van leestekens en iconen zoals smileys – typisch zaken die in een gesproken tekst niet voorkomen). Soms wordt ook bewust afgeweken van de CGN-praktijk. Dat is onder meer het geval bij de syntactische annotatie, waar de ervaringen met @nnotate geen uitzicht gaven op een vergaande automatisering van dit type annotatie.

Teksten die in het corpus worden opgenomen, worden geconverteerd naar een basis XML-formaat. Vervolgens wordt de tekst opgeschoond: foutieve woordafbrekingen, spelfouten en dergelijke worden gecorrigeerd. Het corpus wordt getagd, gelemmatiseerd en voorzien van een syntactische annotatie. Een deel van de annotaties wordt handmatig geverifieerd.

Om exploitatie van het corpus mogelijk te maken, wordt de COREX-software aangepast, die oorsponkelijk voor het CGN werd ontwikkeld. Tot slot worden twee verkennende studies uitgevoerd naar de mogelijke annotatie van semantische rollen en van spatio-temporele aspecten.

Het project wordt uitgevoerd door een Vlaams-Nederlands consortium dat bestaat uit zes kennisinstellingen en één industriele partner.

D-coi is een project uit de eerste ronde en loopt nog tot eind december 2006.
Zie: http://lands.let.ru.nl/projects/d-coi.

NELLEKE OOSTDIJK
COÖRDINEERT HET
D-COI-PROJECT

Corea
(COreference Resolution for Extracting Answers)

Stel, je wilt weten wat Yevgeni Kafelnikovs beste prestatie was op Wimbledon. Als je googelt op “Kafelnikov” en “Wimbledon”, krijg je 150.000 resultaten. Dat schiet niet zo op. Voor het beantwoorden van dit soort vragen in natuurlijk taal, het zogeheten *Question Answering*, kan de computer goede diensten bewijzen.

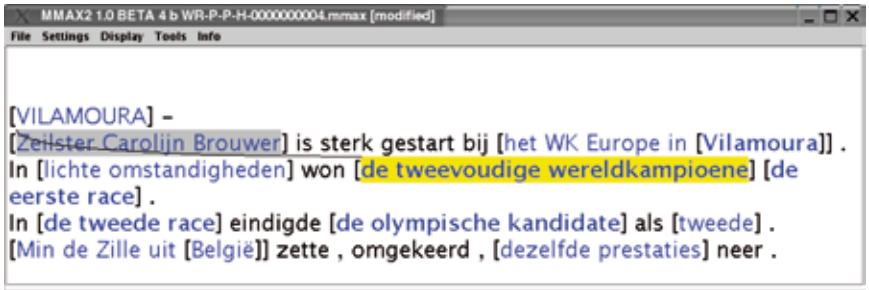
| GOSSE BOUMA |

Uit de volgende tekst kun je opmaken dat Kafelnikov ooit de kwartfinales van Wimbledon heeft bereikt:

Yevgeni Kafelnikov wist dat hij de favoriet was, in de derde-ronde-partij tegen de Argentijn Guillermo Canas. Die wetenschap deed hem verstijven. De mentaal onevenwichtige Rus reikte op Wimbledon nooit verder dan de kwartfinales.

> “Door de verwijzingen op te lossen wordt de samenhang binnen een tekst in kaart gebracht.”

Die conclusie vraagt wat denkwerk. Wil je de puzzel kunnen oplossen, dan moet je weten dat “de mentaal onevenwichtige Rus” verwijst naar “Yevgeni Kafelnikov”. Of, anders gezegd: dat die twee woordgroepen coreferentieel zijn. Het project Corea richt zich erop om uit te zoeken welke uitdrukkingen in een tekst coreferentieel zijn.



annotatie-tool

GOSSE BOUMA COÖRDINEERT HET COREA-PROJECT

Behalve identiteitsrelaties zoals in bovenstaand voorbeeld, waar twee woordgroepen gebruikt worden om naar hetzelfde individu te verwijzen, worden ook andere relaties geannoteerd. In het volgende voorbeeld verwijst “het vrouwentoernooi” naar “Wimbledon” in de voorgaande zin.

De uitschakeling van Kafelnikov was de grootste verrassing bij de mannen op Wimbledon. In het vrouwentoernooi struikelde Mauresmo over Tanasugarn.

Door die verwijzingen op te lossen wordt de samenhang binnen een tekst in kaart gebracht, en de bruikbaarheid sterk vergroot.

De onderzoekers, afkomstig van de Universiteit Antwerpen, de Rijksuniversiteit Groningen en het bedrijf Language and Computing, hebben twee doelen. Het eerste doel is het systematisch markeren van de coreferentierelaties binnen een corpus van zo’n 100.000 woorden. Er wordt gebruik gemaakt van software die het annoteren van teksten vergemakkelijkt (zie figuur). De gegevens worden opgeslagen in XML, een publieke standaard voor het vast-leggen van data.

Het tweede doel is het ontwikkelen van software die automatisch kan bepalen wat de coreferentierelaties binnen een tekst zijn. Daarvoor wordt gebruik gemaakt van alle beschikbare informatie: de herhaling van patronen, de grammatica van de zinnen, de betekenis van de woorden, en zelfs de afstand tussen uitdrukkingen. De onderzoekers combineren taalkundige kennis met statistische methoden, en leveren zo maximale prestaties.

Evaluatie van de nieuwe software in bijvoorbeeld Question-Answering-toepassingen en informatie-extractie zal aantonen hoe groot de winst van de nieuwe methode is.

Corea is een project uit de eerste ronde. Het project loopt tot eind maart 2003. Zie: http://www.cnts.ua.ac.be/~hoste/corea.html.

Een lexicaalsemantische database voor taaltechnologie

Cornetto
(COmbinatorial and Relational NETwork as TOolkit)

Het doel van Cornetto is het bouwen van een lexicale database voor het Nederlands met zowel semantische relaties als combinatorische informatie die nodig is om woorden te combineren. Semantische relaties vindt men onder andere in wordnets, waarin groepen synoniemen door middel van voornamelijk verticale subtyperelaties met elkaar worden verbonden: *waakhond* > *hond, koffie* > *drank*. Daarnaast zijn er ook horizontale relaties: *waakhond* > *bewaken, school* > *lesgeven*.

| PIEK VOSSEN |

Een netwerk van dergelijke conceptuele relaties vormt een krachtige kennisbron om te redeneren over teksten. Toch is taal niet alleen een kwestie van conceptuele relaties. Zo kunnen we op grond van het feit dat koffie en thee dranken zijn toch niet voorspellen dat we koffie en thee *zetten*, maar limonade *maken*. Hetzelfde geldt

voor voorzetsels bij werkwoorden zoals “behandelen *aan* zijn verwondingen” maar “behandelen *voor* een ziekte”.

> “Taal is niet alleen een kwestie van conceptuele relaties.”

PIEK VOSSEN COÖRDINEERT HET CORNETTO-PROJECT

Er is een heel scala aan dergelijke combinatorische informatie die typisch is voor het Nederlands. Deze informatie moet gekoppeld worden aan de conceptuele informatie om het computers mogelijk te maken de betekenis van woorden in teksten te herkennen, maar ook om vloeiende teksten te genereren in toepassingen.

De methode die in Cornetto wordt gehanteerd, is het samenvoegen en verder verbeteren van twee bestaande databases, namelijk het Nederlandse Wordnet (*Dutch WordNet*, DWN) en het ReferentieBestand Nederlands (RBN). Het DWN bevat verticale en gedeeltelijk ook semantische horizontale relaties. Het RBN bevat horizontale relaties en combinatorische informatie. Een van de eerste doelstellingen van Cornetto is van iedere woordbetekenis in DWN te bepalen met welke woordbetekenis in RBN die correspondeert en vice versa. Dit gebeurt door middel van een programma dat overlappende informatie uit beide bestanden vergelijkt.

Vervolgens moet worden bepaald of de samengevoegde informatie semantisch valide en coherent is, en in hoeverre uit de samengevoegde informatie verdere relaties kunnen worden afgeleid. Om dit te bereiken wordt iedere woordbetekenis gekoppeld aan een formele ontologie (SUMO, DOLCE), die bepaalt welke relaties wel en niet mogelijk zijn. De relaties in het semantisch netwerk kunnen dan worden “doorgerekend” om te zien of er inconsistenties optreden.

Om het samenvoegen en vergelijken van de verschillende bronnen en stukken informatie mogelijk te maken, is er een speciale database ontwikkeld waarin alle informatie efficiënt aan elkaar gekoppeld is en met een editor kan worden gecontroleerd en bewerkt. Ook worden er automatische extractiemethodes ontwikkeld, die worden toegepast op een juridisch domein. De uiteindelijke data komen beschikbaar via de TST-centrale.

Cornetto is een project uit de eerste tenderoproep. Het project loopt tot eind maart 2008. Zie: http://www.let.vu.nl/ onderzoek/projectsites/cornetto/.

Taaltechnologie in een juridisch informatieportaal

Rechtsorde.nl

Taaltechnologie is de sluitsteen van zoektechnologie. De laatste jaren is flink geïnvesteerd in de optimalisatie van zoektechnologie. Daarbij is vooral ingezet op aspecten als robuustheid en snelheid van zoektechnologie. Huidige zoeksystemen kunnen enorme hoeveelheden data aan en komen in een fractie van een seconde met resultaten op iedere willekeurige zoekvraag. Aan kwantiteit geen gebrek. Maar hoe staat het met de kwaliteit?

| MICHEL MOOREN |

De gebruiker van zoektechnologie is klaar voor intelligentere systemen die ook echt antwoord geven op de vraag die is gesteld. Taaltechnologie kan daar een belangrijke rol in spelen. Om dit aan te tonen is in het kader van het project Rechtsorde.nl de zoekfunctionaliteit van het juridisch informatieportaal rechtsorde.nl verrijkt met taalbewustzijn.

> “In oudere systemen moet de gebruiker vaak goed nadenken over de zoektermen die hij kiest.”

De Nederlandse overheid is er de laatste jaren meer en meer toe overgegaan om elektronische informatie op het gebied van wet- en regelgeving publiek toegankelijk te maken. Helaas wordt deze informatie gepubliceerd over vele (niet gestandaardiseerde) websites van de overheid. Dit maakt het haast onmogelijk voor een professionele gebruiker om de gezochte informatie snel boven water te krijgen. Er is daarom grote behoefte aan één centrale ingang waar alle openbare informatie over wet- en regelgeving volledig en snel doorzocht kan worden.

MICHEL MOOREN
COÖRDINEERT HET
C-CONTENT-PROJECT

Het bedrijf C-CONTENT (bouwer van onder andere de elektronische woordenboeken van Van Dale en detelefoongids.nl) heeft daarom het systeem rechtsorde.nl gebouwd dat dagelijks (geautomatiseerd) informatie over wet- en regelgeving vergaart van tientallen vrij toegankelijke overheidssites. Deze informatie wordt vervolgens in één grote databank opgenomen, met een overzichtelijke inhoudsopgave. Alle verwante documenten zijn verder met elkaar gelinkt. Rechtsorde.nl is gericht op de professionele eindgebruiker en bevat onder andere wetten, jurisprudentie, CAO’s, ministeriële regelingen, officiële publicaties, en verordeningen van lokale overheden.

Omdat de Rechtsorde-databank vele honderdduizenden documenten bevat, is het belangrijk dat de gebruiker snel tot die paar documenten kan komen die daadwerkelijk zijn informatiebehoefte vervullen. Daartoe werkte C-CONTENT in dit demonstratieproject samen met Polderland Language & Speech Technology.

In oudere systemen moet de gebruiker vaak goed nadenken over de zoektermen die hij kiest, en kost het veel tijd om een goede zoekopdracht te formuleren. In rechtsorde.nl wordt dit omgedraaid: één zoekveld, gestructureerde resultaten en de mogelijkheid om door te klikken aan de hand van suggesties.

Met dit zogenaamde “suggestiegestuurd zoeken” heeft rechtsorde.nl het *bedoelde u…* van Google geperfectioneerd. Het zoeksysteem kan hele zinnen aan. Aan de basis verbetert de zoekmachine spelfouten en suggereert het synoniemen: *wiplash* (sic) wordt *zweepslag*. Onder water worden samenstellingen ontleed en vervoegingen en verbuigingen herleid naar de woordstam, waardoor de kwaliteit van het zoekresultaat (precision & recall) verbetert.

Daarna kan de gebruiker gaan verfijnen op domein, door bijvoorbeeld aan te geven dat hij een Nederlandse wetstekst zoekt, een arrest, of een Europese verordening. Ook kan hij verfijnen op trefwoordsuggesties afkomstig van een krachtige thesaurus.

Rechtsorde.nl is een demonstratieproject uit de eerste oproep. Het project loopt tot begin april 2007. Zie: http://www.rechtsorde.nl/.

Welkom bij

GemeenteConnect

de automatische vraagbaak van de gemeente Gilze en Rijen

... zo luidt de welkomstboodschap van de nieuwe telefonische ingang van het gemeentehuis in Rijen (Noord-Brabant). Bellers kunnen voor alle gemeentelijke producten (bijvoorbeeld een Nederlands paspoort) hun vraag stellen aan het systeem en krijgen na een korte vraag-antwoorddialoog een samenvatting met informatie over het betreffende product. Als hiermee de vraag nog niet beantwoord is, wordt de beller doorgeschakeld naar een medewerker van de gemeente.

| HANS JONGEBLOED |

In totaal zijn er circa 150 producten in het systeem verwerkt. Over elk product kunnen bellers een grote variatie aan vragen stellen, waarvan ze het antwoord in de samenvatting kunnen beluisteren. In de figuur staat een voorbeeld van een dialoog die iemand kan voeren met GemeenteConnect.

> “De medewerkers hebben dan tijd om de complexe vragen te beantwoorden.”

Hoe werkt dit allemaal? Gemeente-Connect is ontwikkeld door Irion en Dutchear. Irion is leverancier van geavanceerde taal-technologieproducten, zoals een vraag-antwoorddialoogsysteem dat op de internetsite van de gemeente gebruikt kan worden. Dutchear levert producten en maatwerkoplossingen voor spraakherkenning toegepast in telefonische dienstverlening. Door de softwaremodules van Irion en Dutchear te combineren werd het mogelijk om telefonisch toegang te krijgen tot het vraag-antwoordsysteem op het internet.

Een voorbeelddialoog

Systeem: Welkom bij GemeenteConnect, de automatische vraagbaak van de gemeente Gilze en Rijen. Wat is uw vraag?
Beller: Ik wil graag weten waar ik mijn bouwafval kwijt kan. Ik ben namelijk aan het verbouwen en heb een hoop puin over.
Systeem: Gaat uw vraag over ‘milieuzaken’ of over ‘bouwen en wonen’?
Beller: Milieuzaken
Systeem: Volgens mij gaat uw vraag over ‘afval storten’ of ‘gevaarlijk afval’. Welke wilt u?
Beller: Afval storten
Systeem: Hier volgt de informatie over ‘afval storten’
Systeem: Voor het storten van afval kunt u terecht bij de milieustraat .....

HANS JONGEBLOED PARTICIPEERT AAN HET GEMEENTECONNECT-PROJECT VOOR DUTCHEAR

Het telefonische systeem neemt automatisch het gesprek aan en stelt een aantal vragen. Spraakherkenningsoftware herkent zo goed mogelijk de antwoorden van de beller. Taaltechnologiesoftware achterhaalt de betekenis van de vraag en relateert die aan de kennis die over de categorieën (gemeentelijke producten) in het systeem is opgeslagen.

Bij kleine en middelgrote gemeentes zijn er vaak maar een beperkt aantal medewerkers die telefonisch vragen beantwoorden. Op het moment dat al deze medewerkers in gesprek zijn, worden andere bellers teleurgesteld. Met het systeem kan een groot deel van de telefonisch gestelde vragen automatisch beantwoord worden. De medewerkers hebben dan tijd om de complexe vragen te beantwoorden, zoals ingewikkelde regelingen rondom bouwvergunningen. Het systeem kan bovendien ook op internet geraadpleegd worden, zodat de informatievoorziening via telefoon en internet tegelijk geregeld is.
Doordat het systeem op elk moment voorzien kan worden van nieuwe samenvattingen kan de gemeente snel inspringen als door actuele issues het telefoonverkeer plots toeneemt. Denk hierbij aan vragen over wijzigingen in het ophalen van vuilnis of tarieven. Ook kan het systeem snel bijgetraind kan worden voor nieuwe producten of nieuwe vragen.

Met het demonstratieproject waarin de gemeente Gilze en Rijen beschikking heeft gekregen over het systeem, proberen Irion en Dutchear ook andere gemeentes in Nederland te overtuigen van de voordelen van GemeenteConnect.

GemeenteConnect is een demonstratieproject uit de eerste oproep. Het project loopt tot begin november 2006.

Dutchear helpt agenten aan voertuiginformatie

Spraakgestuurde Nummerbord Retrieval Tool

Dutchear realiseert voor Politie Utrecht de Nummerbord Retrieval Tool (NRT). Agenten kunnen lopend, op de mountainbike, in de auto en op de motor met hun gsm bellen met de NRT. De agent spreekt het kenteken in en krijgt vervolgens alle beschikbare informatie over het betreffende voertuig (naam eigenaar, APK, verzekering, gestolen) voorgelezen via text-to-speech. De NRT zorgt ervoor dat agenten van Politie Utrecht altijd op een snelle, gemakkelijke en veilige manier voertuiginformatie kunnen krijgen.

| ELS NACHTEGAAL |

GROTE WINST

De NRT levert grote winst op ten opzichte van de huidige situatie. Momenteel belt een agent met zijn gsm of C2000 naar de meldkamer of naar de infodesk, wanneer hij een kentekenplaat wil natrekken. De snelheid waarmee hij geholpen wordt, is geheel afhankelijk van de beschikbaarheid van medewerkers op de meldkamer of bij de infodesk. De beschikbaarheid van de meldkamer is echter beperkt voor deze informatieverzoeken, waardoor de wachttijd voor de agent oploopt.

> “Agenten van Politie Utrecht zijn direct betrokken bij de ontwikkeling van de dialoog.”

Hierbij kan de druk op de meldkamer (te) groot worden waardoor de primaire taak van de meldkamer, het het coördineren van acties bij calamiteiten, in gevaar komt. De huidige situatie is daarom onwenselijk.

De NRT zorgt ervoor dat een agent sneller over de relevante informatie beschikt en dat de meldkamer wordt ontlast.

DIALOOGONTWERP SAMEN MET AGENT

Om de wensen van de agenten goed te laten aansluiten bij de manier waarop de NRT functioneert, zijn agenten van Politie Utrecht direct betrokken bij de ontwikkeling van de dialoog. De dialoog heeft tot doel om snel de juiste informatie uit alle gegevensbronnen van de politie te kunnen presenteren aan de agent in alle denkbare situaties: op de fiets, in de auto, op de motor en te voet. Door de inzet van de agenten in de ontwikeling van de NRT is deze geheel ingericht volgens de wensen van de agenten.

Wilt u meer weten over deze toepassing of over de inzet van spraaktechnologie voor telefonische diensten, neem dan contact op met Dutchear.

POLITIE Utrecht

DUTCHEAR

ELS NACHTEGAAL
PARTICIPEERT AAN HET
NRT-PROJECT VOOR DUTCHEAR

Spraakgestuurde Nummerbord Retrieval Tool is een demonstratieproject uit de eerste oproep. Het project loopt tot eind september 2006. Zie: http://www.dutchear.nl/.

NICTP 2006:

Netwerkdag ICT & Handicap

Op donderdag 14 september 2006 vond in het mooie Visio-gebouw in Huizen de eerste editie plaats van NICTP 2006, een netwerkgelegenheid voor ICT-projectleiders en -medewerkers die projecten voor mensen met een beperking initiëren en begeleiden.

Het was een heel interessante bijeenkomst met 70 deelnemers, voornamelijk uit de zorg en het bedrijfsleven. Er werd een twintigtal presentaties en workshops gehouden over verschillende onderwerpen zoals:

- De toegankelijkheid van educatieve software
- Het gebruik van ICT bij de opleiding en begeleiding van personen met een verstandelijke beperking
- Het digitaal aanbieden van lectuur aan mensen met visuele beperkingen.
- Informatiesystemen over hulp-middelen in Nederland en België.
- Het begeleiden van meervoudig gehandicapte kinderen in taal- en communicatieontwikkeling.

Taal- en spraaktechnologie was voornamelijk vertegenwoordigd door spraaksynthese, die wordt ingezet voor allerlei voorleesfuncties. Andere mogelijkheden van TST bleven vrijwel onbesproken. Aangezien alle deelnemers een exemplaar van het rapport “Taal- en spraaktechnologie en communicatieve beperkingen” hebben ontvangen, is te hopen dat in volgende edities meer aandacht zal zijn voor de bijdrage die TST kan leveren aan het verbeteren van de positie van mensen met een beperking.

Op deze dag is ook de “Jacko van Dijk Stimuleringsprijs” uitgereikt aan Solutions Radio B.V. voor de Orion Webbox, een webradio waarmee blinden en slechtzienden kunnen luisteren naar gesproken lectuur zoals boeken, kranten en tijdschriften.

Symposium Spraak- en Taaltechnologie

ten behoeve van de Spraak- en Taalpathologie

Vrijdag 15 december 2006
09.30 - 17.00 uur
Sint Maartenskliniek
Hengstdal 3, Nijmegen

De Sint Maartenskliniek heeft in 2005 van het Ministerie van Volksgezondheid, Welzijn en Sport een erkenning gekregen als Ontwikkelcentrum voor Spraak- en Taaltechnologie ten behoeve van spraak- en taalpathologie en revalidatietechnologie in het algemeen.

Het symposium zal in het teken staan van de spraak- en taaltechnologie ten behoeve van diagnostiek, therapie en ondersteunende ommunicatiemiddelen. Er zullen presentaties gehouden worden over lopende projecten en toekomstige ontwikkelingen. Daarnaast wordt er een informatiemarkt georganiseerd van bedrijven die zich richten op toepassingen ten behoeve van mensen met communicatieve beperkingen.

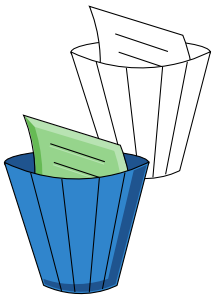
DOELGROEP:
Belanghebbenden op het gebied van spraak- en taaltechnologie en op het gebied van spraak- en taalpathologie; zorgverzekeraars; therapeuten; onderzoekers; patiëntenorganisaties; bedrijven aangesloten bij de Nederlandse Organisatie voor Taal- en Spraaktechnologie.

Voor nadere informatie kunt u per e-mail contact opnemen met mevrouw Pétri Holtus, projectleider Ontwikkelcentrum voor Spraak- en Taaltechnologie (OSTT).
E-mailadres: p.holtus@maartenskliniek.nl.
www.ostt.eu

De kleur van spelling

In augustus werden de nieuwe officiële spellingregels van kracht. Er is inmiddels een nieuw Groene Boekje en verschillende uitgevers geven producten uit die zich conformeren en daarmee een spellingkeurmerk krijgen van de Nederlandse Taalunie. Ik zal deze spelling verder aanduiden als de *groene* spelling.

| THEO VAN DEN HEUVEL |



De laatste tijd zijn er regelmatig verhalen in de pers verschenen, waaruit moet blijken hoe slecht de officiële spellingcommissie haar werk gedaan zou hebben. Er wordt meteen een alternatief naar voren geschoven: omdat de alternatieve regels beschreven gaan worden in het nieuwe Witte Boekje van het Genootschap Onze Taal, ligt het voor de hand deze variant aan te duiden als de *witte* spelling.
Centraal in de berichtgeving staat de absurditeit van een aantal regels en de inconsistentie ervan, waarbij een kleine verzameling voorbeelden keer op keer genoemd wordt. Ik wil het niet over die voorbeelden hebben en ik wil ook graag beargumenteren waarom ik dat geen acceptabele benadering vindt. Sterker nog, ik wil graag aantonen dat de Nederlandse pers enorme steken heeft laten vallen en haar boekje (van welke kleur dan ook) behoorlijk te buiten is gegaan.

> “Spelling moet een benadering van de uitspraak zijn.”

WAAROM SPELLINGREGELS?

Misschien bent u van mening dat de overheid zich niet met spelling moet bezighouden. Dit is bijvoorbeeld het geval in Groot-Brittannië en de Verenigde Staten. Met name vanuit het onderwijs is echter wel behoefte aan eenduidigheid. Wilt u dat uw kinderen leren spellen op school? De Nederlandse Taalunie is met name opgericht om de spelling van Vlaanderen en Nederland op een lijn te brengen. Dat leidde in eerste instantie tot het ontstaan van het Groene Boekje van 1954. Dat was weliswaar een eerst officieel spellingvoorschrift, maar liet nog ruimte doordat het een voorkeur uitsprak. In de praktijk werd de voorkeursspelling vooral in Nederland gebruikt en de nakeursspelling in Vlaanderen. Dit leidde tot veel onzekerheid, vooral in het taalonderwijs. Bedenk, een sollicitatiebrief vol spelfouten heeft een verlaagde kans om tot een gesprek te leiden. Maar wat is dan een spelfout? De Taalunie kreeg de opdracht van de Vlaamse en Nederlandse Ministers van Onderwijs om een eenvormig spellingvoorschrift op te stellen, en elke tien jaar zonodig een bijstelling te doen. Deze bijstellingen zijn bedoeld om eventuele foutjes in het voorschrift op te lossen en vooral om recht te doen aan het veranderingsproces van taal. Taal verandert zichzelf tamelijk snel, zeker in een maatschappij als de onze waarin massamedia zo’n belangrijke rol spelen.

DE GROENE PROCEDURE

Pas in 1996 werd er een officiële spelling ingevoerd. Deze werd zonder veel weerstand geaccepteerd, al waren er kleine hindernissen, zoals dat de woordenboekuitgevers soms moesten kiezen tussen slaafs volgen van de regels en consistentie, iets wat tot verwarring leidde bij de gebruikers. De Taalunie heeft daarom de eerste herziening gebruikt om een andere werkwijze in te voeren, waarbij organisaties waarvoor spelling een belangrijke rol speelt in een vroeg stadium uitgenodigd werden om zitting te nemen in een Platform. Zo kregen alle uitgevers en producenten van taalproducten de kans om te reageren op de nieuwe versie van het spellingvoorschrift en hulp bij het omspellen van eigen materialen. Ook de grootste criticasters van de spelling zaten in dit overleg, al kwamen ze niet altijd opdagen. Sommigen ontkennen nu zelfs dat ze lid waren van het spellingplatform.

HOE BOUW JE EEN SPELLING OP?

Als je een spellingvoorschrift opzet moet je een evenwicht vinden tussen drie randvoorwaarden.

- Spelling moet een benadering van de uitspraak zijn.
- Spelling moet consequent zijn.
- Spelling moet recht doen aan de geschiedenis.

Deze drie overwegingen staan ongelukkigerwijs op gespannen voet met elkaar. Je komt een heel eind als je één van de drie randvoorwaarden mag schrappen. Dat hebben de Italianen gedaan, bijvoorbeeld. Die houden geen

"Spelling moet consequent zijn."

rekening met de oorsprong van het woord en de spellinggeschiedenis. Als wij dat principe zouden overnemen, zouden we misschien 'fotui' gaan schrijven in plaats van 'fauteuil'. Dit zal niet snel gebeuren, omdat de meeste mensen die zich druk maken over spelling neigen naar een behoudend standpunt: we willen liefst zo weinig mogelijk veranderen. Een ander voorbeeld is 'product'. De eerste randvoorwaarde dwingt tot een spelling met een 'k', maar dan ontstaat er een breuk met het gerelateerde woord 'produceren', zeker als dat ook nog eens met een 's' geschreven zou worden. Het heeft dan ook helemaal geen zin om een discussie te voeren met niks anders als inzet dan de afbreekstreepjes en accentjes in steeds dezelfde excentrieke voorbeeldwoordjes, waar wij Nederland ons zo graag druk over maken.
Spelling is hoe dan ook een compromis.

Ook de witte spelling zal een compromis zijn. Ik kan me niet voorstellen dat de tegenstanders van de groene spelling het eens kunnen zijn over een witte. Dat kan alleen als de sentimenten niet over de regels gaan, maar over iets heel anders.

Wie de discussies in de Nederlandse kranten volgt zal misschien denken dat er een groot verschil is tussen witte en groene spelling. Ik kan u verzekeren, dat het voor u onmogelijk zal zijn van een gemiddeld document te bepalen of het in witte of groene spelling geschreven is. Zelfs als dat document een dikke roman is of een dik wetenschappelijk werk. Als over de witte spelling even grondig is nagedacht als over de groene, kan het me niet veel schelen welke van de twee spellingen ik moet toepassen. Het zou wel heel fijn zijn als iedereen dezelfde kleur gebruikt.

HOE BREEK JE EEN SPELLING AF?

De Nederlandse pers heeft zich afgekeerd van de groene spelling en omarmt de witte spelling. Dat komt niet omdat de witte spellingregels beter zijn, want die zijn nog niet gepubliceerd. Dat komt ook niet omdat de groene regels niet goed zouden zijn, want dan hadden ze wel van de gelegenheid gebruik gemaakt om daarop te reageren. Na verschillende gesprekken met betrokkenen stel ik vast, dat ik er nooit achter zal komen. Feit is dat de NOS en de Nederlandse kranten vakkundig de groene spelling hebben ondermijnd en daarbij geen instrument geschuwd hebben. Daarbij maken ze op onacceptabele wijze gebruik van hun machtspositie. Van "hoor en wederhoor"is geen sprake. Bij alle berichtgeving wordt gesuggereerd dat de Taalunie stoffige, stokoude en wereldvreemde kamergeleerden aan het werk heeft gezet en zonder enige vorm van inbreng van derden een decreet heeft

"Spelling moet recht doen aan de geschiedenis."

uitgevaardigd. Dit beeld is volstrekt onjuist. Het lijkt soms wel dat wat het witte kamp verenigt vooral de afkeer van de Taalunie is.
Ook het beeld dat er veel weerstand zou bestaan tegen de groene spelling klopt in het geheel niet. De media verslaan geen nieuws, maar creëren het. Dat is de grootste journalistieke misdaad. Het is ook heel eenvoudig: je interviewt een aantal voor- en tegenstanders, knipt bij de voorstanders alle argumenten weg en monteert het geheel alsof de voorstanders met nietszeggende *one-liners* reageren op valide bezwaren van de tegenstanders. Het feit dat u dit artikel niet in een krant leest is dus geen toeval.
De rol van het Genootschap Onze Taal in dit verband is opmerkelijk. Het is lid van het spellingplatform, voor zover ik weet nog steeds, maar heeft ergens in het proces toch gekozen voor het witte kamp. Bij ontbreken van formele witte regels kan dat alleen een politieke keuze zijn.
Vlaanderen kijkt met verbazing toe.

WAT NU?

De kleurenstrijd is dramatisch voor de geloofwaardigheid van de Nederlandse spelling. Veel Nederlanders (en misschien ook wel veel Vlamingen) denken nu dat de geleerden het kennelijk nog niet eens zijn en dat het hun tijd wel zal duren. Kortom, het beleid van de regeringen om tot een eenvormige spelling te komen staat onder druk.
Ik ben erg teleurgesteld over de wijze waarop de discussie gevoerd is en ben erg bang voor de toekomst van de Nederlandse spelling. En mocht er echt een breed wit front blijken te bestaan, dan vrees ik ook voor een hernieuwde spellingsbreuk tussen Nederland en Vlaanderen. Ik wil U, beste lezer, dan ook uitnodigen om na te denken over het belang van spelling. De media wil ik vragen te doen wat ze moeten doen en omzichtig om te gaan met berichtgeving over een zaak waarin ze zelf zo overduidelijk partij zijn. Pas dan kan blijken dat de opgeklopte verschillen niet de moeite van een spellingstrijd waard zijn.

THEO VAN DEN HEUVEL
IS DIRECTEUR VAN
POLDERLAND LANGUAGE &
SPEECH TECHNOLOGY
TE NIJMEGEN

Verslag van een tussentijdse beoordeling

IMIX Mid-Term

De eerste dag van de IMIX Mid-term review in het Amsterdamse Felix Merites, was vooral interessant door de presentaties van drie buitenlandse reviewers, de presentatie van het onderzoek naar interactieve vraag-antwoordsystemen bij LIMSI in Orsay en een presentatie van het Nederlandse bedrijf Q-Go. Hieronder een samenvatting van de vijf boeiende presentaties en enige achtergrondinformatie.

| LOU BOVES |

SMARTWEB

Norbert Reithinger van DFKI (Duitse Onderzoeksinstituut voor Kunst-matige Intelligentie) gaf een over-zicht van het Duitse SmartWeb: een onderzoeksprogramma dat ruim vijf keer zo groot is als IMIX, in ieder geval in termen van geld. Het doel van SmartWeb is in wezen hetzelfde als IMIX: maak een systeem dat *begrijpt* wat mensen willen weten, en dat dan precies de goede informatie geeft. Niet in de vorm van een aantal documenten waar het antwoord op de vraag waarschijnlijk in staat, maar gewoon meteen door het geven van het antwoord. Dat antwoord moet dan op de meest effectieve manier gepresenteerd worden in een combinatie van tekst, spraak, plaatjes en eventueel video en uiteraard aangepast zijn aan de terminal - meestal een PDA of een geavanceerde mobiele telefoon.

> “We willen een systeem dat begrijpt wat de mensen willen weten.”

Op dit moment concentreert het onderzoek in SmartWeb zich op het wereldkampioenschap voetbal, maar op de langere termijn moet de technologie geschikt zijn voor willekeurige domeinen. SmartWeb steunt voor een belangrijk deel op hergebruik van standaard componenten, en waar er nog geen standaard bouwblokken zijn, wordt gestreefd naar standaardisering. Reithinger liet filmpjes zien van een aantal live demo’s van SmartWeb tijdens de laatste CeBIT. Ondanks de beruchte rumoerige omgeving verliep de spraakgestuurde interactie probleemloos, dankzij hergebruik van technologie voor ruisonderdrukking die ontwikkeld is in de tv-industrie. De technologie die ontwikkeld wordt voor informatie over het WK kan ook gebruikt worden om up-to-date verkeersinformatie te krijgen. En dan niet alleen in een auto, maar ook op een BMW-motor.

BLOGGERID

Jon Oberlander van de School of Informatics in Edinburgh presenteerde de nieuwste resultaten van een lopend onderzoek waarin geprobeerd wordt om persoonlijkheidskenmerken van bloggers af te leiden uit de kenmerken van de teksten die ze schrijven. Het hogere doel van het onderzoek is om uit te vinden of mens-systeeminteractie effectiever en bevredigender wordt als het systeem zich kan aanpassen aan de persoonlijkheid van de gebruiker, op dezelfde manier waarop mensen hun communicatiestrategie en hun taalgedrag aan elkaar aanpassen.
Het blijkt dat het gebruik van specifieke woorden, uitdrukkingen en syntactische constructies inderdaad correspondeert met de scores van bloggers op veelgebruikte persoonlijkheidsschalen. Andere experimenten hadden al laten zien dat veel mensen het gebruik van woorden en constructies inderdaad aanpassen aan de stijl van hun gesprekspartner. De uitkomsten van het onderzoek worden momenteel geïntegreerd in een systeem dat berichten met een bepaalde “persoonlijkheid” kan produceren. De stap naar berichten die passen bij het imago van een bedrijf is dan relatief klein.

TAALWETENSCHAP

De presentatie van Ed Hovy, van het Information Sciences Institute van de Univer-

sity of Southern California in Marina de Rey, was meer gericht op de toekomst van het onderzoek in de taalwetenschap, en op de wijze waarop taalwetenschap maximaal

“Hoe krijgen we een goede semantische annotatie?”

kan bijdragen aan de oplossing van allerlei praktische problemen rond de verwerking van informatie. Hovy onderstreepte het belang van het meenemen van het semantische niveau in beschrijving en analyse. Daarbij deelt hij het standpunt dat modern taalwetenschappelijk onderzoek gebaseerd moet zijn op de verwerking van grote en goed geannoteerde corpora, maar daar ligt ook meteen de crux: hoe krijgen we een goede semantische annotatie? Volgens Hovy is het belangrijker om een annotatie te hebben die consistent is, dan een annotatie die heel gedetailleerd is, maar waarbij verschillende annotatoren het moeilijk met elkaar eens kunnen worden.

RITEL

Sophie Rosset, Olivier Galibert, Gabriel Illouz en Aurélien Max van LIMSI in Orsay presenteerden hun onderzoek in het RITEL-project, dat net als IMIX gericht is op het maken van een vraag-antwoordsysteem waarin de vraag van de gebruiker in een dialoog met het systeem gepreciseerd en vervolgens beantwoord kan worden. Ook RITEL is gebaseerd op hergebruik van bestaande technologie. Omdat RITEL in zijn huidige vorm alleen spraak-gebaseerde interactie over een telefoonverbinding ondersteunt, hechten de onderzoekers groot belang aan het voorkomen van gaten tussen het einde van een vraag van de beller en het antwoord van het systeem. Vertraging in de spraakherkenner wordt voorkomen door meteen als het begin van de spraak gedetecteerd is met herkenning te beginnen en een herkenner te gebruiken die sneller werkt dan real-time.
De output van de spraakherkenner wordt geanalyseerd en geïnterpreteerd met behulp van een groot aantal (op dit moment nog met de hand gemaakte) templates. Dat maakt het systeem snel, maar ook weinig flexibel. Als het systeem ontdekt dat een vraag ambigu is, stelt het een wedervraag om uit te vinden wat de beller bedoelt. Van sommige dingen weet het systeem dat ze niet uniek zijn. In andere gevallen ontdekt het systeem het probleem doordat er verschillende, niet met elkaar overeenstemmende antwoorden terugkomen. Een interessante maatregel om vertraging te voorkomen is het achterwege laten van een dialoogmanager die gaat redeneren over de betekenis van een vraag. Feitelijk is RITEL een dialoogsysteem zonder dialoogmanager. Wat in andere systemen door een dialoogmanager gedaan wordt, wordt in RITEL afgehandeld door op een slimme manier een keuze te maken uit een aantal vaste scenario’s voor de voortzetting

“Er schemert soms onzekerheid door in de antwoorden.”

van het gesprek. Tenslotte onderscheidt RITEL zich nog van de meeste vraag-antwoordsystemen (maar niet van IMIX) door in het antwoord soms onzekerheid te laten doorschemeren. Op de vraag *Van wie is de song Yesterday?* zou het antwoord iets kunnen zijn als *Waarschijnlijk van de Beatles, maar er zijn ook tientallen uitvoeringen van andere artiesten.*

Q-GO

Uiteraard werd het ontwerpen van vraag-antwoordsystemen in de presentatie van Q-go op de eerste plaats benaderd vanuit de behoeften van bedrijven die informatiediensten aanbieden en hun klanten. Diensten die gebruikmaken van technologie van Q-go, handelen per maand zo’n 3.5 miljoen vragen af, waarvan 600.000 in het Nederlands. In die praktijk blijkt dat veel mensen alleen een klein aantal trefwoorden intypen, in plaats van een complete vraag. Daardoor wordt het lastig om de query te genereren die de behoefte van de klant het beste weerspiegelt. Als de vooralsnog experimentele versie van het Q-go-systeem merkt dat het geen complete query kan vormen, of dat er te veel verschillende antwoorden terugkomen, trekt het het initiatief naar zich toe door om de ontbrekende input te vragen.

LOU BOVES IS VERBONDEN AAN HET CLST (RADBOUD UNIVERSITEIT NIJMEGEN)

’n Reële mogelijkheid of toekomst muziek?

Automatisch beantwoorden van E-mails

Achtergrond

Bedrijven en organisaties krijgen steeds meer e-mails: logisch want het is makkelijk, je kunt het doen wanneer je wilt en het is (bijna altijd) gratis. Bij nadere bestudering van de inhoud van de e-mails, blijkt een grote overeenkomst met de gesprekken in het call center: 80% van de e-mails gaan over 20% van de onderwerpen. Dat houdt in dat, wanneer je er in slaagt om voor de 20% meest gestelde vragen een algemeen, passend antwoord te maken, je voor 80% van de binnen komende e-mails een kant-en-klaar antwoord hebt.

| MICHEL BOEDELTJE,
ARJAN VAN HESSEN |

Dit is ook wat er gebeurt in de zogeheten contact centers. Medewerkers lezen de e-mail en beantwoorden hem middels een voorgedefinieerd antwoord.

STEEKWOORDEN

Er zijn twee manieren waarop het juiste antwoord gezocht kan worden: door de medewerker of door de computer. In de praktijk gaat het echter anders: de computer “leest” de e-mail en schotelt de medewerker op basis van trefwoorden in de e-mail een aantal suggesties voor. De medewerker bekijkt de suggesties en kiest het juiste antwoord. Zolang het aantal suggesties dat de computer voorschotelt klein is (≤5, anders moet de medewerker te veel lezen) en het juiste antwoord er meestal (>80%) bij zit, werkt deze aanpak goed en er zijn verschillende software pakketten te koop die dit zo doen.
Anders wordt het wanneer je meer dan 10 suggesties moet tonen om in slechts de helft van de gevallen er het juiste antwoord uit te kunnen halen: dan wordt het lezen van de suggesties en het alsnog zelf zoeken van het juiste antwoord te tijdrovend en had je het antwoord beter zelf kunnen schrijven.

UITDAGING

Het afstudeeronderzoek van Michel Boedeltje bij Em@ilco in Amersfoort is gestart als een soort wedstrijd: laat zien dat IR (Information Retrieval) technologie in combinatie met taaltechnologie een beter resultaat kan opleveren dan de op steekwoorden gebaseerde methode. Bij de “steekwoorden methode” maakt een mens de keuze om mails met bepaalde woorden aan een bepaald antwoord te koppelen. Staan in de e-mail bijvoorbeeld de woorden “opzeggen” & “internet” dat wordt het standaardantwoord geselecteerd voor mensen die hun internetabonnement willen opzeggen.
Zolang het niet om te gevarieerde mails gaat, werkt dit aardig, maar de resultaten bij Em@ilco lieten zien dat er in de loop van de tijd veel vervuiling optreedt waardoor de resultaten langzaam terug lopen. Mensen zijn blijkbaar niet in staat om voor een verzameling van meer dan 10K (=10.000) e-mails de juiste steekwoorden aan de verschillende standaardantwoorden te koppelen.

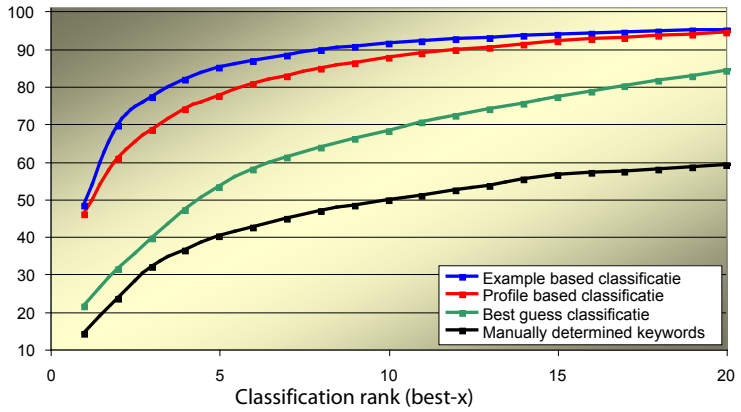
DATA

De gekozen aanpak was als volgt: verdeel de hele verzameling van 17K (=17.000) e-mail s van de Nationale Postcode Loterij (van iedere e-mail was in theorie het juiste antwoord bekend) in een trainingsdeel (80%) en een testdeel (20%). Laat vervolgens de computer zelf bepalen van welke woorden de aan- en afwezigheid relevant is voor een bepaald antwoord (=klasse). Verander daarbij de test en trainingsgroep een aantal keren zodat er geen toevalligheden optreden en kijk of het eindresultaat beter is. Vermeld moet worden dat de database suboptimaal geclassificeerd was. Van veel e-mails waarvoor het bestaande systeem geen juiste suggestie gaf, was het juiste antwoord op een andere manier aan de afzender gestuurd zonder dit juiste

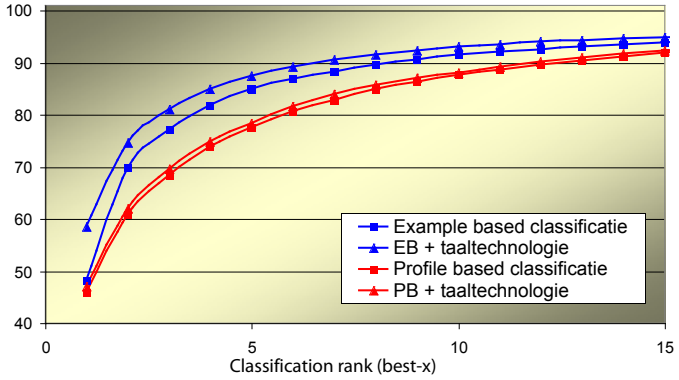
antwoord in de database te noteren. Hierdoor was niet van alle e-mails het juiste antwoord bekend. Ook bleek dat medewerkers vaker dan gedacht verkeerde antwoorden gaven. Tenslotte bleek dat er nogal wat overlap zat in de verschillende categorieën waardoor het, ook voor de medewerkers, niet altijd duidelijk was of een e-mail in de ene of de andere klasse viel.

AANPAK

Deze aanpak klinkt iets simpeler dan het in werkelijkheid is, maar in de basis kwam het hier op neer. Om de rekentijd te verlagen, is het zinvol om de zogeheten stopwoorden (woorden die voor het betekenisonderscheid niet of minder relevant zijn zoals “de”, “het”, “wil”, “mogen”, “wij” etc.) eerst te verwijderen. Vervolgens kan het zinvol zijn de overgebleven woorden te stemmen (fietsen, fiets, fietsje _fiets) om zo de variantie te verminderen. De applicatie moet dan op basis van de resterende woorden bepalen hoe de grenzen tussen de verschillende klassen zo getrokken moeten worden opdat de meeste e-mails in de juiste klasse zouden komen en dus het beste antwoord zouden krijgen.



Figuur 1: de resultaten van de verschillende classificatiemethoden vergeleken met de oorspronkelijke, op steekwoorden gebaseerde aanpak. Wanneer we 5 suggesties op het scherm zetten, stijgt de kans van slagen van 40% naar 85%: meer dan een verdubbeling!



Figuur2: de resultaten van de twee gebruikte IR-methoden met en zonder gebruik van taaltechnologie. Hoewel het verschil afneemt wanneer veel suggesties worden gegeven, is het gebruik van Taaltechnologie zinvol bij volledige self-service waarbij slechts één of twee antwoorden worden gegeven.

RESULTAAT

De resultaten op de 17K e-mails van de Postcode Loterij waren verbazingwekkend goed. Voor ieder aantal suggesties was de kans op succes (d.w.z. het juiste antwoord zit bij de suggesties) meer dan verdubbeld. De kans dat het juiste antwoord erbij zit is bij de nieuwe methode met slechts 2 suggesties al beter dan met 20 suggesties bij de oude methode.

De eerste resultaten (slechts gebaseerd op IR-technologie) waren zo goed, dat het toepassen van allerlei taaltechnologie eigenlijk niet zinvol meer leek. Toch hebben we het gedaan en de resultaten werden er alleen maar beter van (hoewel de stijging natuurlijk minder spectaculair was). Zoals uit figuur 2 blijkt is het eerste antwoord in bijna 60% van de gevallen ook het juiste antwoord. Als we dit percentage nog iets kunnen verhogen, dan komt echte self-service (je stuurt een e-mail en het systeem geeft je het (waarschijnlijk) juiste antwoord) binnen bereik.

VERVOLGONDERZOEK

Hoewel Michel Boedeltje met dit onderzoek zijn studie zeer succesvol heeft afgesloten en nu bezig is de zelfde techniek bij Telecats op gesproken vragen toe te passen, ligt het vervolgonderzoek voor de hand: kun je iets zeggen over de betrouwbaarheid waarmee een antwoord gesuggereerd wordt. Het is waarschijnlijk dat e-mails die qua woordgebruik erg lijken op reeds beoordeelde e-mails, een hoge betrouwbaarheid zullen krijgen. Als dit zo is, wat wordt dan het slagingspercentage als functie van de betrouwbaarheid. Stel dat voor zeer betrouwbare antwoorden het slagingspercentage (de suggestie is juist) 90% is, dan kan overwogen worden om (al dan niet buiten kantoortijden) de e-mailers automatisch antwoord te geven. Dit moet dan uiteraard gepaard gaan met de mededeling dat het antwoord automatisch gegenereerd is en dat men, als het antwoord niet goed is, de vraag nogmaals kan sturen zonder dat men in een soort “loop” terechtkomt.

CONCLUSIE

Het hier gepresenteerde afstudeerwerk laat ‘n aantal zaken duidelijk zien.

- De methode werkt goed, ondanks het feit dat de verzameling waarmee het systeem getraind is (e-mails-met-antwoord) niet 100% correct is.
- De combinatie van IR en taaltechnologie biedt zeer veel mogelijkheden voor het geheel automatisch beantwoorden van (een deel) van de binnenkomende e-mail
- Het samenwerken van zowel grote als kleine bedrijven met universiteiten kan zeer lonend zijn. Enthousiaste studenten kunnen op deze manier de op de universiteiten aanwezige kennis direct voor bedrijven geschikt en toegankelijk maken.
- Volledig automatische selfservice op een deel van de binnenkomende e-mails onder bepaalde omstandigheden mogelijk is.

MICHEL BOEDELTJE IS VERBONDEN AAN TELECATS, ARJAN VAN HESSEN IS VERBONDEN AAN DE UNIVERSITEIT VAN TWENTE EN TELECATS

NOTaS

Stichting NOTaS
Postadres: Postbus 31070,
6503 CB NIJMEGEN
T: 024 – 352 88 88
W: www.notas.nl
E: notas@malta-online.nl

NOTaS behartigt de belangen van bedrijven en kennisinstellingen die actief zijn op het terrein van Taal- en Spraaktechnologie (TST) .
Dit doet zij o.a. door middel van bijeenkomsten, lobby-activiteiten en het tijdschrift DIXIT.

Deelnemers Stichting NOTaS

CLST
Adres: Erasmusplein 1,
6525 HT NIJMEGEN
T: 024 – 361 16 86
W: www.ru.nl./clst.
E: clst@let.ru.nl

Het Centre for Language and Speech Technology (CLST) onderzoekt en adviseert op het gebied van taal- en spraaktechnologie. Tevens is CLST gespecialiseerd in de productie en validatie van taal- en spraakdatabases (SPEX).

Comsys B.V.
Adres: Laan van Blussé van oud Alblas 2a, 3769 AT SOESTERBERG
T: 033 – 445 22 00
W: www.comsys.nl
E: info@comsys.nl

Comsys levert, integreert en onderhoudt voice response systemen (IVR's), spraakgestuurde self-service oplossingen, multimedia contact centers en netwerk services. Hiermee richt Comsys zich voornamelijk op handelsondernemingen, service organisaties, financiële instellingen en telecom operators.

Dedicon
Adres: Traverse 175, 5361 TD GRAVE
T: 0486 – 48 64 86
W: www.dedicon.nl
E: info@dedicon.nl

Dedicon is in Nederland dé specialist in het maken van aantrekkelijke alternatieve leesvormen. Van luisterboeken tot multimedia-uitgaven, Dedicon maakt het nieuwe lezen mogelijk.

Dutchear
Adres: Brassersplein 2,
2612 CT DELFT
T: 015 – 219 11 11
W: www.dutchear.nl
E: info@dutchear.nl

Dutchear staat voor inleving in uw business, creativiteit, gebruikerskennis en state-of-the-art technische oplossingen. Dutchear is een jong bedrijf met veel ervaring in het toepassen van spraaktechnologie voor selfservice en routeringsoplossingen.

Em@ilco.nl BV
Adres: Koningin Wilhelminalaan 1,
3818 HN AMERSFOORT
T: 033 – 460 02 00
W: www.emailco.nl
E: cornelisse.em@ilco.nl

Em@ilco is sinds 1999 succesvol actief met de ontwikkeling van producten en diensten op het gebied van elektronische klantcommunicatie.

Gridline BV
Adres: Keizersgracht 520 sous,
1017 EK AMSTERDAM
T: 020 – 616 20 50
W: www.gridline.nl
E: info@gridline.nl

Gridline is een ICT-bedrijf dat zich specialiseert in taaltoepassingen. Wij zijn expert in terminologie, en bouwen daarmee bedrijfsspecifieke toepassingen. Gridline is van oorsprong een database- en webapplicatiebedrijf, met veel harde IT-kennis. Wij werken graag samen met universiteiten en collegabedrijven, en willen dat via ons NoTaS-lidmaatschap uitbouwen.

Human Inference Enterprise B.V.
Adres: Velperweg 8,
6824 BH ARNHEM
T: 026 – 355 06 55
W: www.humaninference.nl
E: info@humaninference.nl

De software van Human Inference optimaliseert uw bedrijfsprocessen door de verbetering en borging van de kwaliteit van uw gegevens.

Intaal B.V.
Adres: Winthonlaan 198,
3526 KV UTRECHT
T: 030 – 750 89 60
W: www.intaal.nl
E: info@intaal.nl

inTAAL is een laagdrempelig en klantgericht bedrijf voor spraak- en taaltechnologie en ondersteunde communicatie met name voor mensen met een handicap of beperking.

Knowledge Concepts
Adres: De Handboog 9,
5283 WR BOXTEL
T: 041 16 – 108 02
W: www.knowledge-concepts.com
E: sales@knowledge-concepts.com

Knowledge Concepts designs and builds solutions for information retrieval and the monitoring and managing of information and knowledge. We do this using a combination of datasets, linguistic tools, search engines (for both text and audio files) and classifiers.

Logica CMG
Adres: Prof. W.H. Keesomlaan,
1183 DJ AMSTELVEEN
T: 020 – 503 30 00
W: www.logicacmg.com

Als internationale ICT-dienstverlener en haar uitgebreide staat van dienst en branchekennis helpt LogicaCMG haar klanten een leiderschapspositie in te nemen. Het bedrijf is actief op het gebied van business consultancy, systeemintegratie en IT- en business process outsourcing.

Maartenskliniek/Ontwikkelcentrum voor Spraak- en Taaltechnologie/Research Development & Education BV
Adres: Hengstdal 3,
6522 JV NIJMEGEN
T: 024 – 365 91 88
W: www.ostt.eu
E: ostt@maartenskliniek.nl

Het Ontwikkelcentrum voor Spraak- en Taaltechnologie maakt deel uit van de afdeling Research Development & Education van de Sint Maartenskliniek. De activiteiten binnen het Ontwikkelcentrum zijn gericht op toepassing van nieuwe ontwikkelingen op het gebied van spraak- en taaltechnologie ten behoeve van mensen met communicatieve beperkingen. Het OSTT fungeert als platform voor alle belanghebbenden op het gebied van TST ten behoeve van communicatieve beperkingen.

Polderland Language & Speech Technology
Adres: Kerkenbos 11-03A,
6546 BC NIJMEGEN
T: 024 – 352 28 66
W: www.polderland.nl
E: info@polderland.nl

Wanneer heeft u zich voor het laatst geërgerd aan en slechte tekst? Of aan het niet kunnen vinden van de juiste informatie? Polderland zorgt ervoor dat u wordt ondersteund in alledaagse activiteiten zoals het schrijven van een brief of het vinden van informatie op een website. Met producten als spellingcontrole, woordenboeken en zoektechnologie maakt Polderland taaltechnologie tastbaar.

Q-go.com B.V.
Adres: Eekholt 40, 1112 XH DIEMEN
T: 020 – 531 38 00
W: www.q-go.nl
E: contact@q-go.com

Q-go is een toonaangevende Europese aanbieder van oplossingen voor Customer Interaction Management. Met deze oplossingen verbeteren banken, (zorg)verzekeraars, pensioenverzekeraars, telecombedrijven en logistieke dienstverleners hun online klantenservice en klanttevredenheid en verlagen zij hun kosten. De applicaties van Q-go zijn het resultaat van meer dan 100 manjaren onderzoek door experts op het gebied van natuurlijke taalverwerking. Q-go wendt 40% van haar opbrengsten aan voor R&D.

Sabel Communicatie bv
Adres: Begijnekade 6,
3512 VT UTRECHT
T: 030 – 299 30 73
W: www.sabelcommunicatie.nl
E: info@sabelcommunicatie.nl

Wat wilt u bereiken? Dat meer mensen uw organisatie kennen? Dat ze uw plannen begrijpen, begrip hebben voor uw beleid of geïnteresseerd raken in uw product? Sabel Communicatie helpt u uw doelstellingen te realiseren. We zijn een bureau voor geschreven communicatie, dat kernwaarden in de juiste kernwoorden vat.

Stichting Studio Taalwetenschap
Adres: Kazernestraat 29,
1018 CC AMSTERDAM
T: 020 – 639 01 15
W: www.studiotaalwetenschap.nl
E: stichting@studiotaalwetenschap.nl

Zoekt u een taalwetenschapper voor een onderzoek, een rapport, een lezing of een advies? Via Stichting Studio Taalwetenschap kunt u de juiste specialist vinden. Bent u taalwetenschapper en heeft u een project waarvoor u een afnemer zoekt? Stichting Studio Taalwetenschap zoekt en denkt met u mee.

Telecats
Adres: Colosseum 42,
7521 PT ENSCHEDE
T: 053 – 488 99 00
W: www.telecats.nl
E: info@telecats.nl

Telecats ontwikkelt en implementeert turnkey oplossingen om de afhandeling van telefoongesprekken geheel of gedeeltelijk te automatiseren. Telecats wijst u de weg in de wereld van interactive voice response (IVR) en spraaktechnologie en VoIP uitgaande van: omzetverhoging, serviceverbetering en backoffice-integratie.

TST- Centrale, p/a Instituut voor Nederlandse Lexicologie
Adres: Matthias de Vrieshof 2-3,
2311 BZ LEIDEN
T: 071 – 514 16 48
W: www.tst.inl.nl
E: tst@inl.nl

Wanneer u op zoek bent naar (informatie over) digitale taalkundige bronnen dan bent u bij ons aan het juiste adres. Of u voor een kennisinstelling werkt of voor het bedrijfsleven, of u geïnteresseerd bent in taal of in spraak, of u een taalkundige bent of een spraaktechnoloog, wij zijn u graag van dienst.

Universiteit Twente
Adres: Drienerlolaan 5,
7522 NB ENSCHEDE
T: 053-4899111
F:053-4892000
W: www.utwente.nl

De Universiteit Twente (UT) is een ondernemende researchuniversiteit. Als enige campusuniversiteit in Nederland verzorgt de UT onderwijs en onderzoek in wetenschapsgebieden die variëren van psychologie en bestuurskunde tot technische natuurkunde en biomedische technologie.

UvT Faculteit Letteren
Adres: Warandelaan 2,
5037 AB TILBURG
T: 013 – 466 91 11
W: www.uvt.nl/communicatie-en-cultuur
E: communicatie-en-cultuur@uvt.nl

In onderwijs en onderzoek ligt het accent sterk op de maatschappelijke toepassingen van taal, informatie, cultuur en communicatie. Belangrijke aandachtsgebieden zijn interculturele communicatie (toegespitst op Nederlands als tweede taal), tekstwetenschap en communicatie, taaltechnologie en kunstmatige intelligentie, en cultuur en literatuur.

Van Dale Lexicografie bv
Adres: St. Jacobsstraat 127,
3511 BP UTRECHT
T: 030 – 232 47 11
W: www.vandale.nl
E: info@vandale.nl

Taal zal altijd een wezenlijk onderdeel van communicatie zijn. Iedereen kent de Grote of Dikke van Dale, maar Van Dale Lexicografie is inmiddels veel meer dan dat ene dikke woordenboek. Met zorgvuldigheid en respect voor de moderne Nederlandse taal, maar ook voor de grote levende talen beschrijft Van Dale de taal zo compleet mogelijk, met alle middelen die de uitgeverij ter beschikking staan.

Viataal - R&D Group
Adres: Theerestraat 42,
5271 GD SINT MICHIELSGESTEL
T: 073 – 558 81 11
W: www.viataal.nl
E: info@viataal.nl

Viataal zet zich in voor mensen die als gevolg van een beperking problemen ondervinden met hun communicatie. Dat doen we door voorlichting, consultatie, diagnostiek en ondersteuning, onderwijs, behandeling en begeleiding op het gebied van wonen, leren, werken en vrije tijd.

Sponsoren/Samenwerking

Ontwikkelingsmaatschappij Oost Nederland
Adres: Hengelosestraat 585,
7521 AG ENSCHEDE
T: 053 – 484 96 49
W: www.oostnv.nl
E: info@oostnv.nl

Ontwikkelingsmaatschappij Oost Nederland is een NV die door middel van allerlei activiteiten en projecten de economie van Oost-Nederland versterkt en daarmee de werkgelegenheid bevordert. Wij werken voor het Gelderse en Overijsselse bedrijfsleven in opdracht van het Ministerie van Economische Zaken en de Provincies Gelderland en Overijssel.

Nederlandse Taalunie
Adres: Lange Voorhout 19,
2514 EB DEN HAAG
T: 070 – 346 95 48
W: http://taaluniversum.org/taalunie
E: info@taalunie.org

De Nederlandse Taalunie is een beleidsorganisatie waarin Nederland, België en Suriname samenwerken op het gebied van de Nederlandse taal, onderwijs en letteren. De Taalunie bevordert onder meer de gemeenschappelijke ontwikkeling van digitale taalmaterialen die nodig zijn voor nieuwe toepassingen, bijvoorbeeld op het gebied van taal- en spraaktechnologie.

NWO
Adres: Laan van Nieuw Oost-Indië 300, 2593 CE DEN HAAG
T: 070 – 344 06 40
W: www.nwo.nl
E: nwo@nwo.nl

Contactpersoon NOTaS:
E: dijkstra@nwo.nl

De Nederlandse Organisatie voor Wetenschappelijk Onderzoek: heeft tot taak het bevorderen van de kwaliteit en vernieuwing van wetenschappelijk onderzoek, alsmede het initiëren en stimuleren van nieuwe ontwikkelingen in het wetenschappelijk onderzoek.

Senternovem
Adres: Juliana van Stolberglaan 3,
2595 CA DEN HAAG
T: 070 – 373 50 00
W: www.senternovem.nl
E: frontoffice@senternovem.nl

SenterNovem levert een bijdrage aan duurzame ontwikkeling en innovatie door een brug te slaan tussen markt en overheid, nationaal en internationaal. Op professionele wijze voert SenterNovem overheidsbeleid uit rond innovatie, energie & klimaat en milieu & leefomgeving. Bedrijven, instellingen en overheden kunnen bij SenterNovem terecht voor het realiseren van maatschappelijke doelstellingen op deze terreinen.

Stevin
Adres: Lange Voorhout 19,
2514 EB DEN HAAG
T: 070 – 346 95 48
W: http://taalunieversum.org/stevin/

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. STEVIN wordt gecoördineerd en financieel beheerd door de Nederlandse Taalunie.

Ondersteuning

Cumlingua
Adres: Dr. Kanterslaan 126,
5361 NJ GRAVE
T: 0486 – 471 554
W: www.cumlingua.nl
E: info@cumlingua.nl

- Advies en begeleiding op het gebied van marktgericht denken
- Vertaal- en correctiediensten
- Taalles op maat

Design Print
A.: De Vlotkampweg 32,
6545 AG NIJMEGEN
T: 024-3774324
W: www.designprint.nl
E: info@designprint.nl

Als geen ander beseft Design Print dat ieder drukwerk een presentatie op zich is. Vanaf het visitekaartje, brochure, magazine, poster, verpakking en zelfs een simpele bijsluiter of flyer wordt bij Design Print een presentatie op zich en de spiegel van uw emotie. Ook deze DIXIT is door Design Print gedrukt.

Deykerhoff Accountants
Adres: Toernooiveld 300,
6525 EC NIJMEGEN
T: 024-352 88 06
W: www.deykerhoff.nl
E: nijmegen@deykerhoff.nl

Deykerhoff/Accountants is een dynamisch accountantskantoor dat eigenzinnigheid paart aan betrouwbaarheid. Onze pro-actieve instelling is gericht op een grote persoonlijke betrokkenheid en een op maat gesneden persoonlijk contact met de klanten.

Malta & de Keyzer
Adres: Toernooiveld 300,
6525 EC NIJMEGEN
T: 024 – 352 88 88
F: 024 – 354 00 90
W: www.malta-online.nl
E: info@malta-online.nl

Een probleem op het gebied van secretariaat of office management? Wij lossen het op. Snel, professioneel en helemaal volgens uw persoonlijke wensen. Kijk op onze website voor meer informatie.

NOTaS nieuws

Wanneer we terugkijken naar NOTaS in 2006 zien we dat onze organisatie op verschillende fronten aan het veranderen is o.a.:

De deelnemersbijeenkomsten.

Geen gewone vergaderingen meer, maar boeiende thema-bijeenkomsten:

Speed dating

In juni waren bijna alle deelnemers van NOTaS aanwezig bij de Volksuniversiteit van Utrecht voor een speed date-bijeenkomst. Meer over deze zeer succesvolle bijeenkomst kunt u lezen in deze DIXIT.

TST-in de Zorg

Onze tweede bijeenkomst "nieuwe stijl" werd georganiseerd bij Viataal, het centrum voor zorg en onderwijs voor mensen met beperkingen in horen en zien en communicatie te Sint Michielsgestel. Na diverse presentaties over de TST-middelen die nu beschikbaar zijn in de zorg kwam er een boeiende paneldiscussie over wat wel of niet haalbaar is en waar de speerpunten liggen. Voor deze gespecialiseerde bijeenkomst waren ook nieuwe deelnemers en gasten uit de zorgsector uitgenodigd en het werd snel duidelijk dat de zorgsector voor zeer veel NOTaS-deelnemers belangrijk is. In het panel zaten vertegenwoordigers uit de zorgsector, het bedrijfsleven en de kennisinstellingen: Joost van den Broek, Intaal; Vincent de Jong, Dedicon; Lilian Beijer, St. Maartenskliniek
Hans van Balkom, Viataal en Catia Cucchiarini van de Nederlandse Taalunie. Deze laatste zorgde ervoor dat er een discussie op gang kwam over de conclusie van het NTU rapport TST en communicatieve beperkingen waarover tijdens de bijeenkomst een presentatie is gegeven. In het kader van een meerjarenprogramma is de NTU bezig met een mogelijk vervolg hierop.
U kunt gratis een kopie van het rapport aanvragen via www.taaluniversum.org.

Beeld en geluid

Begin 2007 zal de ledenvergadering van NOTaS plaatsvinden in het splinternieuwe mediacentrum van Beeld en Geluid op het omroepterrein in Hilversum. Behalve de noodzakelijke (korte) vergaderzaken, zal speciale aandacht besteed worden aan de rol van Taal en Spraaktechnologie bij het beheren en ontsluiten van audiovisuele archieven. De gastsprekers zullen aandacht besteden aan de verschillende facetten die hierbij komen kijken. De bijeenkomst zal worden afgesloten met een rondleiding door het indrukwekkende, nieuwe gebouw.

Een groei in aantal en in diepte

Ketenbenadering

Met de komst van steeds meer afnemers van TST en hun vertegenwoordigers mag NOTaS nu van een ketenbenadering spreken. Onder de deelnemers van NOTaS kan men onderzoekers van kennisinstellingen, ontwikkelaars en applicatiebouwers van technologiebedrijven, eindafnemers en gebruikers vinden. Vooral bij deze laatste groep willen we steeds meer de discussie aangaan over waar de marktbehoeftes liggen en welke rol TST hierin kan spelen. Heeft u klanten of partners voor wie taal- en spraaktechnologie een rol speelt in hun producten, processen en/of diensten? Laat het ons even weten (notas@malta-online.nl). We sturen graag een informatieset toe.

Samenwerking

Niet alleen tijdens de speeddate-bijeenkomst, maar ook daarbuiten weten NOTaS-deelnemers elkaar steeds vaker te vinden op zowel het gebied van onderzoek en nieuwe ontwikkeling als op het commerciële vlak. Helaas blijkt toch steeds opnieuw dat er technisch veel meer mogelijk is dan de meesten denken. Universiteiten besteden uiteraard veel tijd en aandacht aan onderzoek naar nieuwe technologieën maar "vergeten" dikwijls de "gewone" stervelingen hiervan op de hoogte te brengen. Om hier iets aan te doen, zijn we van plan om in 2007 een speciale bijeenkomst voor TST-onderzoekers en ontwikkelaars te organiseren. Tijdens deze bijeenkomst zullen de kennisinstellingen onder ons de kans krijgen zich te presenteren en hun laatste onderzoeksprojecten toe te lichten. Het doel hiervan is dus om bedrijven en potentiële klanten te stimuleren samen met de kennisinstellingen deze nieuwe mogelijkheden op het gebied van (toegepaste) Taal en Spraaktechnologie te gaan "uitontwikkelen" en om onderzoeksresultaten eventueel om te toveren tot verkoopbare componenten en/of applicaties. U hoort hier t.z.t. meer over.

In onze communicatie:

Jaarboek

Last but not least, zijn we natuurlijk trots om voor het eerste keer een TST-Jaarboek aan u te overhandigen, in de vorm van een gesponsorde DIXIT waarvoor we onze dank geven STEVIN. Dank ook aan onze gastredacteur, Peter Spyns van de Taalunie en iedereen die een bijdrage heeft geleverd aan deze extra dikke editie.

2007: nog meer kansen

In 2007 komen er nog nieuwe uitdagende calls vanuit het STEVIN programma. NOTaS en de Taalunie blijven u hierover informeren. Zorg ervoor dat u regelmatig met mogelijke partners in contact komt door de diverse bijeenkomsten bij te wonen die gedurende het hele jaar door NOTaS, STEVIN en diverse NOTaS-deelnemers, o.a. de TST-centrale worden georganiseerd. Een belangrijke bron van alle events is ook de nieuwsbrief van de Taalunie. Indien u deze nog niet ontvangt kunt u hem gratis aanvragen via de website van de Taalunie:www.taaluniversum.org of via info@taalunie.org.

Namens het Bestuur wens ik u een gelukkig, succesvol en inspirerend 2007 toe.

DEBBIE KENYON-JACKSON
VOORZITTER

U wilt graag weten waarom uw klanten telefonisch contact opnemen met uw organisatie. Om dit te achterhalen is het niet langer noodzakelijk om met verschillende telefoonnummers te werken of om keuzemenu's te gebruiken. CALLERSVOICE *is de oplossing. Met* CALLERSVOICE *stelt u uw bellers een open vraag: "Geef kort en bondig aan waarvoor u belt". Met behulp van geavanceerde spraaktechnologie en zoekfuncties, wordt direct duidelijk waarvoor uw klant belt. Door deze informatie te koppelen aan het klantprofiel en business rules, kan elk klantcontact optimaal worden afgehandeld.*

TELECATS

Luister naar de stem van uw klanten: hear the Caller's Voice!

Telefonische dienstverlening is gebaseerd op:

- **Identificatie van de beller: wie belt?**
- **Classificatie van de wens: wat is de vraag, het verzoek of de klacht?**
- **Autorisatie: is de wens van deze beller via dit kanaal toegestaan?**
- **Serviceverlening: welke dienst bied ik deze beller mede op basis van business rules?**

In het streven naar kostenreductie, omzetvergroting en toename van de klant-tevredenheid, is het essentieel om elke klant de best passende service te verlenen. Hierbij is het van belang om de bellers bij de juiste selfservice applicatie te krijgen en om de bellers waarvoor geen selfservice mogelijk of wenselijk is, door te verbinden met een gekwalificeerde medewerker. Hiervoor is dus een goede *classificatie* vereist.

Bij complexe organisaties is classificatie door middel van hiërarchische keuzemenu's niet eenvoudig. De IVR menu's worden al gauw te breed (te veel opties in één vraag) of te diep (te veel vragen achter elkaar). Daarbij hebben bellers veelal moeite met het maken van de juiste keuze of het gevoel dat hun vraag zo specifiek is, dat deze niet in het menu voorkomt. Deze bellers kiezen voor de categorie "overig".

CALLERSVOICE geeft bellers de mogelijkheid om hun vraag of probleem in hun *eigen woorden* kenbaar te maken. Hiertoe wordt de beller een open vraag gesteld: "Welkom bij de bank. Spreek alstublieft kort en bondig in waarvoor u belt".

Het is wenselijk om de classificatie te combineren met identificatie. Het systeem vraagt de klant hiervoor eerst een klant-, polis- of rekeningnummer in te voeren. Ook is het mogelijk om op basis van postcode en huisnummer een beller te identificeren. De identiteit en het classificatieresultaat kunnen worden aangevuld met kennis over de beller, zoals het klantprofiel (b.v. goud, zilver of brons) en de status (b.v. een betalingsachterstand). Op basis van deze gegevens en de business rules wordt de juiste dienstverlening voor de desbetreffende klant bepaald.

De voordelen van CALLERSVOICE op een rijtje:

- **Snellere en betere classificatie t.o.v. keuzemenu's**
- **Meer gesprekken per telefoonlijn**
- **Betere benutting van bestaande en nieuwe selfservice applicaties**
- **Verbetering van de (eerste) routering / vermindering herroutering**
- **Positief effect op de One Call Resolution**
- **Positieve invloed op de klanttevredenheid en werknemerstevredenheid**
- **Snel inspelen op ad hoc ontwikkelingen en calamiteiten**
- **Innovatief imago**
- **Kostenbesparing**

CALLERSVOICE is een product van TeleCats. Voor meer informatie bel: 053 488 99 00 of stuur een e-mail naar info@callersvoice.nl

WWW.CALLERSVOICE.NL