Luisterend leren met een hybride boek

Een boek op je computerscherm. Een boek dat door een professionele voorlezer wordt voorgelezen terwijl op het scherm een kleurbalk meeloopt over de tekst. Dat is het nieuwe hybride boek dat Dedicon heeft ontwikkeld.

Ik begrijp zo veel meer

Met een hybride boek kunnen dyslectische leerlingen makkelijker lezen. Dat bleek uit een onderzoek dat Dedicon begin 2008 uitvoerde. “Ik kan me beter op de tekst concentreren.” En “Ik begrijp zo veel meer van de tekst.” oordeelden de proefpersonen.

Nog beter

Maar het kan allemaal nog beter, dus ontwikkelt Dedicon door. Samen met Telecats wordt gewerkt aan automatische synchronisatie van gesproken en weergegeven teksten. Zodra dit project is afgerond, kan Dedicon op korte termijn starten met de geautomatiseerde productie van hybride schoolboeken.

dedicon

Over Dedicon

Dedicon ontwikkelt en produceert lectuur en informatie in aangepaste leesvormen voor mensen met een leeshandicap. Hierbij gaat het om mensen met een visuele handicap of dyslexie, maar ook om mensen met een lichamelijke beperking, zoals afasie, Parkinson of reuma. Voor hen maakt Dedicon boeken, kranten, tijdschriften, schoolboeken, muziek, folders en brochures in de vorm die bij hen past.

Dedicon werkt in opdracht van partijen als de Vereniging Openbare Bibliotheken, OCW en alle andere bedrijven en instellingen die hun informatie in toegankelijke vorm willen aanbieden. Door voortdurende ontwikkeling, onderzoek en internationale samenwerkingsprojecten maakt Dedicon innovatieve producten die ervoor zorgen dat iedereen kan lezen wat, wanneer en hoe hij dat wil.

Meer informatie is te vinden op www.dedicon.nl

DIXIT

TIJDSCHRIFT OVER TOEGEPASTE TAAL- EN SPRAAKTECHNOLOGIE

STEVIN en ONDERWIJS

Spraak- en Taaltechnologische Essentiële Voorzieningen in het Nederlands

Het klaslokaal van de toekomst?

Met TST extra motiverend!

JAARGANG 5 NUMMER 2 NOVEMBER 2008

INHOUD

STEVIN

Onderwijs

Voorwoord

Alle facetten van Taal- en Spraaktechnologie in een notendop

Er is de laatste jaren veel gebeurd in de Nederlandse TST-gemeenschap. De eerste STEVIN onderzoeks- en ontwikkelingsprojecten zijn afgerond en hun resultaten, samen met die van de demon-stratieprojecten, komen nu beschikbaar.

Samenwerking centraal

Het STEVIN-programma heeft, behalve tastbare resultaten (producten, (wetenschappelijke) artikelen, conferentiebijdragen, media-aandacht), vooral de onderlinge samenwerking van kennisinstellingen en bedrijven in zowel Vlaanderen als Nederland sterk verbeterd.

Steeds meer (potentiële) klanten komen, mede dankzij het STEVIN-programma, in aanraking met zegeningen van taal- en spraaktechnologie (TST). En dit heeft al tot de oprichting van een nieuw bedrijf geleid! Net als in andere gebieden (bijv. Nano- en Biotechnologie) zien we ook bij TST dat samenwerking tussen kennisinstellingen, technologie-ontwikkelaars en eindgebruikers tot zeer succesvolle, innovatieoplossingen leidt voor een breed spectrum van gebruikers: van tieners tot ambtenaren; van politieagenten tot advocaten.

Taal- en Spraaktechnologie in het onderwijs

Als vanzelfsprekend vinden maatschappelijke kwesties als vergrijzing en gebrek aan docenten hun weerslag in de activiteiten van NOTaS-deelnemers. De steeds complexer wordende maatschappij eist nu eenmaal steeds vaker 'intelligente' ondersteuning in het oplossen van knelpunten en naarmate toepassingen meer in het domein van de talige mens komen, moet ook de manier waarop ze reageren, "menselijker" en dus "taliger" worden.

De ontwikkeling van TST en het toepassen daarvan in gebieden zoals onderwijs en zorg worden op verschillende manieren door de Vlaamse en Nederlandse overheden gestimuleerd. Hoewel de volledig computergestuurde docent vooralsnog een ijdele wensdroom zal blijven, zijn er wel steeds meer toepassingen die zowel docenten ondersteunen als kinderen in staat stellen om ook zonder de docent de leerstof op een interactieve manier tot zich te nemen. Dat NOTaS deelnemers hierbij een cruciale rol zullen spelen, spreekt eigenlijk vanzelf!

De Nederlandse economie heeft baat bij goede TST-oplossingen uit eigen keuken

Hoewel technologie in principe grenzeloos is, is dit voor TST net iets anders. Het herkennen van een woord in een geluidsfragment is universeel: het herkennen van een Nederlands woord in een Nederlands gesproken fragment vereist toch specifieke kennis van het Nederlands. Het is daarom belangrijk dat TST (ook) lokaal wordt ontwikkeld. NOTaS hoopt daarom dat de Nederlandse en Vlaamse overheden de TST-gemeenschap zullen blijven stimuleren met zowel financiële als organisatorische bijdragen om te vermijden dat we straks te horen krijgen: "het kan wel maar helaas niet voor het Nederlands!"

Het is weer een uitdaging geweest zoveel verschillend TST-nieuws in één DIXIT te krijgen. Naast de vaste redactie zijn we daarom vooral de gastredacteuren van de Nederlandse Taalunie (Peter Spyns en Catia Cucchiarini) dankbaar voor hun inspirerende bijdrage aan zowel de organisatie als de inhoud van deze nieuwe DIXIT.

Namens NOTaS wens ik u dan ook veel leesplezier!

Debbie Kenyon-Jackson
Voorzitter

Voorzijde gemaakt door Folkert de Vriend. Met dank aan Jonas Beskow van het Center for Speech Technology, KTH, Stockholm, voor toestemming voor het gebruik van "Kattis".

NOTaS

Stichting NOTaS
Secretariaat: 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, lobbyactiviteiten en het tijdschrift DIXIT.

Deelnemers Stichting NOTaS

CLST
Postadres: Postbus 9103,
6500 HD NIJMEGEN
Bezoekadres: Erasmusplein 1
6525 HT NIJMEGEN
T: 024-3611686
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. Belangrijke speerpunten zijn robuuste ASR en de inzet van TST ten behoeve van het onderwijs en voor mensen met communicatieve beperkingen. Tevens valt onder CLST het Speech Processing Expertise Centre (SPEX), dat is gespecialiseerd in de productie en validatie van spraakdatabases.

Data Archiving and Networked Services (DANS)
Adres: Postbus 93067,
2509 AB Den Haag
T 070 3494450
W www.dans.knaw.nl
E info@dans.knaw.nl

DANS is een instituut van de Koninklijke Nederlandse Akademie van Wetenschappen (KNAW), dat mede wordt ondersteund door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). DANS zorgt voor de opslag en blijvende toegankelijkheid van onderzoeksgegevens in de alfa- en gammawetenschappen. Daartoe ontwikkelt DANS zelf duurzame archiveringsdiensten, bevordert het dat anderen dat doen, en werkt samen met databeheerders om ervoor te zorgen dat zo veel mogelijk data vrij beschikbaar komen voor gebruik in het wetenschappelijk onderzoek.

Dedicon
Postadres: Postbus 24,
5360 AA GRAVE
Bezoekadres: Traverse 175 5361 TD GRAVE
T: 0486 – 48 64 86
W: www.dedicon.nl
E: info@dedicon.nl

Dedicon, voorheen FNB, is in Nederland dé organisatie die lectuur en informatie toegankelijk maakt voor mensen met een leeshandicap. Binnen de diensten van Dedicon vormt de moderne taal- en spraaktechnologie een belangrijke rol.

Dialogs Unlimites BV
Adres: Takkebijsters 17 sub 11,
4817 BL BREDA
T 076-5724777
W www.dialogsunlimited.com
E info@dialogsunlimited.com

Dialogs Unlimited levert turnkey oplossingen en werkt op niet exclusieve basis samen met vooraanstaande technologiebedrijven, applicatie-ontwikkelaars, systeemintegratoren en service providers.

Dutchear
Postadres: Postbus 5050,
2600 GB DELFT
Bezoekadres: Brassersplein 2
2612 CT DELFT
T: 015 – 219 11 11
W: www.dutchear.nl
E: info@dutchear.nl
Dutchear past spraaktechnologie toe voor telefonische selfservice, open vraag spraakherkenning en audio mining. Dutchear ontwerpt en realiseert de best-of-breed oplossing op basis van deze technologieen, die voor elke organisatie de beste prijs-kwaliteit oplevert.

GridLine B.V.
Adres: Keizersgracht 520,
1017 EK AMSTERDAM
T: 020-616 20 50
W: www.gridline.nl
E: info@gridline.nl

GridLine is een Amsterdams IT-bedrijf dat zich specialiseert in de toepassing van Nederlandse taaltechnologie. Wij zijn expert in het toegankelijk maken van bedrijfsinformatie via thesauri en andere systemen voor kennismanagement. We bereiken dit doel door de inzet van hulpmiddelen voor de automatische opsporing van trefwoorden, vaktermen en betekenisrelaties in domeinspecifieke tekstverzamelingen. In onze projecten werken we graag samen met universiteiten en collegabedrijven. NOTAS biedt ons een platform om met elkaar in contact te komen.

inTAAL B.V.
Adres: Winthontlaan 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. In nauwe samenspraak met gebruikers en begeleiders vertalen wij technische ontwikkelingen op ons gebied in oplossingen voor gebruikers

Knowledge Concepts
Adres: De Handboog 9,
5283 WR BOXTEL
T: 0411– 61 08 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
Postadres: Postbus 159,
1180 AD AMSTELVEEN
Bezoekadres: Prof. W.H. Keesomlaan 14, 1183 DJ AMSTELVEEN
T: 020 – 503 30 00
W: www.logica.com
E: info@logica.com

Samen met onze klanten streven we ernaar het beste in hun organisatie naar boven te brengen. Logica maakt veranderingen mogelijk die leiden tot verbeterde efficiency, versnelde groei en risicobeheersing.
Door onze diepgaande branchekennis, technische kwaliteiten en expertise op het gebied van wereldwijde levering van diensten, helpen we onze klanten leiderschapsposities in te nemen.

OSTT/Sint Maartenskliniek
Adres: Hengstdal 3,
6522 JV NIJMEGEN
T: 024 – 365 97 18
W: www.ostt.eu /www.maartenskliniek.nl
E: ostt@maartenskliniek.nl

Het "Ontwikkelcentrum voor Spraak- en Taaltechnologie ten behoeve van spraak- en Taalpathologie en revalidatie algemeen" is erkend als expertisecentrum door het Ministerie van het VWS aan de Sint Maartenskliniek. In het kader van het OSTT werkt de Sint Maartenskliniek samen met de afdeling Taalwetenschap van de Radboud Universiteit en met het UMC St. Radboud in Nijmegen.

Polderland Language & Speech Technology b.v.
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 een 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.

PonTeM
Adres: Wijchenseweg 122,
6538 SX NIJMEGEN
T : 024 - 366 7466
F : 024 - 366 7461
E : info@pontem.nl
W : www.pontem.nl

PonTeM is een stichting, geïntieerd door Viataal en de Koninklijke Effatha-Guyot Groep. Zij functioneert als een eigenstandig centrum voor onderzoek, ontwikkeling en innovatie met als doel het kennisbestand en de product - dienstcombinaties in zorg, onderwijs, diagnostiek en dienstverlening voor mensen met auditieve en communicatieve beperkingen te vergroten, te vernieuwen en te valideren.

Q-go.com B.V
Adres: 'Diemercircle' Eekholt 40,
1112 XH DIEMEN
T: 020 – 531 38 00
W: www.q-go.nl
E: info@q-go.com

Q-go is een internationale aanbieder van Natural Language Search technologie. De natuurlijke-taaltechnologie wordt ingezet in applicaties waarmee onder andere banken, verzekeraars, telecombedrijven en logistieke dienstverleners hun online klantenservice verbeteren, hun kosten verlagen, en hun inzicht in de klant vergroten. Q-go wendt 40% van haar opbrengsten aan voor R&D.

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 uitgaande van: kostenverlaging, serviceverbetering en backoffice-integratie.

TNO Defensie en Veiligheid
Postadres: Postbus 23,
3769 ZG SOESTERBERG
T: 0346-356205/211
W: www.tno.nl
E: secretariaat-HI@tno.nl

De afdeling Human Interfaces van TNO Defensie en Veiligheid in Soesterberg benadert spraaktechnologie op een integrale manier. TNO past de technologie niet alleen toe, maar helpt ook deze verder te ontwikkelen. Daarbij richten we ons op automatische spraak-, emotie- en taalherkenning en werken we ook aan evaluatie en implementatie van spraaksynthese en gesproken dialoogsystemen.

TST- Centrale, p/a Instituut voor Nederlandse Lexicologie
Postadres: Postbus 9515,
2300 RA LEIDEN
Bezoekadres: 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 - HMI
Postadres: Postbus 217,
7500 AE ENSCHEDE
Bezoekadres: Drienerlolaan 5,
7522 NB ENSCHEDE
T: 053 – 489 91 11
W: www.utwente.nl
E: info@utwente.nl

De HMI-groep van de Universiteit Twente (UT) is een ondernemende onderzoeksgroep die zich op het gebied van mens-machine interactie zowel met fundamenteel als met meer toepassingsgericht onderzoek bezig houdt. Rondom de campus van de Universiteit Twente bevinden zich een groot aantal spin-off bedrijven waarvan enkele ook op enigerlei wijze bezig zijn met Taal & Spraaktechnologie en Mens-Machine Interactie.

UvT Faculteit Geestesweten-schappen
Postadres: Postbus 90153,
5000 LE TILBURG
Bezoekadres: Warandelaan 2,
5037 AB TILBURG
T: 013 – 466 91 11
W: www.uvt.nl/communicatie-en-cultuur
E: uvt@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
Postadres: Postbus 19232,
3501 DE UTRECHT
Bezoekadres: St. Jacobsstraat 127
3511 BP UTRECHT
T: 030 – 232 47 11
W: www.vandale.nl
E: info@vandale.nl

Iedereen kent de Grote of "Dikke" Van Dale. De eerste uitgave van dit woordenboek stamt uit 1864. Bijna anderhalve eeuw later is Van Dale Lexicografie nog steeds bij de tijd en heeft het bedrijf zich ontwikkeld tot de meest toonaangevende en gezag hebbende uitgeverij van woordenboeken in Nederland en België. Met zorgvuldigheid en respect beschrijft Van Dale de taal zo compleet mogelijk, niet alleen het Nederlands, maar ook de grote vreemde talen.

Sponsoren/Samenwerking

Nederlandse Taalunie
Postadres: Postbus 10595,
2501 HN DEN HAAG
Bezoekadres: 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.

Stevin
Postadres: Postbus 10595,
2501 HN DEN HAAG
Bezoekadres: Lange Voorhout 19,
2514 EB DEN HAAG
T: 070 – 346 95 48
W: http://taalunieversum.org/stevin/

STEVIN 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.

NWO
Postadres: Postbus 93138,
2509 AC DEN HAAG
Bezoekadres: 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.

Ontwikkelingsmaatschappij Oost Nederland
Postadres: Postbus 5518,
7500 GM ENSCHEDE
Bezoekadres: Hengelosestraat 585, 7521 AG ENSCHEDE
T: 053 851 68 51
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.

Senternovem
Postadres: Postbus 93144,
2509 AC DEN HAAG
Bezoekadres: 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.

Ondersteuning

Cumlingua
Adres: Bronkhorstweg 48,
5363 TZ Velp (N.B.)
T: 0486 – 471 554
W: www.cumlingua.com
E: info@cumlingua.com

- Commerciële en juridische vertalingen Nederlands - Engels
- Correctiediensten
- Engelse teksten schrijven
- Engelse les
- Stimulerende taalbegeleiding- voor dementerende
- Advies en begeleiding op het gebied van marktgericht denken

Bloembergen Santee
Adres: Ambachtsweg 31,
6541 DA NIJMEGEN
T: 024 – 378 2643
W: www.bloembergensantee.nl
E: info@bloembergensantee.nl

Bloembergen Santee,is gespecialiseerd in het vervaardigen, organiseren en beheren van uw papieren communicatiestroom. Dat betekent voor u gemak, efficiency en meer rendement in tijd en geld.

Deykerhoff/Accountants
Postadres: Postbus 31070,
6503 CB Nijmegen
Bezoekadres: Toernooiveld 128,
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
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.

Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands

Het STEVIN onderzoeks- en stimuleringsprogramma (deel II)

De persoon Simon Stevin (1548-1620)[1] en het programma STEVIN zijn wellicht geen onbekenden meer voor u. In 2006 stelden zij zich reeds uitgebreid aan u voor in het decembernummer van dit tijdschrift. Toen had ik het ook over een spraakgestuurde zeilwagen. Die heb ik nog niet voorbij zien flitsen over het Scheveningse strand. Toch trappelde het wereldje van taal-en spraaktechnologie (TST) voor het Nederlands niet ter plaatse.

Toekomstige zeilwagenbestuurders zullen hun reisroute inspreken. Ook andere voorzieningen (radio, CD-speler, gsm, …) zijn SPRAAKgestuurd. En MIDAS filtert allerhande storend lawaai (wind, golven, meeuwen, …) voldoende weg. Tijdens de rit worden de namen van nabije steden en bezienswaardigheden correct uitgesproken dankzij Autonomata (Too). Daeso en DAISY maken samenvattingen van de beschrijvingen van de bezienswaardigheden en DuOMAn leert welke hip zijn. Onderweg krijgt de bestuurder een weerberichtenselectie via Easy Info. Hij kan zelfs op CD een complete (Audio)krant beluisteren. Eventueel schakelt hij over op een andere taal dankzij DPC en Paco-MT. Moet hij nog even bij het gemeentehuis langs om een vergunning op te halen, regelt hij dit met GemeenteConnect. Maar pas op, bij een te hoge snelheid zou hij via de Kentekenlijn wel eens snel een verkeersboete kunnen krijgen...

Wat ver gezocht? Toch niet helemaal. Alle vermelde projecten zijn STEVIN-projecten. En dit is slechts een greep uit het programma. Ik had ook andere projecten kunnen noemen om een verhaal over een intelligent huis of kantoor te illustreren. Domotica, en een intelligente omgeving (ambient intelligence and ubiquituous computing) in combinatie met TST leiden ongetwijfeld tot vele nieuwe gebruiksvriendelijke toepassingen.

Leuke dromen voor onderzoekers? De ver-van-de-bedrijven-hun bed-show? Met dit nummer (en het eerste STEVIN-DiXiTnummer) willen we bedrijven laten zien wat er gebeurt binnen STEVIN. Naast een onderzoeksprogramma is STEVIN immers ook een stimuleringsprogramma voor bedrijven. Vandaar dat alle STEVIN-demonstratieprojecten aan bod komen. Binnen zo'n demonstratieproject integreren bedrijven, eventueel samen met kennisinstellingen, onmiddellijk inzetbare TST-technologie in nieuwe toepassingen voor bestaande klanten.

Ook binnen de overheid kan TST nuttig ingezet worden. De rol van de overheid beperkt zich niet langer tot financierder van(onderzoeks)programma's, maar breidt zich uit tot aanbesteder en klant voor innovatieve toepassingen. Dit opent interessante mogelijkheden voor TST-bedrijven.

Taal in Bedrijf

Dit themanummer werd samengesteld met het oog op Taal in Bedrijf 2008. Het biedt alvast een staalkaart van actuele TST-projecten voor het Nederland. Na de ontwikkeling van taalbronnen, en toepassingen - nodig om de positie van het Nederlands in de digitale maatschappij veilig te stellen - is het nu zaak dat deze hun weg vinden naar innovatieve applicaties. Deze applicaties kunnen bedrijven een competitief voordeel opleveren. Wat uiteindelijk de eindgebruikers ten goede komt. Nu wordt nagedacht over een vervolg op het STEVIN-programma. Wij nodigen u uit mee te denken. Graag horen wij uw opinie via www.taalinbedrijf.org/forum.

En wie weet zoeft in 2020 een moderne Simon Stevin met zijn spraakgestuurde zeilwagen over het Scheveningse strand….

Prototype en eigenlijke zeilwagen op het strand van Scheveningen ontworpen in 1601/1602 door Simon Stevin voor Prins Maurits.

STEVIN is een onderzoeks- en ontwikkelingsprogramma, gezamenlijk gefinancierd door Vlaanderen en Nederland, waarbij wetenschappers basismaterialen en –resultaten creëren en ter beschikking stellen zodat bedrijven TST-toepassingen kunnen bouwen die het Nederlands als taal gebruiken. Het loopt van 2005 tot eind 2011. Zie ook www.stevin-tst.org.

Peter Spyns

Taal in Bedrijf wil uitgroeien tot dé bedrijfsbeurs voor TST voor het Nederlands. Ontwikkelaars komen er in contact met potentiele nieuwe klanten, onderzoekers leren waar bedrijven behoefte aan hebben, overheden ontdekken mogelijkheden voor innovatief aanbesteden. Zie ook www.taalinbedrijf.org.

Peter Spyns is de STEVIN-programmacoördinator bij de Nederlandse Taalunie gedetacheerd vanuit de Vlaamse overheid - departement Economie, Wetenschap en Innovatie.

De TST-Centrale, STEVIN en u

De TST-Centrale werd in 2004 op initiatief van de Nederlandse Taalunie (NTU) opgericht. Ze is, gefinancierd door de NTU, als project ondergebracht bij het Instituut voor Nederlandse Lexicologie (INL). De missie van de TST-Centrale is het stimuleren van hergebruik van digitale (basis)taalmaterialen die met (Vlaams en Nederlands) overheidsgeld zijn gefinancierd. Zo wordt kapitaalvernietiging tegengegaan. De taalmaterialen worden in beheer genomen, onderhouden en gedistribueerd. Daarnaast biedt de TST-Centrale diverse diensten aan, zoals gebruikersondersteuning via een servicedesk. Ten slotte stimuleert de TST-Centrale het hergebruik door de taalmaterialen op allerlei manieren onder de aandacht van gebruikers te brengen – dit is er één van.

Remco van Veenendaal

Steeds meer producten

De handvol producten van vlak na de oprichting is in vier jaar tijd vertienvoudigd. Van de meeste producten van het eerste uur werden in samenwerking met het veld verbeterde versies uitgebracht. Niet alleen in kwantiteit groeide de catalogus, maar ook in diversiteit: naast modern-Nederlandstalige bestanden nam de TST-Centrale ook historische Nederlandse materialen, zoals corpora, bijbelvertalingen en woordenboeken in beheer. Naast tekst-, spraak- en videocorpora worden woordenlijsten, lexica en bilinguale bestanden gedistribueerd. Ook wordt software via de TST-Centrale beschikbaar gesteld en krijgt het internet steeds meer aandacht. Van het Corpus Gesproken Nederlands werd met partners een onlineversie ontwikkeld en er wordt samengewerkt aan een Europese infrastructuur voor het geïntegreerd ontsluiten van taalmaterialen (CLARIN).

Een belangrijke leverancier van taalmaterialen is het Vlaams-Nederlandse STEVIN-programma. STEVIN-projecten dragen hun resultaten over aan de NTU, waarna ze via de TST-Centrale worden beheerd en gedistribueerd. De TST-Centrale is daarom nauw betrokken bij het STEVIN-programma. Eigendomsrechtenkwesties worden samen met een commissie en gespecialiseerde juristen geklaard, er werd beleid opgesteld voor het opleveren van opensourceresultaten en de TST-Centrale verzorgde enkele eigen STEVIN-themadagen. Via STEVIN laat de taal- en spraaktechnologiesector zien dat ze vooroploopt als het gaat om de valorisatie van onderzoekresultaten. Zowel de Vlaamse als de Nederlandse overheid leggen hier steeds meer nadruk op.

Remco van Veenendaal is projectleider van de TST-Centrale.

Steeds betere dienstverlening

De TST-Centrale en haar belanghebbenden investeren continu in verbetering en professionalisering. Er wordt aansluiting gezocht bij standaarden en best practices. Zo heeft de TST-Centrale ITIL geadopteerd. Ook op allerlei andere gebieden proberen we niet zelf het wiel uit te vinden: er wordt samengewerkt met DANS, ELDA, een prijzencommissie en een commissie voor pr&communicatie. Daarnaast onderhoudt de TST-Centrale een netwerk als deelnemer en bestuurslid van NOTaS, via persoonlijke contacten en dus ook als lid van diverse commissies. Evenementen als CLIN en LREC worden bezocht en gesponsord, maar ook draagt de TST-Centrale bij aan het organiseren van evenementen. Leveranciers, partners en gebruikers worden actief opgezocht: de TST-Centrale komt naar je toe.

Iets voor u?

De TST-Centrale is het centrale loket voor Nederlandstalige taalmaterialen. Voor zover we de materialen niet zelf beheren, kunnen wij u in contact brengen met collega's. U kunt voor commercieel en wetenschappelijk gebruik taalmaterialen bij de TST-Centrale aanvragen. Onze website met onder andere productsheets en online demo's geeft u een goede eerste indruk van onze producten. Voor commerciële partijen gelden marktconforme tarieven – de markt moet niet verstoord worden. Eventueel kunt u eerst een sample van een product aanvragen of komen we het bij u demonstreren. Via onze servicedesk leveren wij support. Kortom: bent u op zoek naar digitale taal- of spraaktechnologische materialen? Bij ons bent u aan het juiste adres!

Toepassingen voor het voetlicht

Een politiekorps wil zijn agenten telefonisch snel kentekeninformatie kunnen verschaffen. De Nederlandse overheid wil alle informatie over wet- en regelgeving op efficiënte wijze publiekelijk toegankelijk maken. Op het gemeentehuis wil men de vragen van burgers over bijvoorbeeld een paspoort adequaat kunnen afhandelen. Voor het bereiken van elk van deze doelstellingen wordt nu met succes taal- en spraaktechnologie ingezet.

Binnen het STEVIN-programma gaat speciale aandacht uit naar projecten waarin succesvolle TST-toepassingen op aantrekkelijke wijze voor het voetlicht gebracht worden met de bedoeling een extra impuls te geven aan de vraag naar TST voor het Nederlands. Deze demonstratieprojecten duren relatief kort en dienen gebruik te maken van technologie die zichzelf al heeft bewezen. Ook wordt met de demonstratieprojecten geprobeerd toegang te krijgen tot nieuwe markten. De eerste ronde voor demonstratieprojecten startte in 2005 en omvatte de projecten *Kentekenlijn, Rechtsorde en GemeenteConnect!* Na de eerste ronde volgde er ook nog een tweede en een derde ronde voor demonstratieprojecten.

Met de *Kentekenlijn* beschikt de Politie Utrecht nu over een nieuwe manier om kentekens op te vragen; door te bellen naar een spraakcomputer. Door de inzet van het systeem wordt de meldkamer ontlast en krijgen de agenten de informatie behorend bij een kenteken precies op het moment dat ze die informatie nodig hebben.

Het *Rechtsorde*-systeem is bedoeld om de grote hoeveelheid overheidsinformatie over wet- en regelgeving op efficiënte wijze te ontsluiten. De zoekmachine van Rechtsorde ontleedt automatisch een ingevoerde zoekterm en voorziet de gebruiker van suggesties voor de zoekterm op basis van een thesaurus met synoniemen. Gebruikers van juridische informatie kunnen met het systeem op beduidend gebruiksvriendelijkere en snellere wijze de gezochte informatie boven water krijgen.

GemeenteConnect! is opgezet als telefonische vraagbaak van een gemeente. Het systeem beantwoordt de meest gangbare informatieaanvragen zoals vragen over de aanvraag van een paspoort of bouwvergunning.

Door de toepassing van spraakherkenning heeft de beller het gevoel op natuurlijke en prettige wijze een dialoog te voeren met het systeem.

De Nederlandse Taalunie ziet voor zichzelf vooral een rol weggelegd binnen de publieke sectoren overheid, zorg en onderwijs. Veel van de demonstratieprojecten vallen binnen een van deze drie sectoren. Zo zijn er naast de projecten uit de eerste ronde ook projecten uit de vervolgrondes die zich specifiek richten op bijvoorbeeld de overheidssector: *Rechtspraakherkenning* en *Klinkende taal*. Voor de zorgsector startten *HATCI*, de A*udiokrant* en *Primus*. De projecten *WooDy, Spelspiek* en *AAP* richten zich op de onderwijssector.
Tot slot zijn er ook nog andere projecten uit de vervolgrondes: *WebAssess, NeOn* en HATCI.

De demonstratieprojecten lenen zich bij uitstek voor het maken van reclame voor TST voor het Nederlands. Het zijn vooral ook deze projecten welke ingezet kunnen worden bij het zogenaamde ‘makelen en schakelen’ dat de Taalunie uitvoert voor TST voor het Nederlands. Enerzijds wil de Taalunie partijen bij elkaar brengen (het schakelen) om samen aan de ontwikkeling van TST te werken. Anderzijds wordt geprobeerd TST voor het Nederlands voor het voetlicht te brengen bij zowel het brede als het meer gespecialiseerde publiek (het makelen). Door te makelen wil de Taalunie meer interesse kweken voor de mogelijkheden van TST. Juist de demonstratieprojecten geven heel goed weer wat voor mogelijkheden dat zijn.

Voor demo’s en verdere informatie verwijs ik u naar http://www.stevin-tst.org/pers/#demo.

Folkert de Vriend

Folkert de Vriend staat bij de Nederlandse Taalunie in voor het `makelen en schakelen’.

‘De AuDioKrant’

De Krant zoals het hoort

Lieve Meers

Sinds juni 2008 zijn de Vlaamse kranten “De Standaard” en “Het Nieuwsblad” verkrijgbaar als AuDioKranten. Een AuDioKrant is een ‘gesproken dagblad’ dat net zoals de gedrukte versie dagelijks bij de abonnees in de bus valt.

Dit product is het resultaat van een gerichte samenwerking tussen drie partners: Kamelego vzw - tot voor kort de Braillekrant genoemd, de K.U.Leuven (DocArch & CUO) en de firma Sensotec. Ze worden hierin gesteund door de Vlaamse overheid, uitgeversgroep Corelio, Cera en het STEVIN-programma.

Een gesproken dagblad

De AuDioKrant wordt elke avond aangemaakt door de medewerkers van de vzw Kamelego, op basis van de redactionele informatie die door Corelio wordt aangeleverd. De eerste stap bestaat uit het zin per zin, of alinea per alinea omzetten van de aangeleverde artikels in mp3 geluidsbestanden via het tekst- naar spraakprogramma Nuance Realspeak.

Om deze kranten dagelijks te kunnen verspreiden koos Kamelego ervoor om te werken met een synthetische in plaats van menselijke stem. Een persoon zou voor het inlezen van alle artikels van 1 krant ongeveer een tijdspanne van 8 tot 22 uren nodig hebben terwijl de computer dit in minder dan 1 uur doet. Bovendien is de kwaliteit van de synthetische stem nu zeker geen contra-indicatie meer, integendeel! De spraaktechnologie is de laatste jaren zo sterk verbeterd dat ze in niets meer te vergelijken valt met de onnatuurlijke computerstemmen van het eerste uur.

Op zich bestaat de AuDioKrant dus eigenlijk uit een reeks mp3-bestanden die via elke mp3-speler of computer kunnen beluisterd worden. Echter om vlot te kunnen ‘bladeren’ of ‘navigeren’ is het noodzakelijk de aangemaakte mp3-artikels te ordenen binnen de structuur zoals we deze ook kennen van de traditionele gedrukte kranten. Hierbij worden artikels geordend naargelang hun inhoud en gegroepeerd binnen mappen zoals ‘sport’, ‘buitenland’, etc.

Lezen van een Daisy-CD

Tenslotte is er nog het omzetten van de inhoud naar een speciale Daisy-standaard, *Digital Access Information System*. Via de Daisy-standaard wordt ervoor gezorgd dat de CD’s niet alleen via een computer met Daisy-software maar zeker ook via autonome en zelfs draagbare Daisy-CD-toestellen gemakkelijk gelezen kunnen worden.

Een Daisy-toestel is een draagbaar CD-leestoestel dat initieel ontworpen is voor personen met een visuele handicap. Het toestel is voorzien van een aantal navigatiefuncties om voor- en achteruit te spoelen in teksten en dit zowel op het niveau van zinnen, alinea’s of bladzijden. Bovendien kan men ‘favorieten’ markeren of bladwijzers plaatsen net zoals bij een gedrukte tekst. Een extra voordeel van de compacte draagbare Daisy-toestellen is dat abonnees de krant ook kunnen lezen zonder gebruik te maken van een computer.

Flexibel lezen

De AuDioKrant is ontwikkeld om kranten ook toegankelijk te maken voor mensen met een functiebeperking. Dit betreft niet alleen personen die problemen hebben met het zien op zich, maar ook mensen met dyslexie, personen die niet kunnen lezen of geen gedrukte krant kunnen hanteren (omwille van een motorische handicap, MS-patiënten enz.), of om welke reden dan ook niet vlot met gedrukte letters om kunnen.

Toekomstperspectieven

Hierbij wordt er o.a. gedacht in de richting van nieuwe doelgroepen, toepassingsvormen, of bijv. het downloadbaar maken van de gesproken krant via een breedbandverbinding.

Contactgegevens

Abonnementsvoorwaarden kunt u vinden op www.audiokrant.be.

De AuDioKrant is een demonstratieproject uit de tweede ronde. Zie http://www.kuleuven.be/audiokrant.

Lieve Meers is projectmedewerkster bij de AuDioKrant.

Klinkende Taal

Een word-plugin die professionals helpt betere brieven te schrijven

De overheid wil burgers beter bereiken. Deze nieuwe openheid betekent dat informatie beschikbaar moet zijn voor iedereen. En sterker, dat de informatie voor iedereen begrijpelijk moet zijn. Betere brieven, tevreden burgers.

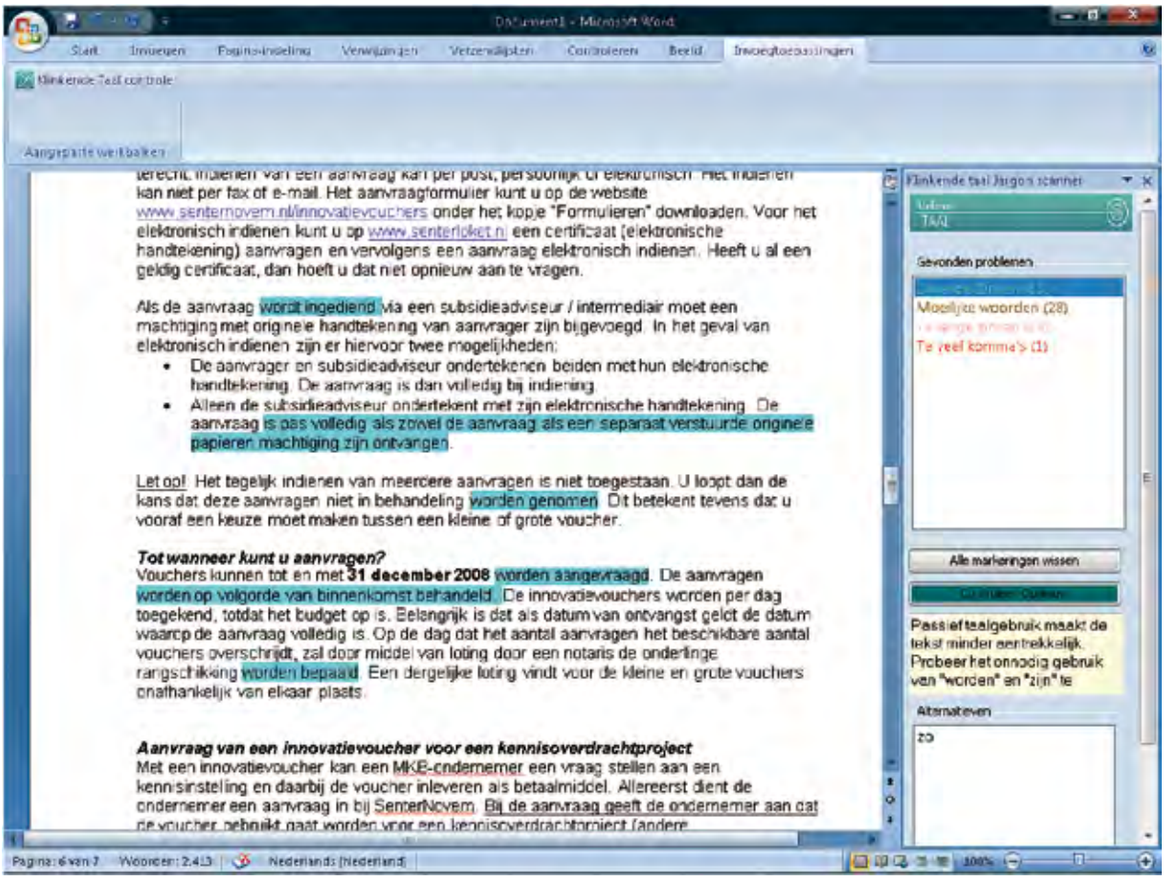
Het bedrijfsleven begrijpt dit al langer. Daar is de business case voor begrijpelijk taalgebruik helder: duidelijke informatievoorziening levert tevreden klanten op, en dus meer omzet en nieuwe klanten. Ook voor de overheid geldt hetzelfde: duidelijk communiceren betekent minder klachten, minder kosten en sneller en prettiger werken. Recente wetten verplichten overheidsorganen om duidelijk te communiceren.
Ambtenaren en andere professionals worden op verschillende manieren geholpen om beter te schrijven: cursussen, schrijfwijzers en eindredactie door tekstschrijvers en communicatiekundigen. Aan dit scala aan producten heeft GridLine nu Klinkende Taal toegevoegd, een elektronisch hulpmiddel dat brievenschrijvers en webredacteuren op ieder moment helpt bij het verbeteren van hun teksten.

Wat is Klinkende Taal

Klinkende Taal analyseert taalgebruik, maar belangrijker, het geeft aanwijzigingen hoe teksten kunnen worden verbeterd. De schrijver wordt niet beoordeeld, maar geholpen. Klinkende Taal vormt daarmee een aanvulling op bijvoorbeeld een schrijfwijzer of schrijfcursus. Een hulp die ieder moment van de dag paraat staat.

Wat zijn de kenmerken van goede teksten? Om die vraag te beantwoorden werkte GridLine samen met de Universiteit Utrecht. Goede teksten bevatten bijvoorbeeld geen onnodig jargon, ingewikkelde constructies, tangconstructies, te lange alinea's en onnodige passieve zinnen. GridLine ontwikkelde, deels in samenwerking met de Universiteieten van Tilburg en Leuven, taaltechnologische modules

Tigran Spaan



Klinkende Taal in actie

die deze en andere tekstkenmerken herkennen. Interessant is het jargon, omdat dat per brance en zelfs per organisatie verschilt. De jargonmodule van Klinkende Taal kan daarom per organistatie worden ingesteld.

Klinkende Taal is opgezet als server-applicatie. Daardoor is Klinkende Taal als Word-plugin, in het contentmanagementsysteem van een website en in redactiesystemen te gebruiken, of bijvoorbeeld in Windows Sharepoint -- steeds met exact dezelfde functionaliteit. Bovendien zijn installaties en upgrades eenvoudig, evenals finetuning met nieuwe schrijfregels, nieuw jargon en aanpassingen aan de doelgroep. De koppeling met de server wordt verzorgd door een razendsnelle REST-webservice.

Bijeenkomsten, klanten en samenwerkingspartners

Klinkende Taal is oorspronkelijk ontwikkeld voor de Gemeente Den Haag en de Provincie Noord-Brabant. Klinkende Taal trekt veel belangstelling van overheidsinstellingen, maar ook van banken, verzekeringsmaatschappijen, woningcorporaties en bijvoorbeeld energieleveranciers. GridLine gaat daarom per branche gebruikersbijeenkomsten organiseren waarvoor zowel oude als nieuwe gebruikers van Klinkende Taal worden uitgenodigd.

Klinkende Taal is een aanvulling op de producten die worden geboden door tekst- en communicatiebureaus. Daarom werken wij graag samen met deze bedrijven. Onze klanten kunnen dus gewoon blijven werken met hun favoriete communicatiebureau, met Klinkende Taal als aanvulling.

Klinkende Taal is een demonstratieproject uit de tweede ronde. Zie http://www.klinkende-taal.nl/

Tigran Spaan
werkte mee aan het Klinkende Taal-project.

klinkende TAAL

Primus

Schrijfhulp voor dyslectische kinderen

In het kader staat een stuk tekst van een dyslectisch meisje van 13 jaar. Dyslexie is een quasi-permanent probleem: vlot spellen of lezen komt nooit binnen het bereik van kinderen met deze leerstoornis. De schriftelijke taalvaardigheid van dyslectische gebruikers verschilt op een aantal belangrijke punten van dat van niet-dyslectische gebruikers.

De spellingsproblemen uiten zich met name in:

- het weglaten, toevoegen, verwisselen of vervangen van letters
- fonetisch spellen (volgens het gehoor)
- het foutief splitsen of aan elkaar schrijven van woorden

Ze spellen bij voorbeeld

'neit'	i.p.v.	'niet'
'zach'	i.p.v.	'zag'
'hadt'	i.p.v.	'had'
'nerends'	i.p.v.	'nergens'
'brugste'	i.p.v.	'beruchtste'
'vander'	i.p.v.	'vader'
'originsatie'	i.p.v.	'organisatie'
'verrukelig'	i.p.v.	'verukkelijk'
'etalatie'	i.p.v.	'etalage'
'eemoscho nele'	i.p.v.	'emotionele'

Spellingcontrole wordt algemeen gezien als één van de belangrijkste hulpmiddelen voor dyslectici. In dit STEVIN-demonstratieproject hebben we een eerste versie ontwikkeld van Primus, een schrijfhulp voor dyslectische kinderen. Primus is geïntegreerd in Microsoft® Office en bevat een aangepaste versie van de standaard Nederlandse spelling- en grammaticacontrole. Primus heeft een eigen interface (zie Figuur), dat eenvoudig te gebruiken is door dyslectische kinderen:

- er is zoveel mogelijk gebruik gemaakt van pictogrammen in plaats van tekst;
- het foute woord wordt altijd in context aangeboden;
- de fout in context en alle suggesties zijn te beluisteren door middel van een ingebouwd tekst-naar-spraak-systeem.

Met behulp van het tekst-naar-spraak-systeem kan de gebruiker van Primus ook zijn/haar eigen tekst terug beluisteren.

Om een goed beeld te krijgen van de fouten die dyslectische kinderen maken is een klein corpus samengesteld met handgeschreven en getypte teksten van kinderen uit Nederland en Vlaanderen. Een gedeelte van de teksten is volledig gedigitaliseerd (ongeveer 30.000 woorden). Van een ander deel zijn alleen de fouten opgenomen in een spreadsheet (3.045 fouten). Op basis van het corpusmateriaal is de spellingcontrole zodanig aangepast dat het suggestiemechanisme in meer gevallen de juiste suggestie bevat. Voor het herkennen van foutief niet aan elkaar geschreven woorden is de grammaticacontrole aangepast.

Primus-interface

Primus is ontwikkeld in een samenwerking tussen Polderland Language & Speech Technology bv (Nijmegen), Technologie & Integratie b.v.b.a. (Gent) en vzw Die-'s-lekti-kus (Leuven). Polderland levert al sinds 1996 de Nederlandse spellingcontrole en grammaticacontrole voor Microsoft Office. Het bedrijf richt zich onder meer op de ontwikkeling van gespecialiseerde schrijfhulpmiddelen voor verschillende talen en geavanceerde zoektechnologie. Technologie & Integratie ontwikkelt technologische hardware en software voor mensen met een communicatiehandicap. Communicatiehulpmiddelen met spraaksynthese als Mind Express en Sprint ondersteunen of herstellen de zelfstandigheid van de gebruiker en bevorderen zijn integratie in de maatschappij. De vzw Die-'s-lekti-kus organiseert en ondersteunt allerhande projecten in verband met dyslexie, dyscalculie, ADHD en andere stoornissen. De vzw treedt bovendien op als het gezamenlijk platform voor de verenigingen en zelfhulpgroepen die in Vlaanderen actief zijn op het gebied van leer- en ontwikkelingsproblemen.

Het resulterende product wordt beschikbaar gesteld aan dyslectische gebruikers in Nederland en Vlaanderen. Het product zal drie maanden gratis gebruikt kunnen worden. De opbrengsten die daarna met de verkoop van het product worden gegenereerd worden gereserveerd voor nieuwe ontwikkelingen voor de doelgroep. Daarbij kan gedacht worden aan aanpassingen van de resulterende demonstrator, maar ook aan nieuwe applicaties. De spellingcontrole zal tevens beschikbaar komen voor integratie in andere tekstverwerkers, waaronder SPRINT.

Primus is een demonstratieproject uit de tweede ronde.

Inge de Mönnink

"maar ze zijn laste gewoorden in te compitzie dat kan de plers en de kots niks zegelen ales maar lol en leuk vinden zij die dat is ook zo ."

Inge de Mönnink
coördineerde het Primus-project.

RechtSpraakHerkenning

Taal en Spraaktechnologie in gesproken rechtbankzaken

In toenemende mate worden verhoren door politie volledig opgenomen. Bij twijfel kan dan altijd de oorspronkelijke opname her-beluisterd worden. Ook Nederlandse rechtbanken experimenteren met geluidsopnamen. De griffier maakt het verslag van de rechtszitting, maar omdat het soms lastig is alles tijdens de zitting correct te noteren, worden voor intern gebruik dikwijls geluidsopnamen gemaakt.

Arjan van Hessen

Door nu iedere spreker op een eigen spoor op te nemen en de opnamen door de spraakherkenner te halen, kan het uitwerkproces efficiënter gemaakt worden. De opnamen worden dan doorzoekbaar op zowel spreker als spraak. Iedereen die straks toegang heeft tot de opnamen kan snel zoeken naar de woorden X, Y en Z, uitgesproken door verdachte A of Rechter B.

De griffier kan de spraakherkenningsresultaten gebruiken om sneller een verslag te maken en rechters kunnen naar een gesproken samenvatting luisteren; bedoeld om hun geheugen op te frissen als ze de zaak weer oppakken na langdurige onderbreking. De Taal- en Spraaktechnologie wordt in het RechtSpraakHerkenningsproject ingezet voor ondersteuning van de rechtbank: niet als vervanging van medewerkers. Rechtspraak blijft vooralsnog echt mensenwerk.

Rechtszaal in Almelo waar de applicatie gaat 'draaien', de spraak van 12 microfoons kan gelijktijdig opgenomen worden.

Arjan van Hessen
coördineert het Rechtspraakherkenning-project.

Gescheiden kanalen

Om bovenstaande te realiseren, moesten er een aantal aanpassingen in de rechtszaal gedaan worden. In de oude situatie werden alle sprekers samengevoegd op één kanaal en dan opgenomen. Hoewel het niet vaak voorkomt dat mensen door elkaar spreken, gebeurt het wel waardoor dan achteraf onduidelijk is wie wanneer sprak. Bovendien spreekt niet iedereen even luid, wat soms lastig is voor de spraakherkenner. In de nieuwe situatie worden maximaal 12 sprekers elk op een eigen kanaal opgenomen. Als een sprekerafhankelijk audioprofiel wordt gemaakt, dan kan op deze manier iedere spreker optimaal herkend worden.

Taalmodel

Spraakherkenning staat of valt met de mate waarin het woordgebruik voorspeld kan worden. Samen met de rechtbank Almelo is er daarom een systeem gemaakt waarmee snel de voor de rechtszaak relevante documenten gebruikt kunnen worden voor het creëren van een taalmodel. Notoire lastige items als "namen van de verdachten", "delicten" en "locaties" kunnen dan wel goed herkend worden.

Spraakherkenning

De combinatie van goede geluidsopnamen, sprekerspecifieke akoestische modellen en een dedicated taalmodel, resulteren in goede herkenning. Dit, gecombineerd met een geavanceerd zoeksysteem dat het mogelijk maakt niet alleen in de 'best herkende' zin maar ook in mogelijke herkenningsalternatieven te zoeken, levert een zeer bruikbaar systeem voor het zoeken in rechtszittingen.

Proof-of-Concept

Als de applicatie draait, zal veel aandacht besteed worden aan de bruikbaarheid: werkt het goed, vindt men het prettig, levert het 'winst' op? Vragen die naast de harde feitelijkheid (hoeveel procent correcte herkenning?) beantwoord moeten worden voordat men kan spreken van een succes.

Het uiteindelijke resultaat is zowel een applicatie die door de rechtbank gezien kan worden als een 'proof-of-concept' als een rapport waarin de waarden en bruikbaarheid van alle verschillende parameters en aanpak zal worden beschreven: men kan zo zien onder welke condities de aanpak zowel feitelijk als menselijk succesvol is.

Rechtspraakherkenning is een demonstratie-project uit de tweede ronde.

Spelspiek

“Als je niet meer zeker weet of je nu met een mens aan het chatten bent of met een machine, dan is die machine intelligent te noemen.” Dat is misschien wel een mooie hedendaagse interpretatie van de Turingtest.

Michel Boekestein

Wellicht kan Spelspiek een argeloze gebruiker heel eventjes foppen, maar slagen voor de Turingtest zal hij zeker niet. Spelspiek, een spellinghulp die je aan je buddylist van MSN kunt toevoegen, gedraagt zich al wel een stukje menselijker dan verschillende andere *chatbots*. Verschillende *saldobots* (waarmee je via MSN onder andere je banksaldo kunt checken) begrijpen alleen wat je bedoelt als je het commando ‘saldo’ geeft.

Spelspiek laat zien dat het best mogelijk is om een chatbot iets te vragen op een manier waarop je dat ook zou doen bij een chatpartner die gemaakt is van vlees en bloed. Spelspiek is gespecialiseerd in spelling, dus je kunt bijvoorbeeld vragen: “Hoe spel je email?” of zoiets als “Is email goed gespeld?”. En we kunnen iets met het antwoord, dat drie suggesties geeft: “*e-mail* (eerste persoon enkelvoud van e-mailen); *e-mail* (electronic mail) en *email* (glazuur, kleursel)”.

De taaltechnologie in Spelspiek gaat niet primair om de verwerking van spellingsvragen in natuurlijke taal, wel om het geven van de juiste spelling van een (meestal) incorrect gespeld woord. Een kruisbestuiving tussen het zogenaamde omspellexicon van Van Dale en de spellingcorrectietechnieken van Polderland zorgde voor een krachtig omspelmechanisme. Van Dale en het Instituut voor Nederlandse Lexicologie zorgden bovendien voor de broodnodige woordenboekinformatie. Naast deze twee componenten (het omspelmechanisme en de woordenboekinformatie) zorgde een semiautomatisch protocol voor een gestaag groeiende dekkingsgraad en precisie: Spelspiek kon (voor hem) onbekende woorden en onoplosbare foute spellingen doorsturen naar een (menselijke) spellingdeskundige. Die mailde het correcte antwoord terug naar de eindgebruiker en bovendien werd de spelfout, samen met de correctie, weer opgenomen in de databank achter Spelspiek. Spelspiek werd in de loop van de tijd dus ‘slimmer’.

Verschillende gebruikers probeerden uit hoe menselijk Spelspiek nu eigenlijk was. En inderdaad, als slagroom op de taart bleek hij ook nog eens te kunnen rekenen, een mop te kunnen vertellen (al is het steeds dezelfde) en zelfs verontwaardigd te reageren als men stiekem een scheldwoord intypte. Een eventueel vervolg op Spelspiek is dan ook makkelijk te bedenken: we zouden ervoor kunnen zorgen dat hij naast spellingskwesties van meer zaken verstand heeft. Een James in je huis of werkomgeving die niet alleen adviseert over de juiste spelling, maar die ook je schrijfhulp is en de laatste stand van je aandelen kan presenteren. Dan merk je al heel snel dat je richting dialoog wilt.

Een nog menselijker chatbot krijg je door hem te laten praten en luisteren. Spelspiek sluit daar perfect op aan: de gebruiker hoeft het woord, waar hij de spelling niet goed van weet, niet meer uit te schrijven. Hij wil het antwoord waarschijnlijk nog altijd in geschreven vorm terugkrijgen, maar gecombineerd met een dialoog is spraaksynthese onmisbaar; voorbeeldzinnen en tegenvragen kan hij dan voor je uitspreken.

Het Instituut voor Nederlandse Lexicologie (INL), Elitech, Polderland Language & Speech Technology bv en Van Dale Lexicografie bv zorgden voor de realisatie van het project, dat eindigde in april 2008. Voor een commerciële doorstart kan men contact opnemen met Michel Boekestein (boekestein@inl.nl) of Katrien Van pellicom (pellicom@inl.nl).

Spelspiek is een demonstratieproject uit de tweede ronde.

Michel Boekestein coördineert het SpelSpiek- project.

WebAssess

Bedrijven besteden veel tijd en geld aan het selecteren van geschikte kandidaten voor call centers: slechts 10% van de aanmelders blijkt daadwerkelijk geschikt te zijn. Gecombineerd met het feit dat de sociaaleconomische realiteit call centers dwingt meer nadruk te leggen op het werven van mensen uit 'moeilijkere segmenten', zal duidelijk zijn dat een applicatie waarmee kandidaten zonder menselijke tussenkomst op onderdelen van hun mondelinge communicatie beoordeeld (en getraind) kunnen worden, zeer wenselijk is. Goede automatische voorselectie geeft bedrijven de mogelijkheid meer tijd en aandacht te besteden aan de geschiktheid van geselecteerde kandidaten.

Arjan van Hessen

In 2006 werd begonnen met VoiceAssess (nu WebAssess) met als doel een applicatie te maken die geheel automatisch een conversatie met kandidaten kan aangaan. Spraakherkenning wordt hierbij gebruikt om te bepalen of bepaalde essentiële woorden wel of niet gezegd werden.

E-Learning

Kandidaten die door het systeem gebeld worden, moeten eerst een webapplicatie met goed gevolg hebben doorlopen. De webapplicatie geeft gedegen uitleg over werken in het call center. De E-Learning module bevat ook een beroepskeuzetest en probeert aan de hand van gegeven antwoorden de motivatie van de kandidaat te bepalen. Als de kandidaten de webapplicatie met goed gevolg doorlopen hebben, kunnen ze het telefoonnummer invullen waarop ze door de applicatie gebeld willen worden.

Dialoog generator

Onderdeel van het project was het maken van een dialooggenerator waarmee niet-experts snel een goede dialoog kunnen maken of bestaande dialogen kunnen aanpassen. De huidige generator is niet-grafisch: iets dat in de toekomst wellicht gaat veranderen omdat grafische intuïtiever werken.

Dialoog

De applicatie belt de kandidaat en start het gesprek. Worden in een gegeven antwoord één of meerdere woorden niet genoemd, dan wordt de vraag op een andere wijze herhaald. Alle vragen en antwoorden worden gelogd en gebruikt om een score te berekenen. De praktijk zal moeten uitwijzen hoe de verschillende onderdelen gewogen moeten worden om tot een bruikbaar score te leiden.

Demonstrator website

In juni 2007 is de website van WebAssess (www.webassess.nl) live gegaan. Op deze website kunnen geïnteresseerden zich aanmelden en min-of-meer direct aan de slag (met username en password). De dialoog op de site is een voorbeelddialoog om te laten zien hoe het concept precies werkt.

Voorbeeld dialoog

Hieronder een deel van deze dialoog waarbij K voor kandidaat en S voor systeem staat. Alles wat onderstreept is dient herkend te worden en heeft invloed op de vervolgprompt. Als K het onderstreepte niet meldt wordt de prompt tussen haakjes afgespeeld. De cursief weergegeven tekst dient te worden herkend maar heeft verder geen invloed op het verloop van het gesprek. Indien de kandidaat deze teksten achterwege laat zal hij/zij echter lager scoren op onderdelen als klantgerichtheid en klantvriendelijkheid.

Conclusie

WebAssess is een aansprekend project gebleken dat niet alleen interesse heeft gewekt bij opleidingsorganisaties voor call center medewerkers maar ook bij organisaties die zich met onderwijs bezighouden. De mogelijkheden zijn talrijk maar mag niet vergeten dat het bedoeld is om eerst een schatting te maken van de mate waarin sprekers conversatie vaardigheden beheersen.

WebAssess is een demonstratieproject uit de tweede ronde. Zie www.webassess.nl.

Belscript WebAssess

K Goedemorgen u spreekt met Geert van HP, waarmee kan ik u *van dienst zijn*?
(S met welke firma spreek ik eigenlijk?)
S U spreekt met meneer Groot, ik heb een printer probleem
K wat is het type nummer van uw printer *meneer Groot?*
(S wilt u niet weten welk soort HP printer ik heb? Normaal vraagt men altijd naar het typenummer)
S ik heb een HP 3250
K kunt u mij aangeven wanneer u het apparaat gekocht heeft *meneer Groot*?
(S heb ik nog garantie of valt mijn printer weer eens buiten de garantieperiode?)
S dat is nu bijna 2 jaar geleden
K *het spijt mij* te moeten vertellen dat een eventuele reparatie buiten de garantie komt te vallen
(S Ja, en heb ik dan nog recht op garantie?)
S dat is vervelend, maar ik wil wel dat het apparaat gemaakt wordt, wat moet ik nu doen?
K Ik *help* u graag verder. Ik ga eerst met u na wat het probleem kan zijn
(S wilt u niet weten wat het probleem is?)
S laten we dat dan maar doen

Arjan van Hessen coördineert het WebAssess- project.

SoNaR

STEVIN Nederlandstalig Referentiecorpus

Het SoNaR-project volgt op het D-Coi (Dutch Language Corpus Initiative) project dat eerder binnen STEVIN werd gefinancierd. Het had tot doel een blauwdruk voor een dergelijk corpus te ontwikkelen en alle noodzakelijke voorbereidende werkzaamheden te verrichten.

Nelleke Oostdijk

Het SoNaR-project beoogt de aanleg van een groot corpus (minimaal 500 miljoen woorden) hedendaags geschreven Nederlands dat als algemene referentie kan dienen voor allerlei onderzoek naar taal en taalgebruik. Daarbij valt te denken aan beschrijvend onderzoek (zoals dat zijn weerslag vindt in bijv. woordenboeken en grammatica's), maar ook aan onderzoek op het gebied van de taal- en spraaktechnologie. Daarvoor dient men te beschikken over grote hoeveelheden tekst met de mogelijkheid deze met eigen software te kunnen bewerken.

Het corpus zal worden samengesteld volgens het ontwerp van het D-Coi project. In het corpus worden enkel (standaard) Nederlandstalige teksten opgenomen van na 1954. Dit kunnen teksten zijn die geschreven werden door moedertaalsprekers van het Nederlands, maar ook teksten die door professionele vertalers uit een vreemde taal werden vertaald naar het Nederlands. Er worden teksten verzameld die afkomstig zijn uit uiteenlopende domeinen en genres, waarbij tevens gekeken wordt naar een brede afdekking van onderwerpen. Voor zover mogelijk worden volledige teksten opgenomen. Dit voorkomt dat op voorhand bepaalde soorten onderzoek worden uitgesloten. In het corpus worden teksten opgenomen van zowel Nederlandse als Vlaamse auteurs.

Bij het verzamelen van teksten gaat speciale aandacht uit naar teksten waar lezers mee in aanraking komen via nieuwe media. Het gaat daarbij onder meer om teksten op websites, sms-berichten, e-mail en chats. Over het gebruik van taal in dit soort teksten is nog relatief weinig bekend.

Het accent van de werkzaamheden gedurende de eerste fase van het project ligt op de acquisitie van tekstmateriaal. Belangrijk hierbij is dat er voor alle teksten een adequate regeling wordt getroffen met de auteursrechthebbenden. Een dergelijke regeling moet garanderen dat de teksten beschikbaar zijn voor onderzoeksdoeleinden, terwijl tegelijkertijd wordt vastgelegd dat teksten niet opnieuw worden gepubliceerd. Toegang tot de tekst als tekst is een essentiële voorwaarde: alleen dan is het mogelijk de noodzakelijke taalmodellen te trainen en allerlei afgeleide informatie (o.a. gegevens over het gebruik en de frequentie van voorkomen van bepaalde woorden of constructies) systematisch te verzamelen. Dergelijke informatie kan vervolgens worden gebruikt om taal- en/of spraaktechnologische applicaties te ontwikkelen. Voorbeelden van dergelijke toepassingen zijn onder meer zoeksystemen, spelling- en grammaticacheckers en dicteersystemen.

Het SoNaR-project wordt uitgevoerd door volgende onderzoeksgroepen:

- Centre for Language and Speech Technology (CLST), Radboud Universiteit Nijmegen
- Induction of Linguistic Knowledge (ILK), Universiteit Tilburg
- Human Media Interaction (HMI), Universiteit Twente
- Departement Vertaalkunde, Hogeschool Gent

Bij verdere financiering wordt het consortium uitgebreid met het Centrum voor Computer Linguistiek (CCL), gevestigd aan de Katholieke Universiteit Leuven en het Utrecht Institute of Linguistics OTS, gevestigd aan de Universiteit Utrecht. In de volgende fase wordt de acquisitie verder voortgezet, maar wordt daarnaast werk gemaakt van de verdere verrijking van het corpus, o.a. door POS tagging en lemmatisering en het toevoegen van semantische annotatie

Oproep:
Graag doen we een beroep op alle eigenaars van Nederlandstalige teksten in elektronisch formaat om deze beschikbaar te stellen voor SoNaR. Met name bedrijven en organisaties nodigen we uit teksten aan te leveren uit hun beroepspraktijk, zoals gebruikershandleidingen, nieuwsbrieven, personeelskranten, jaarverslagen en websites.

Contactpersoon: Nelleke Oostdijk (n.oostdijk@let.ru.nl)

Het SoNaR-project is een project uit de tweede tenderoproep. Het voortraject loopt tot eind 2008.

Nelleke Oostdijk coördineert het SoNaR-project.

Dit is
Autonomata Too

Henk van den Heuvel

Met spraakherkenning kun je alle kanten uit. Dat geldt zeker voor navigatiesystemen. Navigatiesystemen worden gekenmerkt door het gebruik van namen, en meer algemeen door het gebruik van Points of Interest (POIs). Namen leveren voor automatische spraakherkenning aparte problemen op. Autonomata Too zal oplossingen voor deze problemen onderzoeken en een demonstrator bouwen waarin deze oplossingen kunnen worden getoond en geëvalueerd.

Het letter per letter ingeven van de bestemming in een navigatiesysteem is niet echt comfortabel te noemen, zelfs indien die bestemming meestal reeds na het intikken van de eerste drie of vier letters op het scherm verschijnt. Het kunnen inspreken van de bestemming is een vooruitgang, tenminste, als die ingesproken bestemmingen voldoende betrouwbaar herkend worden. Automatische spraakherkenning (ASH) van namen is om een aantal redenen lastig. Vaak is er een andere relatie tussen schrift en klank dan bij gewone woorden, zodat de gangbare verklankingsregels niet opgaan. De spellingwijze van namen is soms verouderd (zoals in *Blaauw, Aelderts*), en er zijn namen van buitenlandse herkomst (*Korlaelçi Demirgökçen, Gillygooley Road*). En soms verbaas je je alleen maar (vergelijk de uitspraak van het woord vulpen met die van de naam in *Gebr. Van Vulpen*). Het komt bovendien voor dat de spreker de juiste uitspraak van de naam niet kent en een alternatieve uitspraak produceert die we een uitspraakvariatie noemen. In elk geval vereist een juiste herkenning dat de werkelijk gebruikte uitspraak in het lexicon van de ASH-machine voorkomt.

Autonomata

Het voorloperproject Autonomata leverde resources op die doelgericht onderzoek naar de ASH van namen toelaten. Ten behoeve van de letter-naar-klank-omzetting (kortweg g2p-omzetting) van namen, zijn goede g2p-omzetters voor persoonsnamen (voornamen, familienamen) en toponiemen (straatnamen, plaatsnamen) gemaakt. En er is een transcriptietoolbox ontwikkeld waarmee men uitgaand van eigen trainingslexica nauwkeurige g2p-omzetters voor andere naamsoorten (bijv. POIs) kan bouwen.
Om de uitspraakvariaties in kaart te brengen is een corpus van voorgelezen namen gebouwd. Daarin lazen 240 sprekers elk 250 namen voor. Naast Nederlandse en Vlaamse namen, zitten er ook namen van Franse, Engelse, Turkse en Marokaanse herkomst bij. Naast Nederlanders en Vlamingen, werden er ook anderstalige sprekers gerekruteerd die het Nederlands wel tot op zekere hoogte beheersen. Alle naamuitspraken zijn fonetisch getranscribeerd[1].

Autonomata Too

Autonomata Too betekent: Autonomata Transfer Of Output. In dit project gaan we de resources van Autonomata inzetten om een demoherkenner voor POIs te maken.
In onze toepassing leggen we ons toe op de ASH van overnachtingsadressen en eetgelegenheden in twee grote steden (vermoedelijk Amsterdam en Antwerpen).
De Autonomata transcriptietoolbox zal worden ingezet voor de bouw van een g2p-omzetter voor POIs en voor het leren modelleren van uitspraakvariaties zoals die te vinden zijn in het Autonomata namencorpus. De bedoeling is om een zo goed mogelijk lexicon van de ASH-machine te krijgen. Aangezien namen vaak delen van vreemde herkomst bevatten zal ook onderzocht worden hoe de ASH verder te verbeteren valt door ook een (minimaal) aantal akoestische modellen voor buitenlandse fonemen te gebruiken.

Uitvoerders

De projectuitvoerders zijn de universiteiten van Nijmegen (CLST), Gent (ELIS) en Utrecht (UiL-OTS) tezamen met de bedrijven Nuance en TeleAtlas.
Het consortium verheugt zich erop om dit project in het kader van STEVIN uit te voeren. We werken aan resultaten waar velen van u alle kanten mee uit kunnen.

Autonomata is een project uit de derde ronde. Het project loopt tot februari 2010. Zie: http://lands.let.ru.nl/projects/AutonomataToo/

1 De resources van Autonomata zijn via de TST-Centrale te verkrijgen (zie www.tst.inl.nl)

Henk van den Heuvel coördineert het Autonomata-Too project.

DAISY

Dutch lAnguage Investigation of Summarization technologY

Het samenvatten van tekst is dikwijls noodzakelijk wanneer we informatie doorzoeken of selecteren in documentarchieven. De actuele technologie voor tekstsamenvatting blijft grotendeels beperkt tot het extraheren van belangrijke zinnen uit een tekst. Er is in het verleden ook weinig onderzoek verricht naar het samenvatten van Nederlandstalige teksten. Het project DAISY wil daarin verandering brengen en wil essentiële technologieën ontwikkelen voor het automatisch samenvatten van Nederlandstalige informatieve teksten. Er worden tijdens het DAISY project innovatieve algoritmen ontwikkeld voor het detecteren van de belangrijkheid van bepaalde inhoud in een tekst, voor de retorische classificatie van de inhoud, de compressie van zinnen en voor tekstgeneratie. Daarnaast wordt een demonstrator ontwikkeld samen met het bedrijf Q-Go.

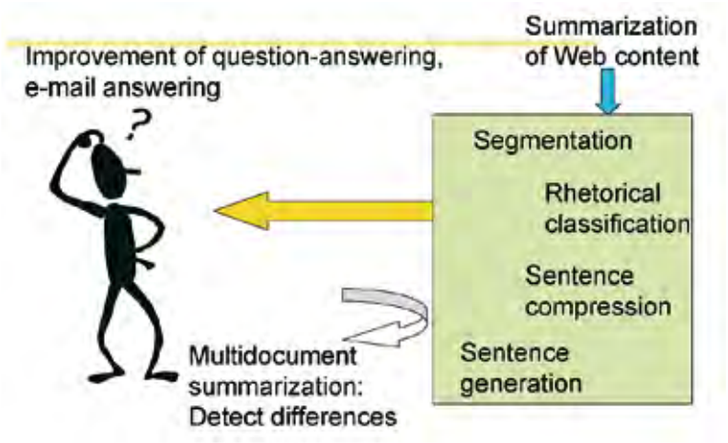
De methoden voor het ontdekken van belangrijke informatie maken gebruik van eigenschappen van het discours en de syntactische en functionele eigenschappen van de constituenten van een zin. Voor de syntactische analyse van de zinnen maken we gebruik van de Alpino-parser. Voor de tekstgeneratie wordt de Alpino parser uitgebreid op basis van abstracties van afhankelijkheidsstructuren zoals ontwikkeld in het Corpus Gesproken Nederlands (CGN), en gebruikt in de D-COI en LASSY projecten. In het onderzoek wordt een sterke nadruk gelegd op het gebruik van technieken van machineleren.

De ontwikkelde technologieën worden publiekelijk beschikbaar gesteld via een demonstrator. Deze demonstrator is toegankelijk via een webgebaseerde interface die toelaat dat gebruikers voorbeeldteksten in de vorm van een opgeladen of een in een tekstvak getypte tekst kunnen laten samenvatten.

We streven ernaar drie types van output te genereren:

1) een samenvatting in de vorm van een kopje met hoofdpunten;
2) een samenvatting die de belangrijkste thema's beschrijft in de vorm van een doorlopende tekst;
3) metatags die de retorische rol van de segmenten in de tekst aanduiden.

De demonstrator wordt op verschillende manieren getest en geëvalueerd door het bedrijf Q-go in een vraag-antwoordsysteem dat vragen beantwoordt over informatieve Webpagina's (in de domeinen van financiën, verzekeringen en luchtvaart). Naast deze testcase hebben de technologieën van DAISY zeer vele toepassingsmogelijkheden zoals het samenvatten van tekst zodat deze gemakkelijker toonbaar is via kleine schermen (bijv. van een smartphone), het filteren en ontsluiten van informatie, en het automatisch beantwoorden van email.



DAISY is een project uit de derde ronde. Het loopt tot eind april 2011.

Marie-Francine Moens

Marie-Francine Moens coördineert het DAISY-project.

DISCO

Nederlands leren spreken met hulp van de computer

Voor anderstaligen die Nederlands willen leren spreken zou het heel handig zijn als zij daarvoor een computerprogramma zouden kunnen gebruiken dat hen corrigeert als ze fouten maken. Om dit mogelijk te maken wordt in het project DISCO ('Development and Integration of Speech technology into COurseware for language learning') gewerkt aan de ontwikkeling van een dergelijk programma dat studenten Nederlands als tweede taal (NT2) feedback geeft over hun uitspraak, woordvolgorde en woordvormen.

Helmer Strik en Catia Cucchiarini

Dit STEVIN-project duurt drie jaar en is een voortzetting van eerder onderzoek (het project Dutch-CAPT), waarin al gewerkt is aan een computerprogramma dat feedback geeft op de uitspraak van het Nederlands. Nieuw aan dit vervolgonderzoek is dat er nu technologie wordt ontwikkeld die ook fouten in de syntaxis (woordvolgorde) en de morfologie (woordvormen) corrigeert. Een zin als "Ik naar huis gaan" moet straks door de computer worden herkend als fout en worden gecorrigeerd naar: "Ik ga naar huis". Om feedback te kunnen geven op morfologische fouten moet je kunnen herkennen dat iemand i.p.v. 'loopt', 'lopete' of 'lopen' zei en voor uitspraak moet je kunnen vaststellen dat iemand 'lopt' zei i.p.v. 'loopt', of dat de 'h' uitgesproken werd als 'g'. Standaard spraakherkenning is voor al deze doeleinden niet geschikt. Daarom zal specialistische spraakherkenningstechnologie ontwikkeld moeten worden die, in combinatie met een geschikt ontwerp van het leersysteem, het mogelijk maakt om dit soort vaardigheden te oefenen op een zo natuurlijk mogelijke manier, bijvoorbeeld in een dialoogsituatie.

Er bestaan al vele computerprogramma's voor het leren van talen, programma's waarbij de student kan lezen, luisteren en reageren via toetsenbord of muis. Maar deze programma's kunnen geen feedback geven over de uitspraak, morfologie en syntaxis van gesproken uitingen van de student. Oefenen van mondelinge taalvaardigheid is belangrijk, omdat spreken een essentiële vaardigheid is voor alle taalleerders, waarvoor in de klas meestal niet voldoende gelegenheid is. Oefenen met schrijven is dan niet voldoende omdat goede schriftelijke taalvaardigheid niet automatisch leidt tot goede mondelinge taalvaardigheid.

Een groot voordeel van een computerprogramma dat feedback geeft aan een NT2-student, is dat die daardoor niet afhankelijk is van de aanwezigheid van een docent. Studenten kunnen dus langer oefenen, wanneer en waar ze zelf willen. Bovendien moeten veel buitenlanders die in Nederland willen komen wonen tegenwoordig in hun eigen land al een taaltoets afleggen. Voor de voorbereiding daarop biedt dergelijke technologie ook een uitkomst. Op langere termijn zijn er ook mogelijkheden voor het ontwikkelen van vergelijkbare technologie voor mensen met een spraakhandicap.

Betrokken bij DISCO zijn het Centre for Language and Speech Technology (Catia Cucchiarini, Joost van Doremalen en Helmer Strik) en het Universitair Taal- en Communicatiecentrum (Ghislaine Giezenaar) van de Radboud Universiteit Nijmegen, het bedrijf Polderland Language & Speech Technology (Peter Beinema en Tanja Gaustad van Zaanen) en Linguapolis (Jozef Colpaert en Frederik Cornillie) van de Universiteit Antwerpen.

DISCO is een project uit de derde ronde. Het loopt tot eind januari 2011.
Zie http://lands.let.ru.nl/~strik/research/DISCO/

Helmer Strik
coördineert het DISCO-project.

DuOMAn

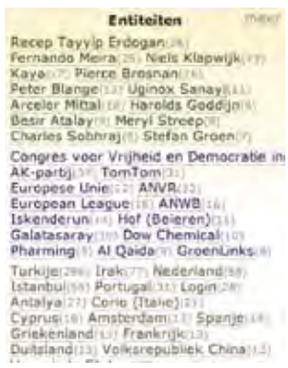
Dutch Language Online Media Analysis

Media-analysten houden zich bezig met onderzoek naar de media reputatie van organisaties. Een belangrijk doel is om klanten te laat zien en begrijpen hoe er door media, journalisten en bloggers over hen wordt geschreven of gesproken.

Er is steeds meer online materiaal, zowel geredigeerd als user generated content in de vorm van blogs of bijdragen aan discussiefora. Hierdoor krijgen media-analysten in principe toegang tot een steeds grotere hoeveelheid opinies, maar efficiënte en effectieve ontsluitingsmethoden zijn nodig om hier werkelijk gebruik van te kunnen maken.

Het DuOMAn project voert een ambitieuze onderzoeksagenda uit die zal resulteren in de ontwikkeling van Nederlandstalige resources en tools voor het identificeren en aggregeren van sentimenten in online bronnen. In het bijzonder zal het project werken aan het herkennen en extraheren van zogenaamde steun- en kritiekrelaties: belanghebbende X steunt/bekritiseert Y met betrekking tot onderwerp Z (hier zijn X en Y actoren, individuen of groepen, en Z kan vrijwel alles zijn). Concrete voorbeelden van dergelijke relaties zijn “Aboutaleb verwijt Verdonk dat ze politieke munt slaat uit de moord op Van Gogh” of “Marcel van Dam hekelde kersvers kandidaat-Kamerlid Mei li Vos, die in het Volkskrant Magazine van afgelopen weekend de opvatting ventileerde dat gepensioneerden met een goed pensioen eigenlijk helemaal geen AOW zouden moeten krijgen.”

Voor het realiseren van deze ambities voert het DuOMAn-project onderzoek uit in een drietal richtingen: het herkennen van sentiment in zowel nieuws als user generated content plus het ontwikkelen van de benodigde lexicale bronnen; het herkennen van



entiteiten (personen, producten, organisaties, etc.) in nieuws en user generated content (zie Figuur links beneden); en het aggregeren van sentimenten, opnieuw zowel in nieuws als in user generated content (zie Figuur boven). Deze laatste stap is onmisbaar om effectief om te kunnen gaan met grote hoeveelheden opinies die geëxtraheerd zijn van het web. Dit behelst onder andere het automatisch ontdekken van thema’s rondom een gegeven onderwerp of persoon: waar heeft men het over als men over dat onderwerp of die persoon spreekt? Ook zal DuOMAn in dit kader werken aan het oplossen van corefererende uitdrukkingen over meerdere documenten. Ter illustratie: in een recente pilotstudie vonden we dat naar Anneke Grönloh in één nieuwsbericht wordt verwezen met Mw. Grönloh, in een blogpost met Anneke G. en in een bijdrage aan een discussieforum met Mevr. Kreunlo.

In een geplande publieke demonstrator zal DuOMAn gebruikmaken van online nieuwsbronnen, blogs en discussiefora om zowel media-analysten als het brede publiek zeer gerichte geopinieerde informatie te geven over mensen, producten en onderwerpen.

DuOMAn is een samenwerkingsverband tussen de Universiteit van Amsterdam, Trend-Light Netherlands B.V., GridLine B.V. de Rijksuniversiteit Groningen en de Hogeschool Gent.

DuOMAn is een project uit de derde ronde. Het loopt tot eind maart 2011. Zie http://ilps.science.uva.nl/Projects/DuOMAn/

Maarten de Rijke

Maarten de Rijke coördineert het DuOMAn-project.

PaCo-MT

Parse- en corpusgebaseerde automatische vertaling

In het PaCo-MT project onderzoeken het Centrum voor Computerlinguïstiek van de Katholieke Universiteit Leuven en de groep Alfa-informatica van de Rijksuniversiteit Groningen, in samenwerking met het vertaalbureau OneLiner bvba, een nieuwe methode om aan automatische vertaling te doen.

Vincent Vandeghinste

Deze methode combineert technieken uit de traditionele regelgebaseerde aanpak voor machinevertaling met technieken uit de recentere datageoriënteerde benadering voor automatische vertaling. Er wordt een prototype geïmplementeerd voor de taalparen Nederlands-Frans, Frans-Nederlands, Nederlands-Engels en Engels-Nederlands.

Vertaalbedrijven maken momenteel niet veel gebruik van systemen voor automatische vertaling, maar maken wel veelvuldig gebruik van vertaalgeheugens, waarin paragrafen en zinnen die reeds vertaald werden worden opgeslagen. Bij het opnieuw voorkomen van deze fragmenten wordt de opgeslagen vertaling uit het geheugen opgehaald. De nieuwe benadering die wij in PaCo-MT onderzoeken maakt eveneens gebruik van reeds vertaalde tekstfragmenten. Waar vertaalgeheugens meestal slechts een exacte match uit het geheugen kunnen oproepen, zal ons systeem een hogere flexibiliteit toestaan en kunnen ook fuzzy matches gebruikt worden.

Het systeem maakt zoveel mogelijk gebruik van reeds bestaande corpora, zowel monolinguale (zoals ondermeer het Corpus Gesproken Nederlands en de Lassy-treebank voor het Nederlands) als parallelle corpora bestaande uit reeds vertaalde teksten (zoals bijvoorbeeld de teksten van het Europees Parlement).

Daarnaast maakt het systeem zoveel mogelijk gebruik van reeds bestaande taaltechnologie, vooral wat betreft de zinsanalyse-software die voor het Nederlands, Frans en Engels gebruikt wordt. Er wordt ook gebruikgemaakt van bestaande aligneringssoftware, die bij eerder vertaalde fragmenten aanduidt welke onderdelen in de bronzin (bijv. welke woorden) vertalingen zijn van welke fragmenten in de doelzin. Deze aligneringssoftware zal na onderzoek verder verfijnd worden. Op deze manier wordt een uitgebreid woordenboek opgebouwd, dat niet enkel uit enkelvoudige woorden bestaat maar eveneens volledige zinssneden bevat. Daarnaast wordt een uitgebreide post-editing component voorzien, waar de uitvoer van PaCo-MT door menselijke posteditors aangepast en verbeterd kan worden: door woorden of woordgroepen te verplaatsen, een ander vertaalalternatief te kiezen (voor het woord, de woordgroep of de volledige zin), of woorden toe te voegen of te verwijderen. Deze post-editing informatie wordt onmiddellijk teruggekoppeld naar de modellen en het woordenboek die in het systeem gebruikt worden om vertalingen te genereren. Op deze manier functioneert PaCo-MT zowel als een vertaalgeheugen als als een machinevertaalsysteem.

Verschillende gebruikers van het systeem kunnen verschillende profielen aanmaken, waarbij elke gebruiker zijn/haar eigen vertaalgeheugen kan activeren, naast een algemeen vertaalgeheugen dat gebaseerd is op publieke informatie, en het systeem zich voor elke gebruiker baseert op de vertaalvoorkeuren van deze gebruiker.

Het systeem zal geëvalueerd worden aan de hand van een gouden standaard van zinnen met referentievertalingen. Naast klassieke academische maten voor MT-kwaliteit zal er ook aandacht zijn voor de mate waarin het vertaalproces versneld wordt, ten opzichte van een controleconditie waarin PaCo-MT niet gebruikt wordt.

Naast het prototype van het vertaalsysteem zelf zal het project leiden tot uitgebreide parallelle en monolinguale treebanks (corpora van geanalyseerde zinnen) en ontwikkelings- en evaluatietestsets met verschillende referentievertalingen voor de ontwikkeling en evaluatie van machinevertaalsystemen.

PaCo-MT is een project uit de derde ronde. Het loopt tot eind januari 2011.

Vincent Vandeghinste coördineert het PaCO-MT-project.

Alfabetisering met een luisterende computer

Veel mensen realiseren zich het niet, maar analfabetisme is een groot maatschappelijk en economisch probleem. Alleen Nederland telt al ongeveer 1,3 miljoen volwassenen die moeite hebben met lezen en schrijven en daarvan zijn er 250.000 volledig analfabeet.

Helmer Strik en Ad Bakker

Een mogelijke oplossing is om meer lessen voor deze mensen te organiseren. Daarvoor zijn meer leerkrachten nodig en omdat veel laaggeletterden zich schamen voor hun achterstand is de drempel vaak erg hoog. Een heel andere oplossing is computerprogramma's voor alfabetisering. Er bestaan al wel computergebaseerde alfabetiseringsmethodes, maar in Nederland is er nog geen alfabetiseringsprogramma dat ook luistert. In het project 'Alfabetisering met een luisterende computer' zal een demonstratiesysteem van een dergelijk programma ontwikkeld worden.

Onlangs verscheen bij de uitgeverij Boom het 'Alfabetisering Anderstaligen Plan' (AAP), ontwikkeld door Ad Bakker. AAP bestaat uit een kerngedeelte 'klanktekenkoppeling', met een voortraject bestaande uit twee delen: luisteren en spreken en schrijven. Het idee is dat de cursist op het moment dat deze eigenlijke alfabetisering begint, de klanken van het Nederlands auditief kan onderscheiden, en de vaardigheid heeft om de lettertekens van het Nederlands te schrijven. Bij de eigenlijke alfabetisering moet de cursist alle energie in de ontcijfering van de alfabetische code kunnen steken en niet meer in de randvaardigheden die in het voortraject behandeld zijn. Voor de auditieve oefeningen zegt de cursist na wat hij hoort, of beantwoordt hij eenvoudige tweekeuzevragen. Aan de hand van de mondelinge reacties kan worden bepaald of de cursist over het register van de Nederlandse spraakklanken beschikt. Aan de ontbrekende klanken kan gericht gewerkt worden.
In een conventionele lessituatie is dit bijzonder lastig. De cursist die een bepaalde klank niet hoort, is afhankelijk van iemand die hem erop attendeert. Als de fout zich blijft voordoen, raakt de omgeving er al snel aan gewend dat de cursist een klank op 'zijn manier' uitspreekt. Een luisterende computer kan echter de cursist de feedback geven die hij nodig heeft.

In het project 'Alfabetisering met een luisterende computer' zullen BEMO (Ad Bakker) en Boom (Mirjam Haasnoot) samenwerken met de taal- en spraaktechnologiegroep van de Radboud Universiteit Nijmegen (Catia Cucchiarini en Helmer Strik) en het bedrijf Polderland Language & Speech Technology (Peter Beinema). Automatische spraakherkenning (ASH) zal ingezet worden om AAP te laten luisteren.

ASH is een techniek die meestal gebruikt wordt om woorden in spraak te herkennen, maar een aangepaste vorm van ASH kan ook gebruikt worden om klanken te herkennen. Hierdoor wordt het mogelijk om de koppeling tussen klanken en tekens te oefenen: de cursist ziet bijvoorbeeld een plaatje van een boom, spreekt de lange klinker 'oo', met behulp van ASH wordt gecontroleerd of het de juiste klank is, en de cursist krijgt onmiddellijk feedback hierover. Ook maakt ASH het mogelijk om te oefenen met (beginnend) lezen: de cursist leest woorden of (korte) uitingen voor, en met behulp van ASH wordt gecontroleerd of die correct zijn voorgelezen.

Het is ook belangrijk om een user-interface te ontwikkelen die geschikt is voor gebruikers die moeite hebben met het lezen, zowel voor het eliciteren van de antwoorden als voor het geven van feedback. Het hoofddoel van dit project is het implementeren, testen, en beschikbaar stellen van een demonstratiesysteem dat duidelijk maakt dat bestaande spraaktechnologie nuttig ingezet kan worden in de alfabetiseringsmethode AAP, en daarmee ook in andere leeromgevingen.

AAP is een demonstratieproject uit de derde ronde. Het loopt tot eind juni 2009.

Helmer Strik en Ad Bakker nemen deel aan het AAP-project.

EasyInfo

Opzet van een volledig geautomatiseerde en gepersonaliseerde nieuwsdienst voor nieuwsberichten in het Nederlands

Nieuwsvoorziening via electronische kranten op internet is een veel voorkomende vorm van publicatie, maar sterk in opkomst zijn ook de zogenaamde news brokers, in sommige gevallen ook knipseldiensten genoemd. Klanten van deze news brokers kunnen een persoonlijk profiel opgeven in de vorm van trefwoorden en dat profiel wordt gebruikt om een selectie te maken uit de actuele nieuwsberichten.

Joop van Gent

Juist dit matchen van nieuwsberichten aan persoonlijke profielen lukt in de praktijk zelden goed. Bovendien komt er heel veel handwerk bij kijken: de profielen worden veelal in de vorm van regelsets en combinaties van trefwoorden opgebouwd, maar de matching is desondanks mager. Automatische methoden falen veelal omdat gebruik wordt gemaakt van simpele zoektechnologie of statistische methodes. Nog een bezwaar tegen de bestaande aanpak is dat de matching zich binnen één taal afspeelt. Wie bijvoorbeeld naast Nederlandstalige ook Engelstalige relevante nieuwsberichten wil ontvangen, moet daarvoor een extra profiel (laten) aanmaken.
Een tweede punt is de lengte van de geboden nieuwsberichten, die vaak te lang is. Iedereen die geabonneerd is op een digitale krant heeft die ervaring: het is van groot belang om het nieuws in enkele seconden te kunnen 'scannen' zodra het in de mailbox wordt ontvangen. Kranten werken natuurlijk al sinds jaar en dag zo: de krantenkop moet de lezer verleiden verder te lezen, en een 'intro' zorgt nog eens voor het nieuws in hoofdlijnen. Maar de news brokers werken zeer vaak niet alleen met kranten, maar eveneens met allerlei andere bronnen, die veel minder goed samengevat worden aangeboden.

In ***EasyInfo*** willen drie partijen aantonen dat het ook anders kan, namelijk met natuurlijke taaltechnogie.

Irion Technologies beschikt over een op natuurlijke taaltechnologie gebaseerd classificatiesysteem voor nieuwsberichten in het Nederlands, die geheel automatisch volgens de IPTC-classificatie kunnen worden geclassificeerd. De IPTC-classificatie is een internationale standaard voor nieuwsrubricering, ontworpen door de International Press and Telecom Council (IPTC).

Carp Technologies beschikt reeds over samenvattechnologie die breed kan worden ingezet en ook specifiek kan worden ingericht op het samenvaten van nieuwsberichten. Daarbij kunnen de zinnen uit de oorspronkelijke tekst bovendien worden geparafraseerd. Dat is iets wat normaliter ook gebeurt als handmatig wordt samengevat.

MD Info is reeds een multimediale aanbieder van gepersonaliseerde zakelijke informatie en heeft reeds een complete infrastructuur die geschikt is voor dit type nieuwsvoorziening.

Kern van het project is: aantonen dat met de bewezen IPTC-classificatie van Irion Technologies, en met de eveneens bewezen samenvattingengenerator van Carp Technologies news content uit diverse bronnen op grote schaal en geheel automatisch kan worden voorzien van de juiste profielen en bruikbare samenvattingen, waardoor een volledig geautomatiseerde, en dus kosteneffectieve, knipseldienst ontstaat, die kwalitatief uitstekend kan concurreren met de handmatige knipseldiensten. Irion en Carp stellen via het internet een classificatiefunctie en samenvatterfunctie beschikbaar, en MD Info koppelt deze functies aan haar standaardplatform. Dit platform verzamelt uit een groot aantal bronnen nieuwsberichten en prepareert ze voor ontsluiting. Slechts een klein deel van de binnenkomende berichten wordt momenteel daadwerkelijk door deze redactie ontsloten, en het is met EasyInfo de bedoeling om de volledige nieuwsstromen te laten classificeren door het classificatiesysteem van Irion en te laten samenvatten door de samenvatter van Carp.

Easy Info is een demonstratieproject uit de derde ronde.

Joop van Gent coördineerde het EasyInfo-project.

HATCI

Hulp bij Auditieve Training na Cochleaire Implantatie

Filiep Vanpoucke

Doofheid of ernstige slechthorendheid wordt vaak veroorzaakt door lokale schade aan de haarcellen in het binnenoor of cochlea. Voor deze vorm van doofheid kan een cochleair implantaat (CI) een uitweg bieden. Deze gehoorprothese overbrugt de beschadigde haarcellen door rechtstreekse elektrische stimulatie van de gehoorzenuw. Een recente evolutie is dat kinderen op jongere leeftijd geïmplanteerd worden.

Na de ingreep moet de patiënt zich aanpassen aan zijn of haar nieuwe manier van horen. Binnen de logopedische behandeling wordt hiervoor o.a. gebruikgemaakt van “speech tracking”, een auditieve oefening waarbij onder begeleiding van een spraaktherapeut de patiënt een voorgelezen tekst zin na zin moet nazeggen. Naast het auditieve aspect zijn deze oefeningen ook belangrijk voor de algemene taalontwikkeling van kinderen. Hun beheersing van zinsbouw, grammatica en woordenschat kan gradueel worden opgebouwd door de complexiteit van de voorgelezen teksten te verhogen.

Deze oefeningen kunnen slechts in gespecialiseerde centra aangeboden worden, wat het voor de patiënt tijdsintensief en eventueel ook vrij duur maakt. In de praktijk gaan de meeste mensen hooguit één of twee keer per week naar de therapeut. Een computerpakket zou in principe ondersteunend kunnen werken doordat de patiënt vaker en in zijn eigen omgeving (thuis of school) zou kunnen oefenen. De uitdaging is uiteraard dat in dit geval de computer de correctheid van het antwoord van de patiënt moet kunnen beoordelen.

Het project HATCI richt zich op de ontwikkeling van een computerpakket dat gebruikmaakt van automatische spraakherkenning (ASH). De doelgroep is zowel kinderen als volwassenen die reeds over een goede articulatie beschikken, maar voor wie de hoorvaardigheid, de woordenschat en de zinsontwikkeling evenals het auditief geheugen verder moeten worden gestimuleerd. In de applicatie zal vooraf opgenomen spraak, al dan niet vergezeld van het mondbeeld, aan de patiënt worden aangeboden. De herhaling van deze uiting wordt opgenomen en beoordeeld d.m.v. ASH.

De applicatie kan ingezet worden als meetinstrument of als therapeutisch instrument. In het eerste geval beperkt de functionaliteit zich tot het registreren van het aantal woorden dat per minuut correct kan worden herhaald. De computer houdt de herhalingsfouten bij en berekent een score foutsoort. In het tweede geval wordt feedback gegeven aan de patiënt. Binnen het project wordt gezocht naar een optimale manier van feedback. Bij gebruik

van automatische spraakherkenning is er verhoogde kans op onjuist gedetecteerde fouten (valse positieven), wat erg storend zou zijn voor de patiënt en mogelijk het leerproces nadelig zou beïnvloeden. De keerzijde van de medaille is dat ook een substantieel aantal fouten ongedetecteerd blijft. Desalniettemin gaan we ervan uit dat dit ondersteuningsmiddel ook onder deze omstandigheden een positief effect zal hebben op het leerproces.

In het acht maanden durende project wordt in eerste instantie leermateriaal aangemaakt, opgenomen en voorzien van een annotatie van verwachte herhalingsfouten. Parallel wordt de gebruikersinterface ontwikkeld en de ASH aangepast aan deze taak.

HATCI is een samenwerking tussen Advanced Bionics N.V., het Onafhankelijk Informatiecentrum over Cochleaire Implantatie (ONICI) en de K.U.Leuven, departement Elektrotechniek (ESAT).

HATCI is een demonstratieproject uit de derde ronde. Het loopt tot eind november 2008.

Filiep Vanpoucke
coördineert het HATCI-project

NEON

NEderlandstalige ONdertiteling

De ultieme droom is een volledig automatisch systeem waarbij alles dat in het TV-programma wordt gezegd, realtime foutloos wordt herkend waarbij m.b.v. kleuren ook wordt aangeven wie wanneer spreekt.

Arjan van Hessen

Realiteit

Naspreken

Wat redelijk werkt is 'respeaking'. In de studio spreekt een getrainde spreker na wat er in de uitzending wordt gezegd en geeft via knoppen aan wie hij naspreekt. Deze spraak wordt met spraakherkenning omgezet in ondertitels. Doordat een getraind persoonlijk akoestisch model beschikbaar is, is de herkenningsgraad hoog. Toch kost één uur uitzending één uur naspreken en nog één uur oplijnen (handmatig aangeven wanneer de herkende spraak op het scherm moet komen). In live-programma's wordt het herkenningsresultaat wel onmiddellijk uitgezonden wat te merken is aan het storende, soms seconden, achterhollen van de ondertitels op de audio.

Oplijnen

In het NEON-project ligt de focus op het oplijnen van bestaande teksten (scripts, autocues) met gesproken audio om automatisch het tijdstip te vinden waarop een ondertitel in beeld moet verschijnen. Het materiaal waarmee gewerkt wordt is eerder 'goed gelijkend' zoals bij soaps ("Onderweg naar Morgen") waarbij het script redelijk overeenkomt met de gesproken spraak.
Kleine afwijkingen tussen geschreven en gesproken woorden vormen voor oplijning geen probleem, bij grote afwijkingen ligt dit anders.

Ontbrekende audio

Als afwijkingen het gevolg zijn van het ontbreken van stukken audio (programma werd ingekort) dan moet de oplijner proberen de synchronisatie tussen tekst en audio te herstellen door een tekst over te slaan: iets dat meestal redelijk goed lukt.

Extra audio

Bij extra audio waarmee geen tekst overeenkomt, is het probleem groter. De spraakherkenner moet dan terugschakelen van oplijningsmodus naar herkenningsmodus. Maar het herkennen van de spraak kan behoorlijk lastig zijn. Naast storende achtergrondgeluiden, kunnen niet-getrainde sprekers niet-grammaticaal met slechte of dialectisch getinte uitspraak spreken (aarzelingen, herhalingen). "State-of-the-Art" spraakherkenners zijn nog niet in staat om nauwkeurige, voor ondertiteling bruikbare herkenningsresultaten af te leveren. Daarom richt NEON zich voorlopig op programma's waarvoor voldoende met de audio overeenkomend tekstmateriaal aanwezig is.

Sprekeridentificatie

"Wie spreekt"?, (nodig om ondertitels de juiste kleur te geven) wordt zoveel mogelijk afgeleid uit aanwezige scripts. Als deze informatie niet aanwezig is, wordt een segmentatiemodule gebruikt die de audio opdeelt in aparte segmenten die overeenkomen met een bepaalde spreker.

Herschrijving

Na oplijning wordt het tekstmateriaal m.b.v. taaltechnologie herschreven (niet te lang, minder moeilijke woorden, geen herhalingen als gevolg van aarzelingen. Met parafrasering worden delen van de tekst vervangen door kortere tekst
"een voortdurend toenemend aantal"
Een regelgebaseerd systeem gaat gebruikmaken van informatie van oppervlakkige ontleding en relevantiematen om te beslissen welke zinsdelen zonder gevaar verwijderd mogen worden.
Dit resultaat wordt dan onder het betreffende geluidsfragment gelegd.

Demonstrator

Het NEON-project moet leiden tot een demonstrator voor (semi-)automatische ondertiteling die omroepen in staat stelt geavanceerde toepassing van spraakherkenning te evalueren. De uitvoerders realiseren zich dat de eerste resultaten ver van de beschreven droomresultaten zullen liggen. Herkenningsfouten leiden tot oplijningsfouten, samengevatte zinnen kunnen onbegrijpelijk zijn en sprekerdetectie zal soms fout gaan. Toch is dit beter dan niets en we verwachten dat met NEON een redelijke productiviteitswinst mogelijk is.

NEON is een demonstratieproject uit de derde ronde. Het loopt tot eind maart 2009.

Arjan van Hessen
coördineert het Neon-project.

WooDy

Een sprekende en zelf corrigerende woordvoorspeller specifiek gericht op gebruik door dyslectische gebruikers

Het gebruik van woordvoorspelling is voor heel wat dyslectische gebruikers de aangewezen manier om spelfouten te voorkomen en op een snelle en efficiënte manier tekst te produceren. Spelfouten voorkomen is bij mensen met dyslexie/dysorthografie gemakkelijker te realiseren dan spelfouten corrigeren m.b.v. spellingscontrole. Zij maken immers voornamelijk inconsequente fouten. Door de combinatie te maken van 'zelfcorrectie' en woordvoorspelling, kan de focus op het begin van het woord geplaatst worden (hoofdaccent op eerste letter, afnemend voor de volgende letters). Het volstaat om de 'zelfcorrigerende' functie te beperken tot de beginletters van een woord om ondanks de foute schrijfwijze toch de correcte suggestie te blijven zien.

Probleem is echter dat de huidige woordvoorspellers geen of onvoldoende oplossingen aanreiken voor het feit dat dyslectici:

- de neiging hebben om fonetisch te gaan schrijven of te kampen hebben met letteromkeringen.
- fouten maken op het niveau van klankzuivere woorden: stemhebbend/stemloos, korte en lange klinkers in gesloten lettergreep, omkeringen, letteromissies, letteraddities, volgordefouten, enz.
- fouten maken op weetniveau: homofonen (hard, hart), tweeklanken (blauw, blouw), c of k, i of ie, -tie/-sie of –cie enz…
- fouten maken op regelniveau: open en gesloten lettergrepen, verdubbeling, ontdubbeling, eind d of t,…
- voor de keuze tussen de aangeboden woordsuggesties extra ondersteuning nodig hebben, bij voorkeur aangepast aan hun kennis- en leesniveau.
- voor de keuze tussen de aangeboden woordsuggesties sterk aangewezen zijn op een perfecte uitspraak bij het voorlezen van de suggesties.
- zeer veel woorden fout schrijven, en daardoor geen gebruik kunnen of mogen maken van het automatisch laten aangroeien van woordenlijsten op basis van de ingetikte woorden.

De kern van een woordvoorspeller zijn de woordenlijsten waaruit de voorspelling wordt afgeleid en de algoritmes die gebruikt worden om te bepalen welke woorden en woordencombinaties uit de woordenlijst zullen aangereikt worden aan de gebruikers als suggestie(s) en in welke volgorde die suggesties zullen geplaatst worden. Binnen dit project zullen we een basisset aan woordenlijsten ontwikkelen (sets volgens lees/spellingsniveau, leerniveau, studierichting, vakgebied, ….), met waar nodig uitspraakcorrecties in functie van de spraaksynthese die de gebruiker wenst in te zetten, en ook de gereedschappen om die woordenlijsten aan te maken en te onderhouden. Daarnaast willen we algoritmes ontwikkelen die de dyslectische gebruiker maximaal de geschikte woorden aanreiken, rekening houdende met de eigenheden van zijn persoonlijke beperking. Hij zal kunnen aankruisen waarmee de woordvoorspeller al dan niet dient rekening te houden (fonetisch schrijven, letteromkeringen, stemhebbend stemloos, …). Het automatisch laten aangroeien van woordenlijsten zal kunnen beperkt worden tot woorden die door een spellingscontrole geraken en/of woorden die voor komen in een te definiëren 'moederlijst'. Manueel aanvullen zal natuurlijk altijd kunnen, ongeacht spellingscontrole en/of moederlijst. De woordenlijsten en algoritmes willen we dan implementeren in en demonstreren met een prototype sprekende woordvoorspeller die binnen het project ontwikkeld zal worden.

Het project richt zich zowel op dyslectische kinderen en kinderen met dysorthografie / dysgrafie als op dyslectische studenten en volwassenen. De resulterende woordvoorspeller zal echter ook perfect inzetbaar zijn in alle situaties waarin individuen worden geconfronteerd met taalbeperkingen, zoals kinderen met leerproblemen, kinderen of volwassenen waarvan het Nederlands niet hun moedertaal is.

WooDy is een demonstratieproject uit de derde ronde. Het loopt tot halverwege mei 2009.

Frank Allemeersch

Frank Allemeersch coördineert het WooDy-project.

Deze DIXIT wordt u aangeboden door de deelnemers van NOTaS:

CLST
Data Archiving en Networked Services (DANS)
Dedicon
Dialogs Unlimited B.V.
Dutchear
GridLine B.V.
InTAAL B.V.
Knowledge Concepts
Logica
Maartenskliniek/Research Development & Education
Polderland Language & Speech Technology
PonTeM
Q-go.com B.V
TST- Centrale
TeleCats
TNO Defensie en Veiligheid
Universiteit Twente
Universiteit van Tilburg
Van Dale Lexicografie bv

En in het bijzonder door:

STEVIN
Nederlandse Taalunie

Bent u al op de hoogte van het Taaluniedebat?

Op 24 november kunt u in Amsterdam debatteren over burger, taal en overheid.

Debatgoeroe Peter de Geer leidt het debat.

Wees er snel bij.

Zie:
www.taalunieversum.org/taalunie/debat

Onderwijs en Taal- en Spraaktechnologie: een vruchtbare kruisbestuiving

Catia Cucchiarini

Taal- en spraaktechnologie (TST) heeft intussen een stadium bereikt waarin ze rijp en betrouwbaar genoeg wordt geacht om ingezet te worden in het onderwijs, mits er goed rekening wordt gehouden met de beperkingen die deze technologie nog steeds heeft. Het ontwikkelen van goede TST-toepassingen vergt inspanningen van deskundigen uit verschillende disciplines. Het is bijvoorbeeld belangrijk dat, afhankelijk van het doel van de toepassing, de juiste technologie wordt gekozen en dat er in het design rekening wordt gehouden met pedagogische criteria. Verder moet de technologie betrouwbaar genoeg zijn en moeten de gebruikers uiteindelijk iets kunnen leren en het liefst daar ook plezier aan beleven.

Dit besef begint gelukkig de laatste tijd door te dringen en de belangstelling voor TST in het onderwijs groeit gestaag. Dit blijkt onder andere uit het feit dat er aandacht is voor dit onderwerp in tijdschriften, conferenties en ook in nationale en internationale onderzoek- en ontwikkelingsprogramma's. Dit is voor een deel ook het geval in het Vlaams-Nederlands programma STEVIN en op Taal in Bedrijf.
Uit het overzicht van de STEVIN-projecten dat in dit nummer wordt gepresenteerd, blijkt dat er een vijftal projecten gerelateerd is aan het onderwijs, of resultaten oplevert die daarvoor ingezet zouden kunnen worden:

1. Spelling- en grammaticacontrole voor dyslectische gebruikers (Primus);
2. Spellinghulp via een chatbot (Spelspiek);
3. Development and Integration of Speech technology into COurseware for language learning (DISCO);
4. Alfabetisering Anderstaligen Plan (AAP);
5. Sprekende zelfcorrigerende woordvoorspeller voor dyslectische gebruikers (WooDy).

Naast deze STEVIN-projecten zijn in het Nederlands taalgebied ook andere projecten uitgevoerd die zich op het snijvlak bevinden van TST en onderwijs. In dit tweede deel van dit themanummer worden enkele van deze initiatieven aan u voorgesteld.

Het gaat om de volgende initiatieven:

1. het SPACE-project, waarbij gewerkt is aan de ontwikkeling van
   a. automatische, diagnostische leestoetsen;
   b. een hulpmiddel bij het leren lezen, een zogenaamde leestutor (zie de twee bijdragen van Van hamme);
2. het MASLA-project, waarbij instrumenten worden ontwikkeld en getest waarmee digitaal lesmateriaal voor Nederlands als tweede taal automatisch kan worden aangepast aan individuele leerders (zie de bijdrage van Van der Werf en Vermeer);
3. het REPETITOR-project, waarbij de Delftse methode voor NT2 verrijkt wordt met een module automatische spraakherkenning die het mogelijk maakt om uitspraak te trainen en daar automatisch feedback op te geven (zie de bijdrage van Blom en Van den Heuvel);
4. het MyWordCoach-project, waarbij een computerspel is ontwikkeld waarmee gebruikers hun Nederlandse woordenschat kunnen uitbreiden (zie de bijdrage van Daelemans en Van den Bosch);
5. het "Schoolglas Twente project", waarbij spraaktechnologie wordt ingezet om schoolboeken op een karaoke-manier te laten voorlezen (zie de bijdrage van Van Hessen).

Dankzij deze recente initiatieven komen er ook voor het Nederlands betere technologie en meer relevante data beschikbaar voor het ontwikkelen van TST-toepassingen voor het onderwijs. Om de samenwerking tussen taal- en spraaktechnologen en onderwijsdeskundigen te stimuleren en daardoor extra impuls te geven aan deze ontwikkelingen, heeft de Nederlandse Taalunie het initiatief genomen om gerichte bijeenkomsten te organiseren waar deze experts bij elkaar komen om uiteindelijk een inventarisatie te kunnen maken van mogelijke en wenselijke TST-toepassingen voor het onderwijs en daarbij benodigde digitale voorzieningen (zie de bijdrage van Bolhuis en Cucchiarini).

Catia Cucchiarini is senior projectleider bij de Nederlandse Taalunie en senior onderzoeker bij de Radboud Universiteit Nijmegen.

Leestoetsen geautomatiseerd met spraakherkenning

In onze maatschappij is leesvaardigheid van cruciaal belang. Reeds op de lagere school kan een leesachterstand leiden tot accumulerende leerachterstand en zelfs tot uitval op sociaal vlak. Daarom ondergaat iedereen op de lagere school minstens één maal per jaar een leestest, met als doel de kinderen die additionele begeleiding nodig hebben te identificeren. In het SPACE-project is nu aangetoond dat een dergelijke test geautomatiseerd kan worden met behulp van spraakherkenning. Dit kan tot meerdere voordelen leiden: allereerst kan de tijd die nu door de medewerkers van de centra voor leerlingenbegeleiding (CLB in Vlaanderen) wordt besteed aan het afnemen van testen, beter geïnvesteerd worden in het remediëren van leesproblemen. Daarnaast wordt de test ook meer objectief en herhaalbaar gemaakt, daar men niet meer hoeft te rekenen op menselijk en ogenblikkelijk oordeel.

Hugo Van hamme

De geautomatiseerde test is gebaseerd op de snelheid en de nauwkeurigheid bij het hardop lezen van een lijst van 40 woorden die één per één op een computerscherm verschijnen. De gehanteerde metriek, die ook in manuele tests wordt gebruikt, is de verhouding van de totale leestijd tot het aantal juist gelezen woorden. Door het inzetten van een automatische spraakherkenner kunnen we beide parameters meten en zo de leesprestatie bepalen.

In een experiment met enkele honderden kinderen aan het eind van het eerste tot en met het vierde leerjaar van het lager onderwijs werd de leesprestatie zowel manueel als automatisch bepaald en werden de kinderen op basis hiervan in vijf prestatiegroepen ingedeeld volgens de voorschriften van het CITO: in groepjes van 25%, 25%, 25%, 15% en 10% van de sterkste tot de zwakste lezertjes. Hierbij hebben we vastgesteld dat de overeenkomst tussen het automatische systeem en een manuele classificatie even groot is als de overeenkomst tussen twee onafhankelijke manuele classificaties. Hieruit besluiten we dat het niet echt nodig hoeft te zijn om personeel van de centra voor leerlingenbegeleiding naar alle scholen uit te sturen, maar dat een test-PC per school zou kunnen volstaan voor een objectieve screening.

De spraakherkenner wordt in deze toepassing meerdere taken toebedeeld. Ten eerste staat hij in voor de vaststelling dat een woord werd gelezen en dat het systeem een volgend woord kan aanbieden. Ten tweede beoordeelt hij de correctheid van de gelezen woorden. De uitingen bevatten veelal haperingen, wat dit oordeel bemoeilijkt, maar toch wordt globaal een voldoende nauwkeurigheid behaald. Ten slotte meet hij ook nauwkeurig welke tijd er verstrijkt tussen het aanbieden van een woord op het scherm en het beëindigen van de laatste leespoging hiervan. Merk op dat deze tijd per aanbieding gemeten kan worden; iets wat erg moeilijk is in een manuele test en dus mogelijk de deur opent voor nieuwe analysetechnieken.

Aangepaste spraakherkenningstechnologie blijkt zich dus goed te lenen voor deze toepassing. Een uitgebreide veldtest dringt zich op, evenals het ontwikkelen van een werk- en distributiemodel waarin de hogere complexiteit van spraakherkenningssoftware aan de scholen en testcentra toevertrouwd kan worden.

De auteur dankt alle medewerkers van het SPACE-project voor hun inzet en creativiteit bij het verwezenlijken van dit resultaat.

Voor meer informatie, zie www.esat.kuleuven.be/psi/spraak/projects/SPACE.

Het SPACE project is een samenwerking tussen K.U.Leuven, U.Gent, V.U.Brussel en U.Antwerpen met de steun van IWT-Vlaanderen en wordt gecoördineerd door **Hugo Van hamme.**

1 SPACE = Speech Algorithms for Clinical and Educational applications. SPACE is een project gesubsidiëerd door het IWT Vlaanderen in het SBO-programma.

De leesvaardigheid aanscherpen met een leestutor

Een programma dat echt luistert bij het hardop lezen

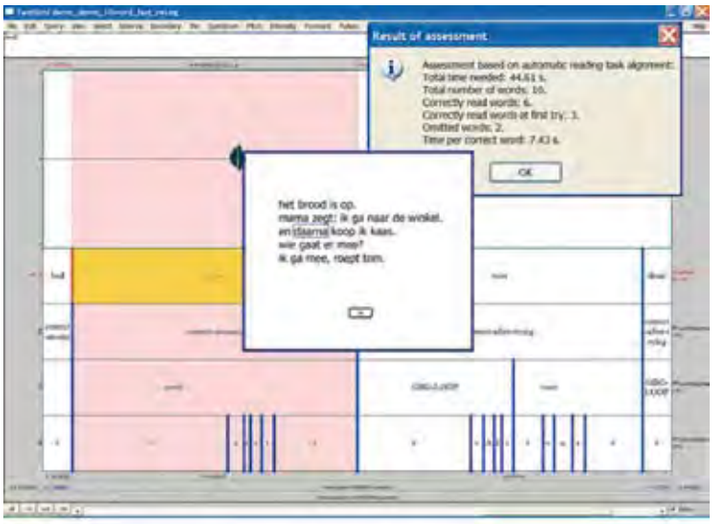
Je kunt op de markt allerlei educatieve software vinden die kinderen kunnen helpen bij het ontwikkelen van hun rekenvaardigheid, hun woordenschat, hun behendigheid, hun luistervaardigheid, enz. Leesvaardigheidsontwikkeling op de beproefde manier van het hardop lezen met terugkoppeling, dat ligt wat moeilijker...

Hugo Van hamme

In het SPACE-project wordt gebruikgemaakt van spraaktechnologie om een leestutor te bouwen die echt naar de gebruiker luistert. De doelgroep zijn kinderen uit de lagere school of lezers met dyslexie. Voor de wat zwakkere lezers kan dit computerprogramma aanvullende oefening naast het klasgebeuren opleveren. Voor dyslectici is het vandaag vaak onhaalbaar om het benodigde aantal contacturen met een therapeut te realiseren en kan deze software uitkomst bieden om het door een therapeut uitgedokterde programma thuis verder af te werken.

In het SPACE-project werd allereerst een batterij teksten van verschillende AVI-leesniveaus aangemaakt. Deze teksten worden op een computerscherm aangeboden in goed gekozen hoeveelheden. Een spraakherkenner volgt vervolgens de leespositie van het kind. Desgewenst kan de leespositie op het scherm aangegeven worden met een cursor, als substituut voor de volgende vinger op papier. Wordt het einde van de tekst op het scherm bereikt, dan gaat het programma automatisch over op de volgende pagina. Het programma volgt de leespositie ook al worden er leesfouten gemaakt door te haperen, woorden over te slaan, woorden meerdere malen te lezen, zinsdelen te hernemen, of zich van alinea te vergissen. Deze fouten worden wel gedetecteerd en kunnen desgewenst aan de lezer teruggekoppeld worden of simpelweg opgeslagen worden als gebeurtenis ter informatie van de taalleerkracht of therapeut. Zijn er problemen bij het lezen van een woord, dan kan het programma hierbij helpen: twijfelt de lezer te lang, of klikt hij/zij op een moeilijk te lezen woord, dan zal dit met behulp van een speciaal hiervoor ontworpen tekst-naar-spraak-module verklankt worden. Hierbij kan gekozen worden voor een (normale) woordgebaseerde leesmode, een splitsing in lettergrepen of een klank-per-klank leesmode.

Diezelfde leesmodes worden ook aangewend voor het geven van feedback wanneer leesfouten optreden. Een interventiestudie bij kinderen met leesproblemen uit de lagere school heeft aangetoond dat alle leesmodes worden goed geaccepteerd. De klanksgewijze leesmode mag echter niet voor al te lange woorden



worden aangewend en ook niet om meerdere fouten in eenzelfde woord te corrigeren. Op dit moment moeten we vaststellen dat het nog erg moeilijk is om per woord nauwkeurige feedback te geven, in het bijzonder voor de minder vlotte lezers. Het risico op foutieve feedback is dan immers erg groot. De analyse van het aantal en het type leesfouten is dan ook eerder bedoeld voor de taalleerkracht of therapeut.

De auteur dankt alle medewerkers van het SPACE-project voor hun inzet en creativiteit bij het verwezenlijken van dit resultaat.

Het SPACE project is een samenwerking tussen K.U.Leuven, U.Gent, V.U.Brussel en U.Antwerpen met de steun van IWT-Vlaanderen en wordt gecoördineerd door **Hugo Van hamme.**

REPETITOR

Een uitspraakoefenprogramma in ontwikkeling

Onder de naam REPETITOR is in 2008 een samenwerkingsproject tussen de TU Delft en de RU Nijmegen van start gegaan, dat tot doel heeft een uitspraakoefenprogramma met automatische feedback te ontwikkelen. Voor het onderwijs Nederlands als Tweede Taal (NT2) betekent een dergelijk programma een grote vooruitgang.

Alied Blom en Henk van den Heuvel

Een correcte uitspraak van om het even welke taal vergt gestage oefening met veelvuldige correctie, maar voor de docent is correctie op uitspraakfouten een tijdrovende bezigheid en slecht in te passen in het stramien van een groepsles. Met automatische feedback op de uitspraak kan dit onderdeel van het leerproces goeddeels via zelfstudie worden aangeboden. REPETITOR ligt in het verlengde van het programma Dutch-CAPT dat in 2007 door de RU Nijmegen ontwikkeld is[1]. Het basisprincipe is dat uitingen van NT2-leerders worden gecontroleerd op een gespecificeerd aantal potentiële uitspraakfouten. Op basis hiervan wordt een goed-/foutmelding afgegeven en gedetailleerdere feedback op gedetecteerde fouten.

Alied Blom en Henk van den Heuvel zijn verantwoordelijk voor het Repetitor-project aan de TU Delft respectievelijk de RU Nijmegen

Het programma REPETITOR, dat bedoeld is voor beginners of gevorderden met elementaire uitspraakproblemen, maakt deel uit van De Delftse Methode: Nederlands voor Buitenlanders aan de TU Delft ontwikkeld. Het leermateriaal van deze methode bestaat uit intensief te bestuderen en te beluisteren teksten, waarop ook het verdere oefeninstrumentarium is gebaseerd. Er is onder meer een geautomatiseerd luisteroefenprogramma, waarbij een deel van een bestudeerde tekst wordt beluisterd en vervolgens wordt ingetypt, waarna automatische score- en foutmelding volgt. Met REPETITOR wordt de Delftse oefenprogrammatuur aangevuld met een geautomatiseerd uitspraakoefenprogramma.

Hoe werkt REPETITOR? De gebruiker kiest een van de bestudeerde teksten, en krijgt een fragment van die tekst zin voor zin te horen, en desgewenst ook te zien. Nadat de zin is beluisterd, start de gebruiker de opnameprocedure en spreekt de zin in. Na korte tijd verschijnt de zin op het scherm, met een indicatie van eventuele gedetecteerde fouten: de letters van een verkeerd uitgesproken lettergreep zijn onderstreept en foute woordklemtonen zijn ook aangeduid. De gebruiker vergelijkt vervolgens de eigen opname met het uitspraakvoorbeeld en probeert het opnieuw, ditmaal hopelijk met een beter resultaat. Nadere advisering ter verbetering van fouten wordt niet nagestreefd, alleen al vanwege de daarvoor benodigde terminologie; bovendien gaan wij ervan uit dat zorgvuldige vergelijking van de eigen opname met het uitspraakvoorbeeld plus een aanduiding van de gedetecteerde fouten de meeste uitspraakproblemen zal doen verdwijnen. Als er geen fout (meer) wordt gedetecteerd, kan de volgende zin worden beluisterd en ingesproken. Na drie pogingen kunnen gebruikers met een hardnekkig uitspraakprobleem eveneens door naar een volgende zin.

Om het detectiegedeelte te trainen zijn inmiddels zo'n 70 lessen van ruim 40 NT2-sprekers opgenomen en geanalyseerd. REPETITOR zal naar verwachting eind 2008 in een testversie op het netwerk van de TU Delft draaien; de eveneens geplande productversie op DVD zal in de loop van 2009 beschikbaar komen.

1 Zie http://lands.let.ru.nl/~strik/research/Dutch-CAPT/index.html

My Word Coach

Taaltechnologie voor een Nintendo brain trainer

My Word Coach is een spel voor de Nintendo DS en Wii, uitgebracht door het Frans-Canadese bedrijf Ubisoft, waarmee je je woordenschat kan vergroten. Nadat je beginniveau is vastgesteld krijg je tijdens zes verschillende spelletjes te maken met nieuwe, onbekende woorden die je op verschillende manieren moet leren herkennen en gebruiken (zowel de spelling als de betekenis). Op die manier wordt de woordenschat spelenderwijs uitgebreid.

Walter Daelemans en Antal van den Bosch

Het idee en de basisopzet voor My Word Coach komen van de Canadese toegepast taalkundige Thomas Cobb; het spel bestond al voor Engels en Frans. Voor de Nederlandse versie vroeg Ubisoft ons om 20.000 woorden te selecteren, verdeeld over 20 moeilijkheidsniveaus. Met behulp van een grote verzameling van teksten (in totaal zo'n 600 miljoen woorden), het Van Dale woordenboek en taaltechnologische tools konden we deze opdracht uitvoeren.

De woorden in het spel zijn ingedeeld in de verschillende niveaus op basis van hun frequentie van voorkomen: woorden die vaak voorkomen in het Nederlands zijn doorgaans makkelijke woorden die de meeste mensen wel kennen, laagfrequente woorden zijn vaak relatief onbekend. Een topwoord als economie komt ongeveer 374 duizend keer voor (of gemiddeld eens per 1.500 woorden), terwijl een woord als xylograaf (houtsnijder) maar één keer voorkomt in de hele verzameling.

Eenvoudig de computer laten tellen hoe vaak ieder woord voorkomt in de verzameling teksten is geen optie. Woorden als grote en grootste zijn namelijk afleidingen van hetzelfde woord groot, maar zouden door de computer als verschillende woorden gerekend worden. Voor de telling is daarom gebruikgemaakt van woordfamilies. Alle afleidingen van hetzelfde woord zijn voor de telling bij elkaar genomen als één woord.

Tagging en lemmatizering

Om die telling uit te voeren hebben we Tadpole gebruikt, een automatische part-of-speech tagger die is ontwikkeld aan de Universiteit van Tilburg en de Universiteit van Antwerpen. Tadpole kan met grote snelheid van alle woorden in een tekst de woordsoort en het bijbehorende lemma (het woord waarvan het woord in de tekst is afgeleid) bepalen. De tagger verwerkt op een snelle computer zo'n 10.000 woorden per seconde – het taggen kostte dus in totaal zo'n 1.000 uur, een taak die we verdeeld hebben over een paar computers.

Nadat de tagger zijn werk had gedaan kon van iedere woordfamilie de frequentie worden bepaald. Uit deze lijst zijn vervolgens 20.000 woorden gekozen voor My Word Coach. Woorden die maar één keer voorkwamen in de hele verzameling van teksten werden geselecteerd voor de moeilijkste niveaus, woorden die vaker voorkwamen zijn in steeds makkelijkere niveaus beland.

Samenstellingen

De hele selectie is tenslotte nog handmatig gefilterd door twee neerlandici, Hanne Kloots en Griet Depoorter, die merknamen, schunnige woorden en andere ongewenste items verwijderden. We hebben ook speciale aandacht moeten besteden aan samenstellingen. Veel samenstellingen komen zelden voor (vangkuil komt maar één keer voor in de hele verzameling teksten), maar door hun samenstelling uit doorgaans bekende woorden is het (te) eenvoudig te begrijpen wat ze betekenen. Op ieder spelniveau is bekeken of de samenstellingen mochten blijven of niet. In principe zou dit probleem ook kunnen worden opgelost door automatische analyse van de samenstellingen.

Walter Daelemans en Antal van den Bosch waren verantwoordelijk voor het MyWordCoach-project aan de Universiteit van Antwerpen, respectievelijk de Universiteit van Tilburg.

Dyslexie te lijf in karaoke-stijl

In de moderne kennismaatschappij is dyslexie in toenemende mate een probleem. Zowel op school als op het werk is het steeds meer noodzakelijk dat men goed begrijpend kan lezen, een ruime woordenschat heeft en dat men fatsoenlijk (zowel begrijpelijk opgeschreven als foutloos neergepend) Nederlands kan schrijven. Betere diagnostiek en grotere alertheid zorgen ervoor dat (mogelijke) dyslexie bij jonge kinderen steeds vaker en ook eerder wordt ontdekt. Het lijkt er soms op dat er een soort dyslexie-epedemie ontstaan is, maar dat is niet zo. Het komt waarschijnlijk net zo vaak voor als vroeger alleen de impact is veel groter. In Twente is afgelopen schooljaar (2008-2009) een bijzonder project van start gegaan om kinderen met dyslexie beter te leren lezen en spellen. Het Enschedese bedrijf Telecats heeft spraaktechnologie ontwikkeld waarmee schoolboeken op een karaoke-manier kunnen worden voorgelezen.

Arjan van Hessen

Woordenschat

Wie leesblind is, heeft moeite met het in hoog tempo verwerken van geschreven informatie. Ook het spellen ofwel het correct schrijven van woorden is voor dyslectici erg moeilijk. Het lukt hen, ook na lang naar de fout geschreven woorden te hebben gekeken, dikwijls niet om aan te geven wát er fout is geschreven in gewone Nederlandse woorden. Hoewel dit een zeer serieus probleem is, is er meer aan de hand. Onderzoek heeft laten zien dat kinderen die moeite hebben met lezen, steeds minder gaan lezen. Hierdoor wordt de woordenschat van deze kinderen steeds kleiner ten opzichte van hun leeftijdgenootjes. Dus los van het slecht kunnen schrijven en het niet zien van de gemaakte spelfouten, hebben dyslectische kinderen een veel geringere woordenschat en begrijpen ze daarom veel informatie minder goed. Het gaat dan niet alleen om geschreven maar ook om gesproken informatie. Hoewel dyslexie dus niet echt genezen kan worden, kan moderne taal- en spraaktechnologie er wel voor zorgen dat kinderen meer gaan lezen waardoor ze een minder grote achterstand in woordenschat krijgen. Bovendien krijgen de kinderen een betere koppeling tussen het gesproken en geschreven woord.

Karaoke

Een mogelijke manier om het lezen ‘leuker’ te maken is door het voorlezen van de teksten. Veel ouders doen dit regelmatig en wijzen daarbij de woorden die ze uitspreken, met de vinger aan. Nu gaat er natuurlijk niets boven voorlezende ouders, maar bij gebrek hieraan (geen tijd, zelf niet kunnen lezen) is een voorlezende computer die bovendien aangeeft welk woord wordt voorgelezen een aardig alternatief. Er zijn twee manieren om dit te realiseren: de computer leest met een computerstem de tekst voor of een ouder/leerkracht/leerling leest de tekst voor en de computer plakt een tijdcode op ieder uitgesproken woord waardoor later het uitgesproken woord kan worden gemarkeerd (karaoke). Uiteraard is dit laatste prettiger omdat het dan om een echte menselijke stem gaat, maar het voorlezen en al het georganiseer eromheen is behoorlijk tijdrovend.

Schoolglas

Telecats is, in het kader van het “Schoolglas Twente project”, samen met de onderwijsadviesdienst Expertis en enkele Twentse basisscholen een proef gestart waarbij beide voorleesmethoden beproefd worden. De spraak en de in het Daisy-formaat gecodeerde teksten en plaatjes liggen opgeslagen op een centrale server en kunnen via supersnel glasvezel makkelijk door tientallen leerlingen tegelijk benaderd worden. Die krijgen dan een soort website met daarop de aantrekkelijk vormgegeven tekst die vervolgens wordt ‘ voorgelezen’ . Iedere zin die wordt voorgelezen wordt nu voorzien van een gele achtergrond. Als de proef slaagt, wordt dat straks instelbaar: op woord- of op zinsniveau.
De komende maanden zullen vooral in het teken staan van het testen van het systeem, het uitzoeken van de snelheid waarmee de teksten moeten worden voorgelezen en van de wijze waarop teksten op het computerscherm getoond moeten worden (kleurgebruik, grootte van de letters, hoeveelheid tekst) om de beste resultaten te verkrijgen. Daarna hoopt Telecats op het gereed maken van grote hoeveelheden lesmateriaal en op een verdere uitrol van het dyslexie-project voor alle (basis)scholen in Twente.

Arjan van Hessen (Universiteit Twente en Telecats) is samen met Wim Rijders (Schooladviesdienst Expertise in Hengelo) en Jessica van Haren (bovenschools taalcoördinator openbaar primair onderwijs Hengelo) verantwoordelijk voor het project.

Het MASLA-project

Individuele verschillen in het taalonderwijs

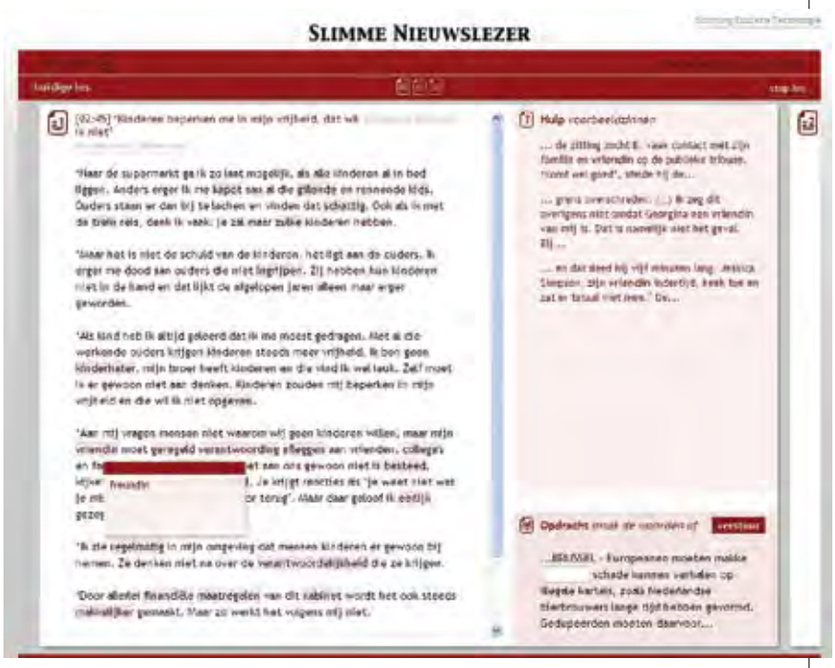
In het onderwijs is vaak sprake van grote individuele verschillen tussen leerlingen binnen dezelfde klas. Leerlingen kunnen sterk verschillen in niveau, onderwijsachtergrond, leerstijl, motivatie en tempo van verwerving van nieuwe kennis en vaardigheden. Als het gaat om taalonderwijs zijn deze verschillen tussen leerlingen vaak nog groter omdat er ook buiten de schoolse, klassikale omgeving veel momenten zijn waar taal wordt verworven. Binnen het onderwijs van Nederlands als tweede taal (NT2) zijn deze verschillen nog groter omdat leerlingen binnen een klas ook op het gebied van (onderwijs)achtergrond en leeftijd sterk van elkaar kunnen verschillen.

Rintse van der Werf & Anne Vermeer

Leren is het meest effectief als materiaal wordt aangeboden dat past bij de leerling. Het moet begrijpelijk zijn en daarnaast genoeg mogelijkheid bieden om nieuwe kennis en vaardigheden te verwerven. Aanpassen aan individueel, zelf geformuleerde, leerdoelen kan daarnaast motiverend werken. Binnen een klassikale situatie met een grote groep leerlingen is het voor een docent ondoenlijk om voor alle leerlingen rekening te houden met deze individuele verschillen en op basis daarvan het beste lesmateriaal te selecteren. Taaltechnologie kan hier helpen.

Binnen het MASLA-project aan de Universiteit van Tilburg zijn methodes en technieken ontwikkeld waarmee digitaal lesmateriaal wordt aangeboden dat past bij de leerling. Er is eLearning-software ontwikkeld en geëvalueerd waarmee de effectiviteit van lesmateriaal kan worden verhoogd. Het lesmateriaal binnen MASLA bestaat uit leespassages, zoals actuele nieuwsberichten, die worden verrijkt met hulpbronnen en opdrachten. Het basismateriaal is afkomstig van het web, uit grote (geannoteerde) corpora of kan door auteurs/docenten handmatig in een online auteursomgeving worden toegevoegd. Van dit basismateriaal worden per leerling lessen gegenereerd. In het voorbeeld in de figuur is bijvoorbeeld een les gegenereerd op basis van enkele actuele krantenberichten (maximaal 16 uur oud) waarbij rekening is gehouden met niveau van de leerling en voorkeur voor een nieuwssoort.

Binnen het MASLA-project is in een aantal deelonderzoeken met leerlingen van verschillende schooltypen (ROC's, talencentra, VMBO) geëvalueerd hoe op basis van iemands woordenschat geautomatiseerd begrijpelijk leesmateriaal kan worden geselecteerd. Daarnaast is het leereffect van het toevoegen van verschillende typen hulp



Een automatisch gegenereerde les op individueel niveau en interesse

onderzocht, zoals woorddefinities in de eerste en tweede taal, voorbeeldzinnen in de tweede taal en morfo-syntactische informatie (woordvormingregels). Tevens is de ontwikkeling van iemands woordenschat door het lezen van grote hoeveelheden begrijpelijke teksten longitudinaal onderzocht. Dit onderzoek heeft bijgedragen aan de ontwikkeling van een model voor adaptief lesmateriaal: hoe kan digitaal lesmateriaal voor het leren van Nederlands (als tweede taal) geautomatiseerd worden aangepast aan de individuele leerder.

Rintse van der Werf en Anne Vermeer zijn verantwoordelijk voor het MASLA-project aan de Universiteit van Tilburg.

Taal- en Spraaktechnologie ten behoeve van het Onderwijs in en van het Nederlands

Maryse Bolhuis en Catia Cucchiarini

De afgelopen jaren wordt in het onderwijs een groeiende behoefte geconstateerd aan leermiddelen die het mogelijk maken om leren op maat aan te bieden, aan leertrajecten die onafhankelijk zijn van tijd en plaats en aan motiverende interactieve leeromgevingen die rekening houden met de context en de netwerken van de gebruiker. Dankzij ICT en de ontwikkelingen daarin kan langzamerhand aan deze behoeftes worden voldaan. Maar als het gaat om taalleren en het ondersteunen daarvan, dan kan taal- en spraaktechnologie (TST) nog een extra impuls geven.

TST maakt het mogelijk om taalproducten, zowel geschreven als mondeling, door computers op een intelligente manier te laten verwerken, te beoordelen en eventueel van feedback te voorzien. TST kan ingezet worden om het onderwijs in en van het Nederlands efficiënter te laten verlopen of om het aantrekkelijker en plezieriger te maken voor leerlingen. Bijvoorbeeld voor vaardigheden waarbij veel oefening nodig is, kan TST gebruikt worden om leerlingen zo vaak als nodig te laten oefenen in hun eigen tempo (het zogenaamde 'kilometers maken'). Dit kan voor alle vaardigheden: lezen, schrijven, spreken en luisteren. Maar TST kan ook ingezet worden om interessantere en interactieve programma's te ontwikkelen waarmee leerlingen deelvaardigheden kunnen leren die ze meestal niet zo aantrekkelijk vinden, bijv. grammatica en spelling. De Taalunie wil zich inspannen om eraan bij te dragen dat het onderwijs in en van het Nederlands (basis- tot en met volwassenenonderwijs) in Nederland en Vlaanderen optimaal gebruik kan maken van de huidige en toekomstige mogelijkheden van TST.

Het bevorderen van TST voor het Nederlands gebeurt al via het huidige STEVIN-programma. In dit kader zijn verschillende TST-onderzoeks- en demonstratieprojecten gerealiseerd (zie overige bijdragen in dit nummer). Dit programma loopt in 2011 af en de Taalunie acht het van belang om in een vervolg van een soortgelijk stimuleringsprogramma de wenselijkheid van TST-projecten ten behoeve van het onderwijs in en van het Nederlands onder de aandacht te brengen.

Met het oog hierop brengt de Taalunie in 2008 (taal)onderwijsprofessionals en taal- en spraaktechnologen bij elkaar. Tijdens deze bijeenkomsten wordt geïnventariseerd waar nu juist (latente) behoeften liggen in het onderwijsveld en hoe TST-toepassingen daar in de nabije toekomst soelaas kunnen bieden.

In 2009 verschijnt een rapport met een overzicht van mogelijke en wenselijke TST-toepassingen en daarbij benodigde digitale voorzieningen, zoals jaren geleden werd gedaan voor de voorbereiding van het STEVIN-programma in het BATAVO-rapport (BasisTaal-en spraaktechnologische Voorzieningen, [1]zie Daelemans en Strik, 2002). Hiermee kunnen de diverse spelers in het TST- en onderwijsveld (overheid, uitgeverijen etc.) aan de slag om te komen tot concrete applicaties.

1 Bibliografie
Daelemans, W. and Strik, H. (2002) Het Nederlands in taal- en spraaktechnologie: prioriteiten voor basistaalvoorzieningen. Rapport geschreven in opdracht van de Nederlandse Taalunie, (Den Haag), http://taalunieversum.org/stevin/documenten/batavo.pdf

Maryse Bolhuis en Catia Cucchiarini zijn senior projectleider bij de Nederlandse Taalunie

NOTaS nieuws

Zoals u in deze DIXIT heeft kunnen lezen gaat het ineens heel hard in onze branche. De resultaten uit de STEVIN-projecten leveren indrukwekkende en succesvolle producten op die toegang bieden tot nieuwe marktsegmenten. Binnen NOTaS zorgen we ervoor dat onze deelnemers op de hoogte blijven van de nieuwste ontwikkelingen door de dialoog te stimuleren tussen bedrijven, kennisinstellingen en sinds dit jaar ook grote industriële partners als de NS, Philips en de Rabobank.

Het TST-in de praktijk zorgt voor boeiende netwerkbijeenkomsten

De succesvolle deelnemersevenementen van vorig jaar krijgen vervolg. Naar aanleiding van de positieve reacties over de laatste deelnemersbijeenkomsten heeft het bestuur besloten om door te gaan met het houden van deelnemersbijeenkomsten op inspirerende en verrassende locaties (zoals de bijeenkomst in november op het NIOD in Amsterdam).

Resonansgroep geeft inzicht in wensen van industriële spelers

Ook dit jaar heeft NOTaS kunnen profiteren van de aanzienlijke hoeveelheid uren die de bestuursleden en andere deelnemers namens onze stichting besteden. We zijn in het bijzonder blij met de komst van een nieuw initiatief “de resonansgroep”. Tijdens de twee bijeenkomsten dit jaar zijn 10 vertegenwoordigers van grote industriële spelers bij elkaar gekomen onder leiding van de NOTaS-werkgroep. Innovatiemanagers en andere sleutelfunctionarissen van organisaties zoals de KLPD, de Rechtbank en de Rabobank komen bij elkaar om de trends en kansen te bespreken en evalueren. De brainstormsessies leveren niet alleen de visie voor de komende jaren en subsidiekansen op, maar ook concrete ideeën voor hun eigen toepassingen en voor mogelijke innoverende consumentenoplossingen. De verslagen van de bijeenkomsten zijn te vinden op een link op de website van NOTaS.

Bedankt en tot ziens

Bij iedere organisatie vinden er bestuurswisselingen plaats. Dit jaar nemen we afscheid van Hans Jongebloed, die o.a. de hele administratie voor mogelijke nieuwe deelnemers heeft opgezet. Binnen NOTaS mogen we met trots melden dat onze kas stabiel is en dat er geen financiële turbulentie te verwachten is. Deze situatie is helemaal te danken aan de enorme inzet van onze penningmeester Nelleke Oostdijk die onze boekhouding in prima staat achterlaat. Het bestuur heeft op een gepaste manier afscheid van te aftredende bestuursleden genomen en hen bedankt voor de tijd en energie die ze aan de activiteiten van NOTaS hebben besteed.

Welkom

Als opvolgers voor Nelleke en Hans hebben Henk van den Heuvel, (CLST) en Tigran Spaan, (Gridline) zich beschikbaar gesteld. Henk neemt het penningmeesterschap voor zijn rekening en Tigran zal de taken van Hans overnemen. Dit jaar treedt Remco van Veenendaal (TST-Centrale) ook formeel tot het bestuur toe, na een periode als waarnemend bestuurslid. We wensen hen en iedereen die zich inzet voor de belangen van onze innoverende en groeiende branche succes!

Debbie Kenyon-Jackson

COLOFON

DIXIT: Tijdschrift over toegepaste taal- en spraaktechnologie – 5e jaargang, speciale editie: STEVIN en Onderwijs. 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 - Henk van den Heuvel - h.vandenheuvel@let.ru.nl - Remco van Veenendaal - veenendaal@inl.nl - René Ouëndag – rene.ouendag@q-go.com - Sylvia Hendriks - sylvia@malta-online.nl - Gwen Segond von Banchet - gwen@malta-online.nl **Gastredacteurs:** Catia Cucchiarini - c.cucchiarini@let.ru.nl - 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) **Druk:** Bloembergen Santee **Verantwoording:** DIXIT is een uitgave van Stichting NOTaS. Overname van de artikelen is alleen toegestaan met bronvermelding en na toestemming van Stichting NOTaS. Stichting NOTaS en de bij deze uitgave betrokken redactie en medewerkers aanvaarden geen aansprakelijkheid voor mogelijke gevolgen die zouden kunnen voortvloeien uit het gebruik van de in deze uitgave opgenomen informatie.