

monitor

4

NIEUWS OVER STIMULERINGSPROGRAMMA'S

JAARGANG 9 | SEPTEMBER 2006

**WINST IN DE KETEN MET
PRODUCTGERICHTE
MILIEUZORG**

**TAALTECHNOLOGIE
SLAAT BRUG TUSSEN
MENS EN MACHINE**

**MEER SAMENWERKEN
VOOR MINDER ENERGIE**

Frisse Scholen VROM-programma pakt slechte luchtkwaliteit aan



Programma STEVIN stimuleert taaltechnologisch onderzoek

Mens en machine dichterbij elkaar

De band tussen mens en computer wordt steeds inniger, maar de communicatie blijft behelpen. Computers hebben moeite met de even complexe als grillige menselijke taal. Het onderzoeksprogramma STEVIN probeert de kloof te overbruggen.

De Taalunie, een samenwerkingsverband tussen Nederland, Vlaanderen en Suriname, vindt: "Iedereen die Nederlands spreekt, moet met zijn of haar taal in zo veel mogelijk situaties terecht kunnen". Daarom houdt de organisatie zich de laatste jaren bezig met spraak- en taaltechnologie. Algemeen secretaris Linde van den Bosch: "De communicatie

tussen mens en computer gaat steeds verder. Wij vinden dat iedereen die toepassingen moet kunnen gebruiken. Niet alleen in het Engels, maar ook in het Nederlands."

Die doelstelling komt onder meer tot uiting in het programma STEVIN, dat loopt van 2004 tot 2009. STEVIN staat voor Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands. Het is

een gezamenlijk initiatief van de Nederlandse en Vlaamse overheid en de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). De Taalunie coördineert de uitgave van 11,4 miljoen euro subsidie en NWO en SenterNovem voeren het bureau.

ONDERZOEK

De Taalunie werpt zich in STEVIN op als kennismakelaar. Een belangrijke troef is een – uit eerdere projecten voortgekomen – overzicht van alle kennis die er nog niet is. Eén van de leemten is een goed corpus, een verzameling teksten. Zo'n corpus biedt een schat aan informatie over de manier waarop het Nederlands op papier gebruikt wordt. Dat kan helpen om slimmere programma's te schrijven, die beter begrijpen wat mensen intypen (in een zoekmachine bijvoorbeeld). Een Nederlands-Vlaams consortium onder leiding van de Nijmeegse Radboud Universiteit is met subsidie uit STEVIN bezig voorbereidingen te treffen voor het samenstellen van zo'n corpus, dat een half miljard woorden moet gaan tellen.

TOEGANKELIJK

Onderzoek is met een budget van 8,5 miljoen euro de belangrijkste peiler van STEVIN (zie ook het artikel over het Autonomata-project). Om de ontwikkeling van praktische toepassingen een impuls te geven, zet de Taalunie daarnaast een voor iedereen toegankelijk loket op. Van den Bosch: "Spraak- en taaltechnologische data en software waren tot op heden erg verspreid aanwezig. Wij maken ze nu makkelijk beschikbaar en houden ze actueel." Om ondernemers op ideeën te brengen, steunt STEVIN enkele demonstratieprojecten (zie het artikel over C-Content). Van den Bosch: "Het is aan de creatieve geesten om toepassingen te bedenken. Wij helpen ze op weg. Zodat ze kunnen denken: hé, maar die techniek zou ik ook kunnen gebruiken in een mobiele telefoon, of in een koelkast." «



Linde van de Bosch

MEER INFORMATIE

taalunieversum.org/taal/technologie/stevin

Álbertlaan, niet Albertlaán

Spraakcomputers hebben moeite met woorden die niet in het gewone woordenboek staan. Vlaamse en Nederlandse wetenschappers zoeken in het onderzoeksproject Autonomata naar oplossingen.

Systemen als TomTom pikken met alle gemak signalen op van 20.200 kilometer hoogte, maar de communicatie tussen mens en machine blijft houtje-touwtje. Wil je veilig een nieuwe bestemming invoeren, dan moet de auto aan de kant. Hoe simpel zou het zijn als je het gewoon kon inspreken? Helaas: makkelijker gezegd dan gedaan. Er zijn toepassingen die letters naar klanken omzetten, maar die verslikken zich nogal eens in straat- en plaatsnamen. Die zijn namelijk vaak nogal eigenaardig: ze bevatten veel oude spelling en delen van vreemde origine. Namen als Enschede, Aarleseweg en Henri Dunantlaan worden stevast verkeerd omgezet.

Oplossingen

Voor een goed werkend programma is een uitspraakwoordenboek nodig waarin correcte uitspraken van namen zitten. Maar het samenstellen van zo'n woordenboek is duur, want het vergt handwerk van gespecialiseerde fonetici. Onderzoekers van de universiteiten van Gent, Nijmegen en Utrecht proberen het anders. In het STEVIN-project Autonomata slaan ze de handen ineen met de bedrijven TeleAtlas (digitale kaarten) en Nuance



(spraaktechnologie, voortgekomen uit Lernout & Hauspie). TeleAtlas en Nuance hebben hun bestanden van persoonsnamen, straat- en plaatsnamen al eerder door fonetici van een goede uitspraak laten voorzien. Deze bestanden vormen de grondstof voor Autonomata. De onderzoekers laten computers naar patronen zoeken en correctieregels opstellen. Professor Jean-Pierre Martens, leider van het project: "Met de standaard uitspraakregels gaat automatische omzetting in de helft van de gevallen mis. Je moet een computer nu eenmaal leren dat het woordje "laan" in Albertlaan op een samengestelde naam wijst, en dat de uitspraak daardoor 'Álbertlaan' is en niet 'Albertlaán'."

Met de automatisch geleerde correctieregels hopen de onderzoekers het aantal gemaakte fouten te halveren. "Uiteindelijk is het niet erg als er in een naam een of twee klanken verkeerd worden uitgesproken. Dan verstaan we het toch wel." De tools worden op dit ogenblik in betaversie uitgetest. Ondertussen sleutelen de Autonomata-onderzoekers ook aan het probleem van de uitspraakvarianten. Ze vragen aan tientallen mensen van verschillende herkomst om woordenlijsten in te spreken, zodat de computer straks met alle accenten uit de voeten kan. "Want als een Engelstalig iemand straks naar Brussel wil, hoe gaat hij dat dan uitspreken?"

Makkelijk zoeken, snel vinden

Het internetportaal Rechtsorde.nl, een STEVIN-demonstratieproject, maakt het vinden van juridische teksten makkelijk. Dankzij geavanceerde zoektechnologie is het niet langer worstelen met zoektermen.

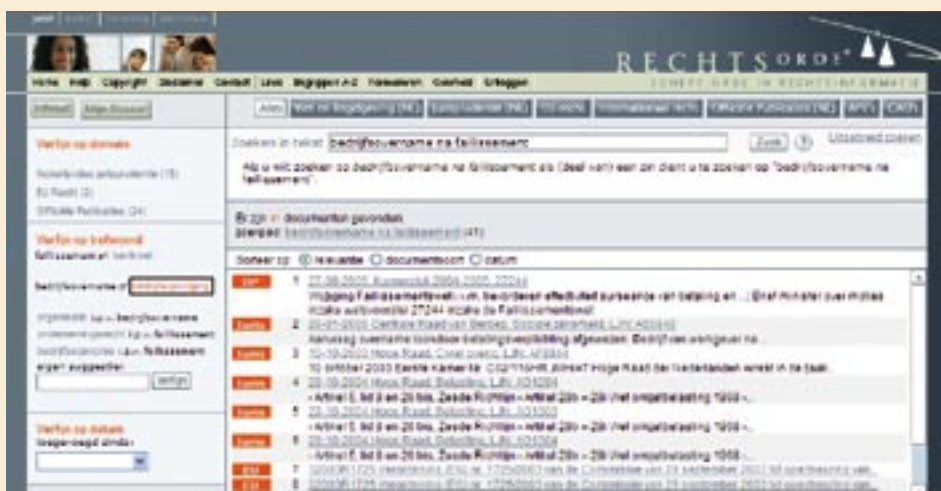
Voor het eerst zit het Bossche bedrijf C-Content op de stoel van uitgever. De specialist in zoektechnologie opereerde sinds de oprichting in 1987 vooral op de achtergrond en leverde de ondersteunende programmatuur voor websites (onder meer De Telefoongids.nl) en portalen (bijvoorbeeld van uitgever Sdu). Met het portaal Rechtsorde.nl gaat C-Content voor het eerst zelf informatie vergaren, verrijken en publiceren. Dat kan omdat de tijden veranderen, zegt Michel Mooren, directeur technologie

van C-Content. "Steeds meer bronnen zijn openbaar beschikbaar op internet."

C-Content zag een gat in de juridische markt. Nationaal en internationaal zijn er talloze sites met wetten, verordeningen, arresten, verdragen, cao's etc., maar de informatie is eindeloos versnipperd en zoeken is een crime. Rechtsorde.nl maakt daarvan een eind. De database bevat nu al meer dan 300.000 teksten van de belangrijkste sites en elke dag komen daar door automatische updates nieuwe bij. De servers van C-Content indexeren het nieuwe aanbod, herkennen verbanden met bestaande teksten en brengen het geheel onder in de nieuwe structuur.

Suggestiegestuurd zoeken

Die enorme bulk moet natuurlijk wel makkelijk te doorzoeken zijn. Mooren: "In oudere systemen moet je goed nadenken over de zoektermen die je kiest, en kost het veel tijd om een goede zoekopdracht te formuleren. In Rechtsorde.nl wilden we één zoekveld, gestructureerde resultaten en de mogelijkheid om door te klikken aan de hand van suggesties." Samen met taaltechnologiespecialist Polderland in Nijmegen werkte C-Content aan een oplossing. 'Suggestiegestuurd zoeken', noemt Mooren het. Het zoekstelsel kan hele zinnen aan. Aan de basis verbeterd de zoekmachine spelfouten en suggereert het synoniemen – 'wiplash' (sic) wordt 'zweepslag'. Het programma ontleedt samenstellingen en herleidt vervoegingen en verbuigingen naar de woordstam. Daarna kan de gebruiker 'verfijnen op domein', door bijvoorbeeld aan te geven dat hij een Nederlandse wetstekst zoekt, een arrest, of een Europese verordening. Rechtsorde.nl is sinds mei in de lucht en geeft nu nog proefabonnementen weg, maar later moeten bezoekers voor de service betalen. C-Content mikt vooral op advocaten, accountants, onderwijsinstellingen en bibliotheken.



MEER INFORMATIE

www.c-content.nl | www.rechtsorde.nl