

tekst Bea Ros

illustratie Carolyn Ridsdale/Artbox
beeld Shutterstock, STEVIN

Computer luistert beleefd en geeft netjes antwoord

STEVIN stimuleert taal- en spraaktechnologie

Zonder moderne taal- en spraaktechnologie redt een kleine taal als het Nederlands het niet in onze digitale samenleving. Binnen het onderzoeks- en stimuleringsprogramma STEVIN zijn Vlaamse en Nederlandse onderzoekers en bedrijven bezig de noodzakelijke kennis en infrastructuur te genereren. Straks kunnen we niet alleen praten met de computer, maar doet hij ook nog wat we zeggen.



*Simon Stevin (1548-1620)
loste praktische problemen
lieft op met behulp van
wetenschap.*

Luister eens naar een taal waarmee je niet vertrouwd bent, zeg Chinees. Daar valt nauwelijks chocola van te maken. Heel misschien kun je nog wat klankpatronen onderscheiden, maar je hebt geen flauw idee waar het ene woord begint en het andere eindigt. Of luister als niet-ingewijde eens naar een discussie onder sterrenkundigen. Je herkent de letters en woorden wel, maar je mist de bagage om ze te snappen. Een computer die taal probeert te begrijpen, kampt met beide problemen: hij ontbeert elk talig of inhoudelijk houvast. Geen wonder dat er nog geen vloeiende vertaalcomputer bestaat en automatische telefoondiensten vooralsnog vooral in beperkte domeinen succesvol zijn. Natuurlijke taal is zo ingewikkeld en vooral zo flexibel dat de systematiek ervan niet makkelijk te vangen is. Des te verbluffender is het dat taal- en spraaktechnologie (TST) al het nodige presteert. Zo zijn er bijvoorbeeld dicteesystemen voor radiologen en is er een volledig geautomatiseerde telefoondienst Openbaar Vervoer Reisinformatie die gebaseerd is op dialogen in natuurlijke taal. Ook scant Google moeiteloos miljoenen internetpagina's op die ene zoekterm.

Deze voorbeelden maken echter ook meteen duidelijk wat de eisen van de moderne samenleving zijn:



steeds meer mensen maken gebruik van digitale informatie. Wil het Nederlands daarbinnen als taal overeind blijven, dan zijn een stevige technologische basis en een toegankelijke infrastructuur nodig die specifiek gericht zijn op informatie in het Nederlands.

Om dat te bewerkstelligen zetten de Vlaamse en Nederlandse overheid gezamenlijk een onderzoeks- en stimuleringsprogramma op. Tussen 2005 en 2010 wordt – onder meer door NWO – 11,4 miljoen euro geïnvesteerd in Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands (STEVIN).

TAALVARIATIE Binnen STEVIN werken alle Vlaamse en Nederlandse TST-onderzoekers samen. 'Het mooie aan het programma is dat elk type

noodzakelijk onderzoek aanwezig is', vertelt Francisca de Jong, bestuurslid van het STEVIN-programma. 'Het is een geslaagde combinatie van fundamenteel onderzoek en ontwikkeling van concrete demonstratiesystemen, infrastructurele voorzieningen en evaluatieprotocollen.'

Een belangrijk taalverschijnsel waarmee een mens uitstekend kan omgaan maar dat een computer moet leren te begrijpen is taalvariatie. Voorbeelden van spraakvariatie zijn dat een Twentse 'a' nu eenmaal anders klinkt dan een Gooise, laat staan een Turks-Nederlandse, of dat 'natuurlijk' in bepaalde zinnen als 'natuluk' of zelfs als 'tuuk' wordt uitgesproken. Ook geschreven taal kent variatie die de computer moet leren te hanteren. Neem bijvoorbeeld 'de oliegi-gigant' en 'Shell': twee omschrijvingen voor een en dezelfde grootheid waar korthedshalve naar verwezen kan worden met voornaamwoorden als 'het', 'die' en 'deze'. Dit laatste verschijnsel wordt coreferentie of semantische overlap genoemd. 'Dat kun je niet verklaren op basis van grammaticaregels alleen,' legt De Jong uit, 'daarvoor is ook kennis nodig van de context en de werkelijkheid.' Systemen voor *information retrieval* – het (automatisch) zoeken naar informatie in teksten – of automatische multidocumentsamenvatters – systemen die teksten samenvatten – zouden enorm vooruitgaan als ze coreferentie beheersten. Daar zijn twee dingen voor nodig, vertelt De Jong. 'Allereerst moet het systeem gevoed worden met een enorme hoeveelheid geannoteerde taaldata. Verder moeten er algoritmes ontwikkeld worden waarmee het systeem patronen eerst leert herkennen, om ze daarna toe te passen op nieuw taalmateriaal.

STEVIN voorziet in het creëren van grote taalcorpora waarmee machines getraind kunnen worden. Voor spraaktechnologie is het bovendien nodig om diverse uitspraken te verzamelen van mannen, vrouwen, kinderen, Limburgers en mensen voor wie het Nederlands niet de moedertaal is.'

Ook informatie over de vorming van meervouden en verkleinwoorden is belangrijk: je wilt niet dat een zoekstelsel een document mist, omdat de zoekterm meervoud is en in die tekst toevallig in het enkelvoud staat.

Tot nu toe werken taal- en spraakherkenners redelijk goed in beperkte domeinen. Het al genoemde informatiesysteem voor NS-reizigers is daar een voorbeeld van. De computer hoeft hier in feite alleen alle plaatsnamen van Nederland te kunnen verstaan. 'Maar wil je bijvoorbeeld radio- en tv-uitzendingen omzetten in tekst, zodat gebruikers in



omroeparchieven kunnen browsen en zoeken naar fragmenten, dan heb je met een veel weidser domein te maken,' vertelt De Jong. 'Dat vraagt om een ander type spraakherkenning, gebaseerd op complexere taalmodellen.'

Binnen het STEVIN-project N-Best worden diverse bestaande spraakherkenners onderling vergeleken en geëvalueerd. Voor onderzoekers betekent N-Best een goede testomgeving en geeft het antwoord op de vraag voor welke parameters van de herkenning verbetering mogelijk is. Voor bedrijven biedt het nuttige informatie over wat diverse spraakherkenners wel en niet kunnen.

Een computer die taal probeert te begrijpen, ontbeert elk houvast

KAPVERGUNNING Zichtbaarheid is een belangrijke doelstelling binnen het STEVIN-programma. Alle resultaten komen straks in de zogeheten TST-centrale beschikbaar voor onderzoekers en bedrijven. Bovendien stimuleert STEVIN mkb-bedrijven om bestaande technologie in demonstratieprojecten te tonen. Zo is er een project waarbij een computer kranten voorleest ten behoeve van blinden en slechtzienden. In Utrecht loopt er een proef met een Nummerbord Retrieval Tool, waarbij agenten na het inspreken per gsm van het kenteken direct informatie over het voertuig en zijn eigenaar krijgen. Joop van Gent van Irion Technologies was betrokken bij de ontwikkeling van GemeenteConnect, een automatische informatieteleservice voor inwoners van Gilze-Rijen.

GemeenteConnect combineert spraaktechnologie (van mkb-bedrijf Dutcheer) met een informatiesysteem in een chatomgeving van Irion. 'Als je kunt chatten met computers, dan moet dat ook kunnen met een digitale telefoonbeantwoorder, was ons idee,' vertelt Van Gent. 'En zo wordt één plus één opeens drie.'



Computers moeten gevoerd worden met enorme hoeveelheid geannoteerde taaldata.

STEVIN

STEVIN staat voor Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands. Het Vlaams-Nederlandse STEVIN-programma wordt uitgevoerd onder auspiciën van de Nederlandse Taalunie. Het programmabureau wordt gezamenlijk gevoerd door NWO-Geesteswetenschappen en SenterNovem.

Budget: ruim 11 miljoen euro, afkomstig van de Vlaamse Overheid (departement Economie, Wetenschap en Innovatie) en de Nederlandse overheid (Ministerie van Onderwijs, Cultuur en Wetenschap, ministerie van Economische Zaken en NWO).

Doel: het stimuleren van de taal- en spraaktechnologie in Nederland en Vlaanderen om de innovatiecapaciteit van de sector te vergroten en de economische en culturele positie van het Nederlands in de moderne digitale informatie- en communicatiewereld te versterken. **Looptijd:** 2004-2010. **Uitvoerders:** Universiteiten van Groningen, Amsterdam, Utrecht, Tilburg, Twente, Nijmegen, Antwerpen, Gent, Leuven, TNO en verschillende grote en mkb-bedrijven waaronder Nuance, Gridline, Irion Technologies, Telecats, Carp Technologies, DutchEar, Polderland, Van Dale, TextKernel Language and Computing, Technologie & Integratie, Sensotec, De Braillekrant.

Meer informatie: www.stevin-tst.org, www.gemeenteconnect.nl



Agenten kunnen via hun mobieltje automatisch gegevens van kentekenhouders opvragen.

ning tot het aangifte doen van geboorte. Dat kan per gemeente naar behoefte worden uitgebreid. Irion is bezig met een consortium van gemeentes om het systeem op grotere schaal in te voeren. Daarbij zijn tal van variaties relatief simpel in te bouwen, bijvoorbeeld het maken van afspraken of het direct doorverbinden naar telefonistes van vlees en bloed. 'Het meest spectaculaire deel zit in het begin van de dialoog, als de computer moet ontdekken wat de beller precies wil.'

TOEWIJDING Zonder subsidie vanuit STEVIN hadden Irion en Dutchear GemeenteConnect niet kunnen ontwikkelen, zegt Van Gent. 'Het is geen nieuwe uitvinding, maar het uitproberen kost ook de nodige tijd. Dat hadden we ons nooit kunnen veroorloven.'

Tijd, weet ook De Jong, is essentieel voor het toepasbaar maken van wetenschappelijke inzichten. Om een corpus voor een bepaald domein samen te stellen, is het louter invoeren van meer data niet het enige wat nodig is. 'Gezond verstand en de nodige toewijding om gericht relevante datasets te selecteren en in te voeren, zijn minstens zo belangrijk. Bij het trainen van een model voor toepassing op een speciale collectie komt veel organisatie kijken. Om van de technologie een succes te maken, moet je daar greep op krijgen.'

Zo zullen wetenschappers ook meer samen moeten werken met andere disciplines. Van cognitief psychologen kunnen ze dingen leren over verwachtingen en wensen van gebruikers. Informatici zijn nodig om goede interfaces en infrastructuur te ontwikkelen. Ligt er een vertaalcomputer of perfect werkende spraakherkenner in het verschiet? Jazeker, belooft De Jong. Zolang we van machines maar niet verwachten dat ze net zo goed zijn als natuurlijke taalgebruikers. 'Bij de toepassing van taal- en spraaktechnologie draait het niet om perfectie, maar om optimale ondersteuning van mensen in de omgang met informatie.' ❏

Bijzonder aan GemeenteConnect is dat het een vrijwel natuurlijke dialoog biedt. 'Het irritante aan veel bestaande telefoonservices is dat het ofwel helemaal menugestuurd is ('toets 1 voor...'), ofwel dat je geen vragen kunt stellen of iets in eigen woorden kunt vertellen. Dat laatste kan in onze chatomgeving wel, de gebruiker kan intikken wat hij wil.'

Dat idee is vertaald naar een telefoongesprek. In Gilze-Rijen begroet de digitale telefoniste je allervriendelijkst, vraagt waarmee zij je kan helpen en

luistert vervolgens beleefd naar het relaas van de beller. Het merendeel ervan zal ze als onbegrijpelijk terzijde schuiven, maar bepaalde sleutelwoorden als 'boom', 'achtertuin' en 'kapvergunning' worden wel herkend. 'Daardoor weet het systeem waar het moet zoeken in de database,'

legt Van Gent uit. 'Binnen twee minuten krijgt de beller antwoord, hij wordt dus niet lang aan het lijntje gehouden.' Het systeem kan 80 procent van de meest gestelde vragen beantwoorden. Alleen bij specifieke vragen als 'kunt u een kadastrale tekening opsturen' moet de computer adviseren tijdens kantooruren terug te bellen.

De database is gevuld met alle standaardproducten en -regelingen van de gemeente, van kapvergun-



'We moeten van machines niet verwachten dat ze net zo goed zijn als wij'