Taaltechnologe Franciska de Jong over multimediale spraaktechnologie

‘Dit onderzoek is een investering in de toekomst’

MARTIJN DE GROOT

‘Technologie veroudert, maar de kennis die er achter zit blijft relevant’. Dat zegt prof. dr. Franciska de Jong, leider van een aantal onderzoekprojecten waarin spraakherkenning een grote rol speelt. De techniek om gesproken woord doorzoekbaar te maken komt steeds verder, blijkt uit een gesprek met de Twents-Rotterdamse taaltechnologe.

WILLIAM HOOGTEYLING

‘Voor onderzoekers is het belangrijk om over data te kunnen beschikken, èn dat die data ergens worden ondergebracht waar anderen er ook iets mee kunnen doen’. Franciska de Jong kan zo’n uitspraak met enig gezag doen want haar eigen onderzoek consumeert enorme hoeveelheden data - zij spreekt glimlachend van tekstcorpussen die honderden miljoenen woorden tellen. Bovendien is juist zij bezig met projecten die belangrijke historische en eigentijdse bronnen in gesproken vorm voor een groot publiek beschikbaar en tekstueel doorzoekbaar maken. De onlangs gelanceerde website *buchenwald.nl*, met 38 volledig doorzoekbare audiointerviews met overlevenden, is er een voorbeeld van.

Toch heeft haar uitspraak niet daar betrekking op, maar op de rol van de Virtual Knowledge Studio (VKS), waarvan zij sinds de zomer van 2007 de Rotterdamse vestiging bestiert. ‘Een goede organisatie van het proces waarbij data worden opgeslagen, bewerkt en gedeeld’, gaat ze verder, ‘is een gemeenschappelijk belang van onderzoekers, beheerders èn gebruikers van dat materiaal, want de beschikbaarheid en uitwisselbaarheid van zulke grote hoeveelheden data is een relatief nieuw fenomeen. De ontwikkeling van nieuwe samenwerkingsverbanden die dat vraagt en de methodologische vernieuwing die erbij komt kijken, dàt ligt op het onderzoeksterrein van de VKS’.

Het beschikbaar komen van grote elektronische datasets heeft het vak enorm veranderd

De Erasmus Studio, zo wordt deze vestiging genoemd, is ondergebracht in een paar kamers op de zevende verdieping van een betonnen studeerkolos van de gelijknamige Universiteit. De lift is traag, de hallen groot, de gangen zijn kaal en grauw. Als je niet wist dat in de jaren zeventig van de vorige eeuw zulke bouwwerken her en der waren verrezen, zou je de stijl voor Rotterdams aan kunnen zien: hier rest de bewoners weinig anders dan studeren en onderzoeken. ‘De Amsterdamse VKS is sterk op de methodologie georiënteerd’ antwoordt ze op de vraag naar het verschil. ‘Dus dan is het logisch dat je niet steeds met hele praktische resultaten kunt zwaaien. Hier in Rotterdam ligt de nadruk misschien meer op de doenerige aspecten van het onderzoeksterrein, èn er zijn hier veel studenten. Wij werken samen met een aantal onderzoeksgroepen, zoals de econometristen en bijvoorbeeld met de medische faculteit in onderzoek naar het ontsluiten van medische informatie met behulp van thesauri.’ En dan vloeit het gesprek automatisch door in de richting waar ook de passie van de hooggeleerde taaltechnologe ligt. ‘Als je zoekt naar bijvoorbeeld de samenhang tussen een gen en een bepaalde ziekte, dan is het meestal zo dat daar bèrgen literatuur over te vinden zijn. Nu is er aan de EUR technologie ontwikkeld waardoor je in die tekstbestanden kan zoeken welke concepten met elkaar in verband worden gebracht die voorheen niet eerder in samenhang zijn besproken. Dan ben je nieuwe kennis of nieuwe inzichten op het spoor. Nu die technologie werkt, gaan we natuurlijk proberen of ze ook voor andere wetenschapsgebieden te gebruiken is.’

Voor dit soort zoektechnieken is naast veel taalkundige data ook veel statistiek nodig en dat is precies de combinatie die de Jong het meeste boeit, vooral als er dan ook nog een bruikbaar resultaat uit voortkomt. Aan de volkswijsheid dat een mens over alfa- òf beta-talenten mag beschikken heeft ze geen boodschap. ‘Je ontwikkelt je op de gebieden waar je aandacht aan geeft. Ik had affiniteit met talen en wiskunde en kwam als student terecht in de taalkunde – een tamelijk exact vak binnen de letteren, met logica, mathematische taalkunde en computationele linguistiek in het pakket. Na mijn studie kon ik bij het Natlab van Philips aan de slag in een onderzoekgroep die zich richtte op machinaal vertalen. Een fantastische tijd. We bouwden met een flink aantal mensen aan een systeem, maar dat werkte vooral op basis van grammaticaregels die in een modulaire opzet over talen heen aan elkaar werden gekoppeld. Daar kwam dus geen statistiek aan te pas. Er kwam een werkend systeem uit, maar een product is het nooit geworden’.

Eigenlijk diende pas in de jaren negentig van de vorige eeuw de kans zich aan op veel betere en krachtiger taalmachines, aldus de Jong, en dat was een direct resultaat van het beschikbaar komen van grote elektronische dataverzamelingen. Dat maakte de opkomst van statistische modellen mogelijk ‘en daardoor is het vak enorm veranderd’. De Jong verwierf in 1992 een leerstoel in de taaltechnologie, die ze nog steeds bekleedt. Ondergebracht tussen Twentse informatici bleek die een mooie basis te bieden om juist op het statistische spoor verder te gaan. In de spraakherkenning, het gebied waar De Jong zich sindsdien steeds verder in ging verdiepen, speelt de statistiek een doorslaggevende rol, licht ze toe. ‘Als je met spraakherkenning bezig bent dan moet je een akoestisch model ontwikkelen waarmee je meer dan honderdduizend woorden van elkaar kunt onderscheiden en in een transcriptie omzetten. En het gaat dan om spontane spraak dus dan moet zo’n model de verschillende uitspraakvarianten van één woord ook kunnen herkennen. Maar woorden lijken soms heel erg op elkaar. Mensen weten dan toch het onderscheid omdat ze de context van die woorden meenemen. Dus dat moet zo’n spraakherkenner ook doen. Daarvoor ontwikkel je een taalmodel, dat is een statistische representatie van al die woorden en hun onderlinge verbanden. Zo’n model levert dan de informatie op dat er bijvoorbeeld is gesproken over een ‘Italiaans restaurant’ en niet over een ‘Italiaans restant’.

De bedoeling is dat we van elkaar kopiëren, want daar worden alle deelnemers beter van

De Jong is wel bereid om een knap stukje werk te demonstreren dat uit de activiteit van haar Twentse onderzoekgroep Human Media Interaction (HMI) is voortgekomen: een showcase op de HMI-website biedt de mogelijkheid om in de acht-uurjournaals van de laatste twee weken te zoeken. Dat gebeurt in feite in de tekst die na afloop van elke uitzending automatisch wordt voortgebracht door de spraakherkenner die haar onderzoekgroep afleverde. Een zoekactie met de woorden ‘kelder’ en ‘Oostenrijk’ levert binnen een paar seconden een reeks verwijzingen op, mèt tekstweergave van de relevante passages èn de mogelijkheid om het betreffende fragment te bekijken. Een overtuigend resultaat, ook met de foutjes die er nog in de weergegeven tekst blijken te zitten – vooral waar de spraakherkenner zich niet ontziet om ook Duits gesproken fragmenten in geschreven Nederlands om te zetten.

Kan zo’n mooi product niet op de markt gebracht worden, om zo wat terug te verdienen van de investeringen die er in de loop van de jaren in zijn gedaan? ‘Dat zou misschien wel kunnen’, denkt de Jong, ‘maar niet kostendekkend. Dan zou het veel te duur zijn want daarvoor zijn de kosten tot nu toe te hoog geweest. De technologie is dan ook in ‘open source’ beschikbaar gesteld, zodat anderen nieuwe toepassingen kunnen ontwikkelingen. Voor ons zelf is het eigenlijk alleen interessant om daaraan te werken als er een nieuw wetenschappelijk perspectief aan verbonden is. En dat is natuurlijk wel het mooie van dit soort projecten. De kennis die je hebt ontwikkeld gaat niet verloren. Daar wordt steeds op doorgebouwd. Wij doen met onze spraakherkennings- en multimediazoektechnologie mee met internationale competities waar je kan vergelijken met anderen, hoe die het hebben gedaan. Wat is hun foutenpercentage? En de bedoeling is dat we van elkaar kopiëren, want daar worden alle deelnemers beter van.’

‘Een investering in de toekomst’ is het vooral, en dat geldt ook voor het *access to oral history*-project CHoral waarvoor de Jong financiering kreeg binnen het cultureel erfgoed-programma CATCH van de Nederlandse Organisatie voor Wetenschappelijk onderzoek. En trouwens voor CATCH als geheel: ‘Doordat binnen CATCH de samenwerking wordt gestimuleerd, krijg je een hele generatie onderzoekers en softwareontwikkelaars met de kennis om de infrastructuur van de toekomst vorm te geven’.

Franciska de Jong werkte na haar studie taalwetenschap als onderzoeker bij het Philips NatLab in Eindhoven. Sinds 1992 is ze hoogleraar taaltechnologie aan de Universiteit Twente. Rond die tijd werd ze ook onderzoeker aan de Universiteit Utrecht en later bij TNO in Delft. In juni 2007 trad ze aan als managing director van de Virtual Knowledge Studio in Rotterdam. Sinds 2004 is zij lid van het NWO-gebiedsbestuur Geesteswetenschappen.