Communiceren machines straks in mensenlijke taal?

Vrijdag 6 juni zijn de Spinoza-premies 2014 uitgereikt, de hoogste onderscheiding in de Nederlandse wetenschap. Taaltechnoloog Piek Vossen was vorig jaar één van de laureaten. *Erica Renckens*

Taaltechnoloog Piek Vossen, winnaar van de Spinozapremie 2013 *foto Studio VU*

Naast eeuwige roem ontvangt de winnaar van de Spinoza-premie ook 2,5 miljoen euro. “Daarmee heb ik vier projecten gedefinieerd met als rode draad de vraag hoe computers natuurlijke taal kunnen begrijpen,” vertelt Piek Vossen, die sinds 2006 werkzaam is aan de Vrije Universiteit. “Het is nog een groot mysterie hoe we dat computers kunnen leren, terwijl wij mensen er ogenschijnlijk geen moeite mee hebben.”
Eerder maakte Vossen voor de Europese Unie *wordnets* in acht talen: netwerken van woorden die op basis van hun betekenis met elkaar verbonden zijn. Soms zijn de woorden synoniemen van elkaar, soms juist tegenstellingen, en soms hebben ze alleen met elkaar gemeen dat ze met dezelfde kleur geassocieerd worden. Met behulp van zulke *wordnets* kan een computer betekenis afleiden en zo spelling controleren, zoekresultaten verbeteren en teksten interpreteren.
Betekenis is voor computers een lastig aspect van taal. “Een woord als ‘slag’ kan in het Nederlands achttien verschillende betekenissen hebben, maar als we een tekst lezen zien we er maar één,” aldus Vossen. Mensen leiden aan de hand van zinsbouw en context moeiteloos af welke betekenis bedoeld is, maar een computer heeft deze kennis niet.

Semantiek centraal

“In het eerste project gaan we het probleem van de meerduidigheid van woorden aanpakken.” Ook in de overige drie projecten staat de semantiek van natuurlijke taal centraal. “Het is de bedoeling dat we meer fundamenteel inzicht krijgen in de manier waarop taal betekenis krijgt,” legt Vossen uit. Daarvoor kijken de onderzoekers in zijn groep onder andere naar de relatie tussen woorden, concepten, perceptie en het brein, naar de invloed van het wereldbeeld van de auteur op teksten en naar de rol van de lezer bij het begrijpen van teksten. “Uiteindelijk hopen we machines te kunnen maken die met mensen kunnen communiceren aan de hand van natuurlijke taal. Daarbij richten we ons niet direct op spraak, maar wel op de betekenis.”

Resultaten delen

Vossen: “Ieder project kijkt naar andere deelaspecten, maar er zijn ook veel verbanden waarbij projecten elkaars deelresultaten kunnen gebruiken. We hebben tien mensen aangesteld, onze groep is verdubbeld. We werken enthousiast samen aan een grote uitdaging, ik hoop dat de goede sfeer en samenwerking die we nu hebben ook in een grotere groep blijft bestaan.”

vossen.info

GELEZEN

Managing and Sharing Research Data. A guide to good practice. Louise Corti, Veerle van den Eynden, Libby Bishop & Matthew Woollard. Uitgeverij Sage. 2014.
René van Horik

Vier auteurs, allen werkzaam bij het UK Data Archive, schreven dit handboek over het beheren en delen van onderzoeksdata. Men richt zich zowel op de onderzoeker (van student tot professor) als op de dataprofessionals, de ondersteuners.
Het handboek biedt een actueel en compleet overzicht. Jammer dat het niet als open access publicatie beschikbaar is. Het zou dan veel meer lezers bereiken. Niet alle elf hoofdstukken zijn van belang voor alle doelgroepen. Dus het is zaak goed de inhoudsopgave te bekijken en te bepalen welk onderdeel bruikbaar is. Elk hoofdstuk bevat een aantal praktische oefeningen en een uitgebreide literatuurlijst. Ook is er een begeleidende website met achtergrondinformatie en databronnen die behandeld worden. Het boek is sterk gericht op de Engelssprekende wereld en mist daarom een aantal belangrijke niet-Engelstalige initiatieven. Er is relatief veel aandacht voor *legal and ethical issues in sharing data*, met goed uitgewerkte oefeningen, bijvoorbeeld manieren om onderzoeksdata te anonimiseren. Voor de onderzoeker zijn met name de laatste drie hoofdstukken interessant.
Ze hebben betrekking op collaborative research, gebruik van onderzoekdata van anderen en het publiceren en citeren van onderzoeksdata.
ukdataservice.ac.uk/manage-data/handbook

COLUMN

Waarom is er nog steeds geen digitaliseringsregister?

In 2004 was ik bij een bijeenkomst in de Koninklijke Bibliotheek over massadigitalisering. Eén van de sprekers was Astrid Verheusen, programmaleider digitale bibliotheek. Met enige gêne stipte Verheusen één van de centrale problemen van massadigitalisering aan, namelijk het ontbreken van een register waarin wordt vastgelegd wie wat aan het digitaliseren is.
Zij liet afbeeldingen zien van enkele reeksen die door diverse bibliotheken en instellingen waren gedigitaliseerd. Steeds dezelfde reeks, steeds met overheidssubsidie.
In de zaal ging een besmuikt gelach op – het probleem was algemeen bekend.
Dat was tien jaar geleden.
Een paar maanden geleden was ik bij een internationale bijeenkomst in de Koninklijke Bibliotheek over textmining. Een Amerikaanse spreker vroeg onder meer: ‘Weet iemand of er een centraal register bestaat waarin je kunt nazien wie wat heeft gedigitaliseerd? Ik heb de indruk dat er veel dubbel werk wordt gedaan en dat lijkt mij zonde van het geld en van de inspanning.’
In de zaal bleef het stil.

Is het werkelijk van belang dat er zo’n register komt? Ja, en niet alleen omdat het zonde is van geld en tijd dat er – nog steeds – zoveel veel dubbel werk wordt verricht. Een andere reden houdt verband met een voortkabbelende discussie in E-data.
In oktober schreef ik in E-data dat ik iemand ken die af wil van een collectie van 13.000 boeken. Het gaat om een met smaak en kennis opgebouwde collectie boeken uit en over de 19de eeuw. Je zou zo’n specialistische collectie aan de KB moeten kunnen schenken, stelde ik, onder voorwaarde dat je ’m binnen een bepaalde tijd in digitale vorm terugkrijgt.
De redactie van E-data legde dit idee voor aan Jan Bos, hoofd collecties van de KB. Zijn antwoord, kort samengevat: de KB zou die dienst graag willen leveren, maar kan dit niet. Bos: “Met name de selectie is veel duurder dan je zou denken. Je moet eerst nagaan wat we al gedigitaliseerd hebben, en van wat er overblijft moet je alle rechten napluizen.”
Inderdaad, het napluizen van rechten kan een tijdrovende en dus kostbare zaak zijn. Maar nazien wat er al gedigitaliseerd is, zou een fluitje van een cent moeten zijn. Niet alleen de KB, maar alle instellingen die digitaliseren, zouden dit met bepaalde technische specificaties moeten vastleggen in een register. In PiCarta of WorldCat bijvoorbeeld. Als niet alleen bibliotheken dat wereldwijd zouden doen, maar bijvoorbeeld ook Google, zou dit heel veel dubbel werk kunnen voorkomen.

Waarom is het in tien jaar niet gelukt om zo’n centraal register op poten te zetten? Desinteresse, eigenwijsheid en onwil, vermoed ik. ‘Eigen instelling eerst’ is het motto van velen, het algemeen belang staat veel lager op de ladder.
Ik heb weleens voorgesteld om instellingen alleen geld voor digitalisering te geven als ze kunnen aantonen dat ze meewerken aan een centraal register. Dat was in 2011, naar aanleiding van een stuk in NRC Handelsblad getiteld ‘Het digitale drama’. Uit dat stuk bleek onder meer dat niemand wist hoeveel geld er tussen 2004 en 2011 aan digitalisering was uitgegeven. De schattingen liepen uiteen van 50 tot 200 miljoen euro.
En waarom was en is daar zo slecht zicht op te krijgen? U voelt ’m al aankomen: mede vanwege het ontbreken van een centraal register.
Stuitend hoe lang het soms duurt voordat een idee dat zo voor de hand ligt, wordt ingevoerd.

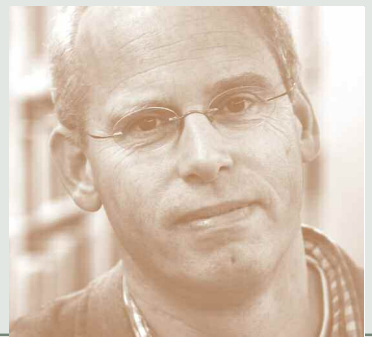

foto Leo van Velzen

Ewoud Sanders
Taalhistoricus en journalist. Sanders is vaste medewerker van onder meer NRC Handelsblad en Onze Taal.