

Alfa-informaticus John Nerbonne:

'Ik vind dat iedereen zijn data beschikbaar moet stellen'

Op het kruispunt van exacte wetenschappen en humaniora. Zo zou je het beste de plek kunnen beschrijven waar John Nerbonne als wetenschapper is neergestreken. Je zou ook kunnen zeggen: Groningen, want voordat de geboren Amerikaan daar hoogleraar alfa-informatica werd had hij al verschillende werkplekken achter zich gelaten.

'Een echte alfa,' noemt Nerbonne zichzelf. Maar, zegt hij er meteen bij, 'je moet in dit vak wel een programma zelf kunnen ontwerpen als je in het bestaande aanbod niet vindt wat je nodig hebt.' En een diepgewortelde belangstelling hebben voor wiskunde, moet daar toch aan worden toegevoegd in het geval van Nerbonne die als veertienjarige jongen een biografie van de Britse getalstheoreticus G.H. Hardy las en daar nu nog een deel van zijn fascinatie op terugvoert. Maar een feit is dat zijn vak computationele taalkunde in Groningen als een van de weinige plekken in Nederland onder de Letterenfaculteit is ondergebracht. 'Bij de meeste universiteiten hier valt het onder informatica, net als in de Verenigde Staten. In Duitsland is het weer meestal onderdeel van taalkunde. Zelf ben ik blij dat ik hier mag werken, maar een informaticus wil niet graag aan de Letterenfaculteit promoveren'.

Duitsland en de Verenigde Staten. Het zijn de twee landen waar Nerbonne verbleef voordat hij in 1993 in Groningen terecht kwam. 'De liefde bracht me naar Nederland' vat hij zijn zwerftocht samen. Hij werd geboren in de Amerikaanse staat Massachusetts en kreeg als 22-jarige de kans om een aantal jaren in Duitsland te studeren: Germanische Philologie aan de universiteit van Freiburg. 'Daar heb ik mijn huidige vrouw leren kennen. We zijn toen samen naar Amerika gegaan, waar ik eerst in Ohio ben gepromoveerd en later in California bij Hewlett-Packard werkte. Maar na een jaar of tien wilde zij graag terug, en ik had nooit met tegenzin in Europa gewoond dus dat hebben we toen gedaan.' Nerbonne nam een baan aan bij het Duitse bedrijf DFKI, dat zich met kunstmatige intelligentie bezighoudt, en dat bleek achteraf het voorportaal van het hoogleraarschap dat hem in Groningen wachtte. Nederland, 'het hoofdland van de logica in de wereld met op dat gebied een topwetenschapper als Johan van Benthem', dat leek hem ook een prima bestemming. De wetenschap was toch wel zijn eerste keus, weet hij nu. 'De vrijheid die je aan een universiteit hebt, dat is uniek. Je vindt iets interessants en dan ga ja daar iets mee doen!'

Die laatste zinsnede moet bij Nerbonne heel letterlijk worden opgevat, zo blijkt. 'Vaak wordt de vraag 'fundamenteel of toegepast' als een polariteit gezien maar ik zie dat niet als een tegenstelling. Computational Linguistics wordt wel een ingenieurswetenschap genoemd. Dat zou ik zelf niet doen maar een feit is dat er veel aan toepassingen wordt gedaan en dat is ook leuk. Het brengt bovendien geld op.' Technieken om hardop te lezen voor dyslectici (?) en andere voorleestoepassingen zijn daar voorbeelden van. Of het werk aan computerondersteund taalonderwijs, bijvoorbeeld met een elektronisch woordenboek achter de tekst in samenwerking met van Dale (GLOSSER: www.let.rug.nl/glosser/, van

1997). Of, ander voorbeeld, software van de hand van Greg Kondrak, een collega van Nerbonne, die producenten van medicijnen helpt om namen uit te kiezen die niet te veel lijken op reeds gangbare namen van andere medicijnen. 'Zanax en Sanex bijvoorbeeld. Onderschat dat niet, dergelijke namen worden in de medische praktijk vaak door elkaar gehaald en daar gebeuren ernstige ongelukken mee'.

Is computationele taalkunde nu eigenlijk een discipline of niet? 'Het is een soort verbond, een federatie van verschillende disciplines,' antwoordt de linguïst, die met gemak voorbeelden uit de mouw schudt van samenwerkingsvormen met andere wetenschappen. 'Ik heb heel vruchtbaar samengewerkt met genetici, om de verbanden te onderzoeken tussen de herkomst van bevolkingsgroepen en uitspraakverschillen. Er zijn ook collega's bezig in de musicologie om in overeenkomsten in de verspreiding van volksmuziek en taalverschillen te zoeken. Er zijn projecten met fysici, met archeologen. En ik kan me voorstellen dat je architectuurpatronen over de Atlantische oceaan heen gaat volgen om die weer in verband te brengen met vestigingspatronen en taalvariatie.'

De laatste jaren is de belangstelling van de Groninger Amerikaan steeds meer de kant van de dialectkunde op gegaan, het vak dat de ontwikkeling, geografische spreiding en variatie van dialecten bestudeert. Hij liet zijn wetenschappelijk oog al vallen op Nederlands, Duits, Noors, Sardijns, Bulgaars en Bantu, maar ook de tweede rijkstaal, het Fries, en een officiële minderheidstaal van de Europese Unie, het Nedersaksisch, mogen zich in zijn belangstelling verheugen. Daarbij gaat het Nerbonne vooral om uitspraakverschillen en de wiskundige formules om die *en masse* te meten. Want dat is ook volgens Nerbonne de 'allerbelangrijkste ontwikkeling' van de laatste jaren: 'De gebruikelijke manier was altijd: je vindt een moedertaalspreker en daar stel je vragen aan. Die vertelt je dan hoe woorden worden uitgesproken en welke woorden wel en niet gangbaar zijn. Maar je was daarmee aangewezen op de tamelijk willekeurige weergave van degenen met wie je in contact kwam. Nu we over steeds meer en grotere corpora kunnen beschikken – en dat neemt echt een steeds grotere vlucht – kunnen we zien of een woord of uitspraak wordt gebruikt en hoe vaak. En je kan natuurlijk veel beter naar patronen gaan zoeken in de verspreiding van taalvariëaties.'

Nerbonne laat met zichtbaar plezier een tweetal kleurrijke kaarten zien van het Nederlandse taalgebied, die een paar jaar geleden het resultaat waren van een samenwerkingsproject met het Meertens Instituut. 'Hierbij hebben we *multidimensional scaling* toegepast, een techniek die het mogelijk maakt om de verschillen in uitspraak en vocabulaire weer te geven als continuüm, in geleidelijke kleurovergangen. Dat was echt een grote stap. Ik denk dat Peter Kleiweg en ik in dit project, waaraan vanuit het Meertens ook Wilbert Heeringa deelnam, de eerste goede visualisatie hebben gemaakt van de geleidelijke verspreiding van dialecten. Op dezelfde manier hebben we een weergave gemaakt (haalt een andere kaart tevoorschijn) waarin de afstand van een bepaalde streektaal tot het Algemeen Beschaafd Nederlands is weergegeven'.

Zulke vormen van onderzoek zouden ondenkbaar zijn zonder de beschikbaarheid van enorme databestanden, weet Nerbonne. 'Mensen die in dit vakgebied specifieke studies willen doen hebben behoefte aan heel veel data. Daarom is de functie van een instituut als

DANS ook van onschatbare waarde. Het zelf verzamelen van data is duur en die kosten moet je besparen voor betere doelen. Maar daarvoor is wel nodig dat onderzoekers hun data ook beschikbaar stellen, en dat gebeurt nog lang niet altijd van harte. Er zijn er nog steeds die daar eigenlijk niet aan willen: 'Ja, we zijn er nog mee bezig, het is nog niet definitief...', zulke argumenten. Ik vind dat iedereen zijn data beschikbaar moet stellen via DANS of een instelling zoals DANS, maar hoe dat in de taalkunde precies in zijn werk moet gaan is nog aan het uitkristalliseren. We horen nog niet zo veel van DANS activiteiten op het gebied van taal en taalkunde. De populatiegenetici hebben het nu bijna zo goed voor elkaar dat ze op websites die hele grote, belangrijke bestanden kunnen zoeken en vinden. Met projecten als CLARIN en DARIAH gaan we hopelijk in de taalkunde ook die kant op.

Martijn de Groot