Marc van Oostendorp ging aan de slag met de data van de KB:

‘Niet álles op zijn kop door opkomst Big data’

Hij was trots op zijn KB-toegangspas en zijn kamer te midden van miljoenen boeken. Maar hij kwam voor de data. Of, liever nog, voor de mensen die hun vak maken van het werken met data. *Inge Angevaare*

INTERVIEW

Fonoloog Marc van Oostendorp was de eerste ‘digitale’ fellow van de Koninklijke Bibliotheek (KB) en het Netherlands Institute for Advanced Study in the Humanities and Social Sciences (NIAS). Zijn voorgangers werkten met het analoge materiaal. “Ik wilde vooral eens heel goed nadenken over de relatie tussen big data en de geesteswetenschappen. Daar wordt vaak zo in extremen over gesproken, dat vind ik jammer. En ik had ook een aantal concrete onderzoeksvragen.”

Goed in saai werk

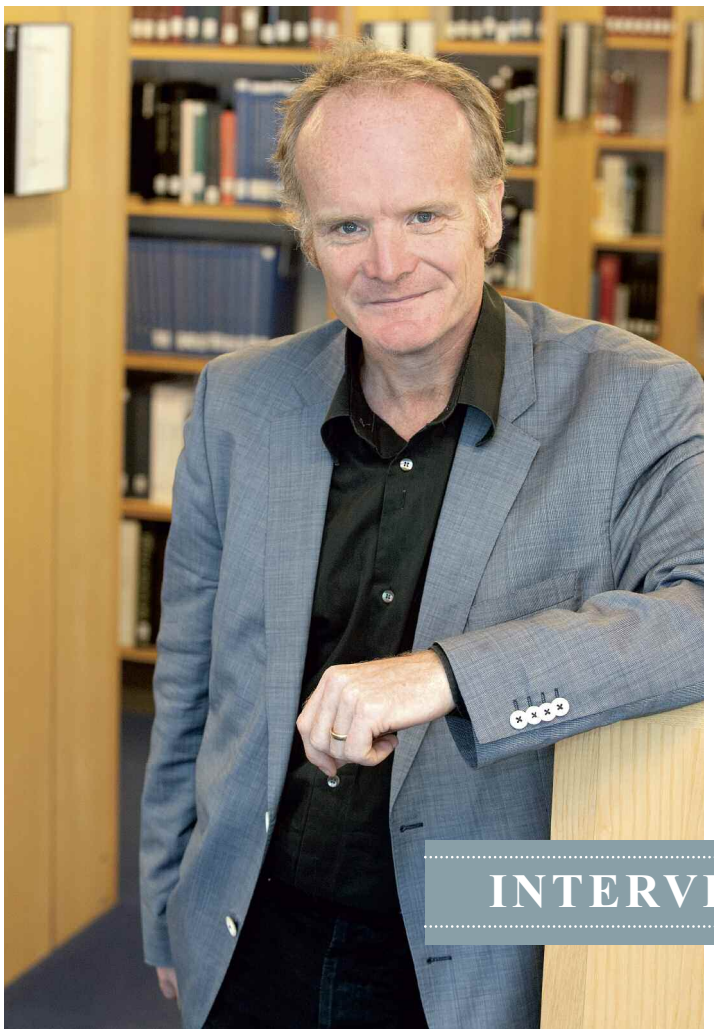
“Nederlanders hebben een voorkeur voor het jambische ritme (tadám tadám tadám tadám tadám), dat vinden we het prettigst. Ik wilde zoeken naar de oudste bronnen met dat ritme in de databestanden van de KB. De Italiaanse dichter Francesco Petrarca zou het ontwikkeld hebben en het zou zich vanuit Italië verspreid hebben over Europa.

In het Nederlands is er slechts één Middelnederlandse bron, Het leven van Sinte Lutgart, en dan minstens honderd jaar niks. Dat is vreemd. Ik wilde in de buurt van die tekst gaan zoeken naar andere voorbeelden. Ritme kun je in die oude teksten alleen aantonen als ze op rijm zijn, en dan moet je nog aannemen dat de uitspraak ongeveer hetzelfde was als nu. Je moet er vele, vele bronnen voor doorspitten. Het is niet helemaal dom werk, het vereist wel enige kennis van zaken, maar verder is het vooral heel veel en heel saai. En dat is precies werk dat een computer goed kan.”

Mensen zijn te slim

“Wat ik heb onderschat, is iets dat op zich heel erg voor de hand ligt. Het Middelnederlands kent geen vaste spelling, er is eigenlijk geen standaard-Nederlands, er zijn alleen maar dialecten. Ik dacht dat op te lossen door de woorden te vervangen door dezelfde woorden uit een (modern) fonetisch woordenboek waar de klemtonen staan aangegeven. Om die vervolgens te gaan mappen. Maar dat bleek een brug te ver voor een logaritme.

Dat is het probleem van het werken met computers. Je denkt als mens in een aantal stapjes die voor jou volmaakt logisch zijn. Maar er is er altijd één bij die een grote gedachtesprong blijkt te zijn, te groot voor programmeertaal. Eigenlijk is dat ook wel het leuke van met computers werken, dat ze je dwingen om heel precies na te denken. Ze moeten letterlijk alles uitgelegd krijgen. Mensen zijn te slim.”



“Werken met computers dwingt je om heel precies na te denken”
foto Jos Uljee / KB

“Big data zijn voor de geestes-wetenschappen wat de telescoop voor de astronomie was”

De onderzoeksvraag

“Die specifieke onderzoeksvraag is dus niet beantwoord in deze periode. Maar we hebben wel heel veel geleerd over het probleem van die spelling. En ik denk dat we het nog wel gaan oplossen in de toekomst.

Een volgende onderzoeksvraag is of je die ritmes ook kunt herkennen in onze moderne zinsbouw. Zulk onderzoek zou je kunnen doen in de databases met kranten en ANP-bulletins van de KB. Samen met een van de programmeurs van de KB ben ik een heel eind gevorderd om dat onderzoek mogelijk te maken. Die samenwerking was heel erg plezierig.”

Ritme van taal zit diep

“Je kunt je trouwens afvragen waarom het ritme van de taal me zo fascineert. In mijn overtuiging is ritme een van de diepste dingen van de taal, letterlijk. We leren het ritme van onze moedertaal al voor onze geboorte, dat is aangetoond. De geluiden die in de moederschoot doordringen zijn te vaag om klinkers en medeklinkers te kunnen onderscheiden, maar het ritme pikken we al op. Pasgeboren baby’s blijven langer wakker als ze hun moedertaal horen spreken dan wanneer het een andere taal is. Dat ritme zit dus heel diep, je hoort het ook in muziek, zelfs als het instrumentaal is. Engelse muziek is anders dan Franse muziek. Daar valt nog zoveel te ontdekken.”

Data geen wetenschap

Terug naar de relatie tussen big data en geesteswetenschappen. Van Oostendorp: “Dat was de andere kant van dit fellowship. Er is een soort polemiek gaande. Tussen mensen die zeggen dat we nu eindelijk écht wetenschappelijk werk kunnen gaan doen, omdat we nu pas objectieve gegevens kunnen verzamelen. Dat het vroeger allemaal subjectief was. Ook is er een tendens om te denken dat verzamelen op zich min of meer vanzelf wetenschap oplevert. Aan de andere kant zijn er mensen die zeggen dat dat allemaal onzin is. Dat alles bij het oude moet blijven, dat er weinig verandert door big data. Het lijkt mij evident dat je tussen die extremen een midden moet vinden.”

Nieuw instrument

“Big data zijn een ontzettend belangrijke ontwikkeling in de geesteswetenschappen, vergelijkbaar met de uitvinding van de telescoop. Maar die laatste leidde niet tot een soort ‘telescoopwetenschap’ met als onderzoeksvraag: wat kun je zien door een telescoop? Het debat bleef gaan over hoe het heelal in elkaar zit. Met het blote oog hadden we al veel gezien. Die kennis werd verdiept en verbeterd door de telescoop. Zo zie ik ook de nieuwe ontwikkeling van de digital humanities. Big data vormen een prachtig nieuw instrument, en er gaat veel veranderen, maar je moet veranderen vanuit wat we al weten. We denken al honderden jaren na over taal, en dat is niet zomaar onzin. Dat is niet alleen maar subjectief, dat is ook getoetst op veel verschillende manieren.”
vanoostendorp.nl

Marc van Oostendorp
Marc van Oostendorp is senior-onderzoeker aan het Meertens Instituut en hoogleraar Fonologische microvariatie aan de Universiteit Leiden. Hij publiceert regelmatig op allerhande (online) fora, waaronder Neder-L, elektronisch tijdschrift voor neerlandistiek. Hij was KB-NIAS fellow van september 2013 tot en met januari 2014.

AGENDA

4 maart • Hilversum
Preservation Metadata in praktijk
Welke gegevens zijn van belang voor de langetermijntoegang tot digitale bestanden? Deze vraag staat centraal in de workshop georganiseerd door het Nederlands Instituut voor Beeld en Geluid en de Nationale Coalitie voor Digitale Duurzaamheid. De workshop is bedoeld voor iedereen die te maken heeft met het duurzaam bewaren van digitale objecten zowel op uitvoerend als op managementniveau: collectiebeheerders, bibliothecarissen, archivarissen, onderzoekers, ICT-beheerders en ontwikkelaars van software.
ncdd.nl

4 - 7 maart • Berlijn (Duitsland)
iSchools conferentie Breaking down walls
Op het forum iSchools kunnen onderzoekers en professionals op het gebied van informatiewetenschap elkaar ontmoeten. In maart organiseert iSchools een conferentie met als thema: Breaking down walls | culture, context, computing.
ischools.org/the-iconference

13 - 15 mei • Rome (Italië)
CRIS2014: Current Research Information Systems
De conferentie richt zich op recente ontwikkelingen in beheer, beschikbaarheid, kwaliteit en gebruik van onderzoekinformatie. De doelgroep wordt gevormd door onderzoekers, managers, financiers, ICT-experts en beleidsmakers.
cris2014.org

13 - 16 mei • Berlijn (Duitsland)
IS&T Archiving Conference
De Society for Imaging Science and Technology organiseert sinds 2004 jaarlijks een conferentie op het gebied van digitale conservering van met name cultureel erfgoedmateriaal.
imaging.org/ist/conferences/archiving

26 - 31 mei • Reykjavik (IJsland)
International Conference on Language Resources and Evaluation
Elke twee jaar wordt de LREC-conferentie georganiseerd. Deze conferentie brengt onderzoekers op het gebied van taaltechnologie bij elkaar.
lrec-conf.org

27 - 30 mei • Istanbul (Turkije)
Conference on Qualitative and Quantitative Methods in Libraries
Deze zesde editie van de QQML-conferentie is bedoeld voor betrokkenen bij het ontwikkelen, uitvoeren en analyseren van kwalitatieve en kwantitatieve methoden om het functioneren van bibliotheken te verbeteren.
isast.org

9 - 13 juni • Helsinki (Finland)
Open Repositories 2014
Er zijn verschillende datamanagementsystemen, met verschillende functies en rollen. Netwerken verbinden de repositories. Vandaar het thema van de conferentie: Towards Repository Ecosystems.
or2014.helsinki.fi

12 - 13 juni • Den Haag
DHBenelux - Conference for Digital Humanities Research
Deze conferentie op het gebied van digitale geesteswetenschappen wordt voor het eerst georganiseerd. De call for proposals meldt dat ook onderzoekers van buiten de Benelux voorstellen kunnen indienen.
dhbenelux.org/dhbenelux-2014-conference