

GEHOORD & BIJGEWOOND

E-humanities ook voor historische taalkunde

Erica Renckens

Oude teksten vormen voor de computer een grote uitdaging, vooral vanwege de enorme spellingvariatie. Dat is echter geen argument om digitale onderzoeksmethoden links te laten liggen in de historische taalkunde, zo betoogde prof. dr. Nicoline van der Sijs tijdens haar oratie op 27 maart jongstleden aan de Radboud Universiteit. Ter ondersteuning van dit standpunt besprak ze verschillende mogelijkheden van de toepassing van computationele technieken in haar vakgebied. Zo bewijst een vergelijking tussen verschillende corpora dat de Statenvertaling al bij zijn verschijnen in 1637 archaisch was: de tekst bevat meer verbogen naamvallen dan andere teksten uit dezelfde periode. Ook kan de computer taalveranderingen als neologismen en nieuwe grammaticale constructies opsporen door twee corpora te analyseren.



Nicoline van der Sijs tijdens haar oratie foto Gert Janssen

Om onderzoeksvragen met behulp van een computer te kunnen beantwoorden, moeten wetenschappers die juist formuleren. Om dat te trainen, organiseren faculteiten cursussen en workshops *e-humanities*. Daarnaast zal Van der Sijs in de loop van volgend jaar met collega's een handboek voor de digitale geesteswetenschappen publiceren. ru.nl/nederlands/wie_wat_waar/medewerkers/nicoline-sijs

Nieuwe tools voor Europees onderzoek

Tom van Nuenen

De universiteit van Würzburg ontving op 1 en 2 april jongstleden een twintigtal jonge wetenschappers uit Europa en Amerika voor de tweede workshop van het *Network for Digital Methods in the Arts and Hu-*

manities (NeDiMAH). Het netwerk, dat in 2011 werd opgericht, richt zich op computationeel onderzoek in de Europese humaniora.

Er was ruimte voor kennismaking met nieuwe tools – zoals *Stylo*, een stilometrische tool die verschillende classificatie- en clusteringalgoritmen combineert. Maar vooral het accent op samenwerking bleef terugkeren. De verschillende sprekers bleken in elk van hun landen bezig met vergelijkbare projecten om grote hoeveelheden literaire teksten te indexeren en analyseren. In Nederland wordt bijvoorbeeld *NederLab* ontwikkeld: een website die Nederlandstalige tekstcollecties onder één dak bijeen brengt en tekstanalyse-software aanbiedt om die collecties te ontleden. In Duitsland, Polen en Amerika bleken soortgelijke projecten gaande. De vraag die daarbij telkens opnieuw aan de orde kwam: hoe zorg je ervoor dat die corpora geformaliseerd en uitwisselbaar zijn tussen die landen en talen?

Een eerste stap werd alvast gezet op de tweede dag, toen een op handen zijnde 'benchmark corpus' van *NeMiDAH* werd besproken. Organisatrice Karina van Dalen-Oskam (Huygens ING) legt uit: "Het gaat om een corpus van honderd Engelstalige romans, waaronder één roman waarvan het auteurschap betwist is en aan sommige van de andere auteurs in het corpus is toegeschreven. Doel van dit corpus is om auteursherkenningssoftware te kunnen testen en de resultaten te kunnen vergelijken. Als iemand een nieuwe tool ontwikkelt, en die op exact dezelfde data toepast, kun je meten hoe goed hij werkt." nedimah.eu

EACL toont ook de muziek van de literatuur

Mike Kestemont

De *Conference of the European Chapter of the Association for Computational Linguistics (EACL)* is een prestigieuze conferentie binnen de computerlinguïstiek, waar onderzoekers de recente vooruitgang in de (vooral Europese) taaltechnologie laten zien. Naast het relatief technische hoofdprogramma kent de *EACL* een vruchtbare traditie van kleinere workshops die aan de conferentie voorafgaan. Deze studiedagen zijn meestal georganiseerd rond een nieuw en uitdagend onderzoeksthema binnen de computerlinguïstiek en haar toepassingen. Zo ook de workshop *Computational Linguistics for Literature (CL4L)*, inmiddels al aan de derde editie toe.



Debora Thomas: "This digital age is like the Wild West: opportunities and unruly" foto Inge Angevaare

Aan deze workshop nam een groep van enthousiastelingen deel die zich interesseren voor taaltechnologische toepassingen binnen de hedendaagse literatuurstudie.

De deelnemers rapporteerden over onderzoek in uiteenlopende gebieden, zoals een sociolinguïstische analyse van personages in het oeuvre van Ibsen, intertekstuele verbanden in middeleeuwse Arabische literatuur of de modellering van het lineair verloop in literaire teksten. Keynotesprekers waren de Hamburgse narratoloog Jan Christoph Meister en de visuele artieste María Mencía. Hoogtepunt van de dag was de innovatieve bijdrage van Hannah Davis en Saif Mohammad, waarin zij *TransProse* presenteerden, een systeem dat automatisch muziek op basis van tekst genereert. Binnen hun intuïtieve aanpak wordt het overheersende sentiment van een tekst eerst benaderd op basis van een emotielexicon. De analyse van deze sentimentanalyse wordt vervolgens omgezet in muzikale patronen waarvan de toonhoogtes en het ritme de gevoelens van de tekst nabootsen. Het verbluffende resultaat van *TransProse* kan online worden geraadpleegd (transprose.weebly.com). Laat u ontroeren door de tragische tonen die het systeem distilleerde uit *Heart of Darkness*. aclweb.org/anthology/W/W14/#0900

Patronen en betekenis in grote tekstcollecties

Peter Boot

Hoe kunnen geesteswetenschappelijke projecten met *big questions* het beste omgaan met de *big data* van de grote tekstcollecties waarmee ze hun vragen willen beantwoorden? Dat was de centrale vraag tijdens

de workshop *Mining digital repositories*, gehouden op 10 en 11 april in de Koninklijke Bibliotheek. De workshop was mede georganiseerd door de projecten *Translantis* en *ASYMENC* (zie ook het interview met *Toine Pieters* op pagina 5).

Debora Thomas toonde de digitale krantencollectie van de Library of Congress (USA). De verzameling op zich is al een klein kunststuk, omdat collecties uit vijftig staten moesten worden geïntegreerd. De kranten kunnen online worden doorzocht en gelezen, maar zijn ook benaderbaar via een *application programming interface* (api): onderzoekers kunnen de krantenteksten analyseren met hun eigen programmatuur en zo onder andere *topic modelling* of *sentiment mining* doen. Zo is bijvoorbeeld bestudeerd hoe de kranten in 1918 verslag deden van de Spaanse griep. Alistair Dunning toonde vergelijkbare mogelijkheden in de *Europeana Cloud*, en Hildelies Balk gaf een preview van het researchlab dat de Koninklijke Bibliotheek aan het bouwen is. Ook de aanbieders van commerciële collecties hebben begrepen dat veel onderzoekers van zulke mogelijkheden gebruik willen maken.

Toch toonden veel geesteswetenschappelijke onderzoekers zich sceptisch. Wat is de status van een grafiek of een *word cloud*? Kunnen geesteswetenschappelijke vragen beantwoord worden met een beroep op getallen? Kunnen ze überhaupt worden beantwoord? En hoe betrouwbaar en representatief zijn de repositories? Organisator Joris van Eijnatten besloot dat er nog heel wat *fooling around* met de data nodig zou zijn voor we hierop een antwoord hebben.

tinyurl.com/ketqa9m

KORT

Maak kans op de Digital Preservation Awards 2014

De Digital Preservation Awards worden uitgereikt aan initiatieven die bijdragen aan het duurzaam toegankelijk houden van digitaal erfgoed. Organisaties, projecten, personen en initiatieven kunnen worden genomineerd voor één van de prijzen. De prijsuitreiking vindt op 17 november plaats in Londen. De finalisten en winnaars begeven zich onder de beste projecten en initiatieven in een groeiende internationale community en zullen veel publiciteit krijgen.

De Digital Preservation Coalition (DPC) en de Nationale Coalitie Digitale Duurzaamheid (NCDD) organiseren samen de Digital Preservation Awards. Nominatie voor de Awards is open tot 28 juli via de website van de DPC. Kijk voor meer informatie op de website van de DPC. (HB) dpconline.org

Beheer Dutch Dataverse Network naar DANS

DANS heeft het beheer van het Dutch Dataverse Network (DVN) overgenomen van de Universiteit Utrecht. Het DVN is bedoeld voor onderzoekers en docenten van universiteiten en andere academische instituten in Nederland. Zij kunnen met deze dienst een grote verscheidenheid aan wetenschappelijke data digitaal opslaan in een online omgeving, op een zeer veilige en duurzame wijze. Anja Smit, directeur Universiteitsbibliotheek van de Universiteit Utrecht: "Het DVN is een wetenschappelijk platform ontwikkeld door Harvard. Na de installatie binnen de Universiteit Utrecht in 2010 werd de dienst in zeer korte tijd een gedeelde dienst van zeven instellingen. Dit succesvolle samenwerkingsverband is nu klaar voor de volgende stap. Wij hebben er alle vertrouwen in dat Dataverse door DANS verder ontwikkeld zal worden." Peter Doorn, directeur van DANS: "DANS bevordert de duurzame toegang tot digitale onderzoeksgegevens. Onze bestaande diensten (zoals NARCIS en EASY) faciliteren onderzoekers al voorafgaande aan en na afronding van hun onderzoek. Met de komst van Dataverse kunnen we onderzoekers ook tijdens hun onderzoek ondersteunen op datamanagementgebied. Bovendien zijn we verheugd met het gestelde vertrouwen in DANS door de deelnemers van Dataverse." (HB) dataverse.nl/dvn