

Een digitale bibliotheek van jaarringen in hout

Al zeventig jaar zijn onderzoekers in Europa bezig met de studie van de groeipatronen van hout uit het verleden. Het gaat daarbij om boomsoorten zoals eik, es, beuk, iep, grove den, zilverspar en fijnspar. Ze onderzoeken cultureel erfgoed zoals scheepswrakken, schilderijen, gebouwen en archeologische vondsten (huisplattegronden, waterputten, grafkisten). Daarnaast richten ze zich op in de bodem geconserveerde resten van oude bosvegetaties. De kennis die dat oplevert is van groot belang om de herkomst van (verhandeld) hout te bepalen.

Een belangrijke onderzoeksvraag van deze dendrochronologen is steeds: hoe oud is een stuk hout precies? Met dendrochronologie kun je namelijk vaststellen in welk jaar elke groeiing in hout is gevormd, dus ook in welk kalenderjaar een boom doodging. Dat kan door een ongedateerd jaarringpatroon te vergelijken met absoluut gedateerde dendrochronologische kalenders. Die datering zegt iets over het moment waarop het onderzochte object is gemaakt. Zo is het hout van een Romeinse rivierkade in Leidsche Rijn omgehakt in het voorjaar van 100 n.Chr. De bouw van deze kade zal niet veel later hebben plaatsgevonden.

Er zijn de afgelopen decennia in Europa grote dendrochronologische datasets opgebouwd. De groeipatronen van in Nederland en Noord Duitsland gedateerde bomen bestrijken de laatste acht millennia,

waarbij elk jaar van dat enorme tijdsinterval wordt gedekt door waarnemingen. Datasets uit andere regio's zijn nog omvangrijker. Deze collecties kunnen een belangrijke rol spelen bij interdisciplinair onderzoek naar bijvoorbeeld veranderingen in het landschap en de menselijke omgang hiermee, houthandel en economie, en klimaat.

Voorwaarde is wel dat de collecties gedigitaliseerd zijn en een structuur hebben die ze doorzoekbaar en vergelijkbaar maakt. Veel gegevens zijn echter alleen vastgelegd op papier. Hier is dus een inhaalslag nodig. Datzelfde geldt voor de veelheid aan digitale dataformats die sinds de jaren tachtig van de vorige eeuw is ontwikkeld. Veel formats zijn intussen onbruikbaar geworden voor uitwisseling. Ook worden aan dendrochronologisch dateringsonderzoek (meestal *contract research*) in het algemeen geen kwaliteitseisen gesteld



BERT VAN AS

Esther Jansma

voor de registratie, verduurzaming en ontsluiting van gegevens. Die bevinden zich daardoor vaak in de pc's van individuele onderzoekers, onderhevig aan digitaal verval.

Nederland heeft, met een subsidie van NWO, het voortouw genomen om deze situatie te verbeteren. Laboratoria in België, Duitsland, Frankrijk, Nederland en Polen werken daarbij samen met de Rijksdienst voor het Cultureel Erfgoed en de Universiteit Utrecht. Dertigdui-

zend dendrochronologische meetreeksen van 6000 v.Chr tot nu en metadata worden geverifieerd, opgewerkt en gecombineerd in dit project met als titel *A digital Collaboratory for Cultural Dendrochronology (DCCD) in the Low Countries*. Het project, dat zich met name richt op de Lage Landen, heeft met hulp van Europese en Amerikaanse wetenschappers en ict-deskundigen een internationale standaard opgeleverd voor digitaal archiveren en uitwisselen: de *Tree-Ring Data Standard* TriDaS. Deze standaard blijkt internationaal aan te slaan. Nog onlangs heeft een belangrijk Amerikaans instituut voor dendrochronologisch onderzoek de eigen collectie naar TRiDaS omgezet, waardoor uitwisseling met het DCCD mogelijk wordt. Andere Amerikaanse instellingen zoeken naar fondsen om hetzelfde te doen. Het TriDaS-model vormt de basis van de dataopslag- en webapplicatie die nu wordt ontwikkeld door DANS, dat de gegevens opslaat volgens de normen van het *Data Seal of Approval*. De DCCD, die eind 2010 gereed zal zijn, biedt een viertalige gebruikersschil waardoor onderzoekers nieuwe gegevens kunnen toevoegen, oude projecten kunnen herwerken en zoekacties uitvoeren. Daarmee wordt het een levend, be-

weeglijk en steeds actueel archief.

In Europa bestaat binnen het vakgebied sterke belangstelling om deze infrastructuur uit te breiden. Een belangrijke reden is dat het DCCD aan dataleveranciers de mogelijkheid biedt om eigen gegevens ten dele af te schermen. Dit is vooral relevant voor onderzoekers in de private sector, die hun gegevens gebruiken als referentie bij dateringsonderzoek en hun gegevens niet willen ontsluiten voor concurrerende bedrijven, maar wel geïnteresseerd zijn in wetenschappelijke samenwerking. Het is voor het eerst dat deze groep, die zich altijd heeft verzet tegen de ontsluiting van gegevens, bereid is mee te denken over de digitale koppeling van collecties.

Esther Jansma is bijzonder hoogleraar in Utrecht en onderzoeker bij de Rijksdienst voor Cultureel Erfgoed.

Lezing: Duizenden jaren houtgebruik: een dendrochronologisch repository voor de Lage Landen (6000 v.Chr. - heden)

www.dendrochronology.eu
www.ncdc.noaa.gov/paleo/tree-ring.html
www.tridas.org

Hoe vrouwvriendelijk is de Tweede Kamer?

MAARTEN MARX

Ruwe tekst als onderzoeksmateriaal voor tekstanalyse door sociale wetenschappers komt in steeds grotere hoeveelheden beschikbaar, bijvoorbeeld via internet. Maar om van tekst in allerlei formaten naar een mooie invoerfile voor het gangbare rekenpakket SPSS te komen is vaak nog een hele stap. Toch zijn de hulpmiddelen daarvoor beschikbaar.

Een recent artikel in *Science*, genaamd 'Computational Social Science', pleit voor een curriculum waarin studenten aan de alfa en gamma faculteiten *tools* leren gebruiken om enorme hoeveelheden tekst te kunnen verwerken. Aan de hand van een voorbeeld kunnen we zien wat die *tools* inhouden, en dat de omgang ermee helemaal niet zo moeilijk is. Computers zijn tegenwoordig zo krachtig en eenvoudig geworden dat ook echte gamma-infobeten kwantitatief onderzoek op basis van gigantische hoeveelheden tekst kunnen uitvoeren. De kennis die daarvoor nodig is kan in een vak 'Tekstanalyse voor de sociale wetenschappers' van bescheiden omvang aan elke student geleerd worden.

Laten we ons richten op de volgende onderzoeksvragen. Er zit nu een recordaantal van meer dan veertig procent vrouwen in de Tweede Kamer. Zijn dat nu de bekende Excuus-Truusen of zijn ze ook evenredig veel aan het woord? En vervolgens: verschilt dat per onderwerp, of tussen de partijen? Hoe zat het vroeger?

Alle data om deze vragen te beantwoorden zijn aanwezig. De Handelingen der Staten Generaal zijn als pdf-bestanden op het internet beschikbaar vanaf 1917. Ze bevatten



ANP

CDA-parlementariër Mirjam Sterk heeft in de Tweede Kamer de aandacht van drie bewindslieden

exact wat iedereen in de Tweede Kamer gezegd heeft. Daarnaast bevat de website *parlement.com* voor iedereen die ooit in de Kamer heeft gezeten een uitgebreide biografie. We zouden dus gewoon kunnen gaan turven hoe vaak elk parlementslid aan een debat heeft meegedaan, hoe vaak hij of zij heeft geïnterrupteerd, aan het woord is geweest en hoe lang dat lid op de spreekstoel heeft gestaan. Tijden staan weliswaar niet vermeld in de Handelingen, maar die kunnen we benaderen door het aantal gesproken woorden te tellen.

Tot zo ver lijkt het dus niet moeilijk. Toch is de uitvoering niet eenvoudig omdat de data niet in het juiste

formaat beschikbaar zijn. Er zijn drie problemen. In de eerste plaats gaat het om heel veel data: 3560 biografieën en meer dan honderd miljoen woorden, gesproken in de Tweede Kamer vanaf 1995. In de tweede plaats is de koppeling van de twee datasets moeilijk omdat er niet consequent met dezelfde namen naar parlementsliden verwezen wordt. Dit probleem is des te erger met data van voor 1995, die zijn ingescand en nog foutjes bevatten door het gebruik van optische tekenherkenning OCR. En ten slotte bestaan de Handelingen uit tekstbestanden in pdf-format, met summier metadata. Voor elk woord in de Handelingen weten we wel op welke

dag het is uitgesproken en op welke bladzijde het staat, maar niet door wie en in welke hoedanigheid: als parlementslid of lid van de regering; als betoog, interruptie of antwoord op een interruptie.

Nadat de Handelingen in een machineleesbaar formaat zijn gebracht kunnen technieken voor tekstanalyse het probleem van de herkenbaarheid en hoedanigheid van de sprekers oplossen. Met *named entity recognition* and *reconciliation* kunnen we sprekers herkennen, hun naam normaliseren en zo de hindernissen voor het combineren van de twee datasets opruimen. Op dat moment kunnen we computers inschakelen om het probleem van de grote hoeveelheid data op te lossen. We hoeven dan niet met dure codeurs en steekproeven te werken, en kunnen de analyse op de gehele dataset uitvoeren. De Universiteit van Amsterdam maakt op dit moment in samenwerking met de Koninklijke Bibliotheek de Handelingen beschikbaar in een XML-formaat waarmee de hiervoor gestelde onderzoeksvraag echt eenvoudig op te lossen is.

Ondertussen hebben we bij wijze van voorproefje al wat ruwe tellingen gedaan om de vrouwvriendelijkheid te meten. Als we de voorzitter niet meetellen is in de periode van 3 februari 2009 tot en met 8 oktober 2009 33% van de spreektijd in de Tweede Kamer door vrouwen gebruikt. Bijna

20% minder dan je zou verwachten op basis van het aantal vrouwelijke leden. Slechts 30% van de tijd staat er een vrouw op de spreekstoel. Dit kan natuurlijk nog van alles betekenen. Misschien zijn vrouwen wel minder langdradig en geven kortere antwoorden op interrupties dan mannen.

Er is enorm veel geïnvesteerd in taal-technologie tools voor het Nederlands. Echter, voor 'gewone' alfa's en gamma's zijn die tools nog vaak erg moeilijk toepasbaar, zeker als ze in combinatie gebruikt moeten worden. Dit is zonde want er is een schat aan ruwe data vrij beschikbaar. Het hiervoor aangehaalde *Science*-artikel beschrijft het reële gevaar dat de industrie (Google, Amazon, etc) het vakgebied 'Computational Social Science' voorgoed voor de neus van de wetenschap wegkaapt. Laten we zorgen dat dat niet gebeurt.

Binnenkort kan iedereen het genoemde onderzoek zelf uitvoeren, want we plaatsen alle Handelingen in een uniform XML-formaat in het EASY archief van DANS. En elke dag wordt het aangevuld.

Maarten Marx is politicoloog en informatiewetenschapper en doceert aan de Universiteit van Amsterdam.

Lezing: Tekstanalyse voor de Sociale Wetenschappen