STEVEN KRAUWER, PETER WITTENBURG, MARTIN EVERAERT

CLARIN maakt teksten op Europese schaal beschikbaar

Een van de projecten met een belangrijke Nederlandse inbreng die afgelopen najaar werden opgenomen in de Roadmap van het EU-forum voor onderzoekinfrastructuren is CLARIN: Common Language Resources and Technology Infrastructure. Drie indieners geven uitleg over de bedoeling.

Vrijwel elke onderzoeker in de geesteswetenschappen en de sociale wetenschappen wordt geconfronteerd met taal in een van vele verschillende rollen: drager van culturele kennis uit heden en verleden, instrument voor communicatie binnen en tussen taalgemeenschappen, onderdeel van de identiteit van een individu, een groep, een cultuur of een natie, etc.

Deskundigen als J.A. Leeghwater werden in de zeventiende eeuw in heel Europa ingeschakeld voor waterwerken. Meertalige intelligente zoekopdrachten kunnen helpen verklaren waarom.

Het volume aan vastgelegde teksten, geschreven of gesproken, is enorm, en groeit exponentieel. Wie als onderzoeker geïnteresseerd is in taalmateriaal kan door de omvang er van niet meer om het gebruik van ICT heen. Computers gebruiken voor dit soort doeleinden is niet nieuw. Wel nieuw is de ontwikkeling om bij dit soort onderzoek niet alleen te kijken naar de vormkant van taal (bijvoorbeeld woordgebruik, zinslengte) maar ook naar de inhoud (wat wordt er gezegd, hoe verschilt de inhoud van de ene tekst van die van de andere).

Hergebruik volgens standaards

Om met behulp van de computer vorm- en inhoudsvragen aan teksten te kunnen stellen moeten teksten in elk geval in digitale vorm beschikbaar zijn, en ze moeten geannoteerd zijn, dus verrijkt met aanvullende informatie over vorm en inhoud. Het verzamelen, digitaliseren en annoteren van teksten in geschreven of gesproken vorm is tijdrovend en duur. Daarom is het belangrijk dat eenmaal gedigitaliseerde en geannoteerde teksten beschikbaar blijven voor anderen die dit materiaal willen onderzoeken. Met name voor dit hergebruik is het van essentieel belang dat de wijze waarop digitalisering en annotatie plaatsvinden in overeenstemming is met bestaande standaards.

Een hypothetisch voorbeeld. Historici willen begrijpen op welke schaal, en waarom al honderden jaren bij waterprojecten in Europa vaak Nederlandse deskundigen werden ingehuurd. Dat betekent er intelligente zoekopdrachten (op namen, concepten) uitgevoerd moeten worden in allerlei historische teksten in een grote hoeveelheid talen (in verschillende taalstadia). Aanvullend zouden er samenvattingen gemaakt moeten kunnen worden van nog onbestudeerd materiaal, om te zien of ze relevant zijn voor nadere, diepere bestudering. En tenslotte zouden alle deze tekstfragmenten aan elkaar gereleerd moeten kunnen worden.

Computernetwerk met expertisecentra

De CLARIN-infrastructuur zal de vorm aannemen van een netwerk van onderling gekoppelde computersystemen. Dat moet op Europese schaal onderzoekers in de geesteswetenschappen en de sociale wetenschappen gemakkelijk digitale toegang geven tot gesproken en geschreven taalmateriaal en tot bepaalde gereedschappen om het mee te kunnen bewerken. Bij het materiaal kan het gaan om bestaande en nog te ontwikkelen verzamelingen; de gereedschappen bieden de onderzoeker de mogelijkheid om het materiaal te verkrijgen in een vorm die het beste aansluit bij het onderzoek dat hij er mee wil doen. Om te garanderen dat de bestanden en de gereedschappen maximaal koppelbaar en combineerbaar zijn wordt binnen CLARIN grote nadruk gelegd op het ontwikkelen en bevorderen van standaards.

Naast het computernetwerk met de daaraan gekoppelde data- en gereedschappenbestanden zal de CLARIN infrastructuur ook een aantal expertisecentra omvatten, waar expertise bijeengebracht wordt over bepaalde talen, talengroepen, of onderzoekstechnieken. Deze centra zullen trainingsactiviteiten verzorgen in het gebruik van de databestanden en gereedschappen door studenten en onderzoekers in de geesteswetenschappen.

Grote en kleine talen

Belangrijk daarbij is dat CLARIN recht wil doen aan de meertaligheid van Europa, een sterkte eerder dan een zwakte van Europa in onze ogen. CLARIN wil in principe alle talen van Europa omvatten, zowel de grote als de kleinere. Het definiëren van een standaard van dataverzamelingen en gereedschappen die voor een taal beschikbaar moeten zijn, is een van de beoogde doelen van CLARIN. Dit moet leiden tot versterking van de positie van de kleinere talen, waarvoor gewoonlijk maar weinig technologische ondersteuning beschikbaar is. En, belangrijker misschien nog, CLARIN wil de faciliteiten aanbieden die het mogelijk maken om over taalgrenzen heen te gaan. Dit moet de onderzoeker in staat stellen om onderzoeksvragen te stellen die teksten in verschillende talen combineren of aan elkaar relateren.

ASTRID VERHEUSEN

Databank Digitale Dagbladen

De Koninklijke Bibliotheek gaat de komende jaren op grote schaal landelijke, regionale en koloniale dagbladen digitaliseren en via Internet toegankelijk maken voor onderzoekers en het brede publiek. De Databank Digitale Dagbladen zal acht miljoen krantenpagina's bevatten vanaf de zeventiende tot en met de twintigste eeuw. Alle worden als afbeelding aangeboden maar zullen ook worden omgezet naar computerleesbare tekst zodat op ieder woord kan worden gezocht. Het project wordt mogelijk gemaakt door een eenmalige subsidie uit het Nationaal Programma Grootschalige Onderzoeksfaciliteiten.

In Nederland zijn in de afgelopen vier eeuwen zo'n zeven duizend landelijke, regionale en koloniale dagbladtitels verschenen. Dagbladen bevatten informatie over de geschiedenis van de samenleving, politiek, economie, kunst, cultuur en wetenschap. Ze vormen een onmisbare bron voor tal van onderzoekers. Kranten worden echter gedrukt voor gebruik op één dag en dus op dun en slecht papier. Door de kwetsbaarheid van het materiaal dreigt een belangrijke bron voor het wetenschappelijk onderzoek verloren te gaan.

Als nationale bibliotheek van Nederland heeft de Koninklijke Bibliotheek een van de grootste collecties kranten in bezit. Vanaf de jaren zeventig heeft de Koninklijke Bibliotheek, in samenwerking met andere erfgoed instellingen, veel aandacht besteed aan de conservering van dagbladen door middel van microverfilming. Miljoenen krantenpagina's werden zo behouden voor het nageslacht. De toegankelijkheid van dagbladen is echter beperkt: onderzoekers moeten het originele materiaal of de microfilm pagina voor pagina doorbladeren en kunnen belangrijke informatie snel over het hoofd zien. Door digitalisering en ontsluiting zal het wetenschappelijk onderzoek in dagbladen veel efficiënter kunnen worden uitgevoerd.

De KB beschikt over de nodige ervaring met het digitaliseren van dagbladen. Eind 2004 werden 350.000 pagina's gedigitaliseerde kranten uit twee projecten beschikbaar gesteld op de website 'Kranten in beeld' (kranten.kb.nl). Op de website zijn de gedigitaliseerde afbeeldingen van de kranten op te vragen en kunnen de kranten fulltext worden doorzocht. 'Kranten in beeld' trok onmiddellijk na de lancering veel belangstelling van wetenschappelijke onderzoekers en het brede publiek. De verwachting is dat door massadigitalisering van dagbladen het gebruik ervan enorm zal toenemen.

De Databank Digitale Dagbladen is niet alleen interessant voor wetenschappelijk onderzoek naar de geschiedenis van de samenleving, politiek en economie. De omvang van het corpus, naar schatting zo'n 25 miljard woorden, maakt de databank ook aantrekkelijk voor onderzoek op het terrein van de taaltechnologie en de taalkunde. De databank kan worden ingezet binnen het vakgebied van de *information retrieval* waarin methoden en technieken worden ontwikkeld om grote, ongestructureerde verzamelingen van documenten toegankelijk te maken.

Voor de selectie van krantentitels is een Adviescommissie ingericht bestaande uit inhoudelijke specialisten. Daarnaast wordt door DANS (Digital Archiving and Networked Services) een onderzoek uitgevoerd naar de wensen van onderzoekers op het gebied van ontsluiting en zoekfunctionaliteiten, zodat optimale onderzoeksmogelijkheden ontstaan. Vervolgens zal geïnventariseerd worden welk materiaal al beschikbaar is op microfilm. Microfilms van goede kwaliteit kunnen worden gebruikt voor digitalisering. Voor complementering van ontbrekende titels zal worden samengewerkt met bibliotheken en archieven die dagbladen in bezit hebben. Ook zal gekeken worden naar elders reeds gedigitaliseerd krantenmateriaal en naar mogelijkheden deze te integreren in de Databank Digitale Dagbladen. Kwesties met betrekking tot copyright zullen worden besproken met de uitgeefsector. In het project zal veel aandacht worden besteed aan het verbeteren van de optische tekenherkenning (OCR) van krantenmateriaal.

Het project is eind 2006 gestart en heeft een doorlooptijd van vijf jaar. In 2007 vindt de voorbereiding van het digitaliseringstraject plaats. Gestart wordt met de inrichting van een projectteam. De Adviescommissie zal zich buigen over de selectiecriteria en DANS is inmiddels begonnen met het onderzoek naar de wetenschappelijke gebruikerswensen. In de eerste helft van 2007 wordt een *Request for Information* gehouden om te inventarisaren welke software beschikbaar is voor de ontsluiting van het materiaal. Door middel van een Europese aanbesteding zal een leverancier worden geselecteerd die de digitalisering ter hand gaat nemen. Eind 2008 worden de eerste resultaten op een website gepresenteerd.

http://kranten.kb.nl