http://nederl.blogspot.nl/2012/09/nederlab-voor-al-uw-diachrone.html

zondag 2 september 2012

Nederlab - voor al uw diachrone onderzoeksvragen

door Nicoline van der Sijs

Op 12 juni is in Neder-L gemeld dat het project 'Nederlab - een laboratorium voor onderzoek naar de veranderingspatronen in de Nederlandse taal en cultuur' 4 miljoen euro subsidie heeft ontvangen van NWO, KNAW, CLARIAH en CLARIN, inclusief matching van wetenschappelijke instituten en universiteiten. Inmiddels is er een projectwebsite in de lucht, http://www.nederlab.nl, waarop de oorspronkelijke aanvraag, de organisatiestructuur van het project en enkele persberichten zijn te vinden. Ik grijp de lancering van de website graag aan om wat achtergrondinformatie te geven over het project en een oproepje te doen.

Het doel van het project is een nieuw onderzoeksinstrumentarium voor de geesteswetenschappen te creëren. Nederlab zal de toegangspoort worden van waaruit studenten en onderzoekers een groot corpus aan gedigitaliseerde teksten onder handbereik krijgen. Op dit moment zijn er wel veel historische teksten gedigitaliseerd, maar ze worden door een groot aantal instellingen – DBNL, KB, universiteitsbibliotheken, onderzoeksinstituten - op verschillende plaatsen aangeboden. Iedere instelling biedt zijn eigen zoekinterfaces en zoekmogelijkheden, er bestaan aanzienlijke kwaliteitsverschillen tussen de verschillende corpora, en iedere instelling voegt zijn eigen metadata toe. Het gevolg hiervan is dat al deze tekstbestanden - en hun metadata - slechts naast elkaar, en niet tegelijkertijd en samen, kunnen worden doorzocht en geanalyseerd.

Nederlab wil aan die versnippering een eind maken, door het bouwen van een gebruiksvriendelijke, algemeen toegankelijke en met tools verrijkte gebruikersomgeving. Binnen die omgeving worden alle gedigitaliseerde teksten die relevant zijn voor de geschiedenis van de Nederlandse taal en cultuur bijeengebracht: van de oudste geschreven periode (circa 800) tot heden, representatief over de hele periode verdeeld. Daarbij worden de metadata die de verschillende instellingen aan de teksten toevoegen, aan elkaar gelinkt en geharmoniseerd. Hierdoor wordt het voor het eerst mogelijk langetermijnveranderingen in de taal en de cultuur te traceren: zo kan de visie op vreemdelingen door de eeuwen heen worden onderzocht, het ontstaan, de verbreiding en het verdwijnen van literaire genres, de veranderende oordelen over literaire schrijvers, geleerden of historische gebeurtenissen, en de ontwikkeling van het gebruik van voorzetsels of voegwoorden.

Iedere onderzoeker krijgt een eigen virtuele werkruimte binnen Nederlab waar hij, alleen of met andere onderzoekers, data kan verzamelen en bewerken. Op basis van de metadata kan een onderzoeker zelf kiezen welk deel van het enorme corpus hij wil gebruiken voor zijn onderzoek: aan iedere tekst wordt informatie toegevoegd over datering, lokalisering, auteur en tekssoort. Een onderzoeker kan op basis van die metadata bijvoorbeeld onderzoeken of er regionale verschillen bestaan tussen rijmteksten uit de 14e eeuw, of in hoeverre het taalgebruik van egodocumenten uit de 18e en 19e eeuw verschilt van de geschreven standaardtaal uit diezelfde periode. Maar ook een onderzoeker die in de 20e eeuw is gespecialiseerd, zal moeiteloos een eigen geschikt subcorpus kunnen samenstellen (waarbij de IPR uiteraard worden gerespecteerd).

Het project start officieel op 1 januari 2013. In het eerste jaar wordt de Nederlab-infrastructuur neergezet, en worden de tekstbestanden en metadata van de DBNL als onderzoekscorpus ingebracht. Begin 2014 wordt een eerste versie van de Nederlab-website

gelanceerd. In de daaropvolgende jaren, tot 1 januari 2018, worden de gegevens gestructureerd uitgebreid: daarbij worden bijvoorbeeld de auteursgegevens van de KB en van universiteitsbibliotheken gekoppeld aan die van de DBNL. Dat moet eenmalig met de hand gebeuren (iemand moet beslissen of Jan Janssen uit de KB dezelfde is als Jan Janssen uit de DBNL). Is de koppeling eenmaal gelegd, dan worden voortaan alle werken van Jan Janssen automatisch aan elkaar gekoppeld, ook werken die hij in de toekomst nog zal publiceren. Ook worden gedurende het project steeds meer tools aangeboden: gebruikersvriendelijke computerprogrammaatjes waarmee teksten en metadata op verschillende manieren kunnen worden doorzocht en waarmee zoekresultaten inzichtelijk worden gemaakt in bijvoorbeeld de vorm van een lijndiagram, staafdiagram of puntenwolk.

Nederlab komt voort uit de wens van de geesteswetenschappelijke onderzoekers. Van begin af aan zullen die onderzoekers zeer nauw betrokken worden bij de inrichting van Nederlab. Onderzoekers, studenten, promovendi en postdocs worden uitgenodigd de infrastructuur te testen en te becommentariëren, en een bijdrage te leveren aan de invulling van de infrastructuur. Zonder data, metadata en tools is de infrastructuur immers niets meer dan een lege dop. Onderzoekers die gecorrigeerde transcripties, geannoteerde teksten of tools hebben gemaakt die geschikt lijken voor opname binnen Nederlab, worden van harte uitgenodigd contact op te nemen. En studenten kunnen overwegen via een stage mee te helpen, met bijvoorbeeld het bouwen van een specifiek subcorpus, het (semi-automatisch met aangepaste tools) corrigeren van leesfouten in ocr-bestanden, het werken aan metadata-schema's, het taalkundig verrijken of annoteren van geselecteerde corpora, of het aanpassen of ontwikkelen van tools voor het doorzoeken, visualiseren en annoteren van historische teksten.

Laat het me vooral weten als u vragen of suggesties heeft: via e-mail (post@nicolinevdsijs.nl) of per slakkenpost: Nicoline van der Sijs, Meertens Instituut Postbus 94264, 1090 GG Amsterdam.