Nederlab

Inleiding

Digitale historische teksten worden nu nog op allerlei verschillende plaatsen en door verschillende instellingen op verschillende manieren beschikbaar gesteld. Nederlab wil een gebruiksvriendelijke webinterface inrichten vanwaaruit onderzoekers al deze losse corpora als eenheid kunnen doorzoeken en analyseren, zowel op tekstniveau als op metadataniveau. Dit maakt het mogelijk langetermijnveranderingen in de Nederlandse taal, letterkunde, cultuur en maatschappij te traceren.

Doelstellingen

Er is en wordt veel overheidsgeld gestopt in het digitaliseren van corpora en in de ontwikkeling van tools om deze te doorzoeken, analyseren en bewerken. Het toepassen van statistische analysemethodes en de introductie van bètawetenschappelijke waarden als verifieerbaarheid en herhaalbaarheid beloven al enige tijd een ingrijpende vernieuwing van het onderzoek in de alfawetenschappen in gang te zetten. Er bestaat echter een kloof tussen de geboden middelen en de onderzoekspraktijk van de gemiddelde geesteswetenschapper, waardoor de beloftes nog niet worden waargemaakt. Het is de ambitie van Nederlab om die problemen in één klap op te lossen.

Het eerste probleem van de diachrone corpora is dat ze op verschillende plaatsen en door verschillende instellingen worden aangeboden. Grote corpora zijn aanwezig bij DBNL, Huygens ING, INL, KB en Meertens Instituut. Daarnaast worden vele kleinere corpora aangeboden op websites van universiteiten of individuele onderzoekers. Al deze corpora kunnen nu slechts naast elkaar – en niet tegelijkertijd en samen – worden doorzocht en geanalyseerd. Bovendien verschillen de zoekinterfaces en zoekmogelijkheden per corpus, bestaan er aanzienlijke kwaliteitsverschillen tussen de verschillende corpora, en voegt iedere instelling zijn eigen metadata toe. Voor longitudinaal onderzoek is het absoluut noodzakelijk dat alle losse corpora als eenheid doorzoekbaar gemaakt worden. De grootste uitdaging van Nederlab zal eruit bestaan te inventariseren welke corpora er zijn en er vervolgens voor te zorgen dat alle (bestaande en nog te vervaardigen) diachrone corpora gedistribueerd doorzoekbaar worden, zowel op tekstniveau als op metadata-niveau.

Het tweede probleem vormen de tools. Momenteel worden er allerlei tools ontwikkeld, die uiteindelijk in principe beschikbaar komen via CLARIN-NL. Deze tools zijn echter vaak niet bruikbaar voor een andere situatie dan waarvoor ze zijn ontwikkeld (bijvoorbeeld voor morfologische verrijking van teksten uit de 14e eeuw die op een bepaalde manier zijn voorbewerkt). Nederlab wil, in nauw overleg met CLARIN, de

bestaande tools op een centrale plaats bijeenbrengen en de gebruikersinterfaces zo aanpassen dat ze binnen de geboden infrastructuur direct bruikbaar zijn voor niet technisch onderlegde onderzoekers, en toepasbaar zijn op diachrone teksten. Daarbij sluit Nederlab aan bij de standaarden in formaten en metadata van CLARIN.

Nederlab bouwt dus voort op de vele initiatieven met betrekking tot corpusopbouw en tool-ontwikkeling die door de Nederlandse overheid en onderzoeksinstellingen als KNAW en NWO zijn ontplooid, maar voegt hieraan belangrijke meerwaarde toe: een comfortabele infrastructuur voor onderzoekers en studenten die automatisch leidt tot samenwerking en synergie binnen de geesteswetenschappen en tot het stellen van nieuwe, veelal interdisciplinaire onderzoeksvragen.

Het INL en Nederlab

Het INL is track supervisor lexicale data en is met name verantwoordelijk voor lexicondata om het historische materiaal mee te verrijken, verrijkt gold standard corpusmateriaal om diverse tools te kunnen trainen en evalueren. Daarnaast worden diverse corpora aangepast en ingebracht.

Projectduur

Nederlab is een NWO-grootproject. Het is gestart per 1 januari 2013 en zal duren tot 31 december 2017.

Website: www.nederlab.nl