Marc Kemps-Snijders en Ineke Schuurman

Naar een betere toegang tot tekst- en spraaktools

Het Nederlands-Vlaamse project 'Taal- en spraaktechnische Tools voor het Nederlands als Webservices in een Workflow (TTNWW)' maakt technische faciliteiten op het gebied van tekst en spraak toegankelijk voor een brede groep onderzoekers in de geestes- en sociale wetenschappen met geringe technische bagage. Daardoor moeten ze hun onderzoeksvragen beter en gemakkelijker aan kunnen pakken en vragen kunnen stellen die tot dusverre niet te beantwoorden waren.

TTNWW sluit aan bij CLARIN, een Europees project om een geïntegreerde en interoperabele infrastructuur te bouwen die onderzoekers eenvoudig toegang geeft tot taalkundige bronnen. Die bronnen kunnen geannoteerde media-bestanden zijn (vooreerst spraak en tekst), lexica en ontologieën, of het kan om technieken gaan zoals verwerkt in spraakherkenners, lemmatizers en parsers. Zulke bronnen, en de technieken om ze toegankelijk te maken, zijn voor het onderzoek van groot belang. Bij het beschikbaar maken ervan zijn vaak aanzienlijke kosten gemaakt voor het verzamelen, digitaliseren en annoteren van bijvoorbeeld grote tekst- of spraakcorpora – werk dat grote vaardigheid en expertise vereist.

Die bronnen en technologie kunnen alleen volledig worden benut als ze op grote schaal eenvoudiger worden ontsloten met een robuuste en duurzame infrastructuur, zodat de onderzoekers niet zelf over (taal)technische knowhow hoeven te beschikken. Zo'n infrastructuur behelst de opslag en toegankelijkheid van data en services op grote schaal, het onderling aansluiten van dataformaten en het vermogen van de gebruikte tools om onderling samen te werken. Hierbij kan gebruik worden gemaakt van middelen die beschikbaar worden gesteld door samenwerkende instituten uit heel Europa. Zo wordt voor onderzoekers de mogelijkheid geopend om een betere keuze te maken uit de beschikbare data en technieken om hun onderzoeksvragen te beantwoorden. Verschillende nationale initiatieven, waaronder CLARIN-NL en CLARIN-Vlaanderen, dragen inmiddels bij aan de realisatie van deze ideeën.

Naast CLARIN-NL wordt het project gefinancierd door het Vlaamse departement Economie, Wetenschap en Innovatie. In de afgelopen tien jaar was er al een nauwe samenwerking tussen Nederland en Vlaanderen op het vlak van taal- en spraaktechnologie. Die kreeg vorm in het project Corpus Gesproken Nederlands en daarna in het bilaterale STEVIN programma voor taal- en spraaktechnologie. Deze samenwerking wordt nu voortgezet in het TTNWW-project. Naast het toegankelijk maken van faciliteiten voor onderzoekers is een belangrijk doel om de in Nederland en Vlaanderen gangbare de facto standaarden voor dataformaten en interfaces tussen tools en protocollen te promoten en te toetsen aan Europese initiatieven zoals CLARIN.

Voor de tekstkant worden veel verschillende tools, waaronder voor alignering, automatische naamherkenning en coreferentie gecombineerd tot standaard workflows voor de beantwoording van onderzoeksvragen over historische teksten, forumdiscussies, vertaalde romans en archeologische data. Voor de spraakkant wordt gebruik gemaakt van spraakherkenners voor het

ontsluiten van gesproken data op basis van toegankelijke en doorzoekbare tijdsynchrone transcripties.

Deelnemers aan TTNWW zijn de universiteiten van Tilburg, Antwerpen, Groningen, Utrecht en Twente, de Katholieke Universiteit Leuven, de Radboud Universiteit, Hogeschool Gent, het Max Planck Instituut voor Psycholinguïstiek, het Instituut voor Nederlandse Lexicologie, het Huygens Instituut, het Katholiek Documentatie Centrum, het Katholiek Documentatie- en Onderzoekscentrum voor Religie, Cultuur en Samenleving en Aletta.

www.clarin.eu
http://taalunieversum.org/taal/technologie/stevin