WP3: Voortgangsverslag

Datum: 2015-11-15

Auteur: Jan Odijk, Sjef Barbiers

Wat doet WP3?

In WP3 wordt gewerkt aan metadata curatie, o.a. in nauwe samenwerking met Oostenrijk, met als doel de vindbaarheid en vooral de 'ontdekbaarheid' van data te verbeteren. Daarnaast heeft het verbeteren en uitbreiden van allerlei zoekengines prioriteit voor: Groningen (PaQu,) Utrecht (grETEL), INL (Webcelex en BlackLab) en het Meertens Instituut (MTAS). Naast het werk aan collectie-metadata is er aan de RU met name gewerkt aan de taalverrijkingssoftware Frog waar met name wordt ingezet op het trainbaar maken van Frog voor nieuwe talen. Verder is er een verzameling aan vocabularies gedefinieerd die gebruikt gaan worden in een pilot voor een semantische infrastructuur.

Voortgangsrapport CLARIAH WP 3: Taalkunde

Na de kick-off van 3 juni zijn er enerzijds verdere besprekingen geweest om de vereisten en specificaties helderder te krijgen, en is er op andere fronten al aan het eigenlijke werk begonnen.

Metadata

Er wordt gewerkt aan metadata curatie, o.a. in nauwe samenwerking met Oostenrijk, met als doel de vindbaarheid en vooral de 'ontdekbaarheid' van data te verbeteren. Het werk aan metadata voor software dat begonnen was in CLARIN-NL is weer opgepakt in Utrecht, en de metadata van de CLARIN-NL Portal voor software zal daar in geïntegreerd worden, de CLARIN NL portal zal verder ontwikkeld worden tot een van de publicatie platformen voor de metadata voor tools en corpora. Er wordt bij RUN gekeken naar het aantal en kwaliteit van de corpus c.q. collectie metadata die op dit moment beschikbaar is en er zal een plan worden opgesteld voor het verbeteren daarvan. Door het Meertens Instituut wordt gewerkt aan een (voorlopig nog NL-only) alternatief voor de Virtual Language Observatory en verder wordt de CMDI2RDF brug die CMDI metadata beschikbaar maakt als RDF verder gestabiliseerd.

Zoekengines

Er wordt gewerkt aan het verbeteren en uitbreiden van allerlei zoekengines. Zo is bijv. in Groningen gewerkt aan een uitbreiding van PaQu met de mogelijkheid zoekresultaten te groeperen op basis van grammaticale data in combinatie met metadata. In Utrecht wordt gewerkt aan een upload-faciliteit voor GrETEL. De Webcelex engine is geporteerd naar het INL. Voor meer informatie en een link naar de applicatie (zie hier). Op het Meertens Instituut werkt men aan een nieuwe zoekengine, MTAS genaamd, waarmee men de hoeveelheid te doorzoeken data significant mee hoopt te kunnen opschalen. En het INL werkt verder aan de BlackLab search engine.

Taalkundige Verrijking

Naast het werk aan collectie-metadata (zie punt metadata) is er aan RU vooral gewerkt aan Frog. De dependency parser in Frog, een efficiënte emulatie van de Alpino Parser, is herschreven naar snellere C++-code. De morfologische analyse-module MBMA in Frog heeft diverse inhoudelijke verbeteringen ondergaan. Er wordt nu gewerkt aan het trainbaar maken van Frog voor nieuwe talen; als test is er een Frog-variant voor het Oud Grieks ontwikkeld in samenwerking met het 'Unravelling the language of perspective' ERC Advanced Grant project van Corien Bary (RU). Onderhoud en voorbereidend

werk is verricht aan FoLiA, CLAM, en (in samenwerking met Tilburg University en WP2) aan TICCL en PICCL. Met De Taalmonsters wordt gewerkt aan de verdere ontwikkeling van Whitelab, momenteel nog voor het CLARIN-NL OpenCGN-project, dat voortbouwt op OpenSONAR.

Semantische lexicons (VU).

We hebben een verzameling aan vocabularies gedefinieerd die gebruikt gaan worden in een pilot voor een semantische infrastructuur. Het gaat hierbij om: Brouwers, Open Dutch Wordnet, Thesaurus Cultureel Erfgoed, en plantennamen in dialecten. Voor deze vocabularies wordt een RDF en OpenSkos formaat gedefinieerd zodat regionale en diachrone informatie kan worden toegevoegd. Daarna worden de vocabularies naar dit formaat geconverteerd en worden ze via het Laundromat platform van de VU gepubliceerd in the LOD. In januari 2016 zal worden gedemonstreerd hoe de verzameling van resources kan worden bevraagd dmv Sparql.

Entiteit detectie en linking (VU).

Een begin is gemaakt met het beschikbaar stellen van een entity detectie en linking module die gebruikt maakt van de entiteiten databases uit WP2. Deze module wordt aanpasbaar gemaakt zodat zogenaamde ‘dark entities’ kunnen worden toegevoegd. Dit is nodig om het aanpasbaar te maken aan historische teksten.