DCS Curatieplannen 2013

Curatie dialectwoordenboeken

Ons is duidelijk geworden dat:

- WLD en WBD binnen COAVA zijn gebruikt voor een demonstratieproject maar niet als zodanig in een standaard CLARIN formaat beschikbaar zijn. De data zijn nu beschikbaar in een SQL-database en voor drie velden in LMF. LMF is wel een standaardformaat
- Nicoline van der Sijs heeft ons gewezen op een grote hoeveelheid dialectwoordenboeken die zij heeft geïnventariseerd en over een inspanning die gaande is om een standaardnederlandse trefwoorden toe te voegen. De DCS zou bij de curatie van deze woordenboeken een belangrijke rol kunnen spelen en het toevoegen van die standaardnederlandse vorm kunnen vereenvoudigen door alvast een automatisch voorstel te doen vanuit de bestaande datavelden.

De DCS zal een uniform LMF-formaat en CMDI-profiel voor dialectale woordenboeken maken en een aantal van de door Van der Sijs genoemde woordenboeken in overleg met R. van Hout en de makers gaan cureren.

Planning:
WGD: maart en april
WBD/WLD: mei en juni
Overig: daarna

CLAVAS organisatienamen voor OpenSKOS

Het CLAVAS-project (CLARIN-NL) wil een overtuigende demonstrator bouwen waarin een aantal elementen in een bruikbare lijst met uniforme namen aan gebruikers wordt gepresenteerd. Die lijsten moeten in OpenSkos worden aangeboden en meer specifiek voor CLARIN-NL ook in Arbil. In dat verband moet er een uniforme lijst van organisatienamen komen. Daarvoor gaan we in eerste instantie uit van de organisatienamen zoals ze nu in de VLO staan (CLARIN's Virtual Language Observatory). Het gaat om organisaties die data(banken) aan de VLO hebben bijgedragen. Later kunnen andere bronnen met organisatienamen worden toegevoegd.
In die VLO-lijst staan dezelfde organisaties in verschillende varianten door elkaar. Voor een deel gaat het om spellingsvarianten, voor een ander deel om semantische varianten van

dezelfde organisatie. Deze moeten alle tot 1 standaardvariant per organisatie worden teruggebracht.

Beoogd CLARIN Centrum: Meertens

Planning: maart en april

Boeken van het Centrum voor Parlementaire Geschiedenis (CPG)

In samenwerking met het CPG worden diverse belangrijke publicaties van het CPG gecureerd. Het gaat om de volgende werken:

1 Jaarboeken van de Nederlandse parlementaire geschiedenis (periode 1999-2009): 11 delen, totaal ca. 2200 pagina's (140 bijdragen) Serie boeken (6 delen) over de Nederlandse kabinetten (in de periode 1945-1956):

- *Het kabinet-Schermerhorn-Drees (1945-1946), 756 pagina's*
- *Het kabinet-Beel (1946-1948), 6 volumes, 4422 pagina's*
- *Het kabinet-Drees -Van Schaik (1948-1951), 3 volumes, 2585 pagina's*
- *Het kabinet-Drees II (1951-1952), 817 pagina's*
- *Het kabinet Drees III (1952-1956), 632 pagina's*
- *Het kabinet Drees IV en het kabinet-Beel (1956-1959), 371 pagina's*

3. Biografieën:

- *Prof.dr. G.M.J. Veldkamp, Herinneringen 1952-1967. Le carnaval des animaux politiques, 229 pagina's*
- *Architect van onderwijsvernieuwing. Denken en daden van Gerrit Bolkestein 1871-1956, 279 pagina's*

De pdf's van deze werken zijn te zien op http://www.ru.nl/cpg/onderzoek/online-cpg/

Curatie omvat het cureren van de text en de metadata.Met de uitgevers is een regeling getroffen v.w.b. de IPR en de werken mogen openbaar gemaakt worden. De boeken staan al online in pdf-formaat bij het CPG en kunnen nu in het kader van CLARIN-NL breder toegankelijk gemaakt worden.
Beoogd CLARIN Centrum: DANS
Betrokken partners: De curatie wordt uitgevoerd door de DCS in samenwerking met het CPG en DANS. Voor de inzet van UCTO en NERD wordt Martin Reynaert (Tilburg University) geraadpleegd (Maarten van Gompel is uitvoerende voor de DCS)..

Planning: in mei wordt geprobeerd of de conversie vanuit pdf zinvol is. Als dat positief uitpakt zal de curatie naar verwachting van juni tm sep. gaan duren.

Data van het Traces of Contacts project

Het betreft hier data die voortkomen uit deelprojecten van het Traces of Contacts project (Muysken; zie: http://www.ru.nl/linc/projects/erc-traces-contact/) Het gaat hierbij om

- Multilingual Netherlands (betrokken onderzoekers: Linda van Meel, Pablo Irizarri van Suchtelen, Suzanne Aalberse, Margot van den Berg
- Multilingual processing (betrokken onderzoeker: Gerrit-Jan Kootstra)
- Suriname
- South America

Deze deelprojecten zijn recentelijk afgerond of staan op het punt afgerond te worden. Dit is dan ook een optimaal moment om de curatie ter hand te nemen.

De werkzaamheden/acties die door de DCS uitgevoerd zullen worden zijn de volgende:
- het regelen van de IPR; er zal een MPI-licentie worden opgesteld die …; de licenties moet vervolgens ondertekend worden door proefpersonen. Daarnaast moet er een gebruikerslicentie worden opgesteld waarin is vastgelegd onder welke voorwarden en voor welke doeleinden de data gebruikt mogen worden (vertaling van dataleverantielicentie naar gebruikerslicentie)
- het DBD-metadataprofiel moet worden aangepast zodat de aanvullende metadata die de nieuwe data met zich meebrengen kunnen worden geaccommodeerd.
-alle data en aanvullende documentatie/informatie moet worden verzameld en overgedragen aan de DCS. De DCS onderhoudt hierover contact met de betrokken onderzoekers.
- voor elk van de datacollecties wordt een curatieplan geschreven.

Planning: tweede helft 2013

Overige

Nadat de DCS eerder was geattendeerd op het LUCL dat over een reeks van resources zou beschikken die wellicht voor curatie door de DCS in aanmerking zouden komen, heeft de DCS bij herhaling gepoogd deze te achterhalen. Aanvankelijk door contact te leggen met Marjan Klahmer. Nadat duidelijk was geworden dat met haar curatiewerkzaamheden in een door CLARIN in Call 3 gehononeerd project in de curatie van alle data van haar en directe collega's (o.a. Mous) voorzien was, hebben we vervolgens op haar suggestie geprobeerd contact te leggen met Ton van Haaften. Dit heeft niet tot resultaat geleid.

De DCS bekijkt met Onno Crasborn de mogelijkheden voor curatie van de transcripten die horen bij het videomateriaal dat reeds CLARIN-conform werd gearchiveerd. Het gaat hierbij om rijke transcripten op papier bij de dataset van Heleen Bos die gericht was op morfosyntactisch onderzoek naar gebarentaal en de dataset van Beppie van den Bogaerde en Anne Baker met kindertaalopnames. Deze twee dataverzamelingen met Nederlandse Gebarentaal vormen samen het enige materiaal buiten het Corpus NGT dat algemeen beschikbaar is voor andere onderzoekers, en daarmee ook het oudst beschikbare materiaal. Eerdere opnames zijn ofwel verloren gegaan ofwel niet gedocumenteerd. De dataset van Bos omvat in totaal 20 uur video. Voor 10 uren hiervan bestaat een handmatige transcriptie (ca. 2000 transcriptiebladen). Voor de

dataset van Van den Bogaerde & Baker (80 video) is voor ca. 20 uren een transcript (ca. 4000 transcriptiebladen) beschikbaar. Doel van de curatie zou zijn om de glos- en vertalingtiers (Bos) resp. de glos- en spraaktiers (Van den Bogaerde & Baker) te converteren naar EAF-files.

Verder overweegt de DCS curatie van de Database Streng n.a.v. het in Call 4 ingediende, maar niet-gehonoreerde LION voorstel (CLARIN-NL-12-017. Het IAP suggereerde hier dat het curatiedeel van het project wellicht door de DCS zou kunnen worden uitgevoerd. De DCS legt hiervoor contact met Ton van Kalmthout en het Huygens ING.

De DCS zal daarnaast nagaan welke van de eerder geïnventariseerde kandidaten concrete mogelijkheden bieden voor curatie op korte termijn.