DK-CLARIN Statusrapport

Aflevering nr. D 2.6.2a

Arbejdspakke, nummer og navn	WP2.6 Parallelt flersprogskorpus
Aflevering, navn	”2 mill. words treated”
Afleveringsfrist	T18
Afleveringsdato	T18 ~30. juni 2009
Projektleder for arbejdspakken	Lene Offersgaard
Deltagere	Claus Povlsen, Dorte Haltrup Hansen
Evt. tilhørende dokumenter på hjemmesiden	

Status
Denne statusrapport omhandler task 6 i arbejdsplanen for WP2.6: ”Collection and treatment of 2 mill words, based on experience from task 3 and 4. No structural alignment done yet. Collaboration with WP5 about metadata, formats and exchange of data“.

Der er taget beslutning om at teksterne der behandles i task 6 primært kommer fra følgende kilder:

- Acquis communautaire, som er betegnelsen for EU’s samlede regelværk. Det vil sige alt lige fra traktater til direktiver, Domstolens retspraksis, erklæringer og internationale aftaler. Aftale ok.
- Søren Kirkegaard Centret. Aftale ok for to SK-journaler på dansk. I gang med at indhente aftale for tilsvarende tyske og engelske tekster.
- Firmaers og institutioners årsberetninger.

De parallelle tekster indsamles primært for sprogene dansk, engelsk og tysk, og det tilstræbes at en dansk tekst så vidt muligt både har en parallel tekst på engelsk og på tysk.

Teksterne formateret til DK-CLARIN WP2 basisformat som er baseret på TEI P5-standarden. Dette format er dog ikke endeligt fastlagt endnu, og ingen tekster er derfor kommet igennem konverteringen til dette format. WP2.6 bidrager til arbejdet med at fastlægge formatet og forventer at formatet kan fastlægges i løbet af august, da ferier allerede nu vanskeliggør kommunikation om de sidste detaljer.

Arbejdet ang. produktion af metadata for WP2.6 tekster er igangsat, baseret på det fælles arbejde i WP2 på dette område. Der er arbejdet med format for aligneringer og forskellige formater har været overvejet, bl.a. TEI-P5, XCESv1 og XCESv2, disse formater har svært ved at dække behovet for aligneringer på en tilstrækkelig fleksibel måde, men det endelige format vil være et xml-format.

Der er således endnu ingen tekster der har det endelige format og er kommet igennem den behandling der hører til task 6. Dette skyldes bl.a. at det endelige DK-CLARIN tekstformat for WP2 ikke er endeligt fastlagt og at task 4 ikke er afsluttet, se evt. Statusrapport D2.6.2.b.

Det forventes at task 6 vil være afsluttet inden T21 ~ 30. Sep. 2009, men er afhængig af at DK-CLARIN WP2 basisformat bliver endeligt fastlagt inden T20.