DK-CLARIN Statusrapport

Aflevering nr. D 2.6.2b

Arbejdspakke, nummer og navn	WP2.6 Parallelt flersprogskorpus
Aflevering, navn	”Tools for alignment, annotation, specification of markup”
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 4 og 5 i arbejdsplanen for WP2.6. Task 4: Pilot study for 0.5 mill words: format issues and tokenization. Given specification of two use cases: test quality of first version of pos- lemma- and entity annotation. Test of sentence and word alignment methods. Study of structural alignment methods” Task 5: “Search possibilities in multilingual resources based on simple CQP interface or DK-CLARIN interface if available.”

Der er etableret et pilot-korpus der består af tekster fra Acquis communautaire og årsberetninger. Der er implementeret konvertere fra Acquis-formatet til en foreløbig version af DK-CLARIN WP2 basisformat, men denne skal tilpasses når basisformatet er endeligt fastlagt. Format-konverteringsflow fra pdf-format til DK-CLARIN WP2 basisformat er overvejet. Som ’use cases’ er i pilotfasen valgt *komparativ sproganalyse* og *maskinoversættelse*, se specifikationsrapporten for beskrivelse af disse ’use cases’.

Der er foretaget en analyse af visse eksisterende pos-taggere for dansk, tysk og engelsk, herunder af egnetheden af deres tagsæt. Der søges stadig en bedre egnet tyske tagger end den undersøgte. Der skal ske en tilpasning af alle de værktøjer, der skal bruges sådan at input kan indlæses i DK-CLARIN WP2-basisformatet, denne tilpasning sker for dansk pos-tagger og for lemmatisere i samarbejde med WP5 og WP2.2 og er påbegyndt. Aligneringsmetoder er afprøvet, kvalitetsmålemetoder er overvejet, men endnu ikke formuleret. Aligneringsmetoderne afprøves både for uddrag af Acquis-korpusset (som har en aligneringsreference, der kan bruges som sammenligningsgrundlag) og for andre tekster.

Arbejdet med at teste og udvælge værktøjer til behandling af teksterne er ikke kommet så lang som planlagt. Dette skyldes både at der er brugt væsentlig færre tidsressourcer, end der var afsat til arbejdet i første halvår af 2009 og at arbejdet med fastlæggelse af formater samt tilpasning af værktøjer til xml-input og til basisformatet har vist sig at være en større opgave end forventet.

Søgefaciliteter i parallelle tekster som er nødvendig til test og til komparativ sproganalyse må forventes at blive løst med et tilpasset CQP-interface, da der ikke synes at være interesse for fælles tekst- og annoteringssøgefaciliteter i projektet.

Det forventes at task 4 og 5 vil være afsluttet inden T21 ~ 30. Sep. 2009, men er afhængig af at DK-CLARIN WP2 basisformat bliver endeligt fastlagt inden T20.