Referencekorpus for dansk: Grundlæggende beslutninger og arbejdsplan

DK-CLARIN WP 2.1-arbejdspapir
Jørg Asmussen og Jakob Halskov

Version 0.9 – 9. oktober 2008
OBS! Oplysninger om resurseforbrug for de enkelte arbejdsopgaver mangler i denne version!

Resumé

Nærværende papir indeholder en status for DK-CLARIN WP 2.1 *Referencekorpus for dansk* ved milepæl T 9. Det indeholder endvidere en arbejdsplan og resurseopgørelse for det videre projektforløb. Arbejdsplanen beskriver de kvartalsvise målsætninger efter T 9 (T 12 – T 36) for projektet.

1 Status

1.1 Grundlæggende beslutninger

Ved milepæl T 9 er der af de medvirkende i projektet, Jørg Asmussen (DSL) og Jakob Halskov (DSN), blevet truffet følgende grundlæggende beslutninger for projektet.

Annotationer på tekstniveau: Der anvendes DSL's etablerede inventar, som er udarbejdet i afdelingen for Digitale Ordbøger og Tekstkorpora (DOT) i forbindelse med *ordnet*-projektet i henhold til [Asmussen, 2008a], og som er dokumenteret i [Asmussen, 2008c].

Annotationer på tokenniveau: Her anvendes ligeledes DSL's allerede etablerede basale inventar, jf. dokumentationen [Asmussen, 2008b]. Tagsættet for POS-taggingen fastlægges dog på et senere tidspunkt.

Tekstflow og opbevaring af korpusmateriale: DSL's eksisterende tekstbankkoncept anvendes, jf. beskrivelsen i [Asmussen, 2008b], dog mangler en evaluering af, hvorvidt DSL's MySQL-baserede tekstbank-applikation skal anvendes eller en anden (XML-baseret) model.

Leveringsformat: De dele af korpus, som måtte være ophavsretligt clearede, vil kunne leveres i et TEI-konformt format, selvom formatet sandsynligvis vil være et andet under den projektinterne processering.

Ophavsret: WP 2.1 betragter det ikke som deres primære opgave at føre principielle forhandlinger om rettighedsspørgmål med tekstleverandørerne og

henstiller derfor til styregruppen og den overordnede projektledelse (WP 1) at anvise en fremgangsmåde, idet det er WP 2.1's opfattelse, at der bør arbejdes henimod en grundlæggende, generel aftale, som omfatter hele DK-CLARIN. Kan der ikke opnås en generel aftale, bør styregruppen anvise en generel rettighedspolitik for hele DK-CLARIN. Indtil da indsamles tekster i overensstemmelse med allerede etableret praksis udelukkende som citerbare tekster, dvs. tekster, der kun kan vises i uddrag, og som ikke kan videredistribueres.

Tekstleverandører: Både DSL og DSN indsamler løbende tekster fra InfoMedia. En fælles tekstregistrant er taget i anvendelse for at undgå tekstdubletter i korpusset. Deruover indsamler DSN i første omgang blog- og forummateriale, mens DSL prøver at supplere med forlagsmateriale. Den oprindeligt planlagte indsamling via *netarkivet.dk* viser sig at være både teknisk og juridisk problematisk, hvorfor den er stillet i bero.

Konkordansværktøj: DSL/JA stiller korpusserveren PyCOCS til rådighed som konkordansværktøj. Der skal dog udvikles en egnet (web-baseret) grænseflade, alternativt kunne man måske få rekonfigureret KorpusDK's eksisterende grænseflade.

DK-CLARIN-samarbejde: WP 2.1 tilstræber et tæt samarbejde med WP 2.2 (fagsprogligt korpus), så redundans i udviklingsarbejdet kan begrænses mest muligt.

1.2 Hidtil udførte opgaver

Grundlæggende beslutninger for projektet blev truffet, de organisatoriske rammer afstukket og en foreløbig arbejdsplan blev udarbejdet.

Tekstregistrant for InfoMedia-tekster blev etableret.

Transducer for InfoMedia-tekster blev udviklet.

Indsamling af materiale fra InfoMedia samt blog- og forumtekster blev påbegyndt.

Potentielle tekstkilder som *netarkivet.dk* og *Wikipedia* blev evalueret.

1.3 Forbrugte resurser

Institution	Kommentar	PM
DSL/JA		0,33
DSL/TT	Transducer-udvikling	0,67
DSN	Afventer tal fra JH	

2 Arbejdsplan

2.1 Resurser

WP 2.1 råder over 1,25 mio. kr. Heraf er 20% (250.000 kr.) institutionel medfinansiering. DSL's andel er 70%, DSN's 30%. Ét årsværk sættes til 450.000 kr.[1] og 215 dage, idet der regnes med 30 feriedage og 8 dage til andet fravær (fx skiftende helligdage, sygdom) per år. Én arbejdsdag sættes til 7,4 arbejdstimer, hvorfra der trækkes 0,5 times frokostpause, hvorefter én arbejdsdag består af 6,9 netto-arbejdstimer. Ét årsværk svarer således til 1483 netto-arbejdstimer. Projektet råder over 2,77 årsværk svarende til godt 33 personmåneder (PM), én PM svarer til 123 netto-arbejdstimer hhv. 17,8 arbejdsdage. DSL bidrager med 23 PM og DSN med 10 PM.

Oveni lønudgifter er der afsat 75.000 kr. til udstyr, hvoraf den institutionelle egenandel udgør 20% (15.000 kr.). DSL's og DSN's andele af denne post er 50% hver.

2.2 Løbende indsamlingsarbejde

Under hele projektforløbet indsamles der løbende tekstmateriale, som behandles automatisk, så det dels kan lægges i en tekstbank, dels siden kan indgå i selve referencekorpusset med tekst- og POS-annotation. Der ses bort fra en ressourcetung manuel processering af tekstmaterialet. Dette kan betyde, at annotationer på tekst- og tokenniveau kan være af skiftende præcision.

2.3 Enkeltstående opgaver

T 12: udgangen af 4. kvartal 2008

Opgave T 12.1: Tekstleverandørregistrant

Beskrivelse: Der etableres en registrant over aktive og potentielle fremtidige tekstleverandører, gerne som integreret del af tekstbanken

Aflevering: Dokumentation af registranten

Opgave T 12.2: Tekstregistrant

Beskrivelse: DSL og DSN registrerer deres InfoMedia-tekster i den fælles InfoMedia-tekstregistrant

Aflevering: Dokumentation af registranten

Opgave T 12.3: Tokenizer

Beskrivelse: DSL's tokenizer evalueres på et udvalg af InfoMedia-tekster. Evt. småjusteringer foretages

Aflevering: Tokenizer-kildekode med dokumentation

[1] Dette er et skøn. En præcis udregning kan muligvis give en mindre afvigelse i forhold til dette tal. En del af arbejdet vil i princippet kunne udføres af (programmeringskyndig) studentermedhjælp. I det omfang der projektorganisatorisk kan allokeres studentermedhjælpsressurser, vil man kunne opnå en besparelse. Denne skal dog afvejes med de resurser, der i givet fald skal bruges til rekruttering og indføring i arbejdet, samt risikoen for, at en medhjælp kan vise sig at være ustabil.

Opgave T 12.4: Tekstbanksystem

Beskrivelse: Der træffes beslutning vedrørende det tekstbanksystem, som skal anvendes til projektintern håndtering af tekstmaterialet. Valget står mellem DSL's MySQL-baserede eller et andet (XML-baseret)

Aflevering: Skriftlig redgørelse for den trufne beslutning

T 15: udgangen af 1. kvartal 2009

Opgave T 15.1: Tekstbanksystem

Beskrivelse: Projektinternt tekstbanksystem incl. brugerinterface klar til ibrugtagning

Aflevering: Dokumentation af tekstbanksystemet

Opgave T 15.2: Processering af InfoMedia-tekster

Beskrivelse: Ophobede InfoMedia-tekster tokeniseres og lægges ind i tekstbanken

Aflevering: Kvantitativ afrapportering

T 18: udgangen af 2. kvartal 2009

Opgave T 18.1: Transducere

Beskrivelse: Transducere udviklet til alle aktive leveranceformater

Aflevering: Dokumentation

Opgave T 18.2: Processering af øvrige tekster

Beskrivelse: Ophobede tekster lægges ind i tekstbanken

Aflevering: Kvantitativ afrapportering

T 21: udgangen af 3. kvartal 2009

Opgave T 21.1: Fuldformsleksikon

Beskrivelse: Fuldformsleksikon klar til ibrugtagning

Aflevering: Dokumentation. Hvorvidt selve leksikonnet kan stilles til rådighed for CLARIN, afhænger af, hvordan det tilvejebringes, og hvilke rettigheder der knytter sig til det

T 24: udgangen af 4. kvartal 2009

Opgave T 24.1.: POS-tagger

Beskrivelse: POS-tagger klar til ibrugtagning

Aflevering: Dokumentation. Hvorvidt selve taggeren kan stilles til rådighed for CLARIN, afhænger af, hvordan den tilvejebringes, og hvilke rettigheder der knytter sig til den

T 27: udgangen af 1. kvartal 2010

Opgave T 27.1: TEI-transducer

Beskrivelse: Formattransducer intern-til-TEI klar til ibrugtagning

Aflevering: Dokumentation

Opgave T 27.2: Downloadservice

Beskrivelse: Downloadmulighed af ophavsretligt clearede eller scramblede tekster etableret

Aflevering: Dokumentation

Da det er uvist, hvorvidt WP 2.1 umiddelbart vil kunne integreres i den infrastrukturløsning, som WP 5.1 skal tilvejebringe, gås der her ud fra, at korpusset kan hostes hos DSL (downloadservice, konkordansværktøj, selve korpusset), men dog tilgås via WP 5.1*s infrastrukturløsning.

T 30: udgangen af 2. kvartal 2010

Opgave T 30.1: Konkordansværktøj

Beskrivelse: Webbaseret konkordansværktøj klar til brugertest

Aflevering: Dokumentation

Opgave T 30.2: Brugerpanel

Beskrivelse: Brugerpanel nedsættes

Aflevering: Rapport

T 33: udgangen af 3. kvartal 2010

Opgave T 33.1: Brugertest

Beskrivelse: Brugertest af konkordansværktøj afsluttet

Aflevering: Rapport

T 36: udgangen af 4. kvartal 2010

Opgave T 36.1: Konkordansværktøj

Beskrivelse: Endelig version af konkordansværktøj

Aflevering: Dokumentation

Opgave T 36.2: Korpus

Beskrivelse: Endelig version af korpus gøres tilgængelig

Aflevering: Dokumentation

Litteratur

[Asmussen, 2008a] Asmussen, J. (2008a). DOT's Sprogteknologiske Drejebog. Udviklingsopgaver i forbindelse med *ordnet*-projektet. Rapport 4, Det Danske Sprog- og Litteraturselskab.

[Asmussen, 2008b] Asmussen, J. (2008b). Udviklingsopgave 1.5: Fastlæggelse og dokumentation af korpusformat og beskrivelse af tekstflowet under korpusopbygningen. Rapport 2.2, Det Danske Sprog- og Litteraturselskab.

[Asmussen, 2008c] Asmussen, J. (2008c). Udviklingsopgave 1.7: Fastlæggelse og dokumentation af headerstruktur. Rapport 1.1, Det Danske Sprog- og Litteraturselskab.