Referencekorpus for dansk: T 9-status og arbejdsplan

DK-CLARIN WP 2.1-arbejdspapir

Jørg Asmussen og Jakob Halskov

Version 1.2 – 21. oktober 2008

Tidligere versioner:

Version 1.1 – 17. oktober 2008

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.

I version 1.2 er oplysninger om det hidtidige resurseforbrug blevet bekræftet og en enkelt slåfejl rettet.

1 Status

1.1 Grundlæggende beslutninger

Ved milepæl T 9 er der af de indtil da 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, jf. opgavebeskrivelse for WP 1 i ansøgningen. Kan der ikke opnås en generel aftale, bør styregruppen eller WP 1 snarest 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. Derudover 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	Møder, administration	0,33
DSL/TT	Transducer-udvikling	0,67
DSN/JH	Møder, tekstindsamling	0,75

2 Arbejdsplan

2.1 Generelt

Arbejdsplanen opererer med en betydelig finere milepæl-opdeling, end projektbeskrivelsen for DK-CLARIN lægger op til. Der er tale om et bevidst valg for at sikre en bedre kontinuitet i projektet.

Resurse-dimensioneringen er forsøgt udregnet så minutiøst som muligt, da de medvirkende også er involveret i andre projekter og en præcis resurseforbrugsafregning derfor er påkrævet. Senere justeringer kan dog ikke udelukkes.

2.2 Resurser

WP 2.1 råder over 1,25 mio. kr. Heraf er 20% (250.000 kr.) institutionel medfinansiering. DSL's andel er 70% (875.000 kr.), DSN's 30% (375.000 kr.). Et DSL-årsværk sættes til 562.000 kr.[1], mens et DSN-årsværk sættes til 450.000 kr.[2]. Et årsværk består af 215 arbejdsdage, 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 1,56 DSL-årsværk svarende til godt 18,5 personmåneder (PM) og 0,83 DSN-årsværk svarende til 10 PM. Én PM svarer til 123 netto-arbejdstimer hhv. 17,8 arbejdsdage. I alt råder projektet over 28,5 personmåneder.

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.

2.3 Administration

Der vil løbende blive brugt resurser til projektadministrative gøremål, som udarbejdelse og opfølgning af arbejdsplaner, afholdelse af statusmøder og koordinering med andre projekter i DK-/EU-CLARIN-regi.

Til administration allokeres følgende resurser for resten af projektets løbetid: DSL/JA: 0,75 PM, DSN: 0,25 PM.

2.4 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 resursetung manuel processering af tekstmaterialet. Dette kan betyde, at annotationer på tekst- og tokenniveau kan være af skiftende præcision.

[1] Seniorredaktør på højeste løntrin i 2009. Oplysningen er indhentet fra DSL's bogholderi.

[2] Dette er et skøn. En præcis udregning kan muligvis give en mindre afvigelse i forhold til dette tal. En del af arbejdet (både DSL-delen og DSN-delen) vil i princippet kunne udføres af (programmeringskyndig) studentermedhjælp. I det omfang der projektorganisatorisk kan allokeres studentermedhjælpsresurser, 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.

Til tekstakvisitionen allokeres følgende resurser for resten af projektets løbetid: DSL: 2,0 PM (JA: 0,5; TT: 1,5), DSN: 2,75 PM.

2.5 Enkeltstående opgaver

T 12: udgangen af 4. kvartal 2008

Opgave 1: Tekstregistrant

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

Aflevering: Dokumentation af registranten

Resurseforbrug: DSL/JA: 0,25 PM

Opgave 2: Tokenizer

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

Aflevering: Tokenizer-kildekode med dokumentation

Resurseforbrug: DSL/JA: 0,25 PM, DSN: 0,25 PM

Opgave 3: 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 redegørelse for den trufne beslutning

Resurseforbrug: DSL/JA: 0,25 PM, DSN: 0,25 PM

T 15: udgangen af 1. kvartal 2009

Opgave 4: Tekstleverandørregistrant

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

Aflevering: Dokumentation af registranten

Resurseforbrug: DSL/JA: 0,5 PM, DSN: 0,25 PM

T 18: udgangen af 2. kvartal 2009

Opgave 5: Tekstbanksystem

Beskrivelse: Projektinternt tekstbanksystem incl. brugerinterface klar til ibrugtagning

Aflevering: Dokumentation af tekstbanksystemet

Resurseforbrug: DSL/JA: 3,25 PM, DSN: 0,25 PM

T 21: udgangen af 3. kvartal 2009

Opgave 6: Processering af InfoMedia-tekster

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

Aflevering: Kvantitativ afrapportering

Resurseforbrug: DSL/TT: 0,25 PM, DSN: 0,25 PM

Opgave 7: Transducere

Beskrivelse: Transducere udviklet til alle aktive leveranceformater. Denne opgave udføres i samspil med WP 2.2.

Aflevering: Dokumentation

Resurseforbrug: DSL/TT: 1,0 PM, DSN: 1,0 PM

Opgave 8: Processering af øvrige tekster

Beskrivelse: Ophobede tekster lægges ind i tekstbanken

Aflevering: Kvantitativ afrapportering

Resurseforbrug: DSL/TT: 0,25 PM, DSN: 0,25 PM

T 24: udgangen af 4. kvartal 2009

Opgave 9: 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

Resurseforbrug: DSL/JA: 2,0 PM, DSN: 0,25 PM

Opgave 10: Lemmatizer

Beskrivelse: Lemmatizer klar til ibrugtagning

Aflevering: Dokumentation. Hvorvidt selve lemmatizeren kan stilles til rådighed for CLARIN, afhænger af, hvordan den tilvejebringes, og hvilke rettigheder der knytter sig til den. Det tilstræbes dog, at den bliver alment tilgængelig

Resurseforbrug: DSL/JA: 2,0 PM, DSN: 0,25 PM

T 27: udgangen af 1. kvartal 2010

Opgave 11: 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

Resurseforbrug: DSL/JA: 2,0 PM, DSN: 1,0 PM

Opgave 12: Downloadservice – **OBS! UDGÅR!**

Beskrivelse: Downloadmulighed af ophavsretligt clearede eller scramblede tekster etableret

Aflevering: Dokumentation

Resurseforbrug: Ingen tilstrækkelige resurser i denne arbejdspakke!

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 (som dog ikke kan etableres i WP 2.1), konkordansværktøj, selve korpusset), men dog tilgås via WP 5.1's infrastrukturløsning.

T 30: udgangen af 2. kvartal 2010

Opgave 13: TEI-transducer

Beskrivelse: Formattransducer intern-til-TEI klar til ibrugtagning

Aflevering: Dokumentation

Resurseforbrug: DSL/TT: 0,25 PM, DSN: 1,0 PM

Opgave 14: Konkordansværktøj

Beskrivelse: Webbaseret konkordansværktøj klar til brugertest

Aflevering: Dokumentation

Resurseforbrug: DSL/NHS: 1,0 PM, DSN: 0,25 PM

Opgave 15: Testbrugerpanel

Beskrivelse: Testbrugerpanel nedsættes

Aflevering: Rapport

Resurseforbrug: DSL/TT: 0,25 PM, DSN: 0,25 PM

T 33: udgangen af 3. kvartal 2010

Opgave 16: Brugertest

Beskrivelse: Brugertest af konkordansværktøj afsluttet

Aflevering: Rapport

Resurseforbrug: DSL/TT: 0,25 PM, DSN: 0,25 PM

T 36: udgangen af 4. kvartal 2010

Opgave 17: Konkordansværktøj

Beskrivelse: Endelig version af konkordansværktøj

Aflevering: Dokumentation

Resurseforbrug: DSL/NHS: 0,5 PM, DSN: 0,25 PM

Opgave 18: Korpus

Beskrivelse: Endelig version af korpus gøres tilgængelig

Aflevering: Dokumentation

Resurseforbrug: DSL/TT: 0,5 PM, DSN: 0,25 PM

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.