Referencekorpus for dansk: T 9-status og arbejdsplan

DK-CLARIN WP 2.1-arbejdspapir
Jørg Asmussen med input fra Jakob Halskov m.fl.

Version 1.5 – 20. januar 2008

Tidligere versioner:

Version 1.4 – 1. december 2008

Version 1.3 – 22. oktober

Version 1.2 – 21. oktober 2008

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.

Version 1.5 indeholder en fodnote til beslutningen om annotationer på tekstniveau i afsnit 1.1, således at det fremgår, hvordan den hænger sammen med projektansøgningen. Korpusindsamlingsarbejdet beskrevet i afsnit 2.4 har fået tilføjet en række kvantitative målsætninger svarende til det, der er beskrevet i ansøgningen. Endelig er der tilføjet et afsnit om kvalitetssikring: 3.

Version 1.4 indeholder visse præciseringer som resultat af skriftligt feedback fra Hanne Fersøe samt typeangivelser for afleveringerne i henhold til typologien under `http://cst.dk/dk-clarin/?q=node/95`.

I **version 1.3** er en misforståelse, om hvorvidt budgettallene for arbejdspakkerne inkluderer institutionel egenfinansiering eller ej, rettet. Hidtidige versioner gik ud fra, at den ikke var inkluderet, tilbagemeldinger fra styregruppemødet den 21.10.2008 viser, at de allerede *er* inkluderet. Dette giver anledning til en revison af resursefordelingen på de forskellige delopgaver. Endvidere er der – ligeledes på baggrund af feedback fra styregruppemødet – foretaget nogle formuleringsmæssige præciseringer. Endelig er lønkvantificeringen af et DSN-årsværk blevet bekræftet.

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 WP 2.1.

Annotationer på tekstniveau: Der tages udgangspunkt i 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]. Det tilstræbes at udtrykke annotationerne vha. TEI P5-specifikationerne.[1]

Annotationer på tokenniveau: Der tages udgangspunkt i DSL/DOT's etablerede token-koncept, jf. dokumentationen [Asmussen, 2008b]. Tag-inventar fastlægges på et senere tidspunkt.

Tekstflow og opbevaring af korpusmateriale: Der anvendes en tekstbank-orienteret fremgangsmåde, jf. beskrivelsen i [Asmussen, 2008b], dog mangler en evaluering af, hvorvidt der skal satses på MySQL-baseret tekstbank-applikation eller XML-baseret model.

Leveringsformat: De dele af korpus, som måtte være ophavsretligt clearede, vil kunne leveres i et TEI-konformt format, selvom formatet måske 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, som er udviklet i tilknytning til OpenCWB-projektet. Der skal dog udvikles en egnet (web-baseret) grænseflade (WP 5.1?), 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.

[1] I ansøgningen går denne arbejdsopgave under betegnelsen *Ontology of text types, genres. XML-based annotation scheme.*

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 1,00 mio. kr til aflønning. Heraf er 20% institutionel medfinansiering. DSL's andel er 70% (700.000 kr.), DSN's 30% (300.000 kr.). Et DSL-årsværk sættes til 562.000 kr.[2], mens et DSN-årsværk sættes til 450.000 kr.[3]. 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 således over

- 1,25 DSL-årsværk svarende til ca. 15 personmåneder (PM)
- 0,67 DSN-årsværk svarende til ca. 8 PM.

Én PM svarer til 123 netto-arbejdstimer hhv. 17,8 arbejdsdage. I alt råder projektet over 23 personmåneder.

Oveni lønudgifter er der afsat 60.000 kr. til udstyr, hvoraf den institutionelle egenandel ligeledes udgør 20%. DSL's og DSN's andele af denne post er 50% hver.

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

[3] 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.

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: 0,50 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.

Til tekstakvisitionen allokeres følgende resurser for resten af projektets løbetid: DSL/TT: 1,25 PM, DSN: 3,00 PM.

Der tilstræbes følgende kvantitaive målsætninger for indsamlingen:

T 18: 15 mio. lbd. ord

T 28: 20 mio. lbd. ord

T 36: 45 mio. lbd. ord

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

Type: Projektrapport

Resurseforbrug: DSL/JA: 0,125 PM

Opgave 2: Tokenizer

Beskrivelse: Tokenizer testes på et udvalg af InfoMedia-tekster. Evt. småjusteringer foretages

Aflevering: Tokenizer-kildekode med dokumentation

Type: Resurse (åbent[4] værktøj) + dokumentation

Resurseforbrug: DSL/JA: 0,125 PM, DSN: 0,25 PM

Opgave 3: Tekstbanksystem

[4] *Åben* betyder, at kildekoden (eller anden resurse) er åbent tilgængelig. Open source-løsninger foretrækkes alt andet lige. Det gælder både programmel og data.

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

Aflevering: Skriftlig redegørelse for den trufne beslutning

Type: Projektrapport

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

Type: Projektrapport

Resurseforbrug: DSL/JA: 0,25 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

Type: Service[5] + dokumentation

Resurseforbrug: DSL/JA: 3,0 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

Type: Statusrapport

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: Resurser (åbne værktøjer) + dokumentation

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

Opgave 8: Processering af øvrige tekster

[5] Typen *service* er ikke forudset i typologien, men skønnes nødvendig. En service er en it-baseret tjeneste der stilles til rådighed for (dele af) CLARIN. Mange af WP 5's afleveringer vil være services.

Beskrivelse: Ophobede tekster lægges ind i tekstbanken

Aflevering: Statusrapport

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. En fuldt offentlig løsning vil alt andet lige blive favoriseret

Type: Resurse (åben resurse) + dokumentation

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. En fuldt offentlig løsning vil alt andet lige blive favoriseret

Type: Resurse (åbent værktøj) + dokumentation

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. En fuldt offentlig løsning vil alt andet lige blive favoriseret

Type: Resurse (åbent værktøj) + dokumentation

Resurseforbrug: DSL/JA: 2,0 PM, DSN: 0,5 PM

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

Beskrivelse: Downloadmulighed af ophavsretligt clearede eller scramblede tekster etableret

Aflevering: Dokumentation

Type: Service

Resurseforbrug: Ingen tilstrækkelige resurser i denne arbejdspakke!

I denne plan gås der ud fra, at korpusset kan hostes hos DSL på eksisterende maskinel, hvor det som minimum vil kunne tilgås af et web-baseret konkordansværktøj. Det bør også kunne tilgås (som web-service) 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
Type: Statusrapport
Resurseforbrug: DSL/TT: 0,50 PM, DSN: 0,5 PM

Opgave 14: Konkordansværktøj

Beskrivelse: Webbaseret konkordansværktøj klar til brugertest
Aflevering: Dokumentation
Type: Dokumentation
Resurseforbrug: DSL/NHS: 0,50 PM, DSN: 0,25 PM

Opgave 15: Testbrugerpanel – **OBS! UDGÅR!**

Beskrivelse: Testbrugerpanel nedsættes
Aflevering: Rapport
Type: Statusrapport
Resurseforbrug: Ingen tilstrækkelige resurser!

T 33: udgangen af 3. kvartal 2010

Opgave 16: Brugertest – **OBS! UDGÅR!**

Beskrivelse: Brugertest af konkordansværktøj afsluttet
Aflevering: Rapport
Type: Statusrapport
Resurseforbrug: Ingen tilstrækkelige resurser!

T 36: udgangen af 4. kvartal 2010

Opgave 17: Konkordansværktøj

Beskrivelse: Endelig version af konkordansværktøj
Aflevering: Dokumentation
Type: Service eller resurse (åbent værktøj)
Resurseforbrug: DSL/NHS: 0,25 PM

Opgave 18: Korpus

Beskrivelse: Endelig version af korpus gøres tilgængelig
Aflevering: Dokumentation
Type: Resurse (tekst)
Resurseforbrug: DSL/TT: 0,25 PM, DSN: 0,25 PM

3 Kvalitetssikring

Arbejdet vil blive udført efter de bedste internationale, praktisk gennemførlige korpuslingvistiske standarder, der skal sikre, at det resulterende korpus' kvalitet fuldt ud svarer til *state of the art* inden for feltet.

En egentlig formaliseret kvalitetskontrol er ikke forudset i planen, da de forhåndenværende resurser er utilstrækkelige. En sådan kontrol kan evt. gennemføres som et særskilt projekt, efter at WP 2.1 er afsluttet.

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, Det Danske Sprog- og Litteraturselskab.

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