Referencekorpus for dansk: Aktuel arbejdsplan

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

Seneste opdatering: 28. januar 2011

Resumé

Nærværende papir indeholder en resurseopgørelse for projektforløbet for DK-CLARIN WP 2.1 *Referencekorpus for dansk*, og det henviser til arbejdsplanen på projektets wiki `http://clarin.dsl.dk`. Papiret giver desuden nogle oplysninger om kvalitetssikring, og det afsluttes med en statusrapport for projektet ved milepæl T 9.

1 Arbejdsplan

1.1 Generelt

Den detaljerede og til enhver tid opdaterede plan over enkelte arbejdsopgaver [1] og afleveringer findes på projekt-wikien `http://clarin.dsl.dk`. Denne plan 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. Justeringer undervejs kan dog ikke udelukkes.

1.2 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] `http://clarin.dsl.dk/wiki/clarin/doku.php?id=deliverables`

[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.

- 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.

1.3 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.

1.4 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 token-niveau kan være af skiftende præcision.

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

Der tilstræbes følgende kvantitative målsætninger for indsamlingen:

T 18: 15 mio. løbende ord

T 28: 20 mio. løbende ord

T 36: 45 mio. løbende ord

De aktuelle tal[4] fremgår af wikien.

1.5 Enkeltstående opgaver og afleveringer

En komplet liste[5] med statusoplysninger kan ses på wikien.

2 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.

[4] `http://clarin.dsl.dk/wiki/clarin/doku.php?id=text`
[5] `http://clarin.dsl.dk/wiki/clarin/doku.php?id=deliverables`

3 Status ved milepæl T 9

3.1 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. Det tilstræbes at udtrykke annotationerne vha. TEI P5-specifikationerne.[6]

Annotationer på token-niveau: Der tages udgangspunkt i DSL/DOT's etablerede token-koncept. Tag-inventar fastlægges på et senere tidspunkt.

Tekstflow og opbevaring af korpusmateriale: Der anvendes en tekstbank-orienteret fremgangsmåde, hvor materialet samles i en særlig database, hvorfra der siden kan udtrækkes et korpus efter nærmere specifikationer. Det skal undersøges, 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 konfigureret 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.

[6] I ansøgningen går denne arbejdsopgave under den misvisende betegnelse *Ontology of text types, genres. XML-based annotation scheme.*

3.2 Udførte opgaver op til T 9

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.

3.3 Forbrugte resurser op til T 9

Institution	Kommentar	PM
DSL/JA	Møder, administration	0,33
DSL/TT	Transducer-udvikling	0,67
DSN/JH	Møder, tekstindsamling	0,75