

Forbedrede annotationer og søgemuligheder for ældre litterær tekst: Arbejdsplan for projektperioden T9 – T36

DK-CLARIN WP 2.4-arbejdspapir
Jørg Asmussen og Stefan Iversen

Version 1.3 – 1. december 2008

Tidligere versioner:

Version 1.2 – 22. oktober 2008

Version 1.1 – 17. oktober 2008

Resumé

Dette dokument indeholder en arbejdsplan for DK-CLARIN WP 2.4 *Ældre litterær tekst* for det videre projektforsøg T9 – T36.

Version 1.3 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 1.2** er DSL-andelen justeret ned, idet egenfinansieringen allerede var med i budgetbeløbene og ikke skulle lægges oveni.

1 Status

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

2 Arbejdsplan

2.1 Arbejdsfordeling

Hovedformålet med JVJ-Centrets del af denne arbejdsplan er at digitalisere en stor del af Jensens forfatterskab med henblik på opmærkning efter TEI P5-standard. 50 længere bogværker forventes digitaliseret. Efter forhandling af rettigheder købes selve digitaliseringen hos KB. Arbejdet med udvælgelsen af og kontrollen med standen af Jensen-tekster vil blive udført af Stefan Iversen, AU med assistance fra Per Dahl, AU.

KB (ved NN) og DSL (ved TT) sørger for konvertering af ADL og JVJ-materialet til det format, som specificeres af WP 5.1 samt (forbedring af) annotation med metadata på teksts niveau. CST (ved NN) udfører lemmatisering og POS-tagging.

Hovedformålet med DSL's del af denne arbejdsplan er at skitsere et prototypisk ortografisk variantleksikon, der i en fuld implementation vil kunne forbedre

søgemulighederne i ældre tekst samt lemmatisering og POS-tagging af ældre tekst.

DSL står endvidere for administration af denne arbejdsplan.

2.2 Resurser

WP 2.4 råder over 0,75 mio. kr i lønmidler. Heraf er 20% institutionel medfinansiering. KU-CST's andel er 50.000 kr., hvorefter der resterer 700.000 kr. fordelt på DSL og JVJ-Centret med 300.000 til hver og 100.000 til KB. CST råder over ca. 1 PM i denne arbejdsplan, som går til lemmatisering/tagging af tekstmaterialet. Antallet af personmåneder (PM) beregnes for DSL's vedkommende på baggrund af en DSL-seniorredaktørårløn på højeste anciennitetstrin, som i 2009 ligger på 562.000 kr. For de øvrige institutioner antages en gennemsnitlig årløn at ligge på 450.000 kr. Nettoarbejdstimetallet for en PM er 123, beregnet på samme måde som i WP 2.1, jf. [Asmussen og Halskov, 2008]. Herefter råder JVJ-Centret over 8 PM, hvoraf 6 udliciteres til KB, idet selve JVJ-digitaliseringen købes hos KB. Der er således allokeret 2 personmåneder arbejdstid fra JVJ-Centrets side samt 6 personmåneder arbejdstid, købt hos KB. KB råder desuden selv over 2,7 PM og DSL over 6,4 PM.

Oveni lønudgifter er der afsat 10.000 kr. til udstyr/drift for JVJ-Centret, hvoraf den institutionelle egenandel udgør 20%.

2.3 Opgaver

Opgave 1: Forhandling af rettigheder med Gyldendal og arvingerne

Resurseforbrug: 0,25 personmåneds arbejdstid fra JVJ-Centret

Aflevering 2.4.1: Kort rapport om afklaringen af rettighedsspørgsmålet.
T 10

Type: Statusrapport

Opgave 2: Digitalisering og opmærkning.

Resurseforbrug: 1,75 personmåneds arbejdstid fra JVJ-Centret og 6 personmåneder arbejdstid, købt hos KB til denne opgave, 1 personmåned fra CST til tekstopmærkning

Aflevering 2.4.2: Tekstsamling bestående af 50 værker af JVJ, digitaliseret og TEIP5-opmærket. T 18

Type: Statusrapport

Opgave 3: Annotering. Det nuværende annoteringsniveau dokumenteres samt ønskelige forbedringer specificeres og udføres i muligt (prototypisk) omfang

Resurseforbrug: Der afsættes 4 PM hertil (2 fra DSL/TT, 2 fra KB/NN)

Aflevering 2.4.3: Specifikation af ADL/JVJ-annotering med prototypiske eksempler. T 28

Type: Statusrapport

Opgave 4: Specifikation af variantleksikon

Resurseforbrug: 2,9 PM (DSL/JA)

Aflevering 2.4.5: Dokumentation og prototype. T 33

Type: Resurse (leksikon, prototype) + dokumentation

Opgave 5: Lemmatisering/tagging

Resurseforbrug: 1,0 PM (CST/NN)

Aflevering 2.4.5: Lemmatiseret/tagget resurse. T 33

Type: Resurse (tekst) + dokumentation

Opgave 6: Integrering. ADL og JVJ integreres i WP 5.1-portal

Resurseforbrug: Der afsættes 1,2 PM hertil (DSL/TT: 0,6; KB/NN: 0,6)

Aflevering: Samling af konverterede ADL- og JVJ-tekster klar til indlæsning i DK-CLARIN's søgesystem. T 33

Type: Statusrapport

Opgave 7: Projektadministration

Resurseforbrug: 0,9 PM (DSL/JA)

Aflevering 2.4.6: Afsluttende projektrapport. T 36

Type: Statusrapport

Litteratur

[Asmussen og Halskov, 2008] Asmussen, J. og Halskov, J. (2008). Reference-korpus for dansk: T 9-status og arbejdsplan. Rapport, DSL/DSN.