# Status for WP3.2-søgemaskinen

En del af målet for WP3.2. ”Tool box for the treatment of spoken materials for va­rious types of research” er at etablere en søgemaskine, der kan søge på tværs af tiers og interviews ud fra et sæt af Praat .TextGrid-filer.

I samspil med workpackagens anden leverance, konverteringsprogrammer kan søgemaskinens input altså komme fra Transcriber™, CLAN™ og Praat™.

Øvrige ’spoken language analysis’-formater, f.eks. ELAN, vil kunne tilføjes med endnu et konverteringsværktøj.

Søgemaskinen er udviklet som en Java Servlet version 2.4 (Java Enterprise Edition 5 biblioteket anvendes) i en såkaldt 3-lags arkitektur, hvor data er adskilt fra hhv. søge og præsentationslogin og kræver en ’Java Servlet Container’ af en art, f.eks. Apache TomCat for at kunne afvikles.

Brugergrænsefladen er web-baseret og for det meste genereret af Java Server Pages (JSP’er) version 1.2.

Funktionaliteten inkluderer et import og vedligeholdsmodul til data-delen.

Brugergrænsefladen kan ses på næste side (Figur 1) og åbner mulighed for søgning:

* i alle, enkelte eller udvalgte sæt af databasens samlede indhold af interviews
* på alle, enkelte eller udvalgte sæt af informanter
* på hele ord, dele af ord og enhver søgeforespørgsel, der kan formuleres vha. regulære udtryk
* på et enkelt søgeudtryk kan kun ske indenfor et valgt tier, men der kan søges efter (andre) forekomster i andre tiers ved at trykke på en Tilføj Tier-knap
* Søgning kan også lede efter forekomster der forekommer et tilvalgt antal ’slices’, typisk ordpositioner såfremt Ortografi-tieret er valgt som primært søge-tier før elle efter det primære søgeudtryk – vi kalder dette for ’kontekst søgning’.

Der åbnes også mulighed for at vælge forskellige visninger af et potentielt søgeresultat

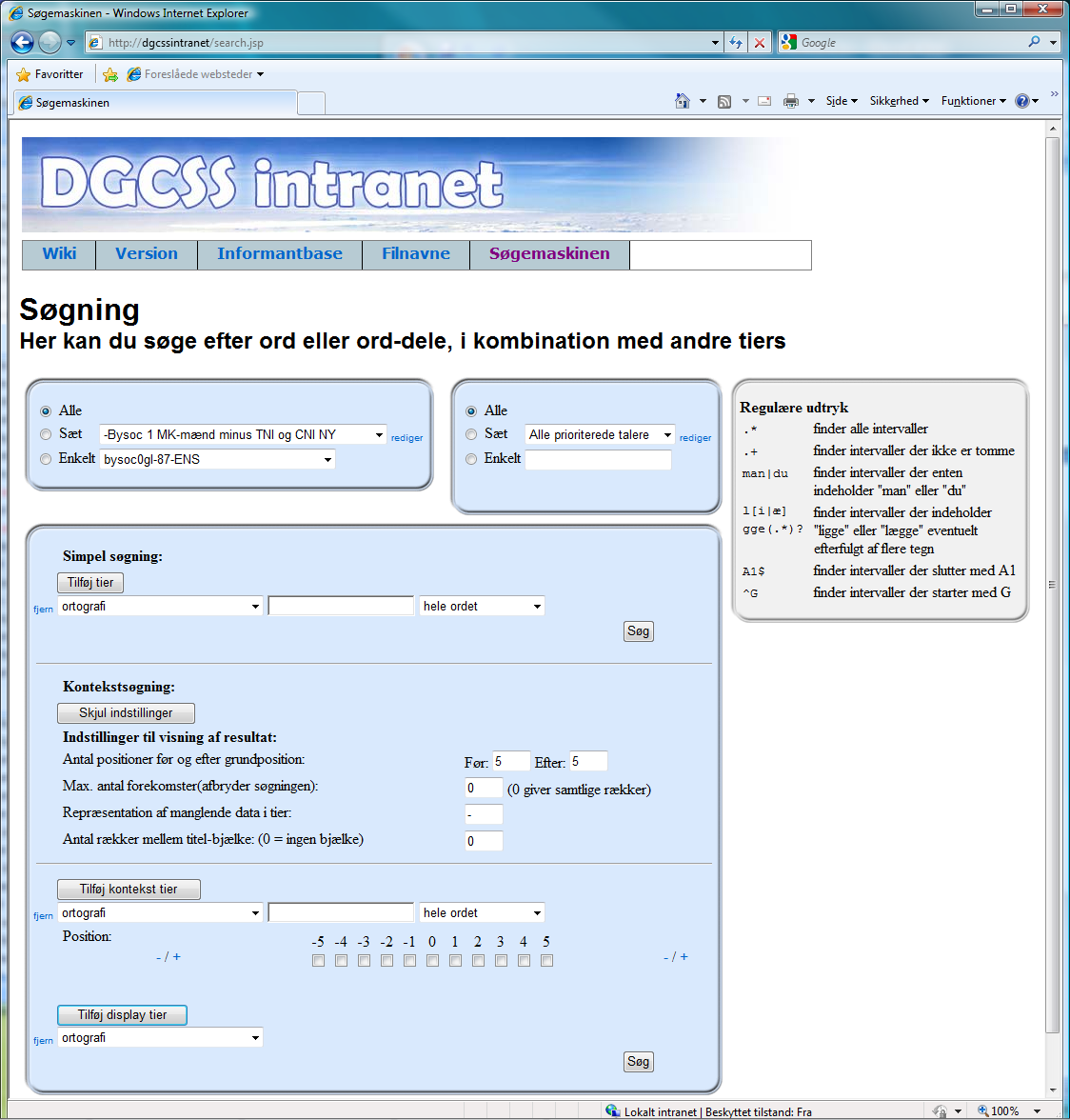
* Enkeltsøgninger kan vises som en simpel liste over, hvilke interviews, der indeholder søgeudtrykket
* som KWIC-konkordanser ved at vælge et eller flere ’Display’ Tiers
* og som CSV-filer (der skal downloades først) hvor alle ’slices’, der opfylder søgekriteriet listes sammen med interviewets metadata samt indholdet af alle slices ved samme position i interviewet.
* Ved kontekstsøgninger kan man angive hvor mange ’slices’ til hver side af det primære søgeresultat, man vil have vist og man har også her mulighed for at få vist et vilkårligt antal tiers indhold side om side med resultatet

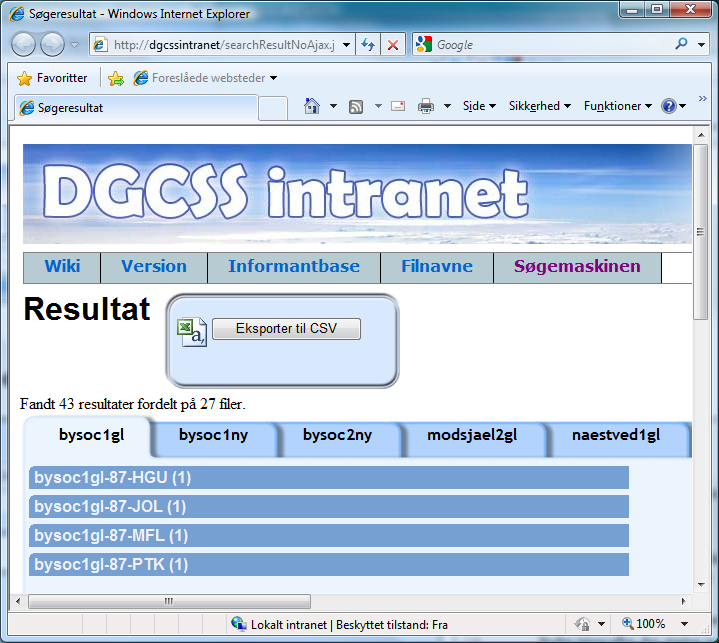
Søgemaskinen er i drift og fungerer. Der rettes løbende på småfejl efterhånden som de dukker op, men må i det store og hele betegnes som færdig.

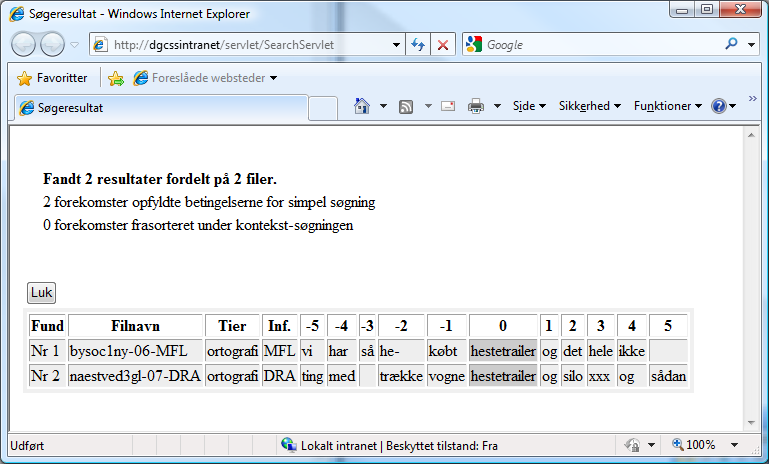
Der er ikke udført nogen formel kvalitetskontrol/afleveringsforretning, men dog et større code-review for ca. et år siden.   
Det skal også bemærkes, at søgemaskinen \*er\* testet ’i drift’ således at brugerne (forskerne) løbende har udført kvalitetskontroller i og med at de har haft holdt søgemaskinens resultater op imod indholdet af .TextGrid-filerne.   
Der er ikke i den forbindelse fundet fejl, der vil kræve store omkodninger eller ændringer i programmet i øvrigt.

Bemærk dog, at ingen dele af systemet er lavet som en webservice og således ikke uden yderligere kodning/programmering kan integreres med DK Clarin’s portal i øvrigt. Integration med DK Clarin kan ikke ske førend DK Clarin’s WP5 leverer snitfladespecifikationer til deres webservice-arkitektur. Der er afleveret uddybende notat (Bilag 1) herom til WP5 v. Lene Offersgaard mandag d. 18.1.2010.

Figur : Søgemaskinens brugergrænseflade



Figur : Simpel resultatvisning med mulighed for dl af csv-fil

Figur : KWIC-visning af resultat

Figur 4: Indhold af CSV-fil for resultat

"filename";"project";"recording";"recordingyear";"participant";"gender";"birthyear";"socialclass";"xmin";"Aktivitetstype";"grammatik";"ANmarkering";"phonetic";"ANRmarkering";"events";"parole\_PoSRED";"DschwaDmarkering";"ordstil";"Samtaletype";"AMmarkering";"ENGmarkering";"lydskrift";"uncertain transcription";"Genre";"emphasis";"Case";"gramma\_II";"W-markering";"IIV SMUL Kommentarer";"parole\_PoS";"SH";"IIV AIG Kommentarer";"Udsigelse";"AJmarkering";"Interaktionsstruktur";"Makro-sproghandling";"generisk";"ortografi";"comments";"translation";"global events"

"bysoc1ny-06-MFL";"bysoc1";"ny";"2006";"MFL";"m";"1959";"MK";"1617,857022";"Asa";"";"";"";"";"";"hestetrailer/N";"";"";"Sieu";"";"";"hestetrailer/h2Esd0tLEJlC";"";"Ggr";"";"";"";"";"";"hestetrailer/NCCPU==I";"";"";"";"";"I8";"Mvi";"";"hestetrailer";"";;

"naestved3gl-07-DRA";"naestved3";"gl";"2007";"DRA";"m";"1991";"MK";"1403,469";"Asa";"";"";"";"";"";"hestetrailer/N";"";;"Sieu";"";"";"hestetrailer/h2Esd0tLEJlC";"";"";;;"";"";"";"hestetrailer/NCCPU==I";"";;"";"";"";"Mvi";"";"hestetrailer";"";"";""

Bilag

1. Notat vedr. leverance af inputintegreret søgemaskineværktøj til DK CLARIN

DGCSS, KUA 12.01.2010/MBR

Notat

## Vedr. leverance af inputintegreret søgemaskineværktøj til DK CLARIN

Der spørges til, om DGCSS kan levere ”Søgemaskine i eksportabelt format” til DK Clarin i T30.

For at svare fyldestgørende på dette spørgsmål , er det nødvendigt at opridse situationen lidt mere detaljeret.

DGCSS’ søgemaskine er udviklet som en Java Servlet der tilgår en MySQL database og kommunikerer med en bruger vha. en JSP/Javascript frontend:



P.t. afvikles denne arkitektur på en Linux/Tomcat/Apache-platform, men i teorien vil en hvilken som helst platform med en MySQL og en Java Application Server kunne servicere funktionaliteten.

Der arbejdes i øvrigt på DGCSS med planer, prototyper og tests med henblik på at migrere til en Windows/IIS/.Net -platform.

Som søgemaskinen har været omtalt i DK Clarin-regi, er den tænkt som et ’input-integreret værktøj’, dvs. den skal modtage sit input fra DK Clarin-infrastrukturen, mens output i form af søgeresultater ikke er tænkt skal have nogen plads i DK Clarins repositorie.

Udvikling kan således først påbegyndes når DGCSS modtager informationer, f.eks. i form af WSDL/WADL-dokumenter, om hvordan DK Clarin vil implementere sin:

* Leveringsservice (2.8.4 i ”Clarin kravspec 051009.doc”)

og

* Service til søgning i metadata (ibid. 2.8.2).

Leveringsservicen skal bruges til at kunne trække tekstressourcer ind i søgemaskinen, mens servicen til søgning i metadata skal bruges til at kunne finde frem til de tekstressourcer og samlinger af samme, brugeren ønsker at søge i.

Dertil skal der træffes beslutninger om

* hvilket afviklingsmiljø DK Clarin kan stille til rådighed. Det skal bemærkes at DGCSS foretrækker .Net-platformen til afvikling, da det er her DGCSS råder over den største ekspertise, men Java kan også gøre det
* om hvorvidt DK Clarins grænseflade til metadatasøgning kan og skal integreres direkte med søgemaskinefunktionaliteten, så det ikke bliver nødvendigt for DGCSS at udvirke sin egen udgave af metadatasøgninger
* om hvorvidt sådanne generede søgemaskiner skal gemmes (egtl. registreres vha. et metadataobjekt) i DK Clarin Repositoriet til senere brug og i givet fald, hvordan en sådan funktionalitet skal integreres i DK Clarins brugerflade og arkitektur i øvrigt

Mest kritisk for at deadline kan overholdes er altså at få webservicespecifikationerne stillet til rådighed, for uden dem kan arbejdet ikke startes[[1]](#footnote-1).

DGCSS kan så dog naturligvis umiddelbart levere en søgemaskine som ekstern ressource til DK-Clarin, der så kun kommer med søgemulighed i DGCSS’ egen leverance af tekstressourcer.

Bemærk at dette faktisk betyder, at såfremt det dynamisk at skabe en søgemaskine over tekstressourcer fundet vha. søgning i metadata fravælges, så kan DGCSS levere (nogenlunde) lige så mange søgemaskiner som eksterne ressourcer, som man måtte ønske.   
Hver sådan søgemaskine kunne så tilbyde søgninger i en én-gang-for-alle udtrukket samling af tekstressourcer og linkes til fra DK Clarins web-brugergrænseflade.

f. DGCSS,   
Michael Barner-Rasmussen  
IT Ansvarlig  
DGCSS, Nordisk Forskningsinstitut (NFI)  
Njalsgade 136, lok. 27.5.61  
2300 København S  
Tlf. Institut: 35 32 83 82  
Tlf. Direkte: 31 25 31 49

1. Billedligt talt, kan situationen altså beskrives sådan at hvis de nævnte webservices er postkasser, så kan DGCSS, som i dette billede er postbudet, ikke levere post til Clarin førend Clarin har sat sine postkasser op. [↑](#footnote-ref-1)