DIGHUMLAB

Brugerundersøgelse – tilgængelige data og brugerbehov

Claus Povlsen og Lina Henriksen

Center for sprogteknologi, Københavns Universitet

Juli 2013

Indhold

Undersøgelsens opgave

Vinklen på denne undersøgelse er dels at få kortlagt blandt forskere på danske universiteter og forskningsinstitutioner hvilket forskningsmateriale (sprogdata og sprogteknologi) der allerede findes og er klar til at blive deponeret i Clarin-portalen, og dels hvilke forskningsbehov hvad angår funktionalitet, der kan afdækkes blandt forskerne i forbindelse med indsamling af nye empiriske data som en del af deres forskningspraksis.

Undersøgelsens fremgangsmåde har været at indkalde alle interesserede til en række møder på de danske uddannelsesinstitutioner. Møderækken (jf. bilag 2) indledtes på Københavns Universitet i juni måned 2012 hvor der på baggrund af en målrettet henvendelse til forskerene på det humanistiske fakultets institutter blev afholdt 3 møder. Derefter blev der efter lignende målrettede henvendelser afholdt møder på Ålborg Universitet, Syddansk Universitet og Aarhus Universitet[1].

Mødedeltagere

En bred vifte af institutter og fagområder var repræsenteret ved møderne. Som eksempler kan nævnes Engerom (KU), INSS (KU), MEF (KU), Design og Kommunikation (SDU), Filosofi (SDU), Institut for Æstetik og Kommunikation – Lingvistik, AAU.

Det betyder at der samlet set har været mange forskellige brugerprofiler repræsenteret ved møderne, både fagligt set og med hensyn til fortroligheden med anvendelse af it-værktøjer i en arbejdsproces. Korte referater af de fleste møder findes som bilag 4.

Mødernes struktur

Strukturen på de enkelte møder var, at repræsentanter for Center for Sprogteknologi (CST) indledte med et oplæg om DIGHUMLAB-projektet efterfulgt af en gennemgang af en første version af den infrastrukturelle platform, clarin.dk udviklet i det danske Clarin-projekt, jf. http://dkclarin.ku.dk/. Målet var at give tilhørerne et indblik i DIGHUMLAB projektets muligheder for at igangsætte eller være behjælpelig med visse initiativer samt give forskerne et overblik over platformens funktionalitet og endelig inspirere dem til at anvende platformen og/eller til at få nye ideer til platformens anvendelsesmuligheder. Derfor indeholdt vores demonstration også en del forskellige eksempler på arbejdsopgaver som nemt og hurtigt løses med værktøjer i Clarin (se Bilag 3 angående beskrivelse af et eksempel på en use case).

Dernæst blev ordet givet til de fremmødte. Vores ønske var at indhente information om dels mængden af deponeringsklare forskningsdata dels om de potentielle brugeres mening om funktionaliteten implementeret i den demonstrerede platform og deres ideer til hvilke nye funktionalitet de kunne ønske sig.

Tilbagemeldinger fra møderne

Deponeringsklare forskningsdata

Det viste sig hurtigt at mængden af data umiddelbart klargjort til import i Clarin-portalen var forsvindende lille. Der er mange årsager hertil, men primært drejer det sig tilsyneladende om at forskningsdata er vanskelige at skaffe pga. ophavsrettigheder, og hvis en forsker endelig har fået stillet nogle data til

[1] Det blev i den sammenhæng udformet en skabelon som værtinstitutionerne gjorde brug af jf. bilag 1.

rådighed, så er forskeren forsigtig med at videregive disse data (selv til forskningsformål!) fordi det er uklart i hvilket omfang man må gøre dem offentligt tilgængelige. Endvidere gav nogle forskere også udtryk for at de ikke arbejder med data i større mængder.

Det overordnede indtryk var dog at de fleste forskere arbejder med data i en eller anden udstrækning og sommetider kunne det være ønskeligt at arbejde med store mængder data, men selv når det er muligt at indhente data så er det stadig vanskeligt at håndtere data, bl.a. fordi afklaring af rettigheder for brug er uklare.

Andre forskningsdata

Parallelt med den foretagne mødevirksomhed blev der tilbudt den mulighed for de fremmødte forskere, at DIGHUMLAB-projektet i form af teknisk bistand kunne yde assistance i forbindelse med klargøring af data til deponering Clarin-platformen. Et integreret element i denne opgave var at forskerne også på denne måde kunne give input til en ændret eller udvidet funktionalitet Clarin-platformen. Resultatet af dette initiativ kan ses i bilag 5 hvor forskellige mulige brugerprojekter er beskrevet. Disse projekter vil blive analyseret med henblik på at fastslå hvor meget arbejde der kræves for at materialet kan gøres tilgængeligt. Planen er at udvælge tre brugerprojekter til videre behandling og deponering.

Tilbagemeldinger om Clarin

En generel erfaring fra den gennemførte møderække har været, at forudsætningerne blandt mødedeltagerne har været meget forskellige. Nogle har som deltagere og bidragydere i det ovenfor nævnte CLARIN-projekt både kompentencemæssigt og videnmæssigt stor indsigt i korpuslingvistik og/eller brug af digitale platforme. I den anden ende af skalaen kunne man observere forskere som fandt det naturligt som empirisk belæg at bruge den introspektive metode. Hele spektret har med andre ord været repræsenteret på møderne hvilket gennemgangen af resultaterne nedenfor også afspejler.

Erfarne brugere involveret i CLARIN-projektet og dermed med et grundigt kendskab til clarin.dk, kunne derfor foruden generelle betragtninger komme med tilbagemeldinger af mere detaljeret art. Andre fremmødte deltagere der blev introduceret for clarin.dk for første gang og som ikke var vante med at bruge begreber som lemmatisering og POS-opmærkning, havde svært ved at kommentere Clarin-platformens forskellige funktionaliteter, men havde i stedet meget generelle kommentarer som i hovedsagen handlede om afdækning af hvad Clarin-platformen egentlig er og hvad de forskellige værktøjer kan bruges til. Det blev på den måde klart at både platformens enkeltdele samt platformens overordnede formål ikke er tilstrækkelig tydeliggjort for uerfarne brugere på nuværende tidspunkt. Hverken i platformens design eller på anden vis.

En særlig gruppe blandt de uerfarne Clarin-brugere er datalingvisterne. Altså brugere med kendskab til begrebsapparatet omkring sproglig opmærkning af tekster, men kun med ganske lidt eller ingen erfaring i brugen af Clarin-portalen. Denne gruppe meldte tilbage at de trods en forståelse for alle de tilgængelige værktøjers formål og funktionalitet, ikke havde fundet løsninger på de opmærkningsopgaver de havde stillet, og at de syntes det var overordentlig svært at finde resurser selvom de vidste resurserne fandtes. Strukturen bag brugen af værktøjerne forekom uforståelig og svært gennemskuelig.

Opsummering af ønsker til funktionalitet

For både erfarne og uerfarne var der enighed om behov for følgende forbedringer til Clarin-portalen:

Deponering af data: Som nævnt ovenfor er køen af forskere der ønsker at få lagt forskningsdata ind på portalen ikke lang. Hvis der i tilgift kræves en specifik viden og kunnen for at en bruger kan deponere sine data, vil køen blive endnu kortere. Der bør derfor tilbydes (semi)automatiske værktøjer der guider en forsker brugervenligt og gnidningsfrit gennem en deponeringsprocedure. Forskere har som ovenfor nævnt behov for tilgang til flere data og derfor vil en interaktiv (semiautomatisk) deponeringsprocedure højst tænkeligt udgøre en særdeles stor forbedring af Clarin som i sidste ende kan være afgørende.

Overskuelighed/gennemsighed: Det er vanskeligt at finde ud af hvordan en søgningsopgave skal gribes an. Mange kom ikke længere end til at trykke på *Find.* Med så store og forskelligartede data i Clarin-portalen er det afgørende at brugeren har forskellige muligheder for at få vist de data som portalen indeholder. Der skal tages højde for at en bruger sommetider skal finde en helt specifik resurse, og sommetider snarere er interesseret i at få et overblik over hvad der findes inden for en specifik type af resurser. Stort set alle brugere gav udtryk for at semantikken bag de valgte metadatanavne bør gøres tydelig og forståelig.

Visning: Det skal være muligt samtidigt at få vist søgningsresultater ikke blot fra flere forskellige medier (ex billede og tekst) men også fra parallelle tekster, ex sætnings- eller afsnitsalignerede dokumenter, og at visionen om også kunne søge på billedmønstre skulle medtænkes i det nye design.

De kyndige clarin.dk-brugere pegede desuden på følgende:

Web-baseret brug af værktøj: De værktøjer der ligger i portalen skal også kunne bruges i selve portalen. Skal en bruger eksempelvis opmærke en tekst, bør dette kunne gøres i portalen således at brugeren ikke behøver at downloade samt installere værktøjet hos sig selv.

Baseret på de gjorte erfaringer jf. ovenfor så blev det hurtigt tydeligt at det af pædagogiske grunde ville være hensigtsmæssigt at give et eksempel på hvordan en forskningsportal som clarin.dk kunne berige ens forskningspraksis, både hvad angår effektivitet (besparelse af tid) og kvalitet (ny erkendelse).

Bilag 1

Skabelon i forbindelse med planlægning af møderækken

Præsentation af DIGHUMLAB – tema 1

Informationsmøde

<dato, tidspunkt, lokale>

Aarhus Universitet, Aalborg Universitet, Københavns Universitet og Syddansk Universitet er initiativtagere til DIGHUMLAB. Initiativet er støttet med midler fra den nationale infrastrukturpulje med henblik på at påbegynde opfyldelsen af det behov der er beskrevet i Dansk Roadmap for Forskningsinfrastruktur 2011 (p. 19): "Både inden for humaniora og samfundsvidenskaberne og på tværs af fagområder er der et øget behov for at sikre datatilgængelighed og databevaring i form af bedre, lettere og mere sikker adgang til registerdata og andre typer af ressourcer".

DIGHUMLAB bidrager til at understøtte forskerne i at forny humanistiske og samfundsvidenskabelige forskningsfelter ved at give bred adgang til digitale kilder og forskningsdata. Man behøver ikke længere vide at der er en bestemt forsker der har digitaliseret et bestemt værk, spørge om man må få adgang, opdage at det ligger på et medium man ikke kan læse eller er skrevet i et format man skal have hjælp til.

DIGHUMLAB skaber derved både nye og tværfaglige muligheder for forskningen inden for litteratur, lingvistik, sociologi, historie, medier, antropologi, arkæologi og journalistik for bare at nævne enkelte discipliner.

På mødet vil KU, Center for Sprogteknologi, indlede med en præsentation af DIGHUMLAB's tema 1: *Sprogbaserede materialer og værktøjer*. Præsentationen vil beskrive ideerne i den del af infrastrukturen og belyse hvordan forskere kan drage nytte af denne infrastruktur inden for humaniora og samfundsvidenskab (tema 1 bliver for tiden udmøntet i den europæiske infrastruktur CLARIN, se fx www.clarin.dk).

Vi inviterer alle universitetets forskere til at komme til mødet og bidrage med deres ønsker til indhold af infrastrukturen . Hvad skal der til for at man som forsker kan bruge infrastrukturen? Hvad skal der til for at man kan bruge den i sin undervisning? Hvilke data og værktøjer vil være mest relevante? Har I allerede noget som er digitaliseret, og som kunne lægges ind i systemet og derved bruges af andre forskere, samt af studerende?

Claus Povlsen, Lina Henriksen, Bente Maegaard

Bilag 2

Oversigt over mødevirksomhed

Kick-off-møde afholdt 8. juni, KUA, 2012

Møde for interesserede på KUA, 18., 21. og 29. juni 2012

Møde med Gyldendals ordbogsafdeling, 2012

Møde med repræsentant for Danske Taler, 2012

Møde i Aalborg, (AU), 13. december 2012, oplæg og diskussion,

Møde med Ulrik Sandberg Petersen, KUA

Møde i Aarhus (AAU) 23. januar 2013, oplæg til Kick-off-mødet

Møde i Kolding (SDU) 27. februar 2013, oplæg og diskussion

Bilag 3

Gesta Danorum – en use case

Eksemplet er evidens eller empiri til en kompositionsanalyse af Saxos *Gesta Danorum (GD)* også benævnt *Saxos Danmarkshistorie*.

Lige et par ord om værket: Værket beskriver i 16 bøger på latin tiden, der strækker sig fra kong Dan til Knud den 6. og afsluttes med vendernes undertvingelse i 1185.

Ud fra en traditionel tilgang er værket opdelt i to hovedsektioner, bøgerne 1-9 der omhandler den sagnhistoriske del (med blandt andet sagnet om Amled) og en historisk anden del der udgøres af bøgerne 10-16. En alternativ opdeling blev publiceret i 1969 i tidskriftet Mediaeval Scandinavia hvor Inge Skovgaard Petersen lancerer en ny opdeling af Gesta Danorum. Hun mener at kunne argumentere for at GD er opdelt i 4 gange 4 bøger, hvor første del (den hedenske) består af 8 bøger og en anden og kristen del bestående af bøgerne 9-16.

Diskussion af disse forskellige tolkninger har siden da været et vigtigt emne inden for Saxo-forskningen. Spørgsmålet er således er det bog 9 eller bog 10 der repræsenterer overgangen fra den hedenske til den kristne periode i værket?

Og vil det være muligt med en på samme tid sproglig og kvantitativ analysetilgang at finde vidnesbyrd der understøtter den ene eller den anden af disse opfattelser?

Vi har taget en digital version af en oversættelse af Saxos Gesta Danorum og strukturelt i teksten opmærket bøgernes placering i værket. Derefter har vi automatisk beriget værket med det vi kalder Part of Speech-annotering, som er opmærkning af de løbende ord i Saxos værk med information om ordklasse og bøjning. Endelig er resultatet lagt ind IMS Open Corpus Workbench © 1993–2006 by IMS Stuttgart .

Lagt ind på denne platform med opmærkning af POS-tagging og lemmatisering er det muligt på en præcis måde dels at formulere et søgemønster der præcist og nemt indfanger kristen sprogbrug i GD og dels at få fremvist søgeresultaterne for bog 8, 9 og 10 på en nem og overskuelig måde.

Antager man at relativt frekvent brug af ord fra et kristent register hænger tæt sammen med beskrivelsens genstand – her altså kristendommens fremkomst i GD – så understøtter undersøgelsens resultat en opdeling af GD i 2*8 bøger idet der relativt set er færre forekomster af kristent sprogbrug i bog 8 sammenlignet med bog 9 og 10, hvilket indikerer at bog 9 repræsenterer overgangen fra hedensk til kristen tematik i GD.

Bilag 4

Referater af møder

Resumé af DIGHUMLAB-møde afholdt 18. juni, KU

Deltagere: Claus Povlsen (ref). Dorthe Duncker, Hanne Ruus

I diskussionen blev der taget afsæt i den funktionalitet som er implementeret i DKCLARIN-portalen.

En generel bemærkning: Hvis DIGHUMLAB ender med en anden dokumentstruktur, så sørg for at der vil være en én-til-en-konversion mellem DK-CLARIN- skemaet og et evt. nyt format.

(1) Det skal være muligt at blive præsenteret for en oversigt over eller fortegnelse for de til enhver tid søgbare data i repositoriet. Hvis man ikke har et overblik over de tilgængelige data, hvordan kan man så forskningsmæssigt set drage nytte af datalageret?

Desuden skal der for hvert menupunkt (eksempelvis i metadata-menuen) være mulighed for at blive informeret om semantikken/indholdet bag det valgte menuelement (evt. ved inventariering af hvert af de forskellige meta-elementers indholdstyper, opdateret til repositoriets aktuelle (be)stand).

(2) Angående brugen af værktøjer: Det må være et brugerkrav at brugen af tilgængelige værktøjer er web-baseret, således at den givne bruger ikke konfronteres med de problemer der uvægerligt opstår når værktøjet skal installeres in-house hos brugeren. Med tanke på at målgruppen i denne sammenhæng er meget bred, kan man ikke generelt set forvente teknisk viden og kunnen til at udføre de krævede installationer.

(3) Visning af flere medier samtidigt i et skærmbillede

(4) Angående: Upload/deponering af egne data. Det bør være muligt at kunne bruge et værktøj (gerne web-baseret) der indlejrer brugerens data i det krævede format, ex TEIP5. Herunder at man får mulighed for at få hjælp når resursen skal beskrives i metadataopmærkningen.

Desuden skal det være muligt at beskytte de af egne resurser som ikke engang CST's vedligeholdelsesstab bør have tilgang til fx data som er belagt med fortrolighedsklausul, eller som indeholder personfølsomme oplysninger. Alternativt at den der uploader filer, har ret til at fjerne dem igen.

Kort referat af introduktionsmøde til DIGHUMLAB-projektet 21. juni 2012, KU

Tilstede: Lisbeth Holtse, Birgit Rønne (tilknyttet og ansat ved Nationalmuseet), Mads Poulsen (psykolingvistik), Henrik Hovmark (NFI – ømålsordbogen) og Claus Povlsen (ref).

CP indledte med en guided tur i clarin.dk

Ud fra en diskussion af de metadata der er fælles for alle medier repræsenteret i clarin.dk gav de to repræsentanter fra Nationalmuseet udtryk for et ønske om foruden den tekstbaserede søgning i data at der også var en mulighed for at søge efter billedmønstre. CP bemærkede at denne facilitet ville indebære at alt billedmaterialet skulle repræsenteres på en anden måde end den nuværende, men måske på længere sigt kunne en billedmotivsøgning være en ekstra funktionalitet i clarin.dk.

Henrik Hovmark gav i forbindelse med gennemgang clarin.dk's uploadsprocedure udtryk for utilfredshed med at grænsefladen kun muliggjorde overførsel af data indlejret i filer. Dette var ikke hensigtsmæssigt når man efterfølgende eksempelvis skulle søge i en ordbogsfil.

Henrik Hovmark så også vældig gerne at der kunne etableres link mellem en ordbogsindgang og en billedillustration af indgangens denotation.

Mads Poulsen havde som udgangspunkt i sin forskning*: Identifikation af den let læste tekst og hvilke træk der karakteriserer denne.* På længere ville de gerne generere et korpus bestående af identificerede let læste tekster med henblik på - via maskinlært modellering - at få automatiseret denne proces.

Referat af DIGHUMLAB-møde 29.6.12, KU

Mødedeltagere: Klaus Bruhn Jensen, Jacob Sidenius, Hanne Jansen, Johan Pedersen, Claus Povlsen, Lina Henriksen (referent)

Claus Povlsen demonstrerede kort clarin.dk mhp. at høre mødedeltagernes reaktioner på funktionalitet, design, formål mv. Disse emner var omdrejningspunktet for mødet.

Vigtigheden af god kontakt ml. projektets temaer blev pointeret. MEF vil typisk være interesseret i det arbejde som primært vil foregå i Århus. I projektbeskrivelsen findes en oversigt over kontaktstrukturen.

Flere kom ind på at ophavsret til data udgør en udfordring.

Metadata blev diskuteret flere gange i løbet af mødet. Essensen af diskussionerne var at metadata som de i øjeblikket præsenterer sig i clarin.dk, er vanskelige at arbejde med. Det er ikke let at gennemskue hvad de betyder, og det er forvirrende at nogle kan være overlappende. Endvidere er nogle perspektiver endnu ikke

repræsenteret i metadata. Det gælder bl.a. det kommunikationsteoretiske synspunkt hvor det vil være interessant at have metadata der beskriver visse forhold vedr. forfatter-læser relationen. Fx hvem/hvor mange har skrevet teksten og hvem/hvor mange har læst teksten, fx ift. facebook-tekster, blogtekster og hjemmesidetekster.

Nogle nævnte at det er for besværligt at lægge data ind som systemet er nu. Andre mente at det muligvis ikke er besværligt, men måske snarere næsten uforståeligt, hvordan man skal gøre. Hvis man skulle lægge data ind i systemet nu, ville det kun blive som flade filer; alt andet er for svært.

Det blev stærkt anbefalet at initiere samarbejde med forskellige data-tunge projekter, for at få sat tingene i gang. På den måde vil DIGHUMLAB delvis kunne udvikle sig på baggrund af data-projekter. Det kan også hjælpe på design af DIGHUMLAB.

Et par af mødedeltagerne vil primært være interesserede i at bruge DIGHUMLAB til at foretage forskellige former for sammenlignende analyser: forskellige sprog, nye-gamle tekster, forskellige teksttyper, osv. Metadata vil være overordentlig afgørende for dette arbejde.

På spørgsmålet om hvad der kan få forskerne til at bidrage med data til DIGHUMLAB kom følgende udmeldinger:

- Man skal vide at der allerede findes en del materiale inden for det emne som interesserer én. Man gider ikke hvis man har en fornemmelse af at man er en af de få, der har lagt noget ind. Kvaliteten af data, metadata og datastrukturen skal styres skarpt for at folk vil være interesserede i at bruge det.
- Det blev nævnt at inden for nogle sprog findes allerede særdeles gode tekstkorpusser som dog ikke er tagget. Det ville være interessant hvis det gennem DIGHUMLAB blev muligt at tagge disse korpusser. (Det blev i den forbindelse nævnt at man ikke så ofte indsamler tekster med henblik på at opbygge eget korpus. Der findes allerede så mange gode korpusser at det kun er nødvendigt i forbindelse med meget specialiserede undersøgelser. Derfor så man det ikke som særligt sandsynligt at mange vil anvende DIGHUMLAB til denne type arbejde.)
- Det vil muligvis være en god ide med en form for automatisk udsendt nyhedsbrev hvor forskeren selv kan afkrydse hvilke emner det vil være interessante at modtage nyheder om, fx i forbindelse med tilgængeligheden af nye data. Nyhedsbrevet kunne evt. være månedligt.

Møde på SDU i Kolding 27. februar 2013

Formålet med mødet var at få en forståelse for hvordan clarin.dk på bedst mulige måde kan bidrage til forskeres og studerendes arbejde.

Mødedeltagere ved dette møde var i meget høj grad oplagte brugere af clarin.dk.

Institut for Design og Kommunikation; Lektor Margrethe Møller: arbejder med maskinoversættelse og multimodalitet. Er interesseret i maskinoversættelsessystemer, multimodale korpusser og bilingvale korpusser.

Institut for Design og Kommunikation; Lektor Nina Bonderup Dohn (hovedfag i filosofi): interesseret i spændingsfeltet mellem erkendelsesteori og læringsteori. Kunne ikke lige sige på hvilken måde clarin.dk evt. kunne bidrage til hendes arbejde, men måske noget med e-læring.

Studerende, specialeskriver: interesseret i publicering af forskningsdata. Var meget afvisende over for tanken om at gøre noget tilgængeligt gennem clarin.dk (ophavsrettigheder mv.).

Institut for Design og Kommunikation; Professor Johannes Wagner: interesseret i multimodale korpora.

Institut for Design og Kommunikation; Videnskabelig assistent Max Eckard: havde brug for et sted at opbevare resurser.

Institut for Design og Kommunikation, Lektor Lotte Weilgaard: arbejder bl.a. med terminologi. Er interesseret i værktøjer der giver mulighed for at arbejde med forskellige typer korpusser, både talte og skrevne. Primære interesse er værktøjer til opmærkning, søgning mv.

Flere mødedeltagere havde afprøvet clarin.dk på forhånd for at forstå hvad platformen kan anvendes til og hvordan man bruger den. Tilbagemeldingen til os var at ingen af de opgaver de havde stillet sig selv lykkedes. For de flestes vedkommende var der ingen problemer med at forstå ord som *metadata, lemmatisering, Tei P5* og lignende (som er et problem for visse potentielle brugere), men de forstod simpelthen ikke strukturen i systemet og havde intet overblik over hvad platformen overordnet er beregnet til.

Bilag 5

Brugerprojekter

Opmærkning af assyriske tekster

Thomas Hertel, post.doc fra Assyriologi på ToRS ønsker mulighed for indlemmelse samt opmærkning af assyriske tekster i DIGHUMLAB regi. Der er ca. 1200 tekster som er translitterationer fra assyriske til latinske bogstaver (med diakritiske tegn) plus deres oversættelse til engelsk.

Assyrisk er en dialekt af det akkadiske sprog og blev talt i det gamle Assyrien. Assyrisk er altså et dødt sprog som der forskes i flere steder i verden. Thomas Hertel vurderer at der på nuværende tidspunkt findes omkring en snes forskere i verden som umiddelbart kunne tænkes at arbejde med digitaliserede, opmærkede assyriske tekster; potentielt set vil dette korpus kunne have forskningsinteresses for et langt større antal assyriologer, semitiske filologer, sproghistorikere og lignende når værktøjet er blevet udviklet og annonceret i internationale forskerkredse

Et brugerprojekt vil involvere:

1) at få teksterne indlejret i TEI P5. CST skal blot lave en skabelon over hvordan en sådan version ser ud – det er blevet at leverandøren i samråd med CST selv vil stå for at få genereret TEI-headeren til teksterne.
2) at få implementeret mulighed for søgning og efterfølgende visning af søgeresultatet, dvs. både kildetekst og måloversættelse. Opmærkning af alignering forventes udført af leverandøren.

Danske Taler

Danske Taler er en website (www.dansketaler.dk) som indeholder en samling af offentlige danske taler. Ressourcen er oprettet af Jesper Troels Jensen (ejer af konsulentfirmaet Rhetorica), som også indtil nu har været den primære drivkraft bag drift og vedligeholdelse af ressourcen. Danske Taler indeholder på nuværende tidspunkt taler fra 98 talere; potentielt kunne der blive tale om langt flere taler. Jesper Troels Jensen ønsker nu et organisatorisk og et fysisk hjemsted således at Danske Taler kan sikres en permanent status der muliggør fortsat indsamling samt studier og forskning i danske taler.

I øjeblikket giver dansketaler.dk mulighed for søgning i talerne med fx *talers navn, talens titel, talens dato* og *nøgleord*, og samlingen er en gratis ressource tilgængelig for alle. De første taler er fra 1500-tallet, og de nyeste er fra nutiden. Danske Taler har endvidere en bestyrelse som bifalder tanken om at DIGHUMLAB bliver hjemsted for ressourcen. Medlemmer af denne bestyrelse er Sabine Kirchmeier-Andersen, Bertel Haarder, Kristian Madsen og Christian Koch.

Danske Taler ligger inde med en del taler som endnu ikke er digitaliserede og som derfor endnu ikke findes i databasen. Digitaliseringsarbejdet vil formodentlig blive udført af bl.a. Institut for Medier, Erkendelse og Retorik, men det er en langsommelig proces eftersom størstedelen af talerne er håndskrevne (og har mange interne henvisninger mv.).

Det er endnu uvist om og hvornår digitaliseringsarbejdet kan fortsætte - og om det vil være interessant for personerne bag Danske Taler at importere dansketaler.dk i DIGHUMLAB når DIGHUMLAB ikke har ressourcer til væsentlig deltagelse i digitaliseringsarbejdet. Kommunikationen om dette er endnu ikke afsluttet. Import af dansketaler.dk i DIGHUMLAB vil involvere oprettelse af metadata på alle taler.

Creolske sprog - deres fællestræk og deres forskellighed i forhold til ikke-kreolske sprog

Kontaktperson: Phd-studerende Aymeric Daval-Markussen**,** Institut for Æstetik og Kommunikation – Lingvistik, Aarhus Universitet

Dette forskningsprojekt har som grundspørgsmål:

a) Er kreolske sprog mere enkle i deres struktur og deres leksikalske inventar?
b) Er der flere fællestræk mellem de kreolske sprog sammenlignet med andre sprog?

Metoden til besvarelse af disse spørgsmål er systematisk udnyttelse af empiri om kreolske sprogdata og data om ikke kreolske sprog. Denne empiriske sammenligning er gjort kvalificeret via manuel annotering af træk der inden for denne forskningsgren er enighed om er typiske for kreolske sprog. Annoteringen blev udført af specialister inden for 18 forskellige kreolske sprog. I forbindelse med den evidensbelagte tilgang til besvarelse af spørgsmål (b) er desuden inkluderet data fra 12 ikke-kreolske sprog taget fra *Word Atlas of Linguistic Structures*. I forskningspraksissen genereres ud fra disse data statistiske modeller primært multipel regression og via implementerede visningsprogrammer repræsenteres resultaterne på en overskuelig og systematisk måde.

CP: Her skal Ulrik Sandborg-Petersen ind – NB: husk at han allerede har givet os en version af det Gamle testamente

Emdros et databasesystem til lingvistisk analyse og opmærkning

Kontaktperson: Ulrik Sandborg-Petersen, ph.d., cand.mag., B.Sc. ,Aalborg Universitet

Ulrik Sandborg-Petersen har implementeret databasen (EMdF) og det tilhørende søgesprog MQL tilsammen benævnt Emdros, for mere information om disse se: http://www.hum.aau.dk/~ulrikp/pdf/petersen-emdros-COLING-2004.pdf

Programmet er open source og kan sammen med dokumentation findes på http://emdros.org/.

De lingvistiske og ekstralingvistiske annoteringer er meget avancerede i hvert fald i forhold hvad der kan gøres automatisk i clarin.dk-sammenhæng og søgesproget er tilsvarende kompliceret og meget lidt brugervenligt.

Emdros er blevet brugt til at opmærke Kaj Munk-tekster på Kaj Munk Forskningscentret ved Aalborg Universitet. Resultatet kan ses på: Det Virtuelle Forskerværksted for Kaj Munk Studier. Det kræver dog at man indhenter et passwrd.

Ulrik Sandborg-Petersen har derudover givet os en digital version af en oversættelse af det Gamle Testamente fra 1871.