DK-CLARIN WP 5.1-arbejdspapir
Jørg Asmussen og Lene Offersgaard

Version 2.0 – 21. september 2008

Resumé

Nærværende papir beskriver et scenarie på en arbejdsgang, som man kan forestille sig ville være af interesse for DK-CLARIN-brugere: Man sammenstiller sit eget korpus på baggrund af tekstmateriale tilgængeligt via DK-CLARIN, annoterer dette særlige korpus på tokenniveau og gemmer sit korpus samt annotationer til senere brug og for andre DK-CLARIN-brugere.

1 Terminologi

En nærmere beskrivelse af den anvendte korpus-terminologi i dette dokument findes i [Asmussen og Offersgaard, 2008]. Begreber, som er defineret her, er i det følgende understreget, første gang de nævnes.

2 Målsætning

I DK-CLARIN skal det blandt mange andre ting være muligt

1. at få et overblik over tilgængelige tekstsamlinger
2. på baggrund af bestemte tekstkarakteristika at oprette et brugerdefineret korpus bestående af tekstenheder fra en tilgængelig tekstsamling
3. at søge i et brugerdefineret korpus
4. at annotere et brugerdefineret korpus på tokenniveau og siden redigere i annotationerne
5. at arkivere det brugerdefinerede korpus sammen med annotationerne, så andre (gennem rettighedsstyring) kan få adgang til både korpusset, udvælgelseskriterierne for det og annotationerne til det.

Det følgende er et oplæg til, hvordan ovenstående punkter kan opfyldes.

2.1 Overblik over og adgang til tekstsamlinger

Beskrivelse

DK-CLARIN-partnerne gør deres tekstsamlingers og allerede eksisterende korporas metadata søgbare via WP 5.1's infrastruktursite. Her kan brugere så få et overblik over, hvilke tekstsamlinger og korpora (oprettet af tekstsamlingsudbyderen selv eller af andre brugere) der er tilgængelige, sammen med en beskrivelse

af dem og de metadatatyper, der knytter sig til de enkelte tekstenheder i dem. Til disse metadata hører også, hvilke annotationer der knytter sig til en tekstenhed på tokenniveau.

Case

En bruger har på DK-CLARIN's infrastruktursite adgang til at se beskrivelser og specifikationer for alle tekstsamlinger, allerede eksisterende korpora og annotationer, herunder hvilke metadata der er specificeret for de enkelte resurser.

Implementering

På infrastruktursitet implementeres en browser, som brugeren kan navigere i resursernes (in casu tekstsamlinger, korpora og annotationer) metadata med.

Udbyderne af tekstsamlinger gør disses metadata søgbare for DK-CLARIN-infrastruktursitet i form af en webservice.

2.2 Sammenstilling af et brugerdefineret korpus

Beskrivelse

En bruger, der ønsker at sammenstille sit eget korpus ud fra det materiale, der er til rådighed i en eller flere tekstsamlinger, kan på baggrund af tekstenhedernes metadata definere en profil, og alle tekster i samlingen, der matcher denne profil, vil da udgøre det ønskede korpus.[1] Såfremt teksterne, som indgår i korpusset, er ophavsretligt clearede og kan distribueres frit, vil brugeren evt. kunne downloade korpusset i sin helhed i et fastlagt format fra udbyderen via infrastruktursitet. Kan teksterne ikke videredistribueres, hvilket vil være standardsituationen, kan brugeren få adgang til dem »virtuelt«. Den virtuelle adgang inkluderer en mulighed for at foretage søgninger i det virtuelle korpus vha. et konkordansværktøj samt at forsyne det med egne annotationer.

Case

Baseret på metadata for den enkelte tekstsamling har brugeren mulighed for at definere et korpus til søgning, annotering eller evt. download. En bruger kan definere et korpus ved at definere en profil på infrastruktursitet og dernæst klikke på *Opret korpus*. Brugeren bliver bedt om at give en beskrivelse af det valgte korpus. Brugeren fastlægger offentlighedskriteriet for korpusset.

Implementering

På infrastruktursitet implementeres et interface til definition af profiler og oprettelse af brugerdefinerede korpora. Profildata stilles til rådighed af tekstsamlingens udbyder i form af en webservice.

På infrastruktursitet implementeres endvidere et interface til input af brugerens beskrivelse af det valgte korpus. Her håndteres og opbevares desuden

[1] Skal et korpus sammenstilles med materiale fra flere tekstsamlinger, kræves, at deres metadata er strukturelt identiske, idet det ellers vil være venskeligt at opstille en profil, der kan anvendes på tværs af forskellige metadatasystemer. I praksis vil et korpus derfor overvejende kun kunne sammenstilles med tekster fra kun én tekstsamling.

en reference til metadata for det brugerdefinerede korpus; korpusmetadata selv sendes til udbyderen og opbevares dér. Korpusmetadata skal indeholde:

1. bruger-id
2. en liste af pointere (indeksreferencer) til tekstbanken; disse udpeger de data (tekstenheder og tokens), som tilsammen udgør korpusset
3. søgeprofilen, som ligger til grund for udvælgelsen af korpusset
4. brugerens beskrivelse af korpusset

2.3 Søgning i et brugerdefineret korpus

Beskrivelse

Brugeren skal kunne lave konkordanssøgninger i det brugerdefinerede korpus vha. et konkordansværktøj (og i andre korpora, han har adgang til).

Case

Brugeren vil lave sprogbrugsundersøgelser i sit korpus som dem, der der beskrevet på KorpusDK's hjemmeside: http://ordnet.dk/korpusdk/hjaelp/teksteksempler/forside.

Implementering

Udviklingen af en grænseflade til konkordanssøgninger er en yderst resursekrævende opgave. Derfor bør WP 5.1 afveje den nøje i forhold til andre infrastrukturopgaver, der også skal løses.

En konkordansgrænseflade i DK-CLARIN kan tage udgangspunkt i KorpusDK's eksisterende grænseflade, som kan ses her: http://ordnet.dk/korpusdk/teksteksempler. Det kan undersøges, hvorvidt den kan konfigureres til DK-CLARIN-brug.

KorpusDK's konkordansgrænseflade samt i øvrigt den langt mere primitive konkordansvisning på http://sproget.dk kommunikerer med Jørg Asmussens korpusserver PyCOCS, som beror på CQP som søgemaskine. PyCOCS skulle med overskuelige midler kunne rekonfigureres til en egentlig webservice.

Problematisk er det at gøre brugerdefinerede korpora søgbare i PyCOCS/CQP, idet det kræver, at teksterne eksporteres fra tekstbanken i et bestemt format, som er nødvendigt, for at det kan indekseres til brug i CQP. Selve indekseringen tager tid, afhængig af korpusstørrelse og platform, skal der regnes med mellem 5 og 20 minutter. Indekseringsprogrammerne er udviklet til en korpusadministrator med særlige rettigheder: Det kan derfor være sikkerhedsmæssigt intrikat at lade alle og enhver bruge dem over nettet til at indeksere egne korpora med.

Et åbent spørgsmål er, om det skal være muligt at foretage konkordanssøgninger i en brugerfastlagt serie af korpora, da det vil være noget mere komplekst at implementere. Forslaget er, at for denne række af korpora kan man lave en selvstændig søgning i hvert korpus, og få præsenteret søgeresultaterne i samme grænseflade. Det foreslås at man let kan vælge at bruge samme søgeforespørgsel i alle de valgte korpora, men også har mulighed for forskellige søgeforespørgsler

for de enkelte korpora. Det primære er, at man har mulighed for at opnå en fælles visning af resultaterne fra søgninger i forskellige korpora. Denne option medfører omfattende ekstra udviklingsarbejde.

Et andet åbent spørgsmål er, hvorvidt det skal være muligt at foretage statistiske undersøgelser af et brugerdefineret korpus (og andre tilgængelige korpora). Dette ville kræve et særligt statistikværktøj, jf. fx KorpusDK's hjemmeside under http://ordnet.dk/korpusdk/naboord – og helst et, der var betydelig mere avanceret. Dette er igen en yderst resursetung udviklingsopgave. Dertil kommer, at statistiske beregninger i korpora kan være yderst tidkrævende: For at nedbringe beregningstiden er en præprocessering af data som regel nødvendig, der dog i sig selv igen kan væres særdeles tidkrævende.

2.4 Annotering af et brugerdefineret korpus

Beskrivelse

Et brugerdefineret korpus skal kunne annoteres på tokenniveau. Dette forudsætter, at tekstenhederne i det én gang for alle er inddelt i tokens, der hvert især har et unikt id, som forbliver konstant over tid. Da en tekstenheds annotationer – også på tokenniveau – opbevares/registreres i tilknytning til tekstbanken (men ikke nødvendigvis på samme maskine som tekstbanken), bør tekstenheder tokenopdeles allerede i forbindelse med indlemmelsen af dem i tekstbanken. Det er hensigtsmæssigt, at DK-CLARIN definerer, hvordan tekst tokenopdeles, jf. under *Token* i [Asmussen og Offersgaard, 2008]. Da hvert token via tekst-id'et og tokenindekset som beskrevet i nævnte arbejdspapir kan udpeges entydigt, kan en ny, brugerspecifik annotering på tokenniveau nemt linkes til de eksisterende tokens i tekstbanken uden selv nødvendigvis at skulle indlemmes heri. Hvis der skulle blive brug for at adressere på sub-tokenniveau, kan der overvejes karaktervis adressering inden for hvert token som supplement til den allerede beskrevne tokenadressering.

Hvis brugeren ikke ønsker at opmærke hele det brugerspecifikke korpus, skal der være en mulighed for at specificere, hvilke dele af korpusset der ikke er opmærket, dvs. der skal skabes en mulighed for nul-annotering.

Annoteringer laves vha. et annoteringsværktøj, enten brugerens eget eller et, der er tilgængeligt fra DK-CLARINS's infrastruktursite.

Case

Brugeren annoterer korpusset vha. et dertil egnet værktøj. Det vil i dette scenarie ske online, direkte ned i infrastruktursitet, der sørger for, at der er forbindelse til korpusset via den webservice, tekstudbyderen har etableret.

Alternativt er det i visse tilfælde tænkeligt, at brugeren kan downloade sit korpus og annotere det på sin lokale computer. Her er det vigtigt, at tokenreferencerne for korpusset bevares, ellers kan man ikke referere annoteringerne til korpusset bagefter, og andre kan ikke få glæde af annoteringen.

Implementering

Annotering kan ske enten manuelt eller maskinelt.

Til den manuelle annotering kræves et annoteringsinterface på infrastruktursitet, hvori brugeren kan fastlægge sine annoteringer, og som han kan anvende til at gå tekstmaterialet igennem med, og for de enkelte tokens i det at sætte de annotateringer, han ønsker. Annotationen gemmes direkte ned i udbyderens tekstbank. Skal der stilles et sådant annoteringsinterface til rådighed i DK-CLARIN-regi, må det udvikles. Det kan evt. bygges oven på konkordansgrænsefladen. Udviklingen af et sådant værktøj skønnes temmelig resursetungt.

Den manuelle annotering kan i princippet også udføres på brugerens lokale maskine. Dette kræver dog, at han har ret til (temporært) at downloade korpusset tekstord for tekstord, tekstenhed for tekstenhed. Han kan så udføre annoteringerne i sit eget annoteringsværktøj og uploade dem i et fastlagt format. Hvis det er muligt at finde en måde at tilgå tekstmaterialet, uden at det kan gemmes lokalt, ville denne løsning måske være interessant. Ellers må man forudse alvorlige ophavsretlige problemer.

Den maskinelle annotering vil typisk foregå efter algoritmer, som brugeren selv fastlægger, hvorfor den medfører de samme problemstillinger som manuel annotering på brugerens lokale maskine. Hvis man kan sikre, at tekstmaterialet forbliver på udbyderens site, kan der muligvis defineres et webservice-agtigt interface, hvorigennem annoteringen kan udføres.

2.5 Arkivering af brugerens korpus og annotationer

Beskrivelse

Få at undgå unødig redundans bør tekstenheder i en tekstbank ikke repliceres, når der tilføjes nye annotationer. I stedet knyttes nye annotationer via tekst-id og tokenindeks til den eksisterende tekst i tekstbanken. Annotationen anses derfor som en selvstændig resurse: en *annoteringsresurse*. Dette koncept gør det muligt at knytte særlige metadata med oplysninger om ejerskabsforhold til annoteringsresursen. DK-CLARIN-annoteringsresurser for tekstmateriale skal være indbyrdes kompatible. Annoteringsresurser gemmes på DK-CLARIN's infrastruktursite.

Case

Efter oprettelsen af det brugerdefinerede korpus »gemmes« det ved at gemme reference-id'erne til tekstbanken, profilen, som fastlægger korpussets indhold, samt brugerens beskrivelse af det (incl. fastlæggelsen af offentlighedskriteriet) på infrastruktursitet.

Efter annoteringen (eller den del af annoteringen man lige nu har tid til at foretage) er udført, gemmes de brugerspecifikke annoteringsdata på infrastruktursitet.

Er der foretaget en lokal annotering på brugerens computer, skal annoteringsdataene være defineret i en annoteringsfil i et fastlagt format, som så kan uploades til DK-CLARIN's infrastruktursite. Der skal også afleveres metadata for annoteringen samt beskrivelsen af det brugerspecifikke korpus – det hele i et fastlagt format. Når annoteringen uploades til infrastruktursitet, registrerer dette annoteringen som en ny annoteringsresurse og korpusbeskrivelsen lagres sammen med annoteringen. Infrastruktursitet oplyser tekstbankværten om eksistensen af den nye annotering. Annoteringen behøver ikke at dække hele kor-

pusset. Den del af korpusset, som ikke er annoteret, kan nul-annoteres, sådan at andre kan se, at brugeren ikke fik behandlet hele korpusset.

Implementering

På infrastruktursitet skal der etableres et interface til håndtering af gemningen og genfinding af gemte annoteringsresurser, og den bagvedliggende funktionalitet skal implementeres.

Der skal endvidere implementeres uploadmekanismer, der gør det muligt at tage imod en brugers annoteringsresurse og ekspedere den på rette vis.

Litteratur

[Asmussen og Offersgaard, 2008] Asmussen, J. og Offersgaard, L. (2008). Korpuslingvistisk terminologi. Rapport, DK-CLARIN.